



HYPERION RESEARCH

Hyperion Research ISC24 Market Update

ISC24

www.HyperionResearch.com
www.hpcuserforum.com

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

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

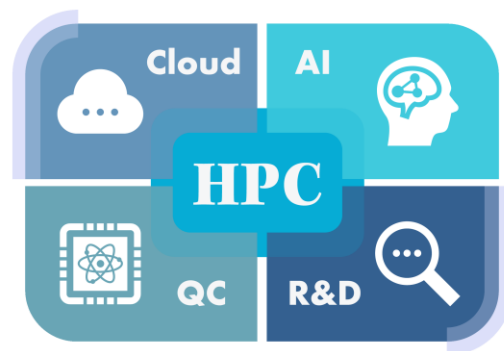
Jie Wu, China and Technology Trends

Mara Jacob, HPC User Forum Support

Major 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**
- **Analyzing All Processor Types**
- **Quantum Computing**
- **R&D and Engineering -- all types**
- **Edge Computing**
- **Supply Chain Issues**
- **Sustainability**



Today's Agenda

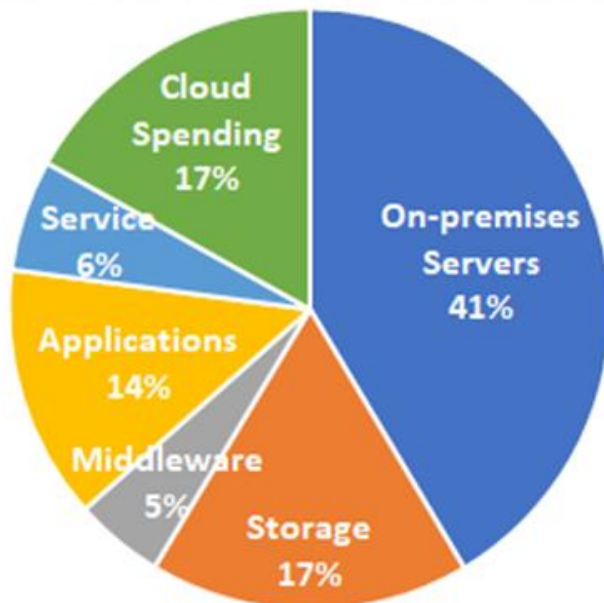
- **HPC Market Update**
- **Navigating the New World of HPC, AI, Quantum and Clouds**
- **Considerations for Storage and Interconnects**
- **Global QC Market: Robust and on the Rise**
- **Considerations for Sustainability**
- **Widespread AI Adoption Creates New Demands**
- **Forecast and High Growth Areas to Keep an Eye On**
- **Considerations for Clouds**
- **Conclusions**

HPC Market Update

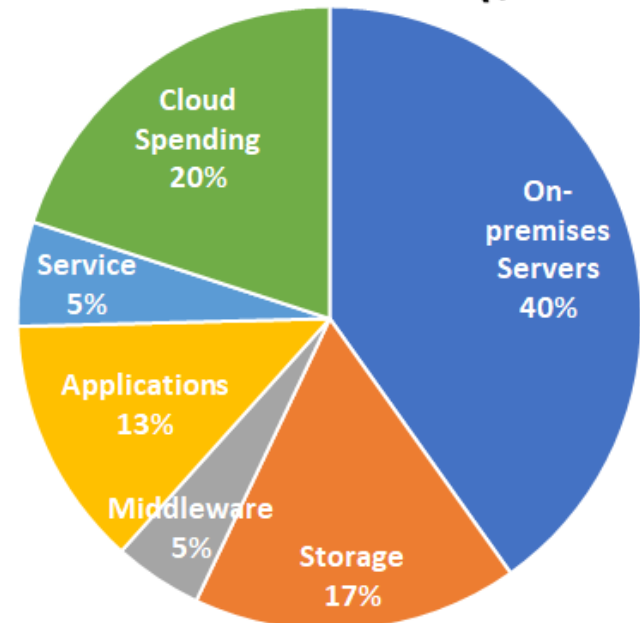
The Overall HPC Market in 2023

*Looking at the overall HPC market, including servers, cloud usage, storage, software and repair services 2023 = \$37.2 billion USD
(2022 was \$37.3 billion)*

