



HYPERION RESEARCH

# Hyperion Research HPC/AI Market Update

October 2024

[www.HyperionResearch.com](http://www.HyperionResearch.com)  
[www.hpcuserforum.com](http://www.hpcuserforum.com)

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

# About Hyperion Research



([www.HyperionResearch.com](http://www.HyperionResearch.com) & [www.HPCUserForum.com](http://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

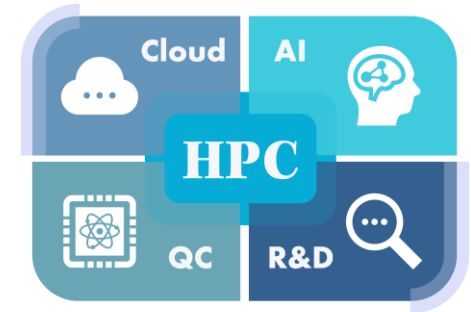
Jie Wu, China and Technology Trends

Mara Jacob, HPC User Forum Support

# Example Research Areas

([www.HyperionResearch.com](http://www.HyperionResearch.com) & [www.HPCUserForum.com](http://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**



# HPC/AI Market Update

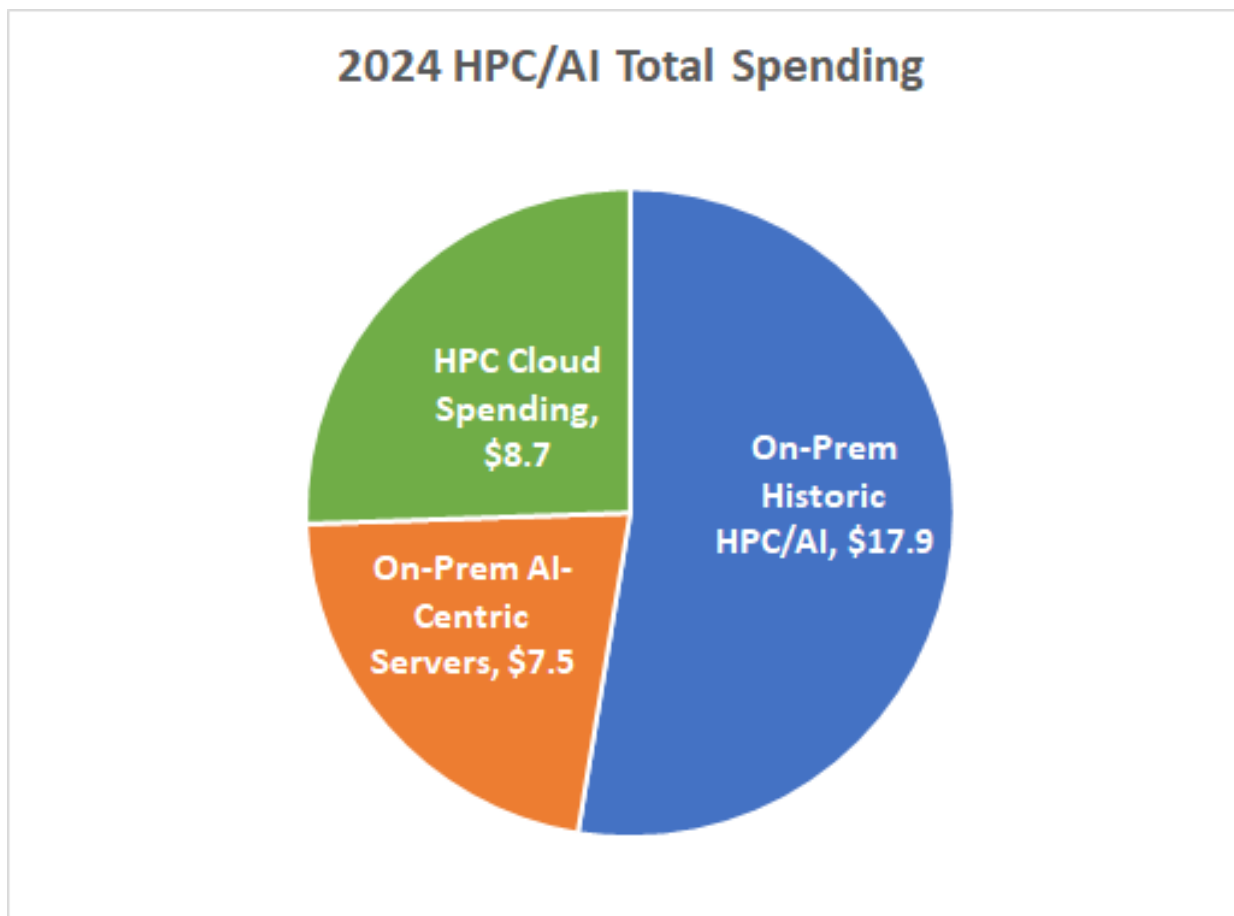
# 2024 Looks Like a Strong Growth Year

*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 20% in 2024**
  - AI-focused on-premises servers are growing at high rate of close to 40% in 2024
- **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

