Hyperion Research Market Update

April 2024

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

www.HyperionResearch.com
www.hpcuserforum.com

About Hyperion Research

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

Hyperion Research mission:

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

HPC User Forum mission:

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

The Hyperion Research Team

Analysts

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

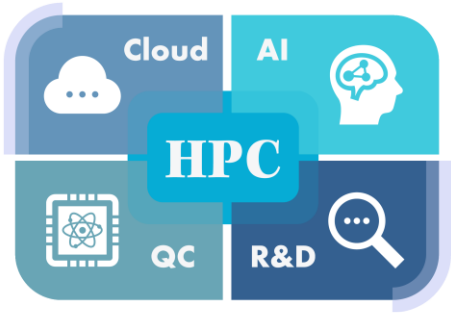
- Cary Sudan, Principal Survey Specialist

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)



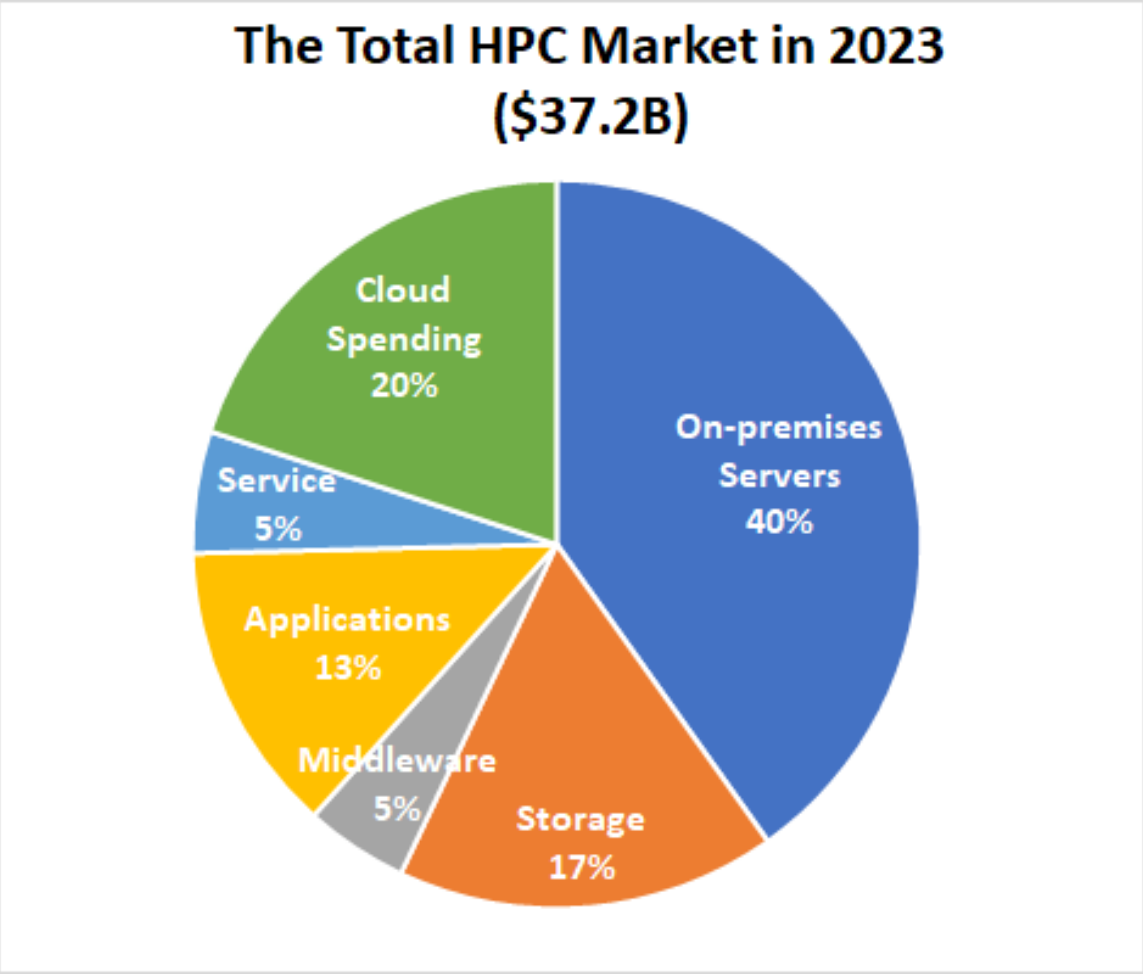
- Traditional HPC
- AI, ML, DL, LLMs, Graph
- Cloud Computing
- Storage & Data
- Interconnects
- Software & Applications
- ROI and Scientific Returns from HPC
- Power & Cooling
- Tracking all Processor Types & Growth rates
- Quantum Computing
- R&D and Engineering -- all types
- Edge Computing
- Supply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

