Hyperion Research ISC24 Market Update

ISC24

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

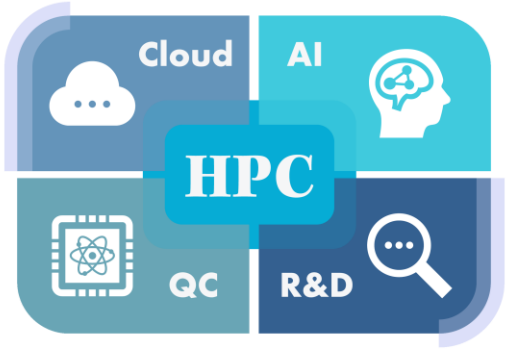
Consultants

- 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

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

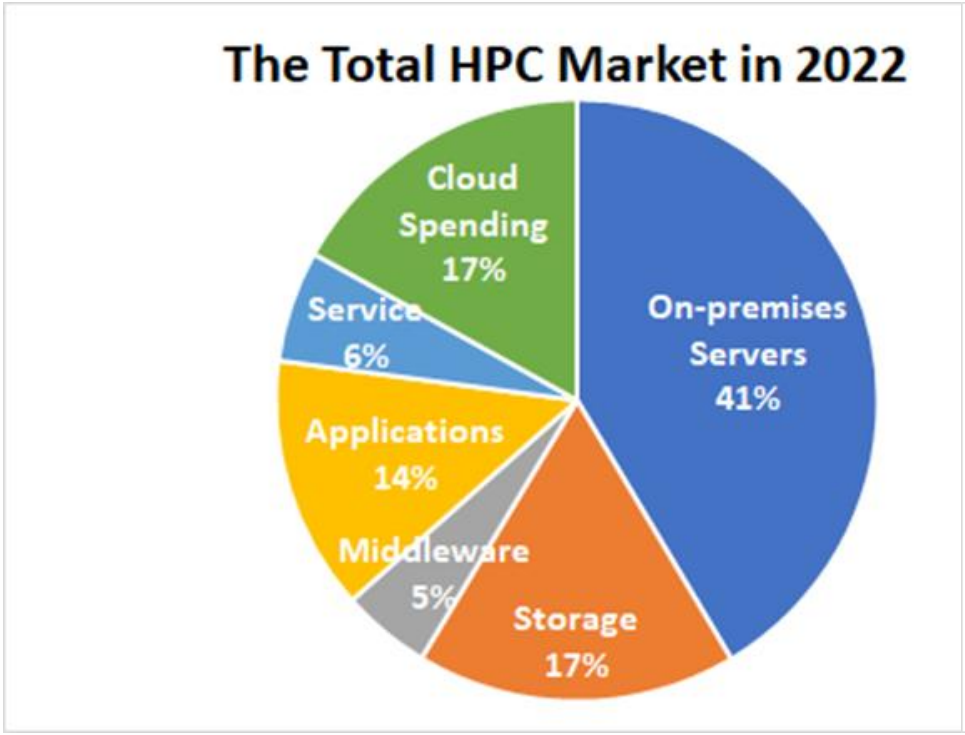
- HPC Market Update
- Navigating the New World of HPC, AI, Quantum and Clouds
- Considerations for Storage and Interconnects
- Global QC Market: Robust and on the Rise
- Considerations for Sustainability
- Widespread AI Adoption Creates New Demands
- Forecast and High Growth Areas to Keep an Eye On
-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

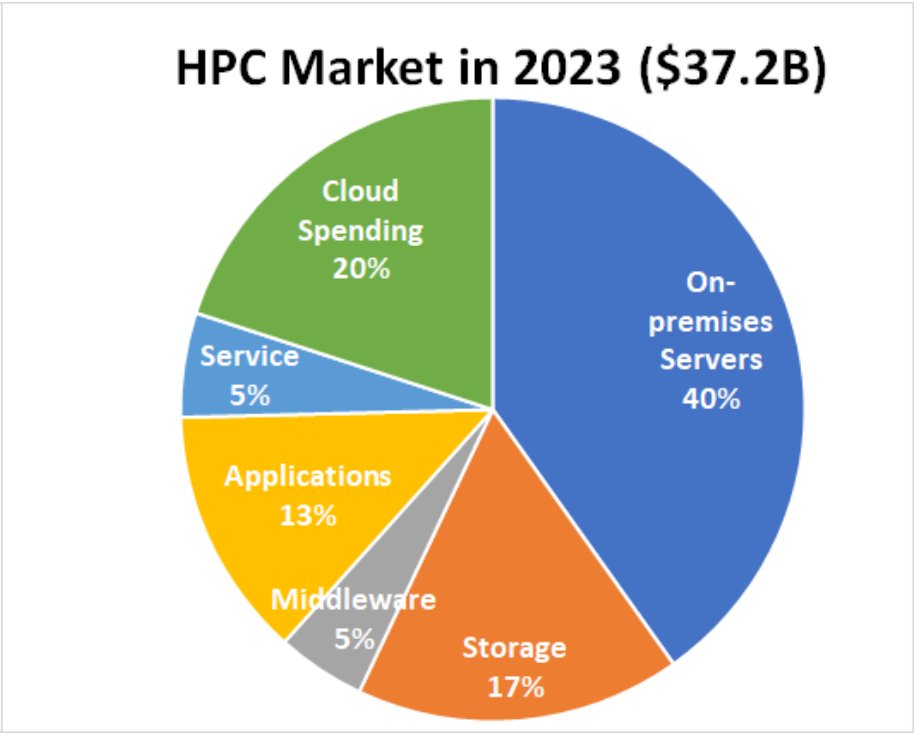
HYPERION RESEARCH

HPC Market Update

The Overall HPC Market in 2023

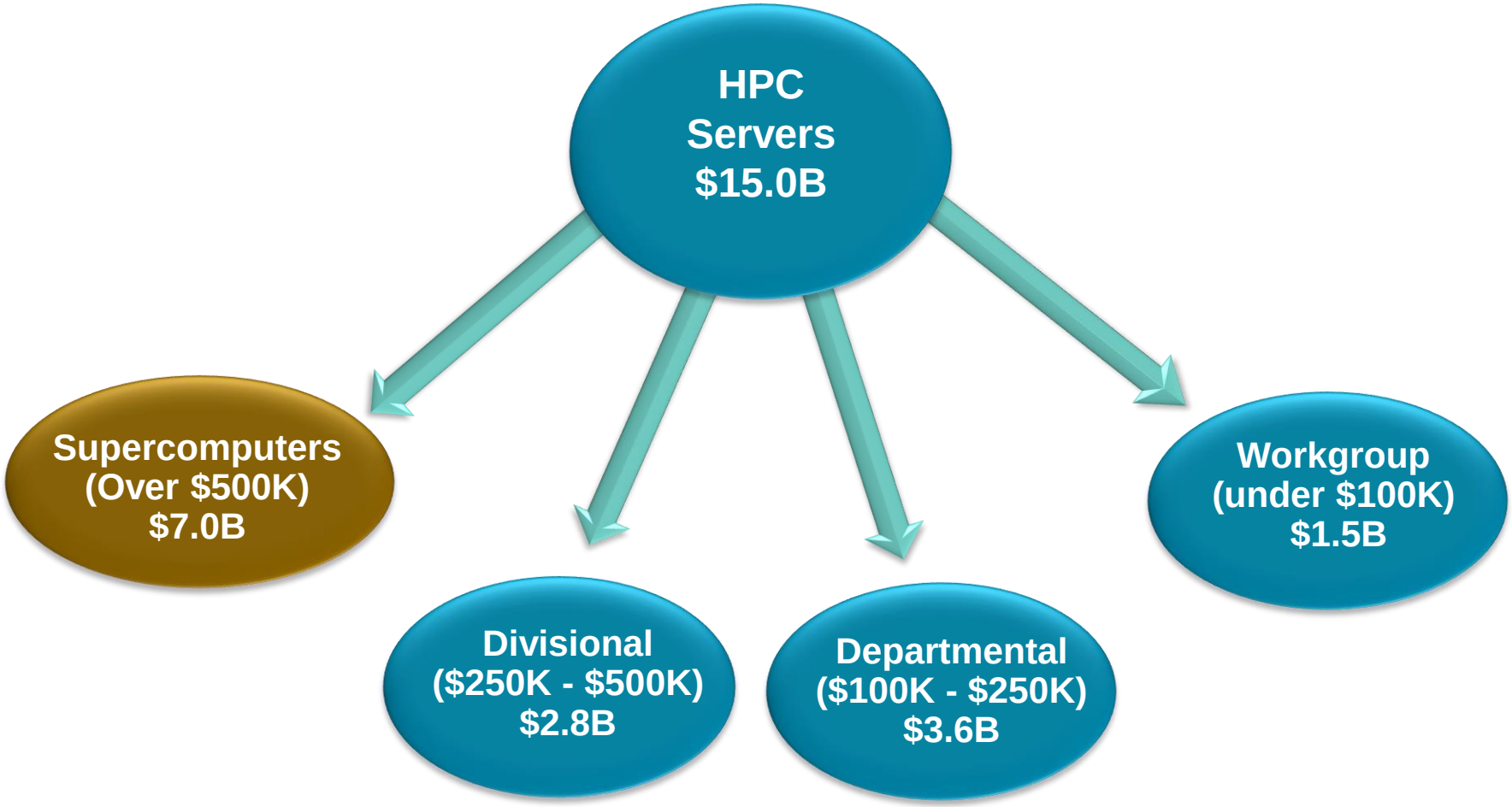
Looking at the overall HPC market, including servers, cloud usage, storage, software and repair services 2023 = $37.2 billion USD (2022 was $37.3 billion)





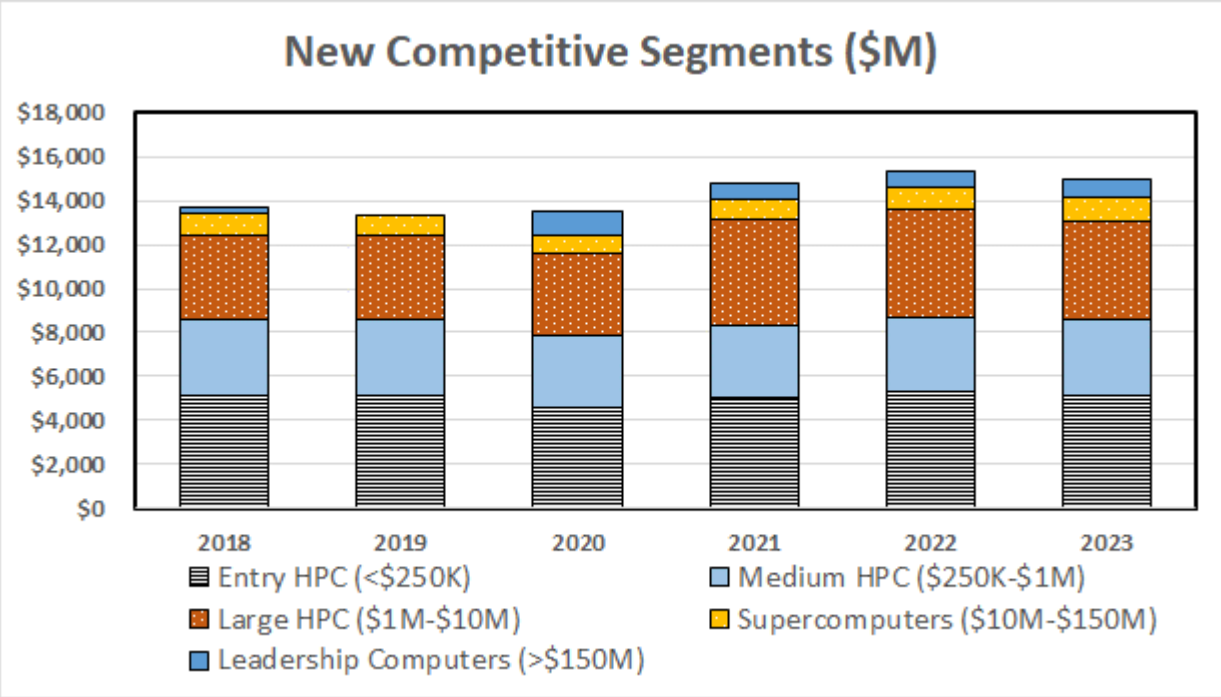
The 2023 Worldwide On-Prem HPC Server Market: $15.0 Billion (down 2.7%)