*The Overall HPC/AI Spending is projected to exceed \$34 billion (\$US)*



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

# Updated View of the HPC/AI Market

*The addition of non-traditional AI servers increases the HPC/AI market by \$7.5 billion (US\$) by 2024*

## **Market Segment Definition:**

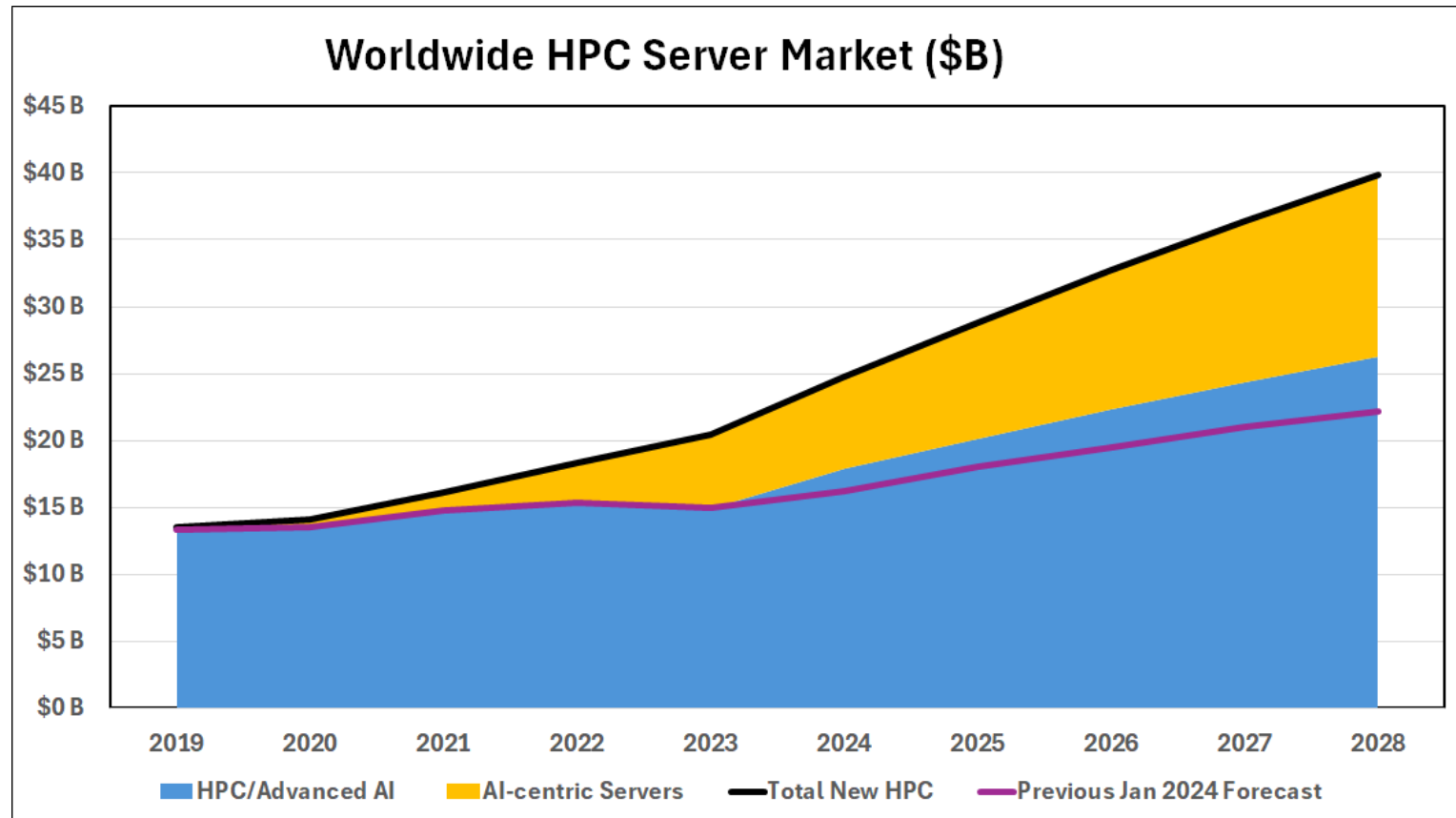
**AI-centric Servers (new revenues added to the previous HPC market sizing)**