HYPERION RESEARCH

HPC Market Update

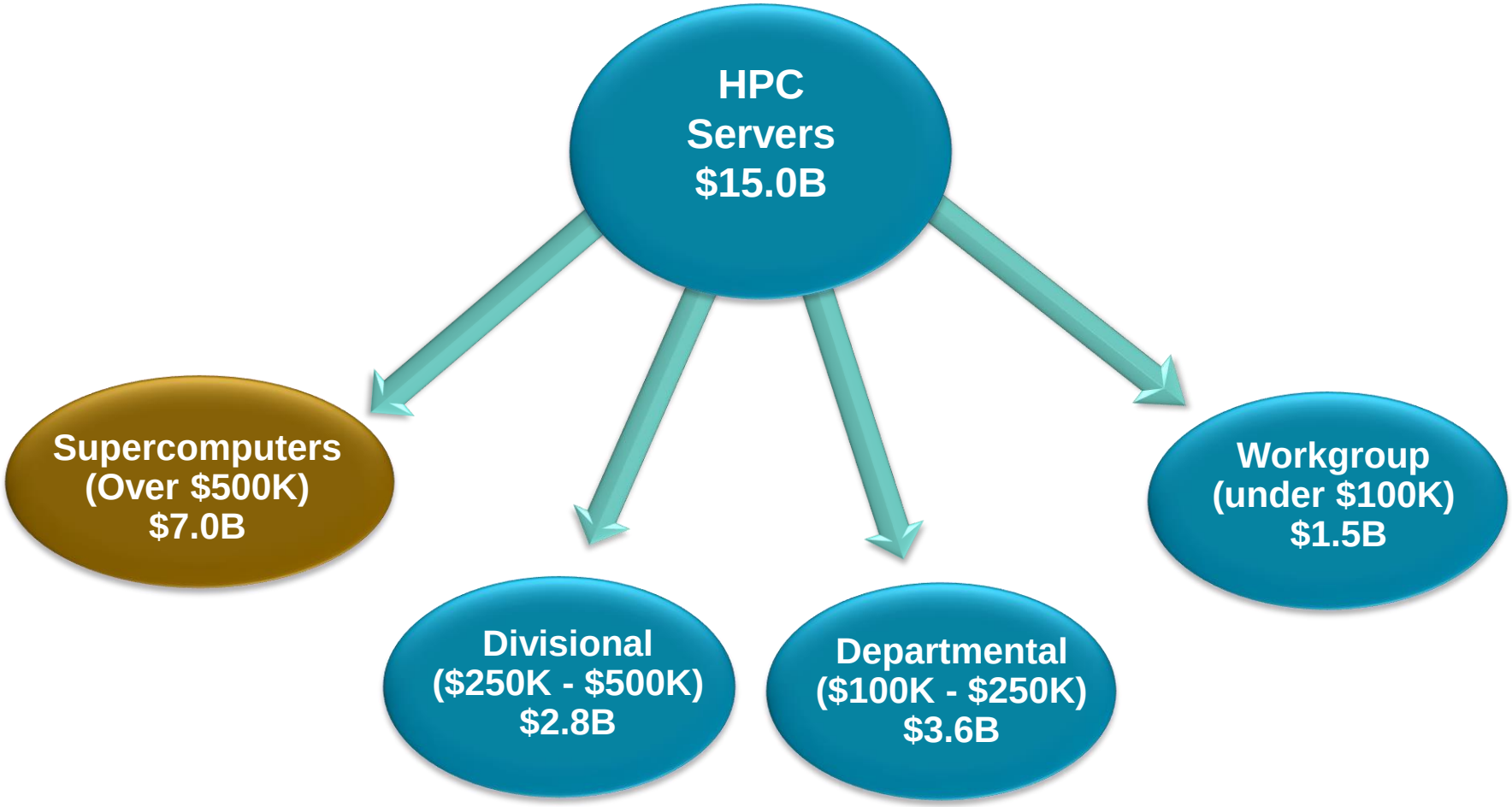
The Overall HPC Market in 2023

Looking at the overall HPC market, including servers, cloud usage, storage, software and repair services = $37.2 billion USD



The 2023 Worldwide On-Prem HPC Server Market: $15.0 Billion (down 2.7%)

2024 is projected to be around $16.3 Billion



2023 WW HPC On-Prem Market by Vendor and Sector ($ Millions)

Vendor	2023 Revenues ($M)
HPE	$4,712
Dell Technologies	$3,659
Lenovo	$1,268
Inspur	$1,041
Sugon	$580
Atos	$414
IBM	$405
Penguin	$385
Fujitsu	$218
NEC	$197
Other	$2,075
Total	**$14,954**

Source: Hyperion Research, April 2024

Vertical/Sector	2023 Revenues ($M)
Bio-Sciences	$1,390
CAE	$1,700
Chemical Engineering	$167
DCC & Distribution	$800
Economics/Financial	$727
EDA / IT / ISV	$841
Geosciences	$970
Mechanical Design	$54
Defense	$1,541
Government Lab	$3,267
University/Academic	$2,567
Weather	$674
Other	$257
Total Revenue	**$14,954**

Source: Hyperion Research, April 2024

The HPC Market Should Grow in 2024

AI and cloud spending are growing quickly
On-premises is expected to grow in 2024

- **2024 is forecasted to reach an all-time high of around $16.3 billion in on-prem HPC servers with $32.3 billion in total on-premises HPC spending**
- **But there are a number of issues:**
  - The overall economy is putting pressure on many buyers
  - The supply chain issues are getting more difficult (e.g., GPUs)
  - Exascale system acceptances are seeing delays
  - The lower end of the on-premises market continues to struggle
- **Growth drivers include:**
  - New use cases especially in AI/LLMs/Generative AI are providing many new areas for users to advance their research
  - Countries and companies around the world continue to recognize the value of being innovative and investing in R&D to advance society, grow revenues, reduce costs, and become more competitive
  - Cloud computing is becoming more useful to a larger set of HPC workloads