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

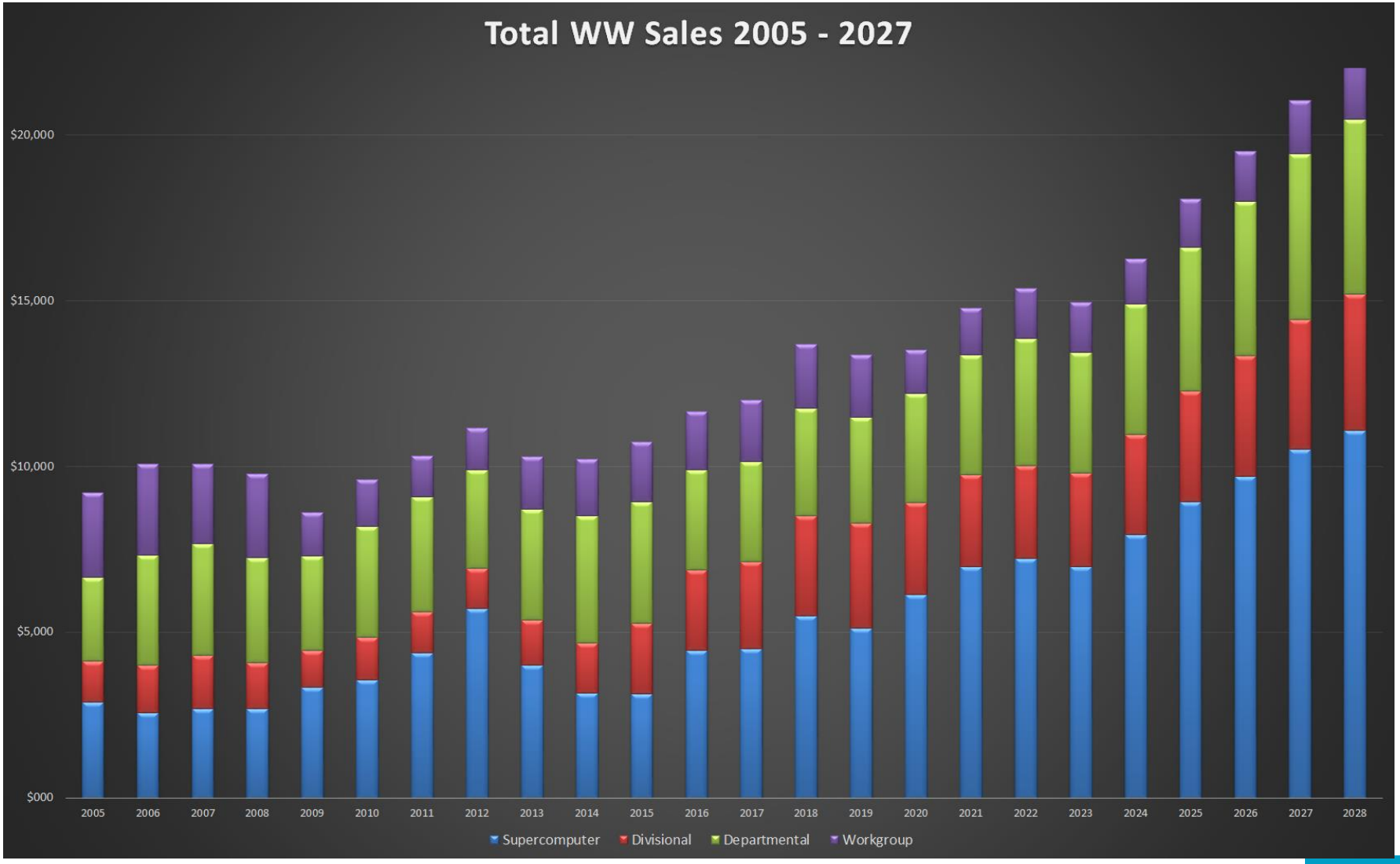
5-Year On-Prem HPC Server Forecast

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**

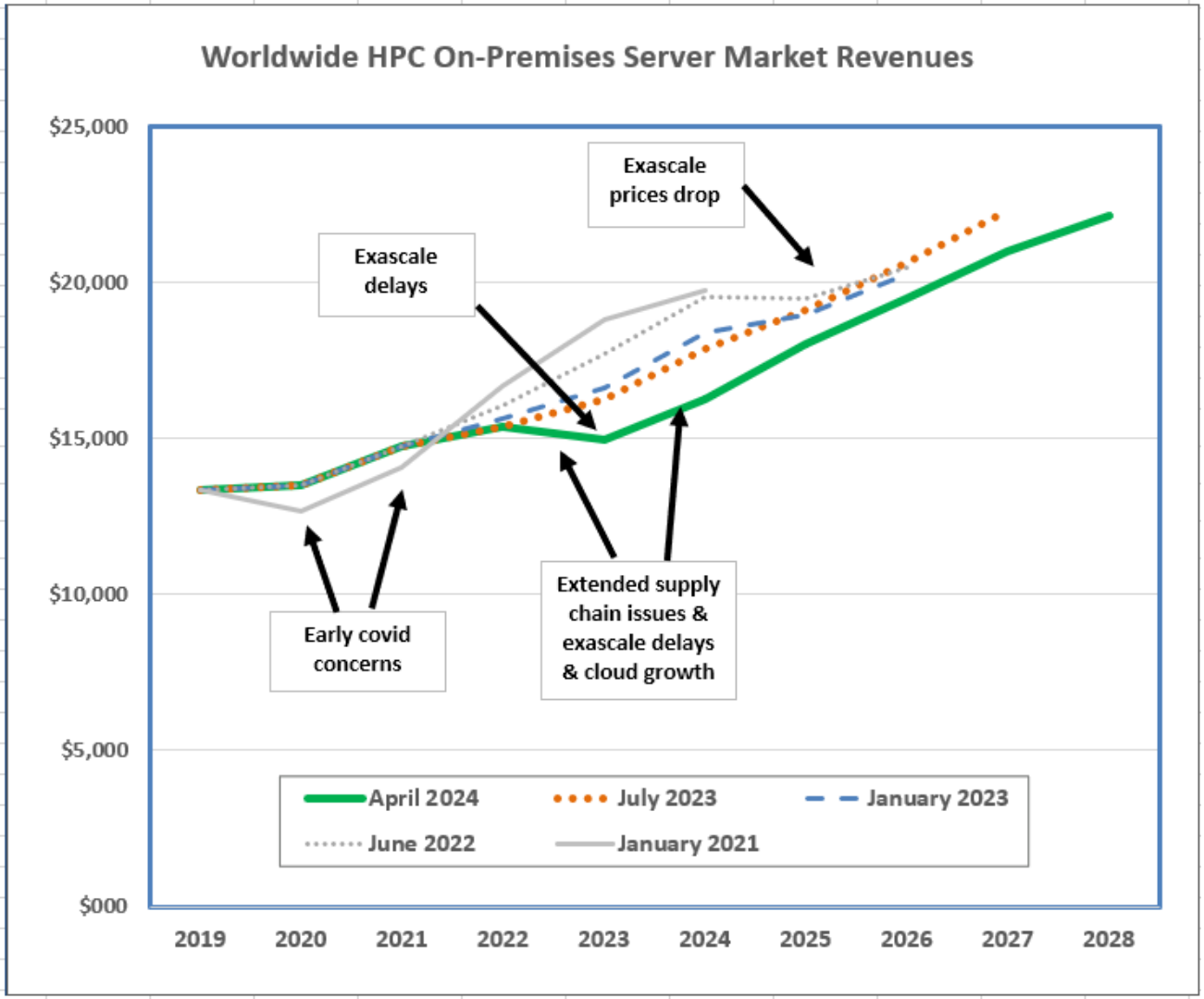
* Includes S. Korea, Singapore, Australia, Russia, Canada, India, Israel, Saudi Arabia, etc.

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%

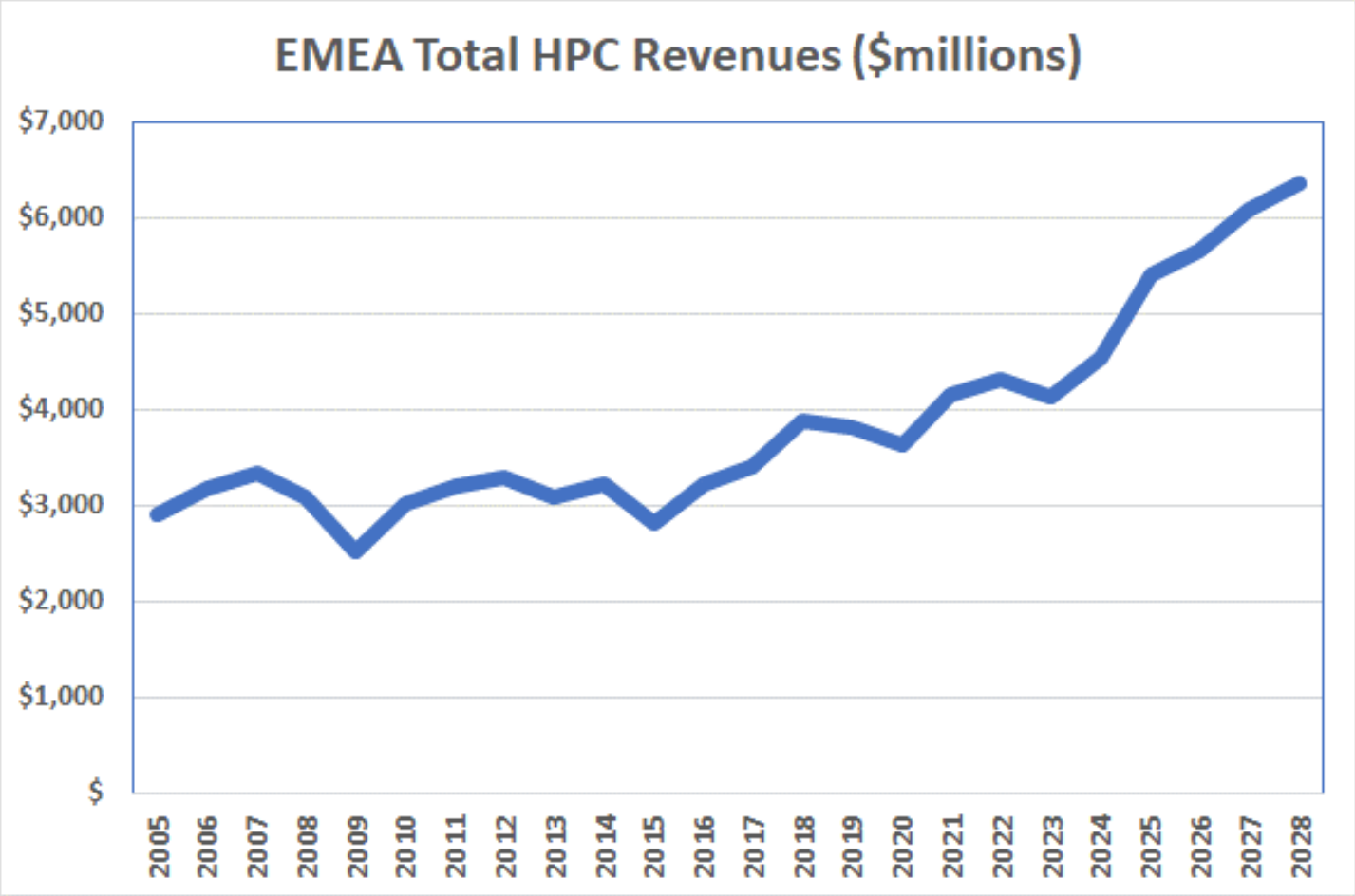
- **On-premises HPC servers: ~8% CAGR**
- **Storage: Over 9% CAGR**
- **GPUs: Over 17% CAGR**
- **Running HPC/AI workloads in the cloud: Over 21% CAGR**
- **On-premises AI-focused servers: Over 30% CAGR**