The Total HPC Market in 2022

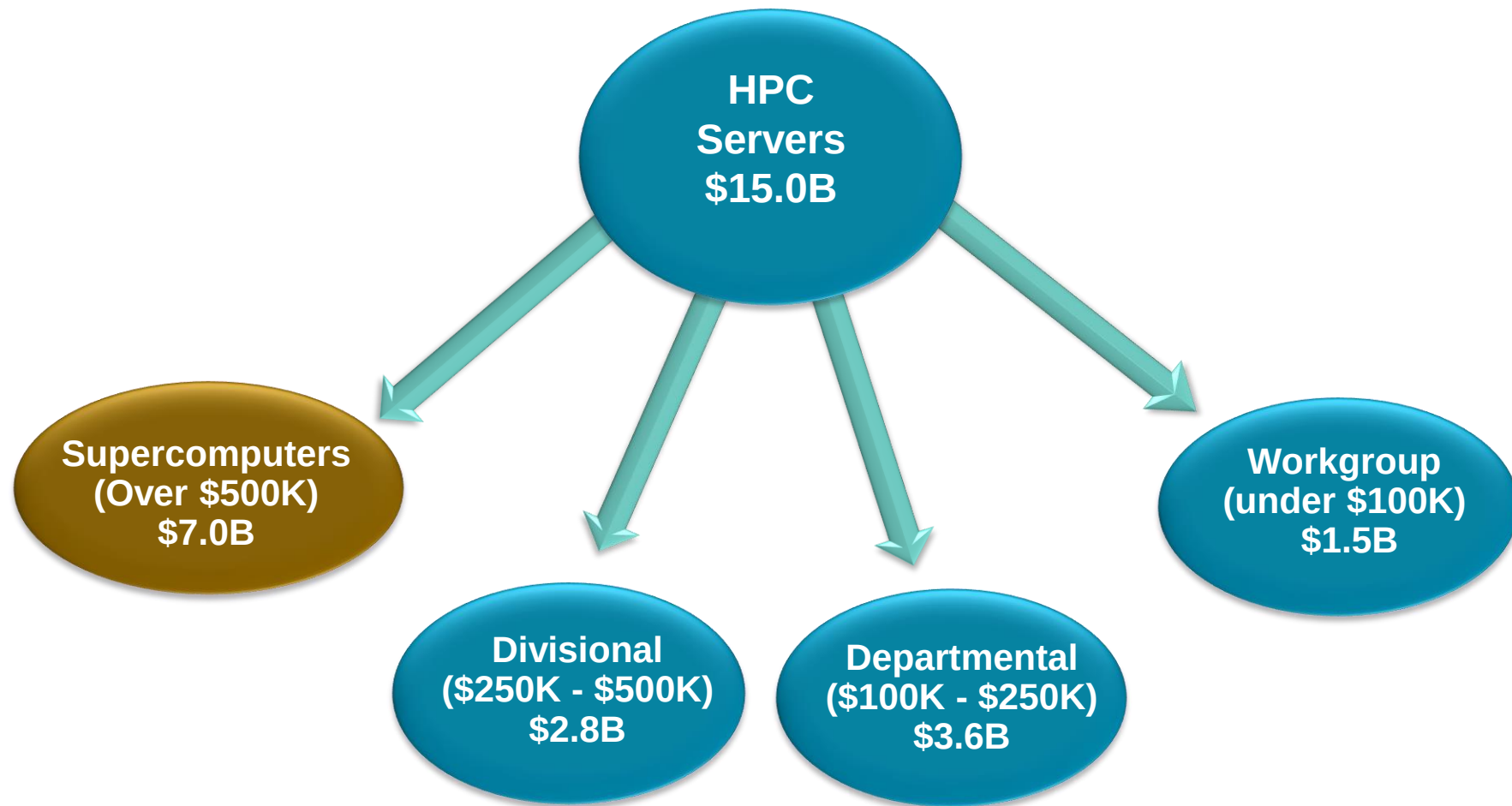


HPC Market in 2023 (\$37.2B)