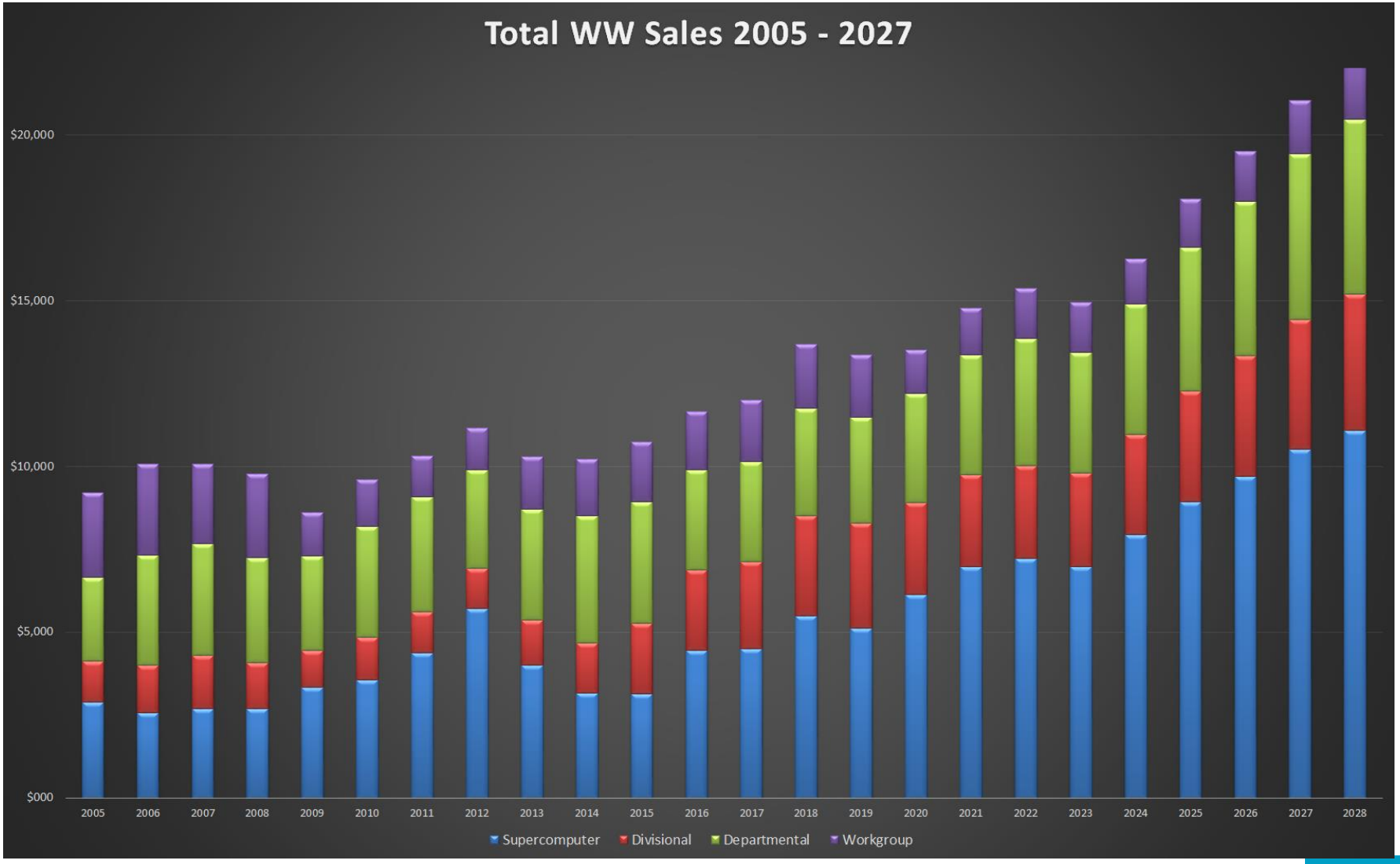
5-Year On-Prem HPC Server Forecast

8.2% yearly average growth over the next 5 years

Worldwide HPC On-Prem Forecast ($M)							
	2023	2024	2025	2026	2027	2028	CAGR 23-28
Supercomputer	$6,979	$7,934	$8,933	$9,698	$10,509	$11,070	9.7%
Divisional	$2,812	$3,028	$3,336	$3,638	$3,915	$4,126	8.0%
Departmental	$3,644	$3,941	$4,343	$4,648	$4,999	$5,269	7.7%
Workgroup	$1,518	$1,351	$1,447	$1,516	$1,601	$1,687	2.1%
Total	**$14,954**	**$16,254**	**$18,058**	**$19,501**	**$21,025**	**$22,152**	**8.2%**