HYPERION RESEARCH

HPC In Europe

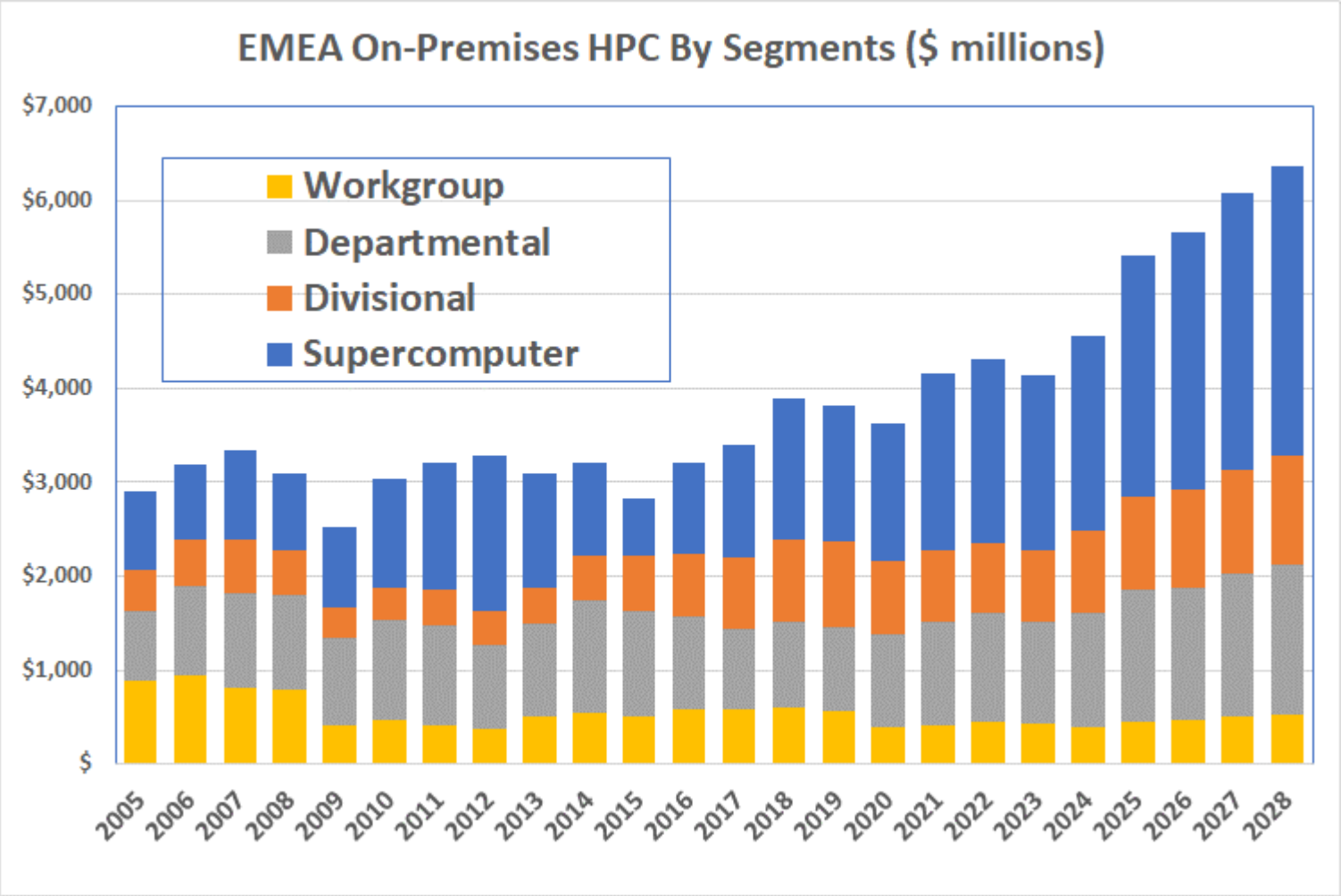
On-Prem HPC In Europe

Healthy growth since 2015



On-Prem HPC In Europe

Strong growth in supercomputers



HYPERION RESEARCH

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**

HYPERION RESEARCH

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, …

HYPERION RESEARCH

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

Navigating the New World of HPC...and AI...and Quantum...and Clouds...

Bob Sorensen
Senior Vice President of Research
Hyperion Research, LLC

The HPC Landscape is Undergoing Significant Changes (Again)

- Multiple developments are significantly altering the overall HPC landscape
  - Applications in machine learning and deep learning
    - Generative AI is a near-term game changer
  - Computing at the edge
    - Supporting real-time decision making
  - CSP's HPC-related impact
  - End user decisions for on-premises vs. cloud vs. hybrid
  - Dark HPCs on the horizon
  - Quantum computing making strides
  - Traditional modeling and simulation remain crucial

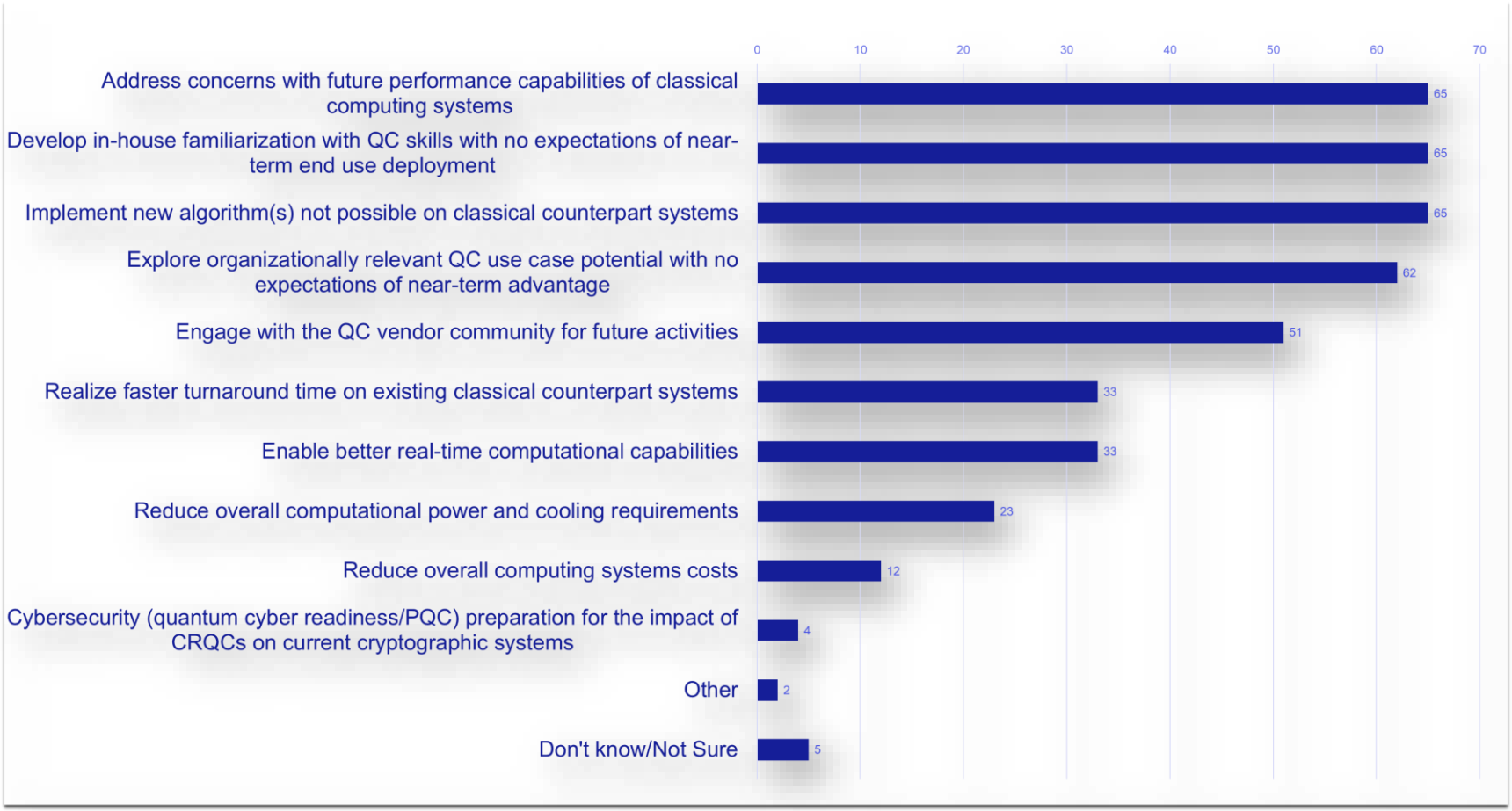
One Size No Longer Fits All

- HPC systems are becoming more complex due to the diverse workloads they are required to address
  - Variety in workloads demand variety in architecture
  - Accurately charactering workloads in an imperative
  - Simple benchmarking is no longer enough
    - Top 500: Perhaps even harmful to the sector?
- Happily, there are technical options to address this evolving environment
  - On-prem trends: moving away from monolithic architectures towards a more flexible, heterogenous design and multiple hardware partitions (or multiple systems)
  - Cloud-based options: more image types, bare metal options, CSP-HPCaaS, experimental/novel hardware access, containers, etc.
- Sampling of options for underlying computational engines
    - Processors: x86, ARM, RISC-V
    - Superchips: Nvidia Grace Hopper, AMD MI300
    - Accelerators: Nvidia Hopper, AMD MI300, and Gaudi, Graphcore, Cerebras, SambaNova, Cambracon, Groq

Different Workloads/Different Solutions

HPC Use Case Characteristics/Use Cases	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, both virtual and bare metal options
Data Storage	Consistent, uniform storage formats, typically with large file size and consistent storage access patterns	Varying data storage 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	Supports multiple instances of varying storage configurations. Can be 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, large data input, matrix operation intensive	Wide job specifics addressable by both virtual and bare metal options
Exemplar Software	C++, Fortran, MPI	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

QC 2026: QC End User Perceptions



Summary of HPC-related LLM-related Activities

	Currently	Next 12-18 months	Change Over Time
Exploring the range of potential performance enhancements by integrating LLMs into existing HPC-based workloads	58%	48%	-10%
Exploring in-house requirements for integrating LLMs into HPC-based workloads	55%	51%	-4%
Testing/assessing LLM-integrated workload performance	34%	45%	11%
Procuring access to necessary LLM software	31%	31%	0%
Reaching out to LLM hardware and software suppliers for information	30%	35%	5%
Passively monitoring LLM technology developments	27%	14%	-13%
Procuring access to necessary LLM hardware	26%	28%	2%
Standing up limited LLM-integrated pilot programs	26%	36%	10%
Porting LLM capability into existing workloads	25%	34%	9%
Running production level LLM-enabled workloads	22%	50%	28%
Standing up a fully funded LLM research efforts	17%	27%	10%
No current activity	1%	0%	-1%
Other	1%	0%	-1%

Thanks

- Questions and comments are most welcome and can be sent to bsorensen@hyperionres.com