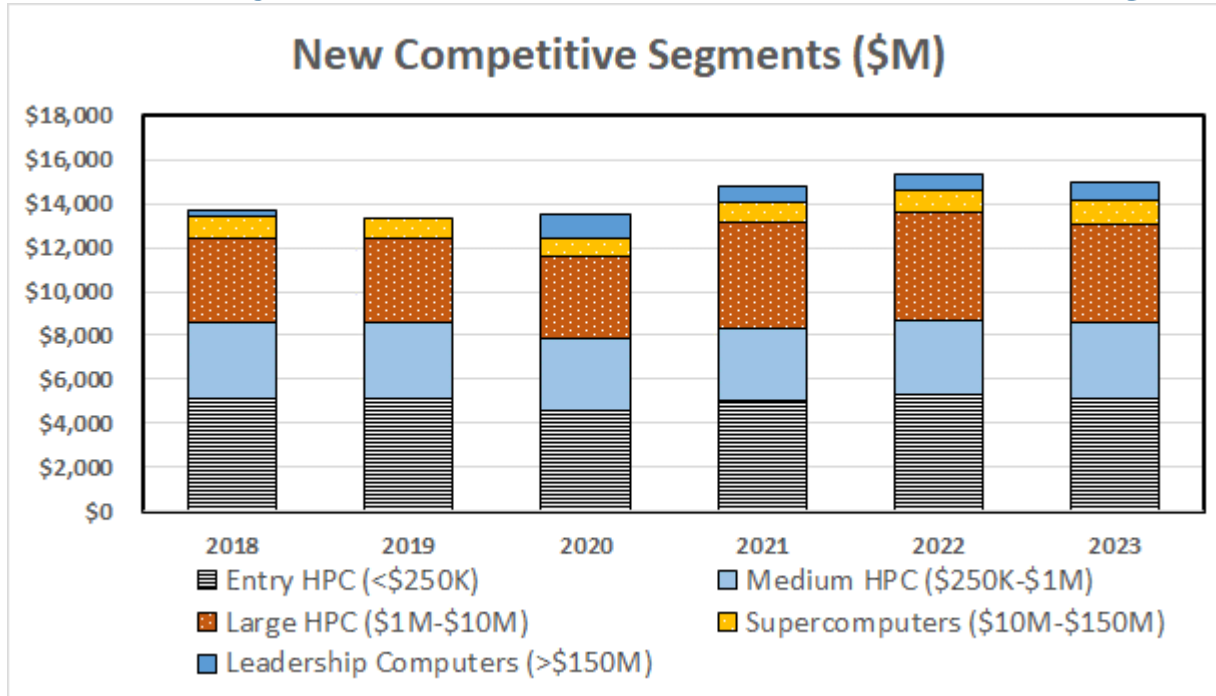
The 2023 Worldwide On-Prem HPC Server Market: \$15.0 Billion (down 2.7%)

2024 is projected to be around \$16.3 Billion



New HPC/AI On-Prem Subsegments

The market for systems under \$10 million is large



2024 New Competitive Segments							
	2018	2019	2020	2021	2022	2023	CAGR 21-26
Leadership Computers (>\$150M)	\$280		\$1,065	\$760	\$750	\$826	24.2%
Supercomputers (\$10M-\$150M)	\$980	\$946	\$862	\$852	\$1,056	\$1,077	1.9%
Large HPC (\$1M-\$10M)	\$3,827	\$3,798	\$3,762	\$4,852	\$4,861	\$4,505	3.3%
Medium HPC (\$250K-\$1M)	\$3,427	\$3,538	\$3,209	\$3,291	\$3,357	\$3,383	-0.3%
Entry HPC (<\$250K)	\$5,166	\$5,087	\$4,622	\$5,026	\$5,345	\$5,163	0.0%
Total	\$13,679	\$13,368	\$13,519	\$14,781	\$15,369	\$14,954	1.8%