These are on-premises AI-centric HPC servers that are provided by non-traditional HPC suppliers like NVIDIA, Cerebras, SambaNova, SuperMicro, etc., frequently at non-traditional HPC user sites like large enterprise sites adding AI capabilities.

- These servers are designed primarily to run AI and AI-related workloads
- These servers are a subsegment of the overall HPC market but haven't historically been accounted for within prior HPC market numbers

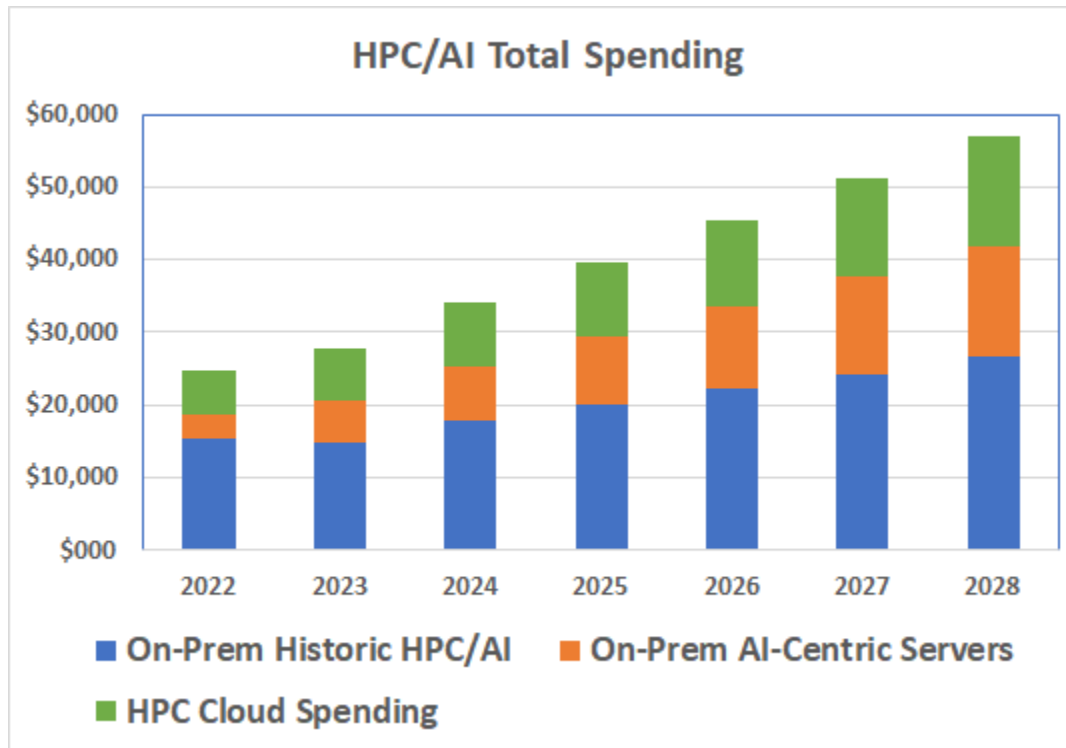
# Updated View of the HPC/AI Market

- *Hyperion Research just announced a 36.7% increase in the HPC/AI market size*
- *Now tracking non-traditional AI suppliers*