Considerations for Storage and Interconnects

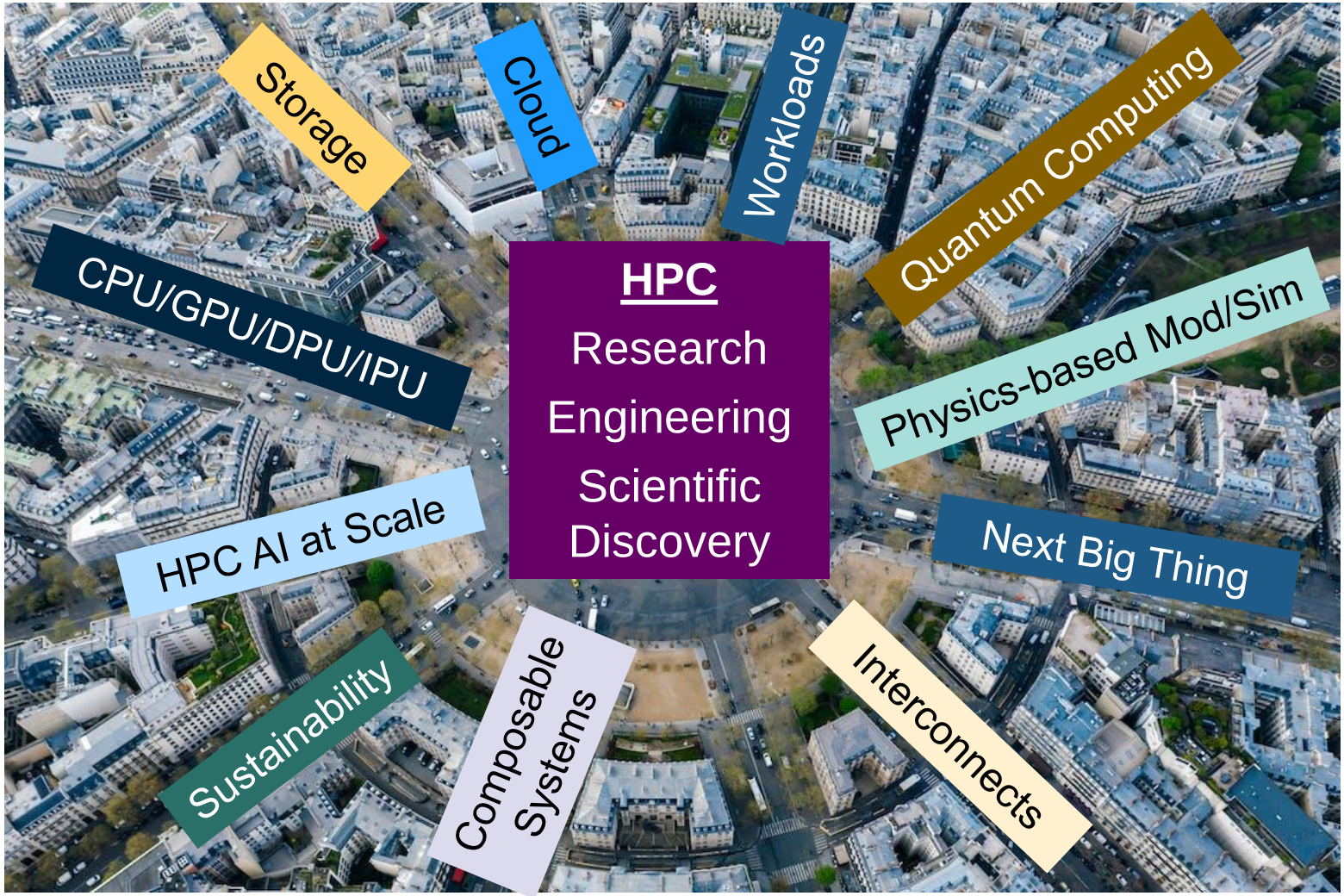
ISC24 Market Update Briefing
May 2024

Mark Nossokoff

www.HyperionResearch.com
www.hpcuserforum.com

Not Your Father's HPC

A Busy Intersection of Complex Challenges



Not Your Father's HPC… …Now Add AI

An Almost Untenable Increase in Complexity

HPC + AI

Research

Engineering

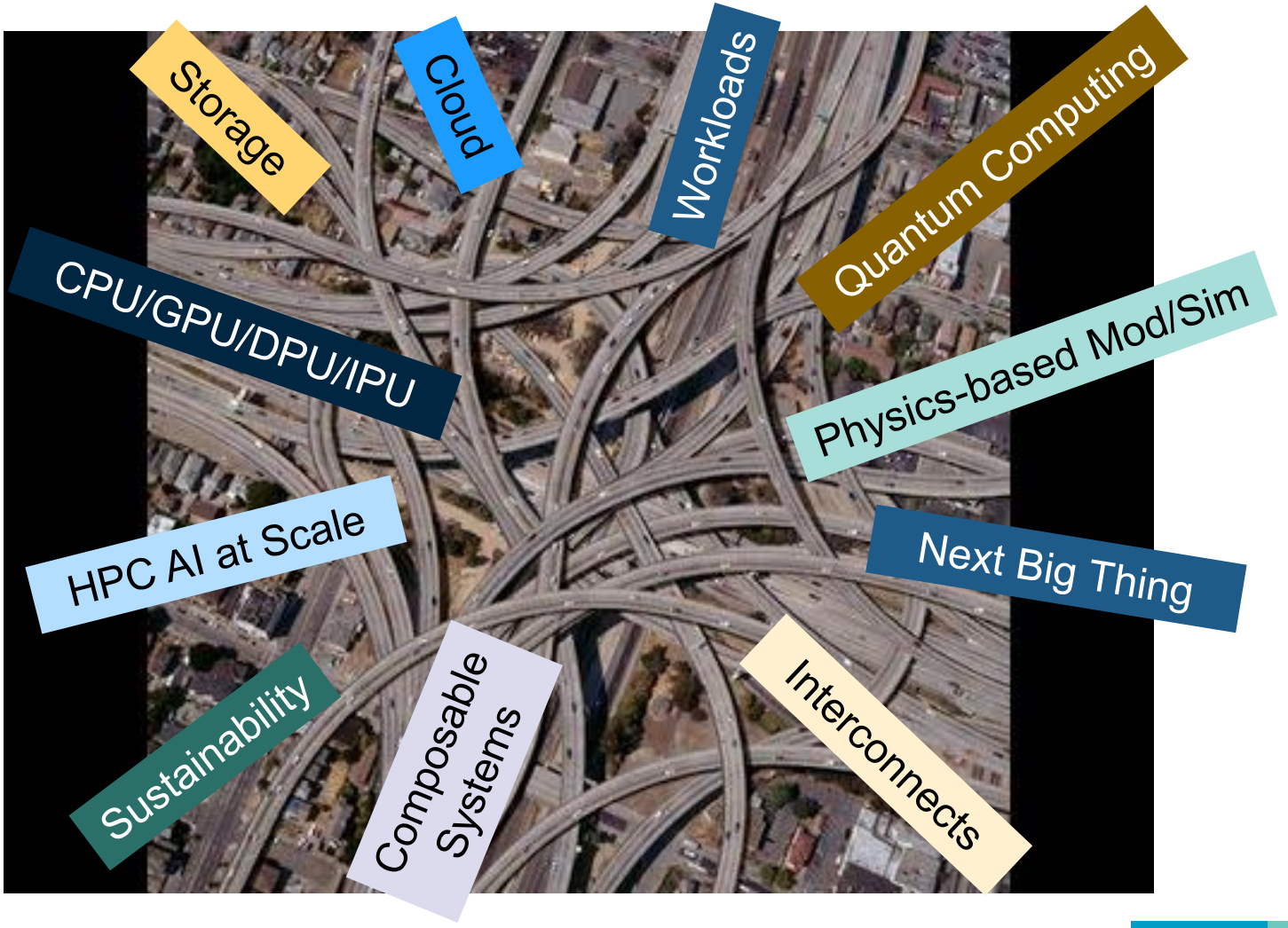
Scientific Discovery

Business Analytics

LLMs

SLMs

Multimodal GenAI



Not Your Father's HPC… …Now Add AI

An Almost Untenable Increase in Complexity

HPC + AI

Research

Engineering

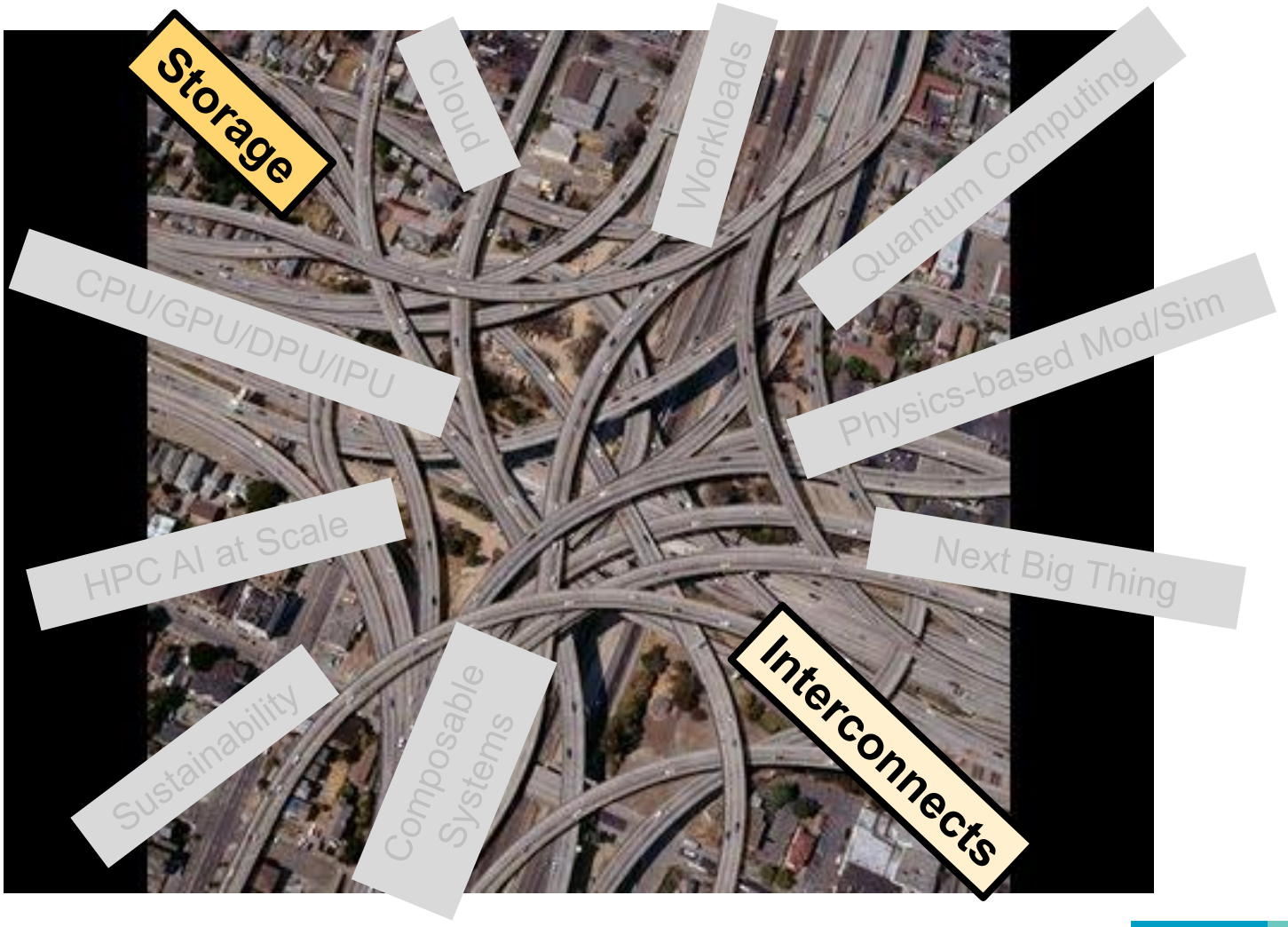
Scientific Discovery

Business Analytics

LLMs

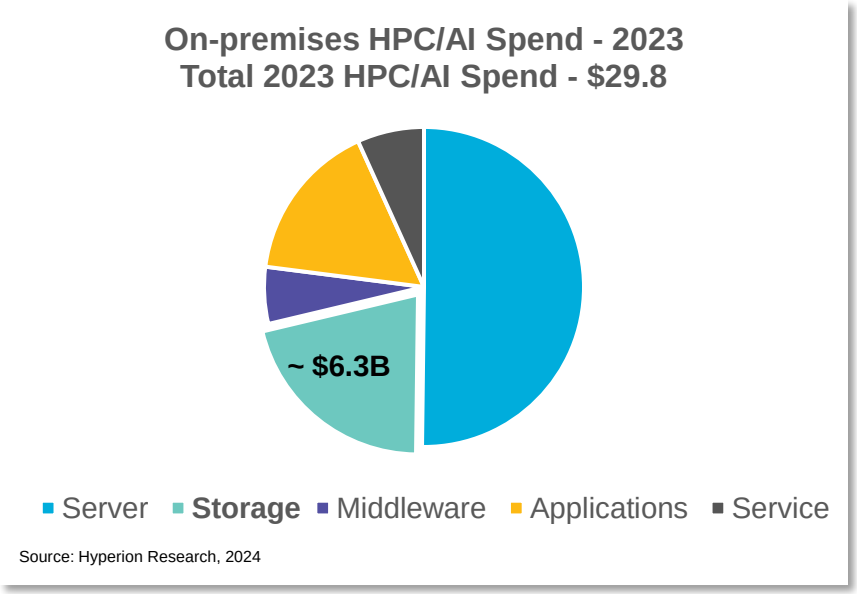
SLMs

Multimodal GenAI



HPC/AI Storage Growth Continues

Demand increasing across all sectors and verticals



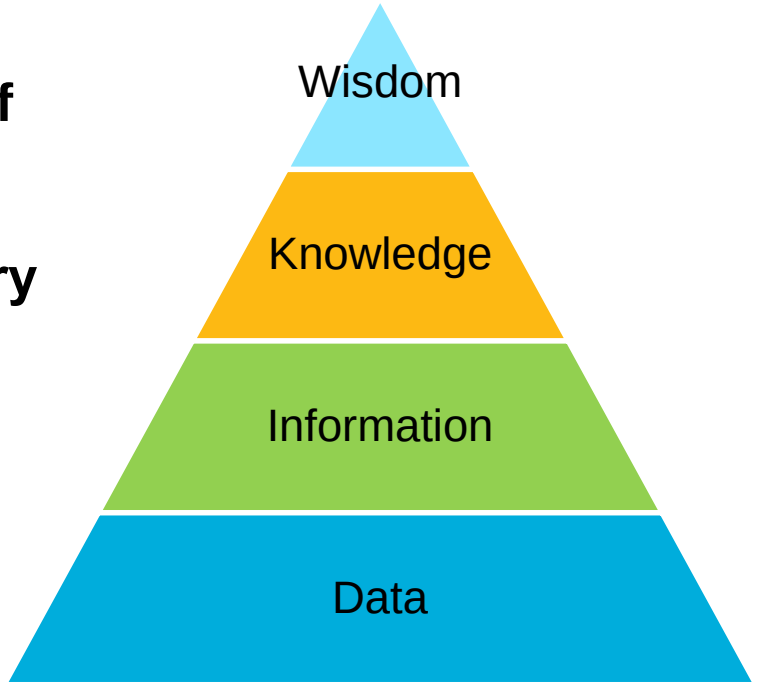
- **Storage continues as highest growth HPC/AI element**
- **Storage represents ~ 21% of on-premises HPC spending and growing**
- **Impact of AI-related requirements and demand still evolving, but anticipated to further increase storage spending**

Area ($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%
Add-on 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%

Data Platform Emerges as Key Storage Value Point

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

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

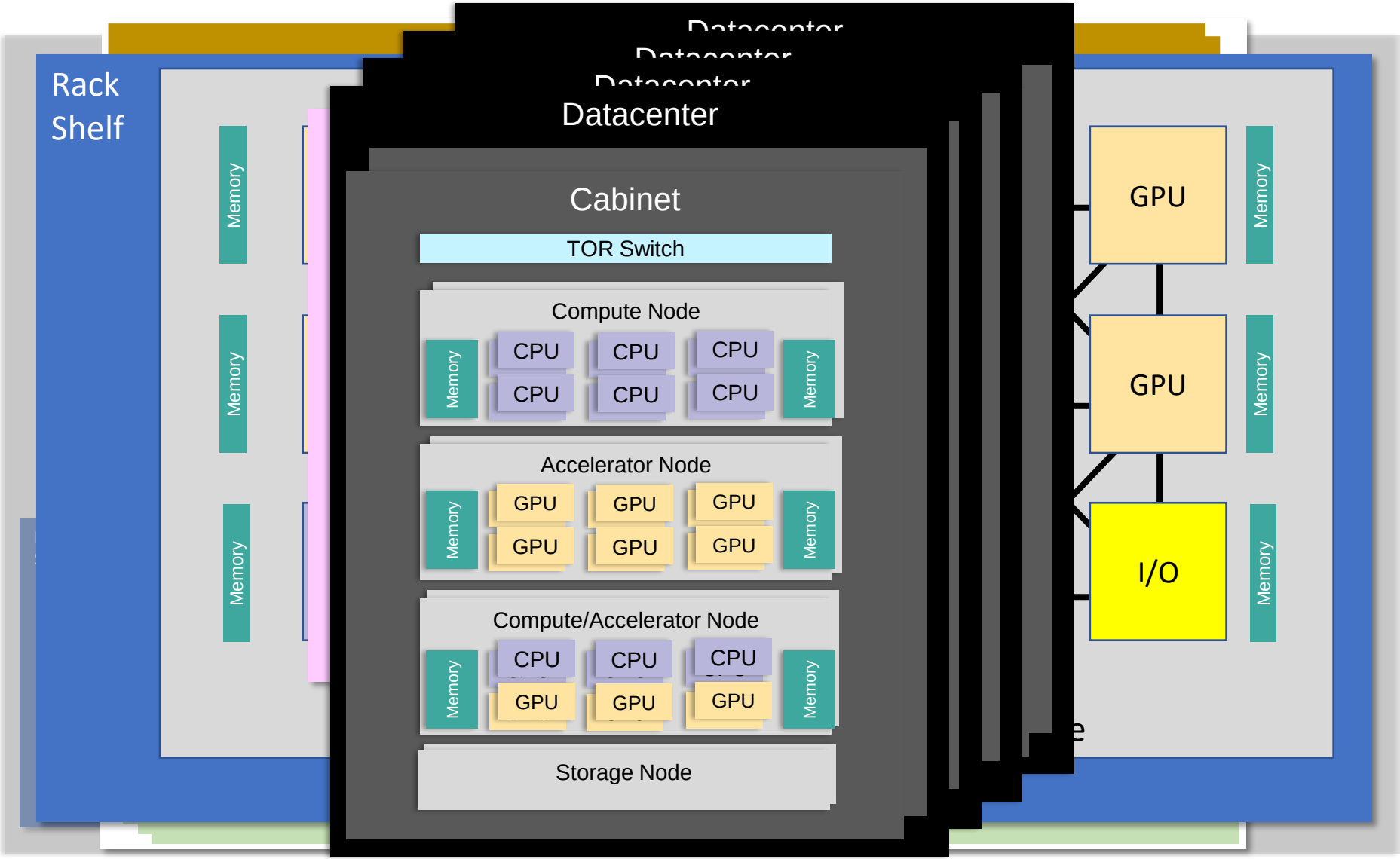


Evolving File System Landscape

AI workflows stressing traditional HPC file systems

- **Incumbent file systems developed for traditional HPC mod/sim being challenged by modern file systems designed with AI in mind from the start**
  - Compute-intensive vs. data-intensive
  - IO profiles (large block sequential vs. small block random)
  - Access methods (file vs. block vs. object)
  - Access frequency (hot vs. archive vs. cold)
  - Locality (centralized datacenter vs. cloud vs. edge)
- **Current file system adoption preferences**
  - Traditional: BeeGFS, Lustre, NFS, OneFS, PanFS, Spectrum Scale
  - Emergent: Hammerspace, VAST, Weka
- **Software business models support broader range of solution choices**
  - Hammerspace, VAST, VDURA (formerly Panasas), Weka

Hierarchy of Networks*



** Network connectivity in diagrams are for conceptual, illustrative purposes only and not meant to necessarily represent architectural realities*

Implications to Interconnect Use Cases

Different interconnect technology options for each use case

- **Ethernet evolving to support current use cases**
- **Protocols critical to satisfying feature and performance requirements**
  - CXL
  - NVMe, NVMeoF
  - Ultra Ethernet