Source: Hyperion Research, 2024

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 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 buyers around the world revealed plans 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 over 21% over the next five years**
 - This strong growth reflects the heavy work that the cloud service providers (CSPs) have done to make clouds more HPC friendly
- **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 around 8% growth a year, over the forecast period

The Overall HPC Market Should See Strong Grow in 2024

AI and cloud spending are growing quickly

On-premises servers are expected to grow in 2024

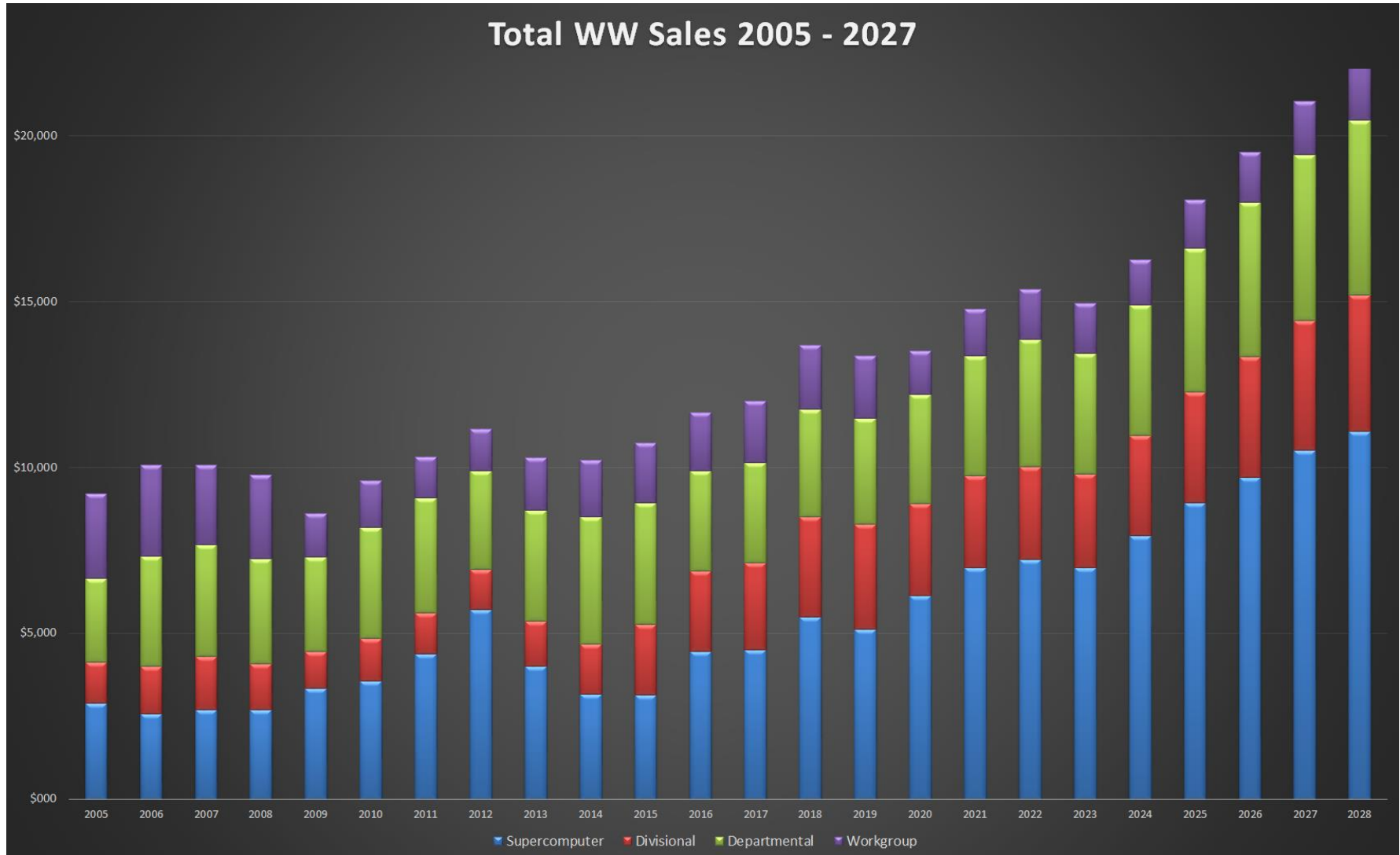
- **2024 is forecasted to reach around \$16.3 billion in on-prem HPC servers with \$32.3 billion in total on-premises HPC spending**
- **But there are a number of issues:**
 - The supply chain issues are getting more difficult (e.g., GPUs)
 - Exascale system acceptances are seeing delays
 - The lower 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 Billion
2028 is projected to reach \$43.7 Billion