# The HPC/AI 5-year Forecast

Projected to reach \$56.9 billion (US\$) by 2028

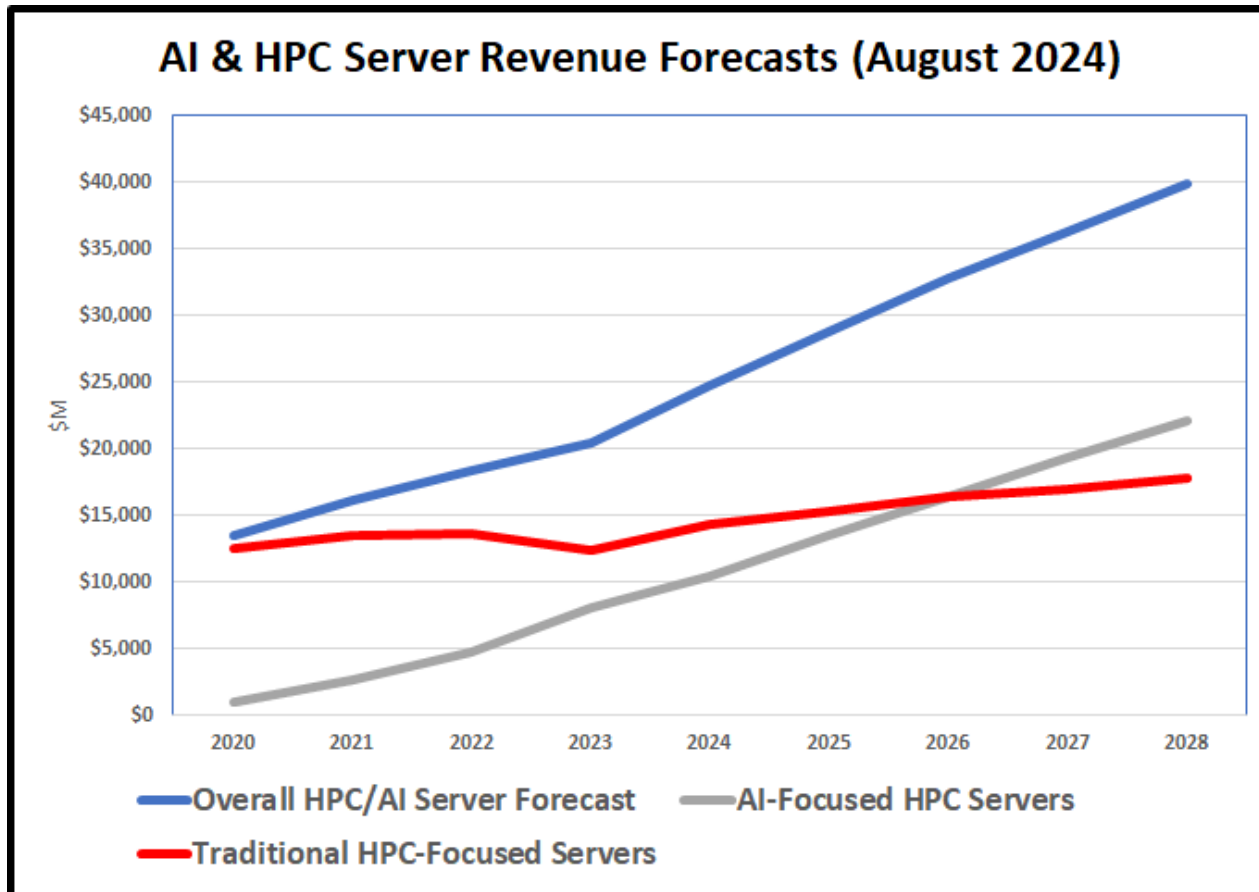


Worldwide Technical Computing/HPC Spending								
	2022	2023	2024	2025	2026	2027	2028	CAGR 23-28
On-Prem Historic HPC/AI	\$15,369	\$14,954	\$17,932	\$20,088	\$22,279	\$24,302	\$26,810	12.4%
On-Prem AI-Centric Servers	\$3,437	\$5,782	\$7,458	\$9,472	\$11,420	\$13,495	\$14,967	21.0%
HPC Cloud Spending	\$6,040	\$7,180	\$8,711	\$10,072	\$11,605	\$13,302	\$15,115	16.1%
<b>Total HPC/AI</b>	<b>\$24,846</b>	<b>\$27,915</b>	<b>\$34,101</b>	<b>\$39,631</b>	<b>\$45,305</b>	<b>\$51,099</b>	<b>\$56,892</b>	<b>15.3%</b>