  - Vendor-unique enhancements
  - RDMA, RoCE
  - iSER
  - iWarp
  - SRP
- **Users increasingly want choice in both technology and vendor**

Interconnect Technologies	Use Cases (current and aspirational)					
	Rack to Rack	Node to Node	Node to Storage	xPU to xPU intranode	xPU to Memory	Chiplet to Chiplet
Ethernet	✓	✓	✓	✓		
Infiniband	✓	✓	✓			
Omnipath	✓	✓	✓			
Slingshot	✓	✓	✓			
NVLink		✓		✓		✓
InfinityLink				✓		✓
PCIe			✓	✓		
UCIe				✓		✓
HBM					✓	
OCP BoW					✓	✓
NuLink				✓	✓	✓

What's next for HPC/AI Storage and Interconnects?

Advanced computing architectures will continue to stress existing interconnect capabilities

- **Continued cadence of speeds and feeds**
- **Optimized data platforms to satisfy evolving AI workflow demands**
  - Locality: on-premises, cloud, edge
  - Business model: increasing adoption of SW-only solutions
- **Standards evolution that address new realities will seek industry adoption**
  - Ultra Ethernet
  - UCIe
  - Sovereign EuroHPC interconnect

HYPERION RESEARCH

Questions?

We welcome questions, comments, and suggestions

Please contact us at:
mnossokoff@hyperionres.com

4th Annual Global QC Market: Robust and on the Rise

QED-C

KISTI www.kisti.re.kr
Korea Institute of Science and Technology Information

Bob Sorensen
Chief Analyst for Quantum Computing
Hyperion Research, LLC

2023 QC Market Dynamics Study Roadmap

Formulate global status and prospects from a data-driven perspective

- **Conducted QC supplier survey to gather data and insights on QC market dynamics**
- **Many thanks for assistance from various QC consortia in reaching out to their respective membership base to encourage participation in this effort:**
  - Quantum Economic Development Consortium (QED-C)
  - European Quantum Industry Consortium (QuIC)
  - Quantum Industry Canada
  - Japan Quantum STrategic industry Alliance for Revolution (Q-STAR)
  - Australian Quantum Alliance
  - UKQuantum
  - Korea Quantum Industry Association
- **Gathered results to span:**
  - Geographic variety, company size (total and QC-related revenues), market concentration, QC industry sentiment, and impressions on general trends in the sector
- **Analyzed results from 133 respondents representing 108 different companies**
  - Individual responses for industry-wide questions
  - Combined single response for organizations

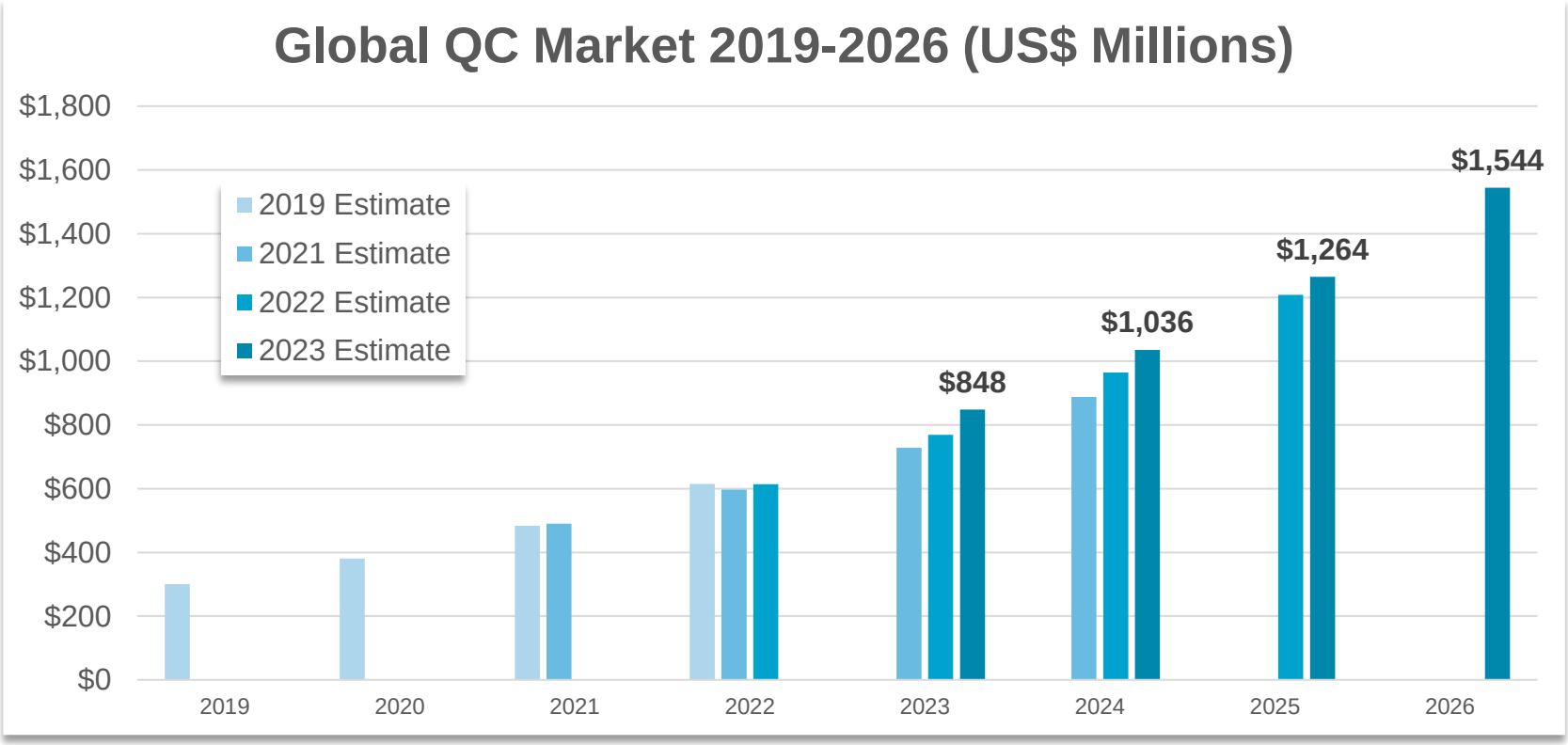
QC Market Highlights

Continued strong and steady progress for the global QC sector

- **The global quantum computing (QC) market is estimated to be worth US$848 million in 2023**
  - A projected annual growth rate of 22.1% brings the global QC market to US$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