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							

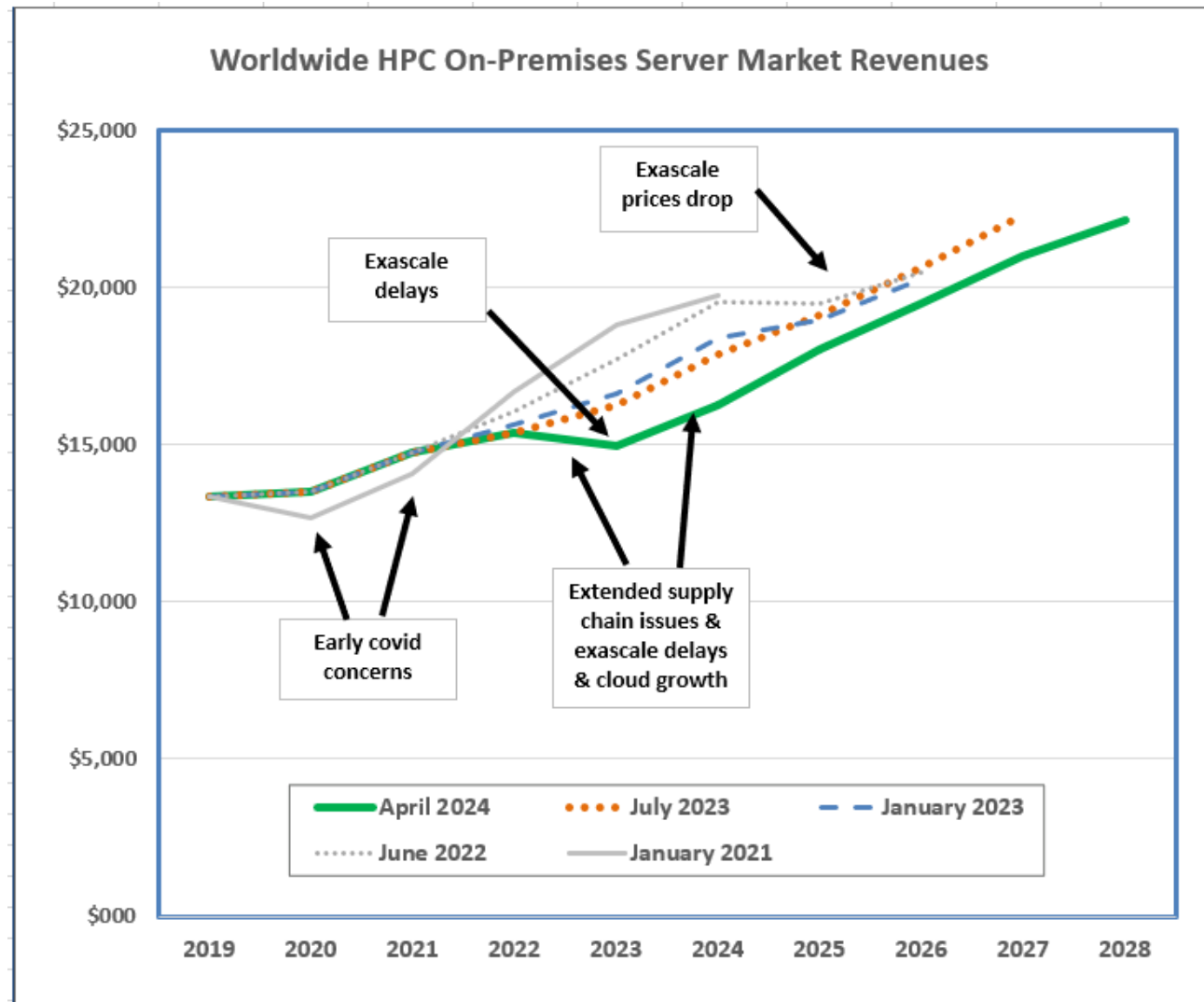
The Exascale Market (System Acceptances)