Source: Hyperion Research, Oct. 2024

# HPC Compared to AI-centric Servers

*Many on-prem servers are running both traditional HPC and AI workloads*



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

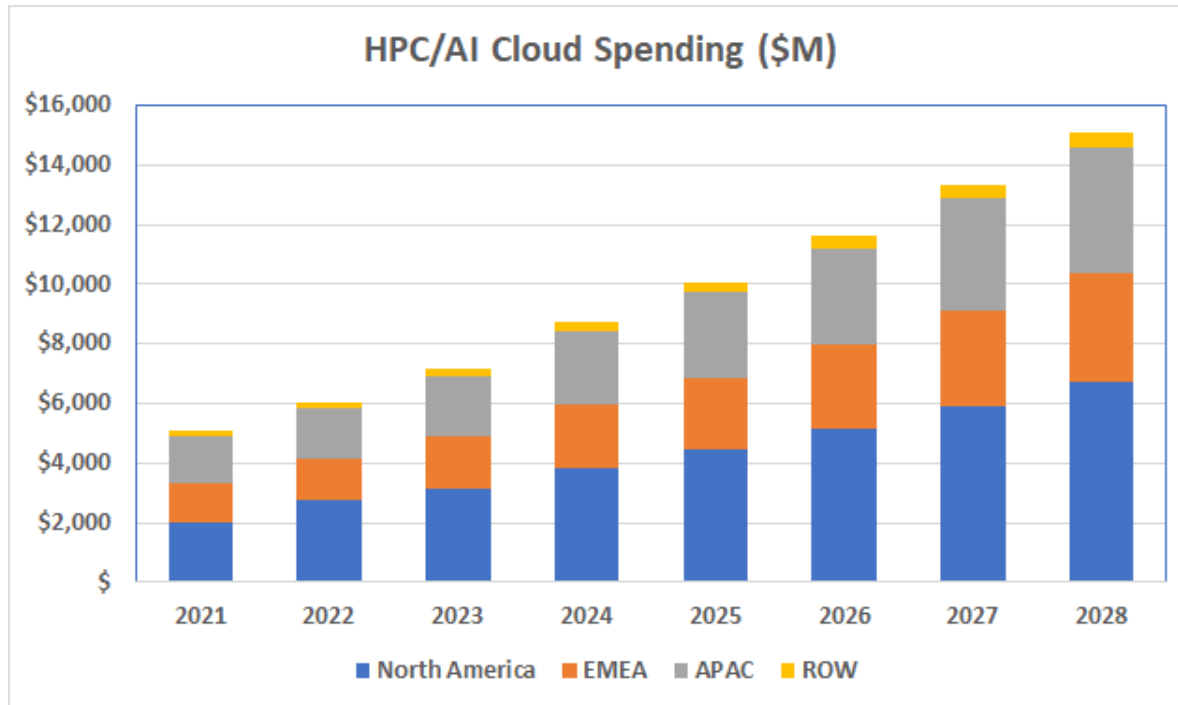
# On-prem 5-year Forecast by Region

Projected to reach \$41.8 billion (US\$) by 2028 (15% CAGR)

Worldwide On-Prem Technical Computer Market Revenue								
	2022	2023	2024	2025	2026	2027	2028	CAGR 23-28
Historic HPC/Advanced AI	\$15,369	\$14,954	\$17,932	\$20,088	\$22,279	\$24,302	\$26,810	12.4%
AI-Centric Servers	\$3,437	\$5,782	\$7,458	\$9,472	\$11,420	\$13,495	\$14,967	21.0%
<b>Total New HPC</b>	<b>\$18,805</b>	<b>\$20,735</b>	<b>\$25,390</b>	<b>\$29,559</b>	<b>\$33,699</b>	<b>\$37,797</b>	<b>\$41,777</b>	<b>15.0%</b>
<i>Source: Hyperion Research, Oct. 2024</i>								
HPC On-Prem Server Sales By Region								
	2022	2023	2024	2025	2026	2027	2028	CAGR 23-28
North America	\$6,876	\$6,580	\$8,067	\$8,666	\$9,821	\$10,719	\$11,893	12.6%
EMEA	\$4,306	\$4,145	\$5,021	\$6,026	\$6,465	\$7,035	\$7,707	13.2%
Asia/Pacific w/o Japan	\$3,308	\$3,385	\$3,784	\$4,218	\$4,701	\$5,152	\$5,684	10.9%
Japan	\$644	\$612	\$756	\$829	\$902	\$964	\$1,042	11.2%
Rest-of-World	\$235	\$232	\$305	\$347	\$391	\$432	\$484	15.9%
<b>Total</b>	<b>\$15,369</b>	<b>\$14,954</b>	<b>\$17,932</b>	<b>\$20,088</b>	<b>\$22,279</b>	<b>\$24,302</b>	<b>\$26,810</b>	<b>12.4%</b>
<i>Source: Hyperion Research, Oct 2024</i>								
AI-Centric On-Prem Server Sales By Region								
	2022	2023	2024	2025	2026	2027	2028	CAGR 23-28
North America	\$1,733	\$2,877	\$3,778	\$4,641	\$5,693	\$6,730	\$7,499	21.1%
EMEA	\$867	\$1,442	\$1,879	\$2,557	\$2,983	\$3,516	\$3,873	21.8%
Asia/Pacific w/o Japan	\$666	\$1,178	\$1,416	\$1,790	\$2,169	\$2,575	\$2,856	19.4%
Japan	\$130	\$213	\$283	\$352	\$416	\$482	\$524	19.7%
Rest-of-World	\$042	\$072	\$102	\$131	\$160	\$192	\$216	24.7%
<b>Total</b>	<b>\$3,437</b>	<b>\$5,782</b>	<b>\$7,458</b>	<b>\$9,472</b>	<b>\$11,420</b>	<b>\$13,495</b>	<b>\$14,967</b>	<b>21.0%</b>
<i>Source: Hyperion Research, Oct 2024</i>								