QC Market Estimate: US$848 Million in 2023

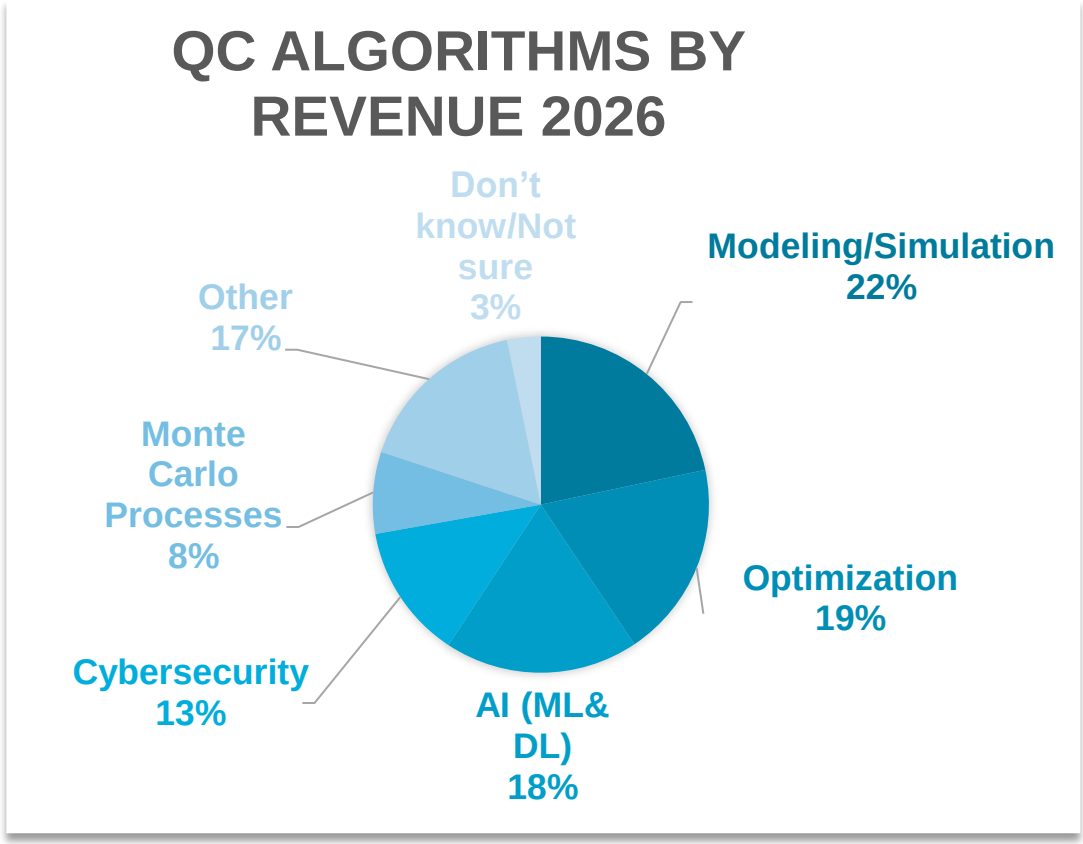
22.1% annual growth rate drives QC global market to US$1.5 billion in 2026



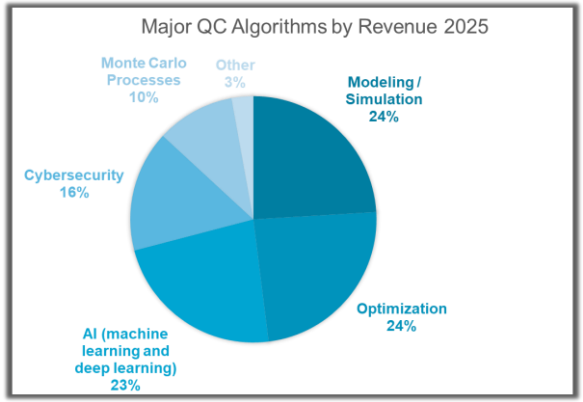
- **Dangerous to project out too far: too many unknowns**
  - But is this the first stages of an exponential curve?

QC Market 2026: Major Algorithms by Revenue

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



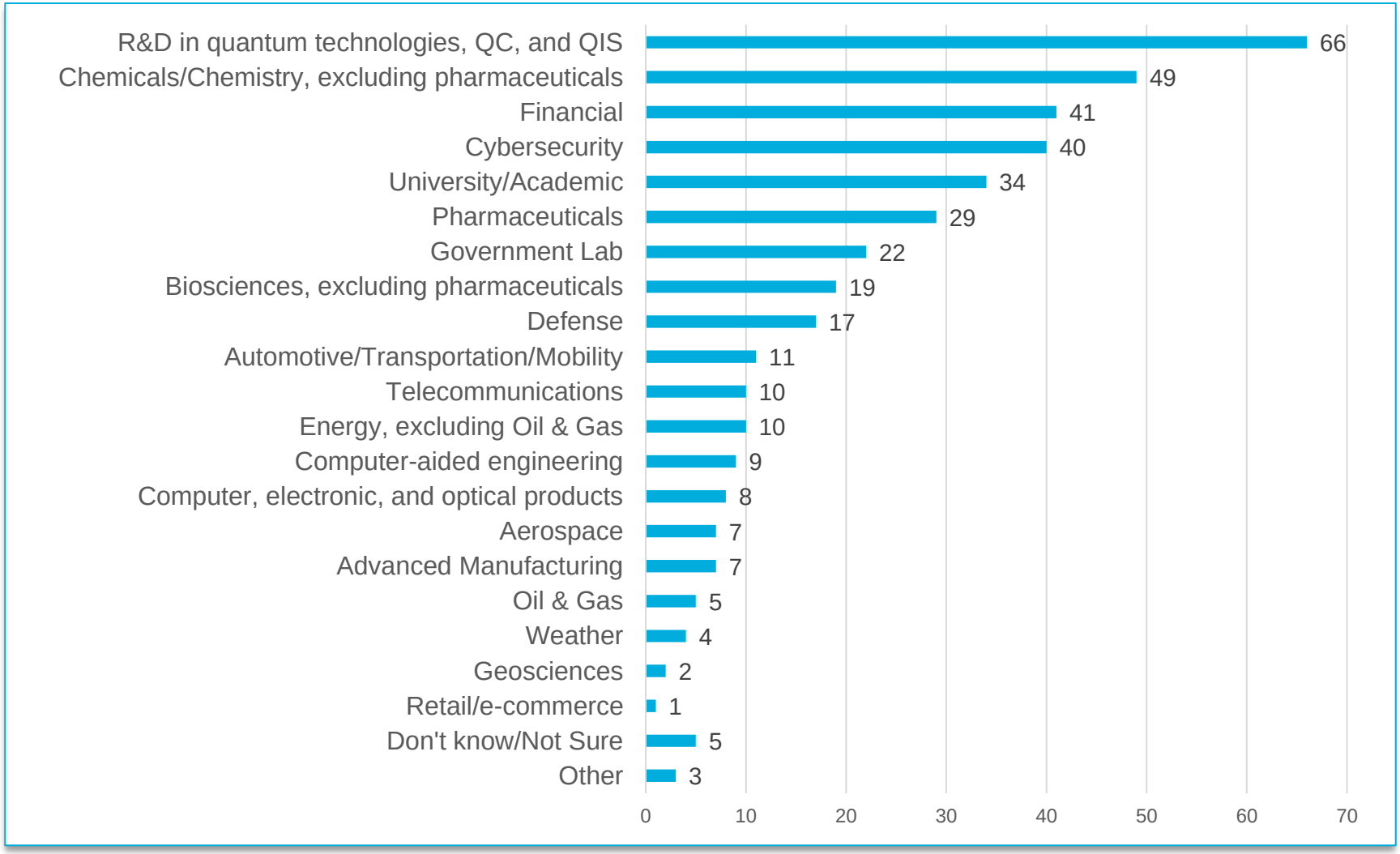
- **Some changes from previous years' studies**
  - **The overall share of the big three is slipping due to major gains in the 'Other" category**
  - **But there was little detail offered about the new alternatives**
  - **Suggests trends towards mixed algorithms use**



N = 133

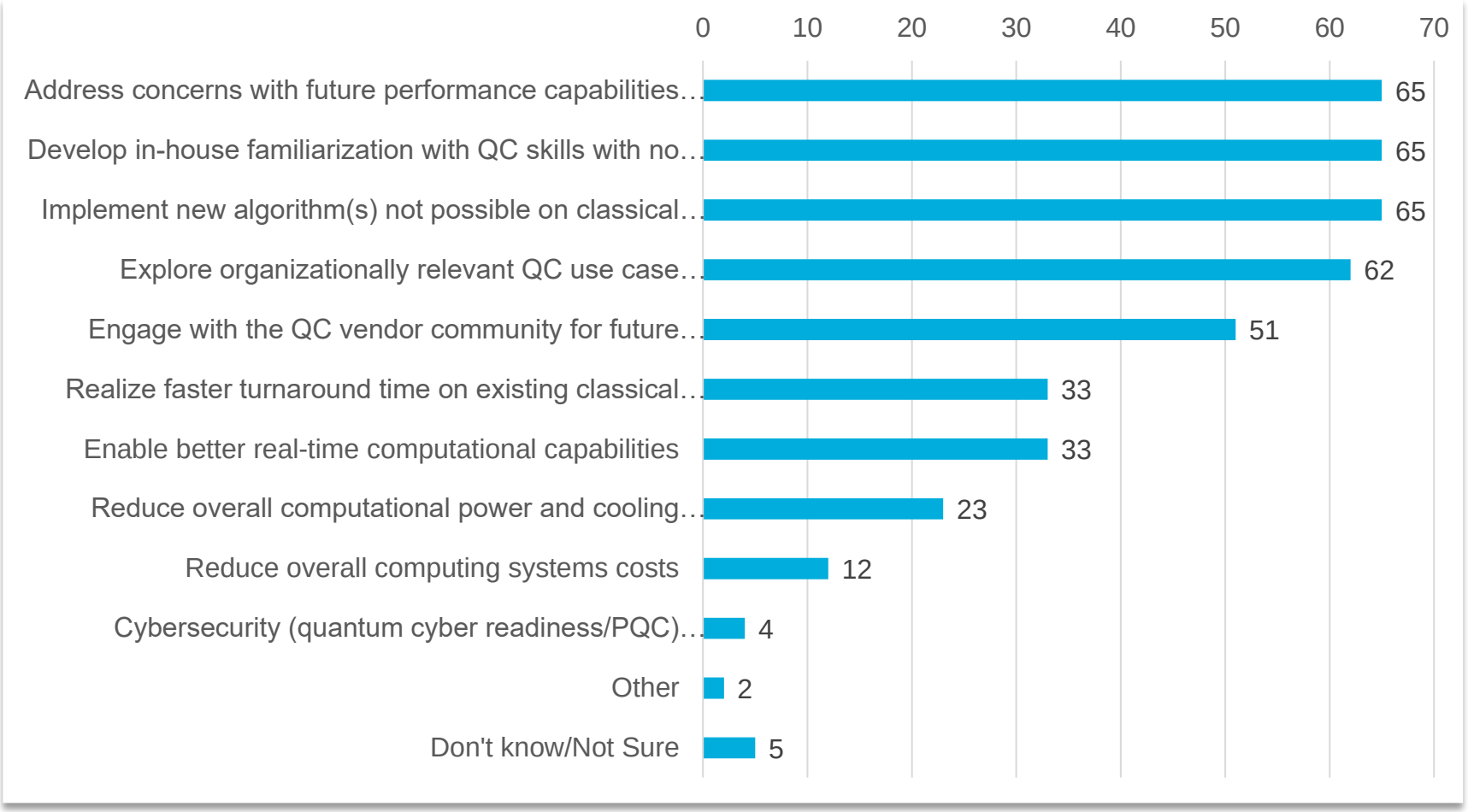
QC Market 2026: Top End User Sectors/Verticals