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							

5-year HPC Server Forecast Changes

Updating quarterly due to so many changes



High Growth Areas

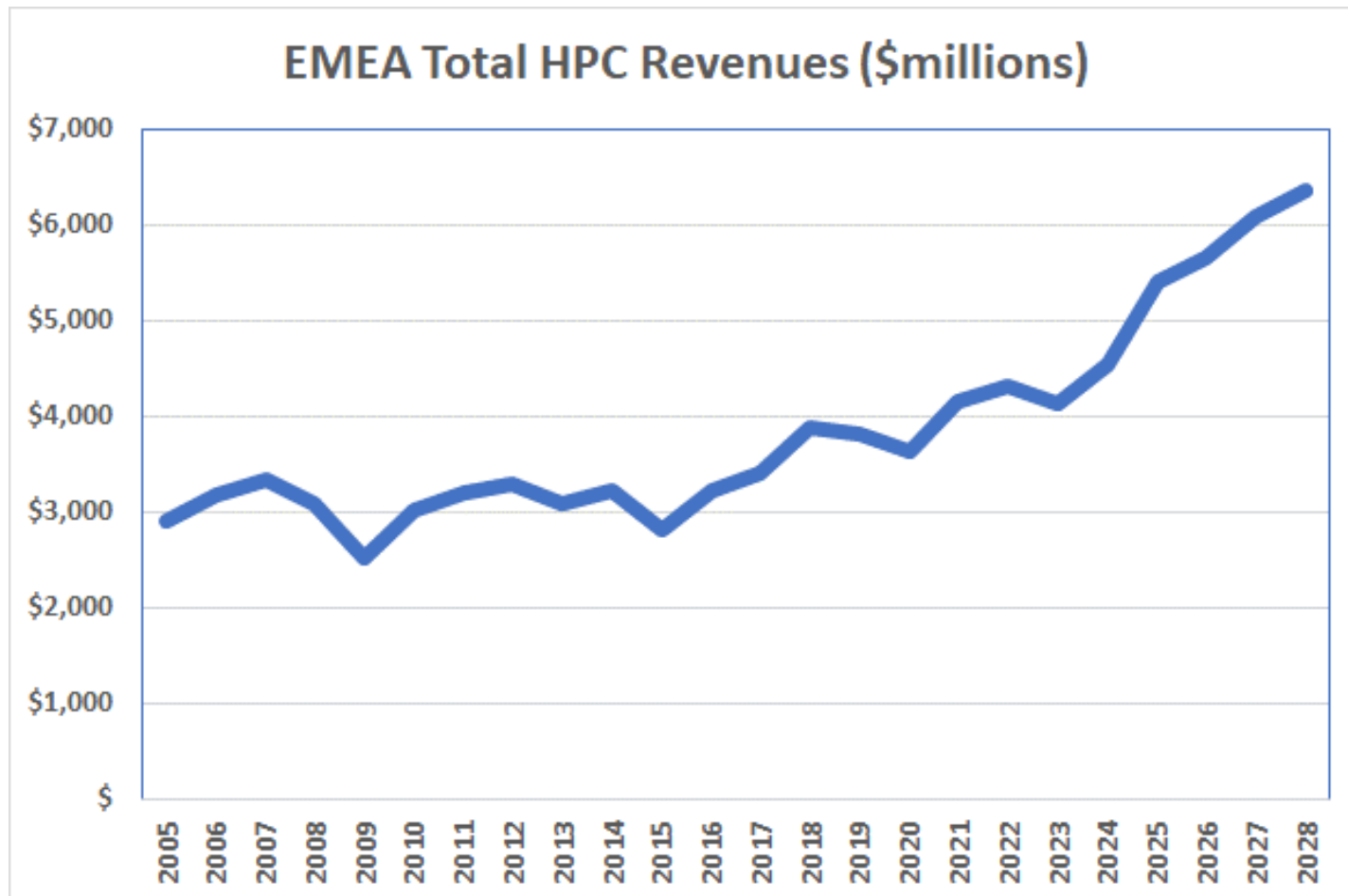
Five-year growth rates vary from 8% to over 30%