Source: Hyperion Research, April 2024

On-prem Historic & Forecasted Revenues



The Broader On-Prem Market

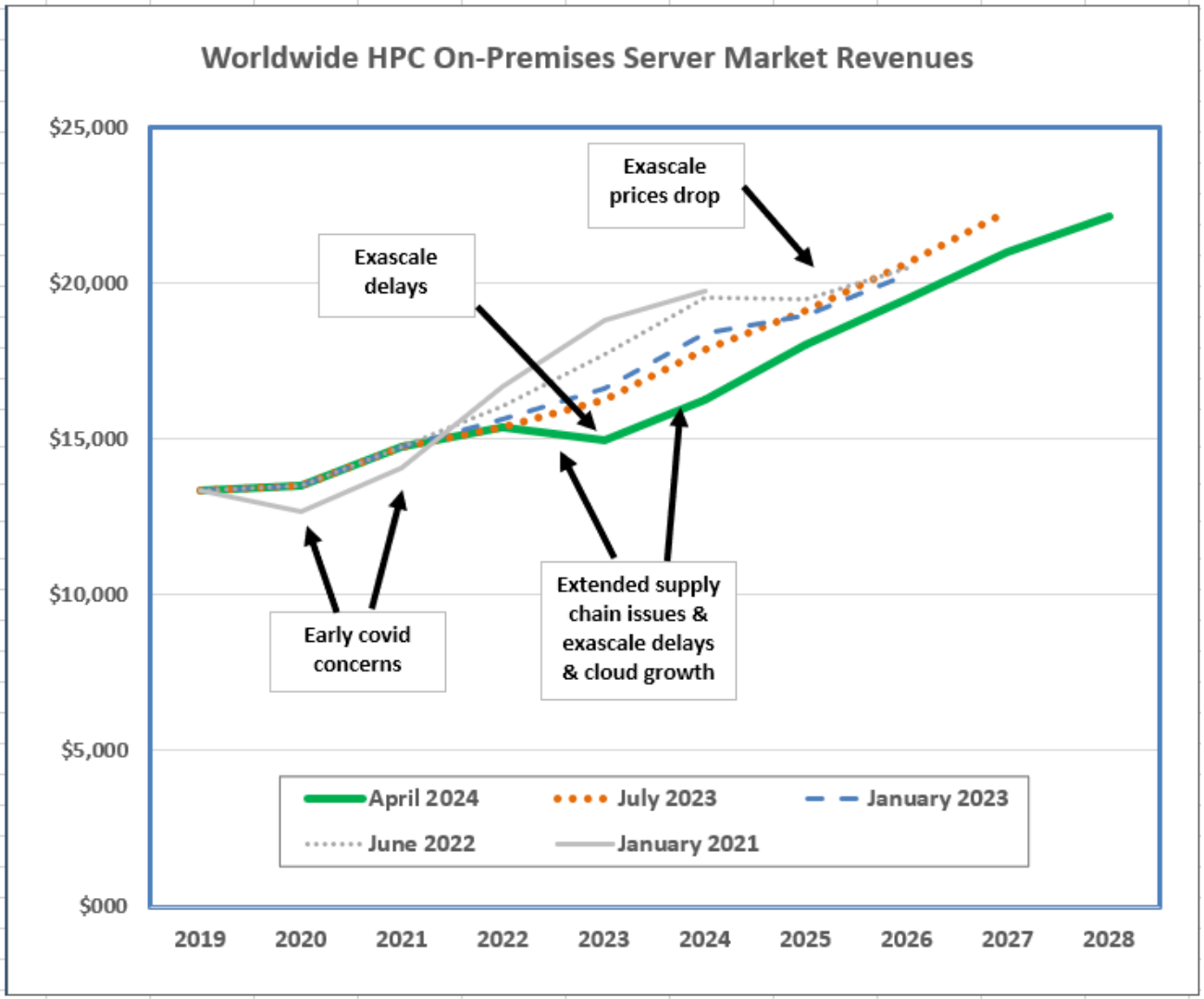
2023 total HPC spending reached $29.7 B
2028 is projected to reach $43.7 B

Revenues by the Broader HPC Market Areas ($M)							
	2023	2024	2025	2026	2027	2028	CAGR 23-28
Server	$14,954	$16,254	$18,058	$19,501	$21,025	$22,152	8.2%
Storage	$6,282	$6,972	$7,862	$8,552	$9,288	$9,786	9.3%
Middleware	$1,711	$1,836	$2,014	$2,160	$2,329	$2,454	7.5%
Applications	$4,830	$5,152	$5,634	$6,020	$6,457	$6,805	7.1%
Service	$2,014	$2,050	$2,167	$2,186	$2,334	$2,458	4.1%
Total Revenue	**$29,791**	**$32,265**	**$35,735**	**$38,420**	**$41,433**	**$43,656**	**7.9%**

Source: Hyperion Research, April 2024

5-year HPC Server Forecast Changes

Updating quarterly due to so many changes



High Growth Areas

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
Total	**12-13**	**14-19**	**5-6**	**8-12**	**7-10**	**47-61**	**$13.4B - $16.8B**