QC R&D and Chemicals on top, but broad applicability envisioned



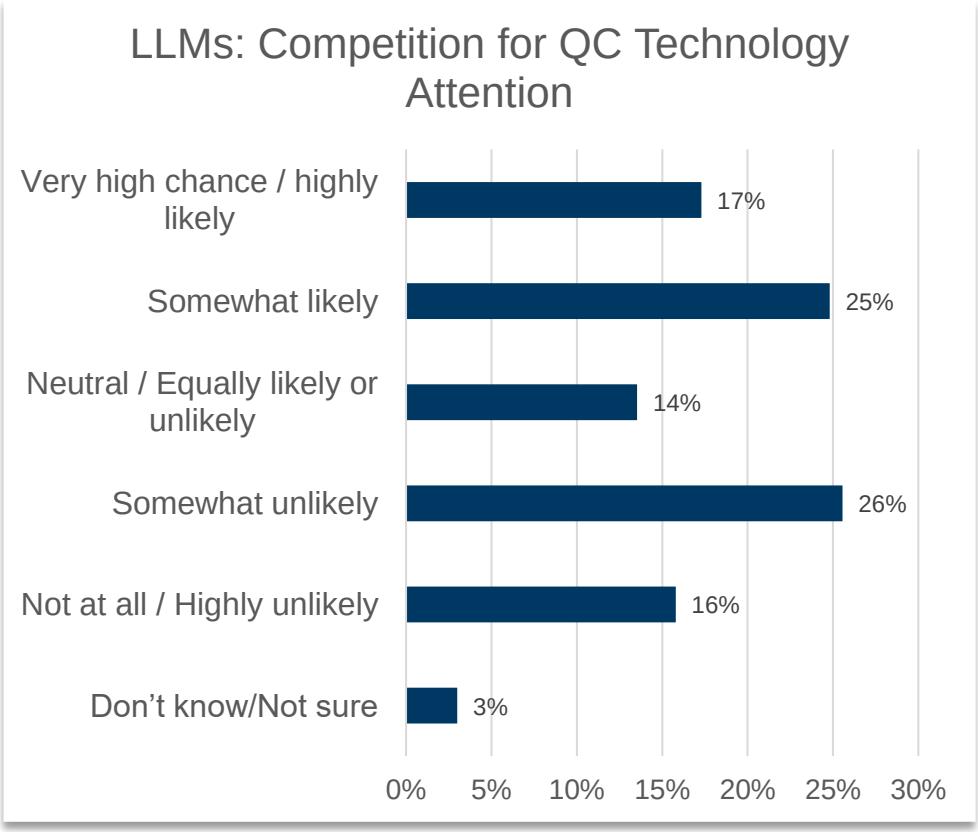
QC 2026: QC End User Perceptions

Mixed drivers for QC end user adoption



QC Distractions and LLMs

How likely is it that the emergence of large language models like ChatGPT and BERT will draw attention away from end user interest in quantum computing?



N = 133

- **LLMs – and likely generative AI in general - seen as near-term competitor for end user interest in QC by 42% of respondents**
- **But equal percentage not overly concerned**
- **Demonstrates need for QC to continue to deliver on technology/performance gains**
- **Highlights perceived end user interest in performance gains no matter how it is delivered**

HYPERION RESEARCH

QUESTIONS?

bsorensen@hyperionres.com

When the great innovation appears, it will almost certainly be in a muddled, incomplete, and confused form. To the discoverer himself, it will be only half understood; to everybody else it will be a mystery.
----Freeman Dyson 'Innovation in Physics', Scientific American, Sept 1958

Considerations for Sustainability

ISC24 Market Update Briefing
May 2024

Mark Nossokoff

www.HyperionResearch.com
www.hpcuserforum.com

Not Your Father's HPC… …Now Add AI

An Almost Untenable Increase in Complexity

HPC + AI

Research

Engineering

Scientific Discovery

Business Analytics

LLMs

SLMs

Multimodal GenAI

Prioritizing Sustainability in Procurements

Sustainability and energy efficiency considerations will become a dominant factor in many procurements

- **New CPUs/GPUs/xPUs have substantially larger power requirements**
  - Maximize performance capabilities for these compute elements, consistent and affordable energy
- **Energy costs**
  - Energy crisis / inflation / geographic concern
- **Energy consumption legislation**
  - Recent legislation on the energy consumption of data centers
  - Implementation: energy-conscious procurements, renewable energy where possible, and optimizing workloads for efficiency
- **Environmental stewardship**
  - Perception of environmental stewardship amongst HPC users
  - Metrics and requirements to identify sustainable vendors
  - Forward-thinking vendors want to respond to the call
- **New technologies are needed to improve energy usage**

Energy Costs to Surpass Maximum Performance in System Tradeoffs

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**

A Datacenter Perspective

Multiple challenges magnified by adoption of AI

- **Power demands and constraints**
  - From 10s of MWs → > 1GW
  - Generation, transmission, or distribution?
  - BYOP (bring your own power)
  - 10s of racks being replaced by a handful of racks for equivalent performance
- **Rack density – from 10KW → 40KW → 400KW**
- **Retrofit or new build for data center expansion?**
  - Access to adequate power for current and planned needs
  - Ability of existing facilities to support internal distribution of more power
  - Readiness for and ability to support current and planned liquid cooling requirements
  - "Time to power" for new builds
- **On-premises vs. co-lo vs public cloud**

What's next in Sustainability?

Is Sustainability sustainable… …yes, but…

- **Advancing AI methodologies to better optimize the energy, performance, carbon footprint algorithm**
  - Being applied in some fashion today
  - Anticipated to increase in use as more sophisticated models are developed
- **Necessity is the mother of invention**
  - Current projections of energy demand outstripping power production are increasingly alarming
  - Research currently being developed likely to advance into production to address both supply and demand
  - ARPA-E COOLERCHIPS*
    - Components relative to secondary cooling loop
    - Cooling systems for modular and edge data centers
    - Data center cooling system software
    - Testing facilities for items #1 & #2

*Cooling Operations Optimized for Leaps in Energy, Reliability and Carbon Hyper-efficiency for Information Processing Systems.

HYPERION RESEARCH

Questions?

We welcome questions, comments, and suggestions

Please contact us at:
mnossokoff@hyperionres.com

Widespread AI Adoption Creates New Demands

Tom Sorensen

May 2024

New Category of Advanced Computing

- HPC users adopting/integrating AI at high rates
- LLMs received most attention among a diverse field of model and method types
- ~90% of HPC users surveyed currently or plan to use AI methods for workloads
- AI methods introduce new demands on sites:
  - Hardware (processors, interconnect, data access)
  - Software (data management, queueing, dev tools)
  - Expertise (procurement strategy, maintenance, troubleshooting)
  - Regulatory (data provenance, privacy, legal)

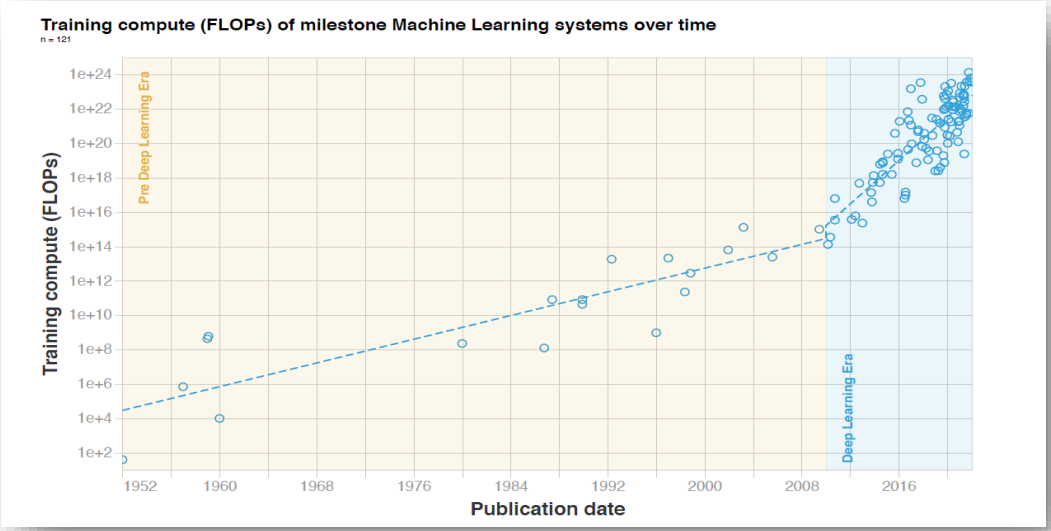
Framing LLM/HPC Requirements

Key elements dominate scaling of LLMs on HPCs

- **Compute: the absolute number of floating-point operations needed to train an LLM to a desired degree of accuracy**
- **Dataset size: input dataset used for training the LLM**
- **Model size: number of tokens or parameters**
  - The larger the number of parameters, the more nuance in the model's understanding of each word's meaning and context
- **Integration: Ability to introduce effective models into a production or experimental environment**
  - Users identify ease of AI integration as the highest consideration when adopting AI tools into production environments
- **LLMs require unique HPC capabilities**

LLMs Consume Significant Flops

LLM flops growth eclipses Top 500 growth



- Pre 2010:
  - On the order of 2 X 10^{12} (200 Tflops)
  - Flops requirements doubling every 21.3 months
  - But not a lot of data points
- Post 2010 to Current:
  - Currently on the order of 6 X 10^{22} flops (60 Zettaflops)
  - Flops requirements doubling every 5.6 months
  - Roughly 11X faster than HPC Top 1 Linpack performance growth rate