# 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%
<b>Total HPC Cloud Spending</b>	<b>\$5,100</b>	<b>\$6,040</b>	<b>\$7,180</b>	<b>\$8,711</b>	<b>\$10,072</b>	<b>\$11,605</b>	<b>\$13,302</b>	<b>\$15,115</b>	<b>16.1%</b>

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	1 exascale ~\$350M, plus 1 exascale (or pre) system ~\$200M	?	2 exascale system ~\$600M	1 pre-exascale system ~\$125M	5-6	~\$1.6B
2025	1 or 2 exascale systems ~\$300M each	2 or 3 exascale systems ~\$350M each	1 exascale system ~\$200M	1 or 2 exascale systems ~\$350M each	1 near-exascale system ~\$125M	6-9	\$1.7B - \$2.7B
2026	2 exascale systems ~\$300M each	2 or 3 exascale systems ~\$325M each	?	1 or 2 exascale systems ~\$325M each	1 or 2 exascale systems ~\$150M each	6-9	\$1.7B - \$2.5B
2027	2 exascale systems ~\$275M each	2 or 3 exascale systems ~\$300M	1 exascale system ~\$150M	1 or 2 exascale systems ~\$275M each	2 or 3 exascale systems ~\$130M each	8-11	\$1.8B - \$2.5B
2028	2 exascale systems ~\$250M each	2 or 3 exascale systems ~\$275M	1 or 2 exascale systems ~\$150M each	1 or 2 exascale systems ~\$275M each	2 or 3 exascale systems ~\$125M each	8-12	\$1.7B - \$2.6B
<b>Total</b>	<b>12-13</b>	<b>14-19</b>	<b>5-6</b>	<b>8-12</b>	<b>7-10</b>	<b>47-61</b>	<b>\$13.4B - \$16.8B</b>

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

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

Source: Hyperion Research, March 2024

# A New Tool For Data Analysis

# The New Hyperion Research Data Tool

*To help find important findings more quickly*



Welcome

Tables

Charts

Questionnaire

This annual study is part of the eighth edition of Hyperion Research's high-performance computing (HPC) end-user-based tracking of the HPC marketplace. It included 107 HPC end-user sites with 2,243 HPC systems.

**This dashboard is provided as a resource to quickly glean insights on the study data and is read-only. To receive an Excel copy of any specific table(s) or chart(s), please contact the email below.**

**“Tables”** tab includes the responses for all multiple-choice questions, overall and split by sector. You can search the question list for any desired key words (e.g., “cloud”) and also filter by countries with sufficient data size.

**“Charts”** tab includes column charts for all multiple-choice questions, overall and by each sector.

**“Questionnaire”** tab includes the full questionnaire text, including possible responses. Hover over any question to see a preview of the associated data table.