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

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

Source: Hyperion Research, March 2024

The HPC Cloud Market Will See Strong Growth in 2024

The growth will build on the fundamental changes in buying behavior seen in 2021 & 2022

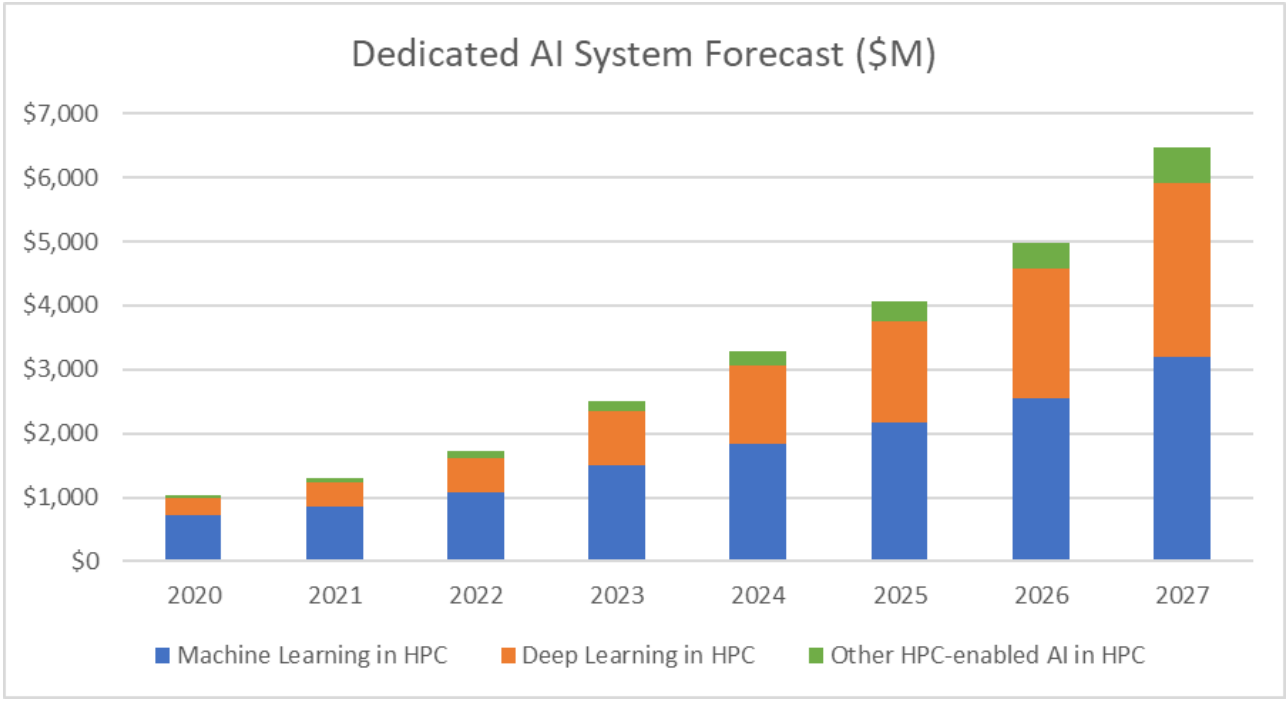
- **In 2021 HPC & AI buyers around the world revealed for the first time that HPC buyers are planning to shift some of their on-premises budgets to spending in the cloud**
- **End user spending on public cloud resources to run HPC workloads is projected to grow substantially at a rate of 18% over the next five years, and will reach US $12 billion in 2026**
  - 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
- **This major shift in buying behavior doesn't mean that on-premises HPC systems are going away**
  - The on-premises HPC server market is anticipated to exhibit healthy growth, 7%-8% a year, over the forecast period