- **On-premises HPC servers: ~8% CAGR**
- **Storage: Over 9% CAGR**
- **GPUs: Over 17% CAGR**
- **Running HPC/AI workloads in the cloud: Over 21% CAGR**
- **On-premises AI-focused servers: Over 30% CAGR**

HPC In Europe

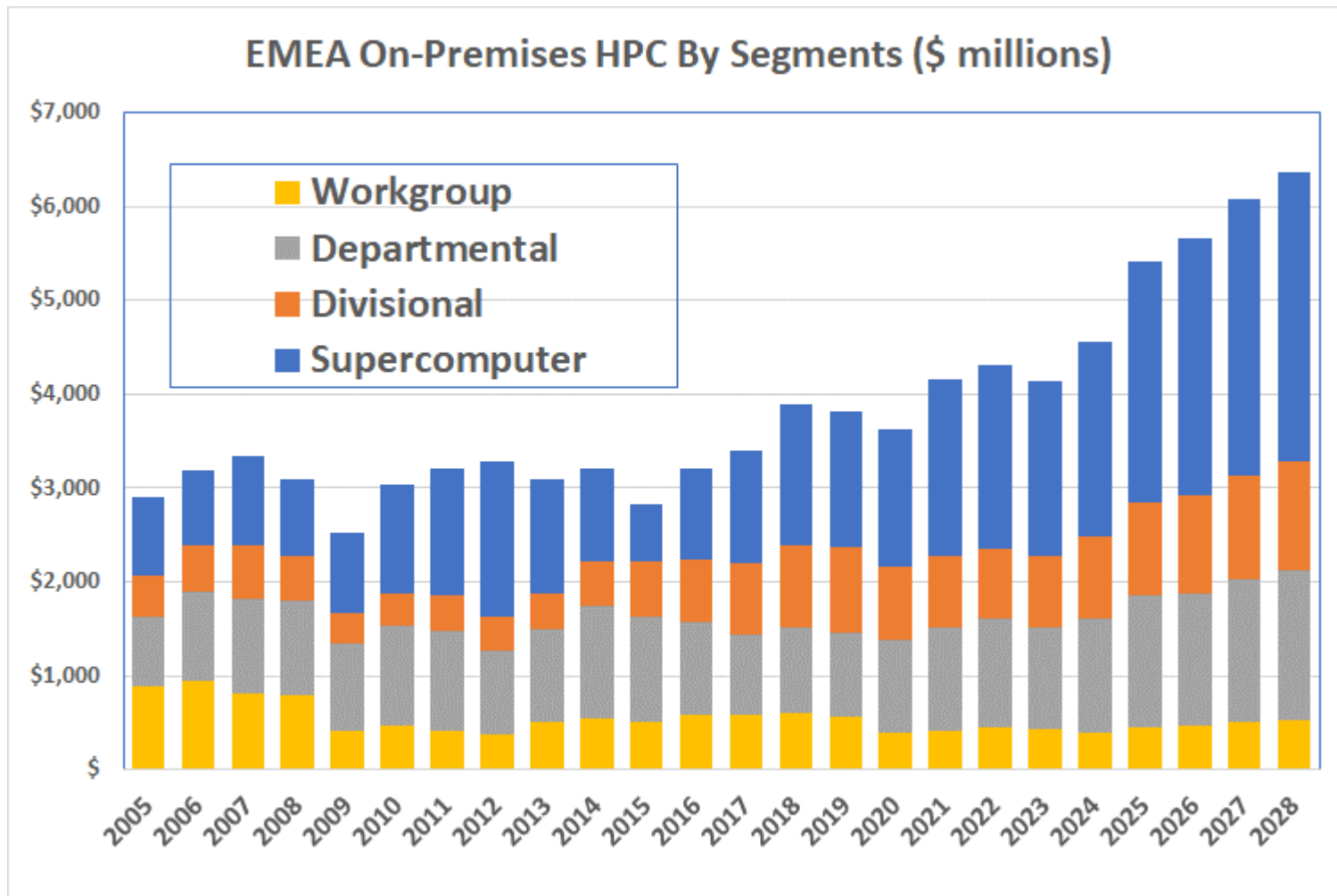
On-Prem HPC In Europe

Healthy growth since 2015



On-Prem HPC In Europe

Strong growth in supercomputers



2024 Predictions

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 Resources

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

- **Hyperion Research's 5-year HPC cloud utilization CAGR of 20% 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**

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**

We Are Announcing An Expanded AI Program And Focus

Building On Ten Plus Years Of Tracking AI

With the changing dynamics, we are adding a more real-time analysis of the issues, what's working & what users are looking for

Adding:

- **Identifying the current key issues, drivers & inhibitors**
 - In hardware, architecture, software, LLMs, other models, new applications, best practices, emerging trends,...
 - Analyzing what is actually being done with AI at user sites
 - Successful use cases in AI & areas that didn't work as expected
 - New opportunities that AI provides to end users
 - Different ways of adding AI to critical workflows
- **Creating an AI Experts Advisory Committee**
 - To drive research on the most critical topics
 - To help direct our reports and study topics