**For questions or data requests, please contact Jaclyn Ludema at [jludema@hyperionres.com](mailto:jludema@hyperionres.com)**

# HPC End User Profile Summary Results 2024

Clear all slicers

Questions

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

Country

All

Multiple Answer

Single Answer

## All Sectors

Questions	Number of Responses	%
<input checked="" type="checkbox"/> 10) How many GPUs/accelerators/co-processors (i.e., GPU, 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%
<b>Total</b>		

Source: Hyperion Research, 2024

## Industry

Questions	Number of Responses	%
<input checked="" type="checkbox"/> 10) How many GPUs/accelerators/co-processors (i.e., GPU, 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%
<b>Total</b>		

Source: Hyperion Research, 2024

## Government

Questions	Number of Responses	%
<input checked="" type="checkbox"/> 10) How many GPUs/accelerators/co-processors (i.e., GPU, 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	%
<input checked="" type="checkbox"/> 10) How many GPUs/accelerators/co-processors (i.e., GPU, 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%
<b>Total</b>		



Welcome

Tables

Charts

Questionnaire

# 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?

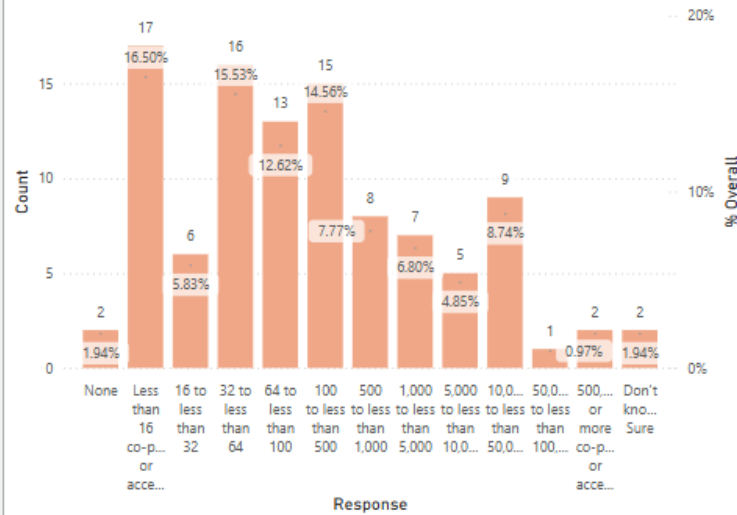
Country

All

Multiple Answer

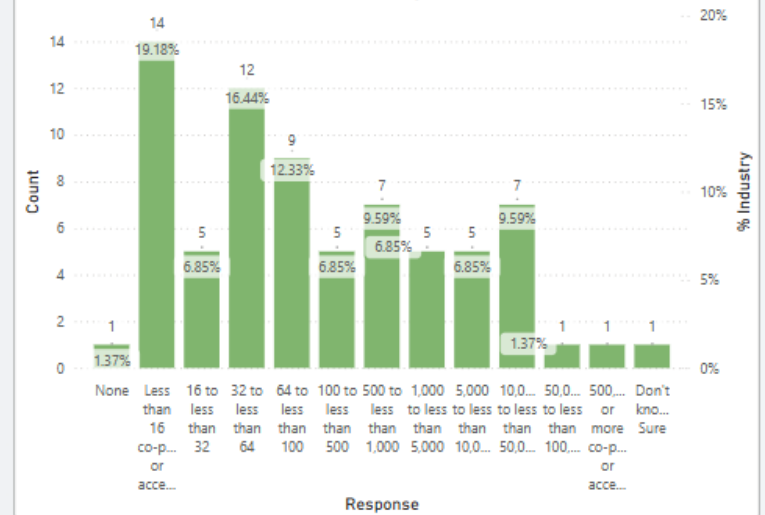
Single Answer

All Sectors



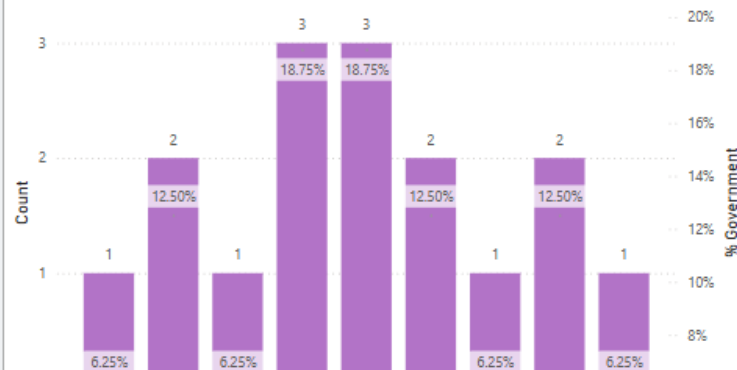
Source: Hyperion Research, 2024

Industry

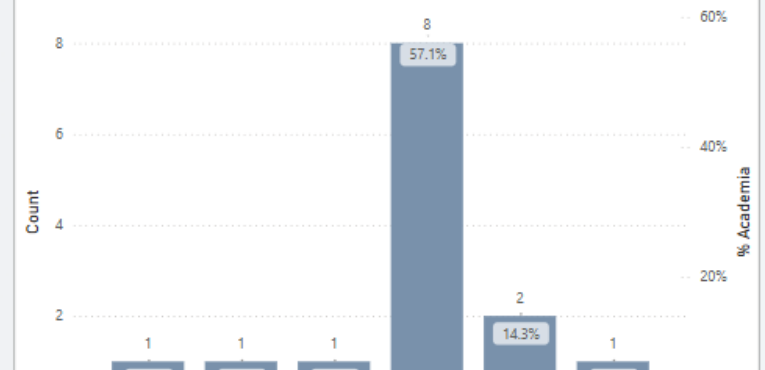


Source: Hyperion Research, 2024

Government



Academia



Welcome

Tables

Charts

Questionnaire

## HPC End User Profile Summary Results 2024

### Questions

Questions	Responses
1) In what sector is your organization in?	<ul style="list-style-type: none"> <li>a) Government</li> <li>b) Academia</li> <li>c) Industry</li> </ul>
1.b) Please specify which industry:	<ul style="list-style-type: none"> <li>i. Bio &amp; Life sciences, pharmaceutical, biological, life sciences, healthcare, drug discovery, bioinformatics, genomics, etc.</li> <li>ii. CAE, manufacturing, e.g., aerospace, automotive, consumer products, etc.</li> <li>iii. Chemical engineering, chemical design, development, and production</li> <li>iv. Mechanical design e.g., CAD</li> <li>v. DCC, entertainment, digital content creation, 3D animation, advanced graphics, gaming, visualization, etc.</li> <li>vi. Financial or economic modeling, pricing, risk management, modeling, business intelligence, etc.