AI Forecast

30.4% growth over the next 5 years

Worldwide HPC-Enabled AI Forecast (ML, DL & Other AI) Server Revenue ($M)									
	2020	2021	2022	2023	2024	2025	2026	2027	CAGR 22-27
ML in HPC	$719	$867	$1,081	$1,500	$1,841	$2,166	$2,553	$3,191	24.2%
DL in HPC	$263	$366	$532	$855	$1,216	$1,584	$2,016	$2,727	38.6%
Other AI in HPC	$57	$75	$104	$160	$230	$312	$401	$548	39.4%
Total HPC-Enabled AI Server Revenue	$1,039	$1,309	$1,718	$2,514	$3,286	$4,062	$4,970	$6,466	30.4%

Source: Hyperion Research, 2023



Hyperion Research's 2024 Predictions

1. Utilization of HPC resources in the cloud will experience accelerating growth as users augment their AI focus on training with inferencing
2. While “speeds and feeds” will continue to be important to buyers of storage systems, the primary value point and competitive advantage for data storage solutions will shift to the *"Data Platform"*
3. System vendors will struggle to a greater degree than hyperscalers in absorbing the accelerated cadence of NVIDIA’s GPU roadmap
4. Interest in procuring on-premises quantum computing (QC) systems will grow, augmenting but not replacing access to QC through the cloud
5. Installations of HPCs built by large end users, not traditional HPC suppliers, will become more commonplace, particularly for those targeting high-end AI computational workloads
6. RISC-V will continue to gain ground as a viable processor option
7. The cost of energy will more frequently outweigh HPC performance increases, as some sites will settle for “good enough” speeds in order to reach energy efficiency goals
8. The importance of monolithic HPC benchmarking will diminish
9. LLM frameworks will gain prominence within the overall trend of LLM usage
10. Generative AI growth will continue, but adoption growth rates will stabilize as applicability and limitations become more well understood
11. Arm-based processors will rise sharply in adoption, with expected revenues for Arm-based HPC systems doubling compared with the previous year

AI Training and Inferencing to Accelerate Utilization of Cloud HPC Resource

1. Utilization of HPC resources in the cloud will experience accelerating growth as users augment their AI focus from training to inferencing

- **Current Hyperion Research 5-year HPC cloud utilization CAGR of 18.1% does not yet fully incorporate all AI workflows**
- **Training models**
  - Often requires light-dimming amounts of computational resources
  - Approaching full utilization for extended lengths of time per training run (e.g., days or weeks)
  - Relatively small number of users and jobs
- **Inferencing**
  - Less computationally intensive
  - Less utilization per job over much shorter time periods (e.g., microseconds, milliseconds)
  - Orders of magnitudes greater # of users and jobs

Accelerated GPU Roadmap Presents System Vendor Roadmap Challenges

3. System vendors will struggle to a greater degree than hyperscalers in absorbing the accelerated cadence of NVIDIA's GPU roadmap

- **NVIDIA shifting its GPU roadmap cadence from 18 months to 12 months may allow nimble vendors and service providers to deliver higher performing solutions to users more frequently**
  - Users and vendors will be looking for successful demonstration over several release cycles of this capability
- **12-month GPU cadence also may introduce some stresses to the ecosystem**
  - Increased competitive pressure on other GPU providers
  - System vendors will need to adapt their budgets, priorities, planning, purchase cycles, and processes to absorb and integrate new technologies more frequently
  - Users who recently have been extending the lifetimes of their on-premises HPC systems from 4 years to 5 or 5.5 years could fall 3 or 4 generations behind leading edge innovations

Increasing Instances of Large End Users Building Own HPC/AI Systems

5. Installations of HPCs built by large end users, not traditional HPC suppliers, will become more commonplace, particularly for those targeting high-end AI computational workloads

- **Development of HPCs by HPC end users and CSPs will likely increase, driven by growing demands for workload specific HPC architecture, particularly those suited to the unique, distinct, and computationally intensive generative AI workloads**
- **Recent examples:**
  - Google
  - Microsoft (Eagle - #3 on Nov 2023 Top500 list)
  - NVIDIA (EOS, #9 on Nov 2023 Top500 list)
  - Tesla
  - Large Chinese sites
- **Equally important is that some of these new HPCs are built using internally developed processors (e.g., Google's TPU v5P) that target both AI as well as traditional scientific and engineering workloads**