 - Conducting studies for the community (in addition to private studies)
 - **Let us know if you would like to be part of the committee!**

AI Research Agenda

Looking at all aspects of advanced AI computing

- **Hardware:**
 - Analyzing and evaluating the fit of the different processors & GPUs for different computationally intensive AI workloads
- **Architecture:**
 - Pro's and con's of different designs
 - Where each approach fits best by application type, job sizes, data types and the trade-offs
 - Where clouds fit well
- **Software:**
 - Analyzing and evaluating the end user software environment, software infrastructure, applications, tools, foundational models, ...
- **Emerging Trends:**
 - Tracking and evaluating leading edge use cases, new models, new approaches, ...

**Join Us For Our
2024 HPC User Forums
And Hyperion Research Meetings**
(www.hpcuserforum.com)

- **September 4 & 5: Argonne National Laboratory, Lemont, IL**
- **October 21-25: HPC User Forums in Europe**
 - **Oct. 21 & 22 at BSC, Barcelona Spain**
 - **Oct. 24 & 25 at HLRS, Stuttgart Germany**
- **November: SC24 Breakfast Briefing**

Next On The Agenda

- **HPC Market Update**
- **Navigating the New World of HPC, AI, Quantum and Clouds**
- **Considerations for Storage and Interconnects**
- **Global QC Market: Robust and on the Rise**
- **Considerations for Sustainability**
- **Widespread AI Adoption Creates New Demands**
- **Forecast and High Growth Areas to Keep an Eye On**
- **Considerations for Clouds**
- **Conclusions**