</li> <li>vii. EDA, electronic design and analysis</li> <li>viii. IT, computers, HPC systems, IT services, ISV, software company, cloud provider, etc.</li> <li>ix. Geosciences, energy, petroleum, oil and gas, seismic, reservoir simulation, alternative energy, power distribution, etc.</li> <li>x. Weather/climate</li> <li>xi. Transportation and logistics, traffic management, pattern recognition, linear programming, etc.</li> <li>xii. Retail, marketing, and related BI</li> <li>xiii. Telecommunications</li> <li>xiv. Other (please specify)</li> </ul>
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.	<ul style="list-style-type: none"> <li>a) None – we only use external clouds</li> <li>b) 1</li> <li>c) 2 to 4</li> <li>d) 5 to 7</li> <li>e) 8 to 10</li> <li>f) 11 to 12</li> <li>g) 13 to 15</li> <li>h) 16 to 20</li> <li>i) 21 to 25</li> <li>j) 26 to 30</li> <li>k) 31 to 40</li> <li>l) 41 to 50</li> <li>m) 51 to 75</li> <li>n) 76 to 100</li> <li>o) More than 100 HPC/AI systems</li> </ul>
3) What is the approximate Peak Performance (in PF) of your LARGEST SYSTEM:	<ul style="list-style-type: none"> <li>a) Less than 0.5 petaflops</li> <li>b) 0.5 to less than 1 petaflop</li> <li>c) 1 to less than 5 petaflops</li> <li>d) 5 to less than 10 petaflops</li> <li>e) 10 to less than 25 petaflops</li> <li>f) 25 to less than 50 petaflops</li> <li>g) 50 to less than 100 petaflops</li> <li>h) 100 to less than 250 petaflops</li> <li>i) 250 to less than 500 petaflops</li> <li>j) 500 to less than 750 petaflops</li> <li>k) 750 to less than 1,000 petaflops</li> <li>l) 1,000 to less than 2,500 petaflops</li> <li>m) 2,500 petaflops or greater</li> </ul>

# 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

# Some Recent AI Study Results

# AI Study Findings

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%

# 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%

# Conclusions

- **2024 is expected to be a strong growth year**
  - GPUs, cloud, AI/ML/DL/LLM are high growth areas
  - Non-traditional suppliers are growing fast
- **New technologies are showing up large numbers:**
  - Generative AI and LLMs 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 and the need for much larger data sets**
- **There are concerns about how the market can adopt to so many changes in GPUs and CPUs**
- **AI is scaling so fast that power & costs are redefining data centers**

# We Welcome Questions, Comments and Suggestions



Please contact us at:  
[info@hyperionres.com](mailto:info@hyperionres.com)