Processors Based On RISC-V Will Grow

6. RISC-V will continue to gain ground as a viable processor option

- **While RISC-V offers some key technology advantages, there are also non-technical considerations that are propelling its adoption**
  - Shifting global policy needs
  - Increasingly stringent trade laws
  - Regional pushes for indigenously sourced technology
- **While proprietary tools are still the industry go-to in this regard, the drive to develop technology in a legally insular environment gives RISC-V an advantage moving forward**
- **Rapid uptake of the RISC-V ISA in the Chinese server community is due, in no small part, to US export controls on counterpart x86 technology**

Energy Costs to Surpass Maximum Performance in System Tradeoffs

7. The cost of energy will more frequently outweigh HPC performance increases as some sites will settle for "good enough" speeds in order to reach energy efficiency goals

- **Many HPC sites already find themselves constrained by the escalating costs of energy consumption due to increasing power requirements of highest performing computational processors**
  - Large AI training models are expected to grow at a rate that will require major increases in power
- **Attention will turn towards a pragmatic tradeoff and balance between affordability and performance**
- **Focus on optimizing performance per unit of energy consumed particularly acute in regions of high energy costs, such as in Europe and Asia**

Conclusions

- **2023 was a down year for on-prem servers**
  - With a 2.7% decline
  - Supply chain issues and exascale acceptance delays along with a greater use of clouds
  - 2024 is expected to be a moderate growth year
  - GPUs, cloud, AI/ML/DL/LLM are high growth areas
- **New technologies are showing up large numbers:**
  - Generative AI and LLMs is fueling a new level of growth
  - Processors, AI hardware & software, memories, new storage approaches, etc.
  - The cloud has become a viable option for many HPC workloads
- **Storage will likely see major growth driven by AI, big data and the need for much larger data sets**
- **There are still concern about the supply chain and growing concerns around power & talent**
- **Diversity in HPC needs to be addressed**

A Concern: HPC Expertise Shortage

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

<u>Two major trends</u>:

1) **A shrinking HPC workforce**
2) **A massive increase in system complexity**

- **HPC experts are an aging workforce**
  - The pipeline of new HPC staff entering the workforce does not adequately match the outflow of retirees
  - Competition for HPC staff will intensify
- **Increasingly complex workloads are more difficult to manage**
  - Augmenting traditional modeling/simulation with AI and big data
  - Incorporating multiple processor types, co-processors, accelerators, and other specialized hardware
  - Balancing on-prem and cloud
  - Enterprise IT users are entering HPC space, and need HPC expertise

- **HPC users will need major improvements in ease-of-use, ease-of-selection, & ease-of-optimization**

HYPERION RESEARCH

Thanks for joining us today!

We welcome questions, comments and suggestions

Please contact us at:
info@hyperionres.com

Next On The Agenda

8:00 – 8:30 am **Welcome and Meeting Overview**

- Welcome & Logistics, Piyush Mehrotra and Earl Joseph
- HPC Market Update, Earl Joseph

(Session Leader: Nick Wright, LBL)

8:30 – 10:45 am **Government Programs Supporting HPC, AI and QC**

- 8:30 – 9:00 am US DOE Office of Science New Programs and Updates, Ceren Susut-Bennett, Department of Energy Office of Science
- 9:00 – 9:30 am Democratizing Access to AI resources through the National AI Research Resource (NAIRR) Pilot, Katie Antypas, NSF
- 9:30 – 10:00 am Chips and Science Program Updates, Stephen Ezell, ITIF
- *Break: 10:00 – 10:15 am*
- 10:15 – 10:45 am New Activities at Sandia National Laboratory, Siva Rajamanickam, SNL

10:45 – 11:15 am **The Making of Data Center Alley – How Loudoun Virginia became the Heart of the Internet,** Buddy Rizer, Executive Director, Dept of Economic Development for Loudoun County, Virginia

11:15 – 11:30 am **Vendor Update, The Need for a New Storage Architecture: Hyperscale NAS**, Rob Renzoni, Hammerspace

11:30 – 12:00 pm **New Research on Quantum Computing,** Bob Sorensen, Hyperion Research

12:00 – 1:00 pm **Lunch**