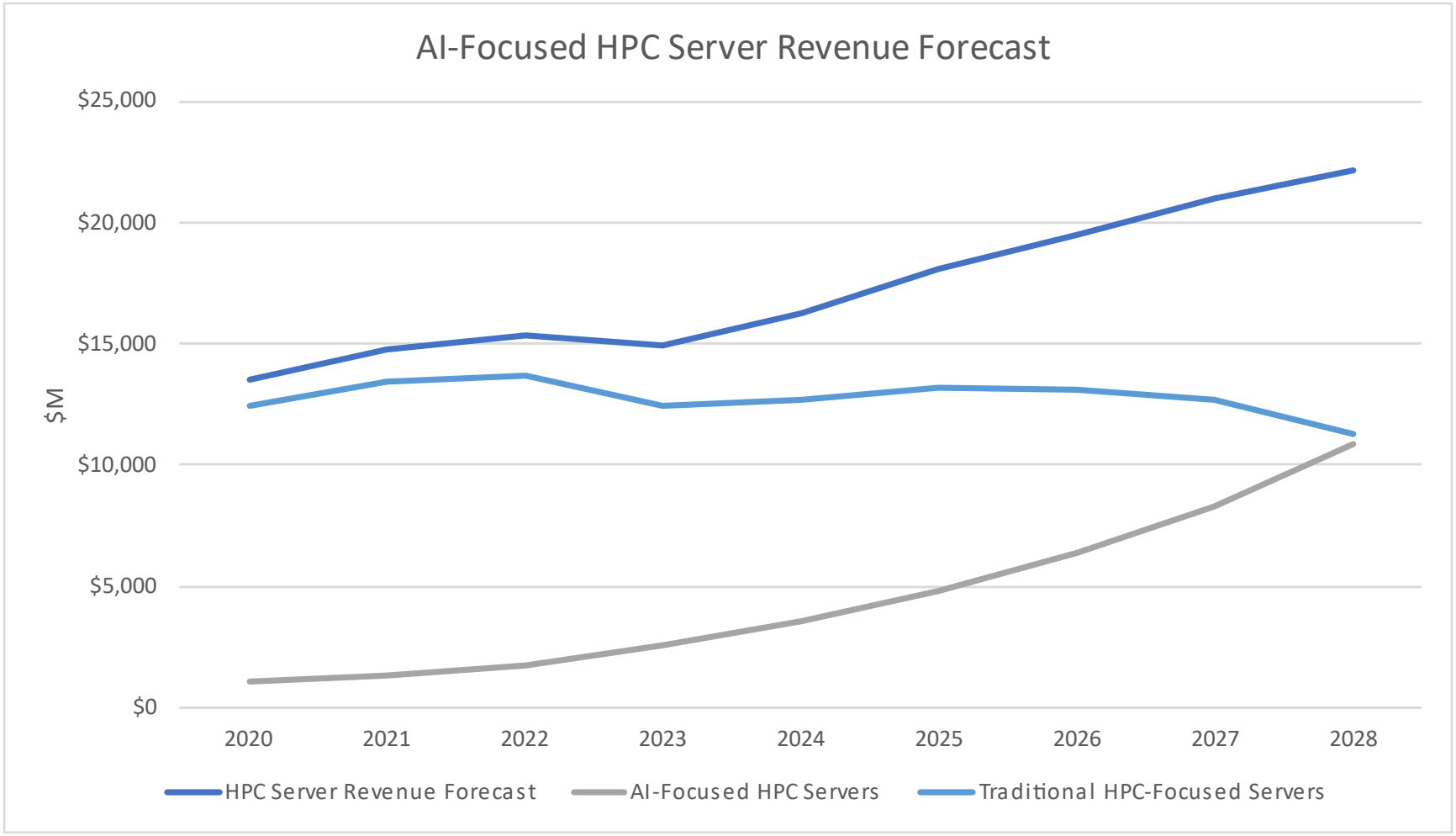
See Compute Trends Across Three Eras of Machine Learning, arXiv:2202.05924

Select Key Findings from AI Use Study

From a survey of HPC users leveraging LLMs

- **1:** LLMs are considered to be an important productivity enhancer asset for both current and planned HPC-related activity in workload management and data analysis, but not necessarily as a computational accelerator.
- **3:** Respondent organizations are exploring a broad set of LLM-related end uses.
- **5:** Adoption may be limited due to the high cost of hardware such as GPU and other accelerators.
- **9**: Open source is currently the most preferred option for accessing LLM software.
- **10:** Survey respondents looked to a wide range of LLM expertise to support the various stages of LLM development spanning foundation model construction, fine-tuning procedures, LLM integration into existing workloads, and supporting inference operations.

AI Server Forecast ($M)



Putting This All Together

Is this (another) new HPC architectural paradigm in the works?

- **Based on a recent LLM analysis by Riken**
- **GPT variant flops requirements**
  - GPT-3.5 (ChatGPT): $3X10^{24}$ flops (estimated)
  - GPT-4.0: $3X10^{25}$ flops (estimated)
- **OpenAI System: Microsoft/Open AI collaboration**
  - Top 5 system when stood up
  - GPU-based BF16 312 Tflop/s x 25,000 = 7.8 Eflop/s TPP
  - GPT-3.5 (ChatGPT): 4.5 days X 2
  - GPT-4.0: 45 days X 2
- **Fugaku:**
  - FP32 6.76 Tflop/s X 158,976 = 1.07 Eflop/s (TPP)
  - GPT-3.5 (ChatGPT): 32 days X 10
  - GPT-4.0 45 days X 2: 328 days X 10 ~= 8.9 years

Distributed Training of Large Language Models on Fugaku, https://t.co/idofa7Tjyu

HYPERION RESEARCH

QUESTIONS?

tsorensen@hyperionres.com
ejoseph@hyperionres.com

Forecasts and High Growth Areas to Keep an Eye on

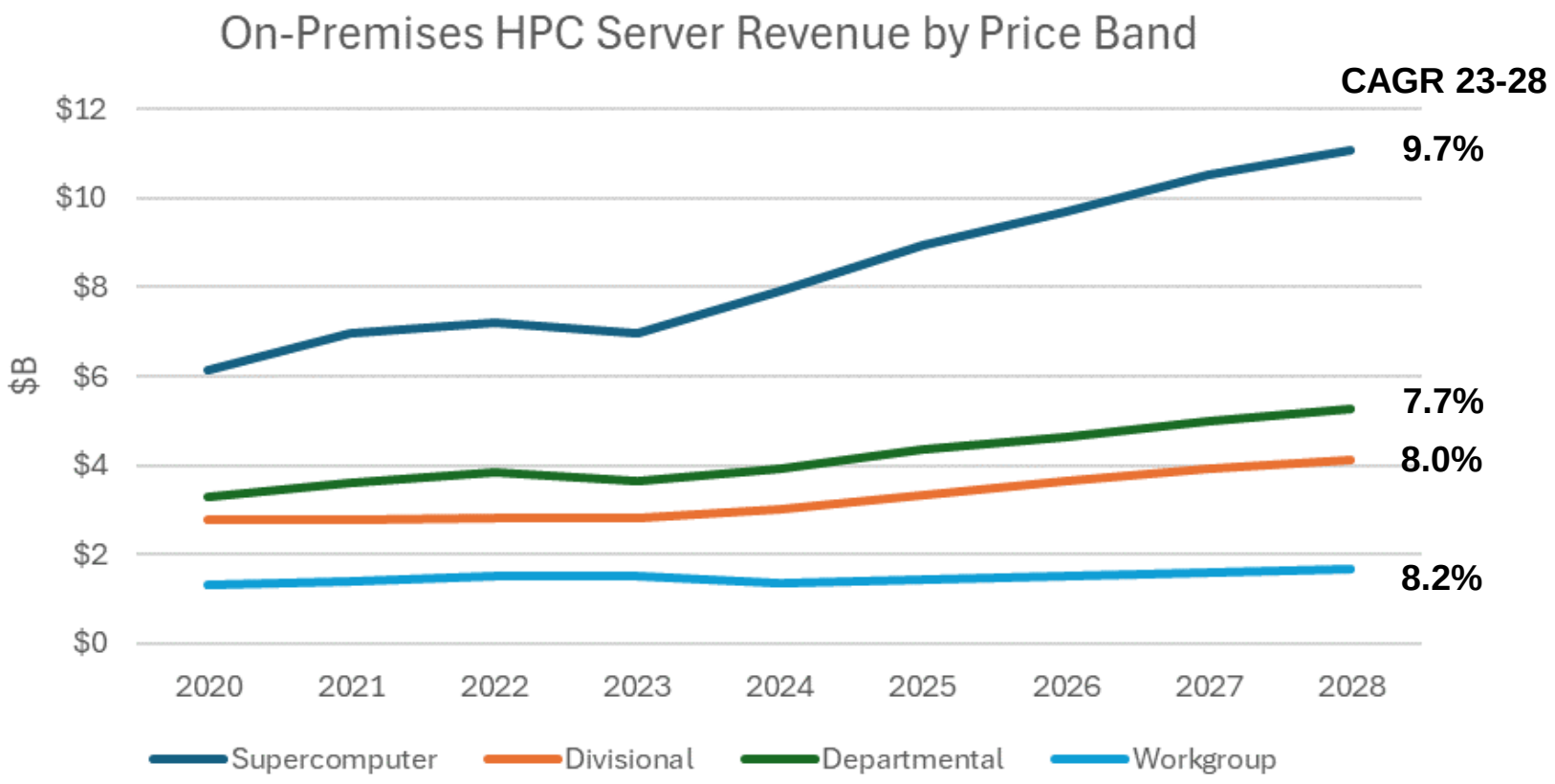
ISC24 Market Update Briefing
May 2024

www.HyperionResearch.com
www.hpcuserforum.com

Melissa Riddle

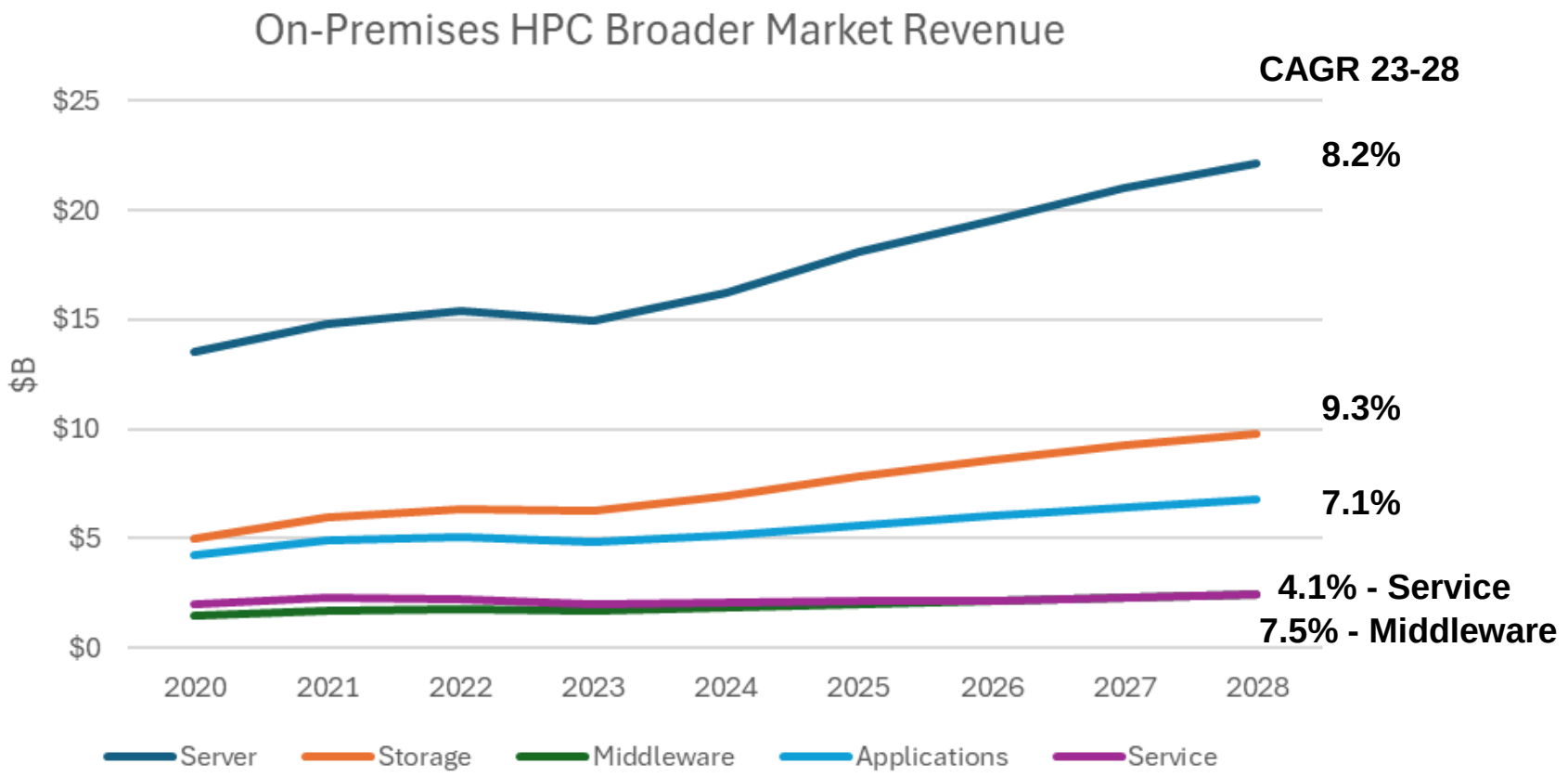
Supercomputers Fastest Growing Server Price Band

Supercomputers are expected to drive the bulk of the growth within the on-premises HPC server market



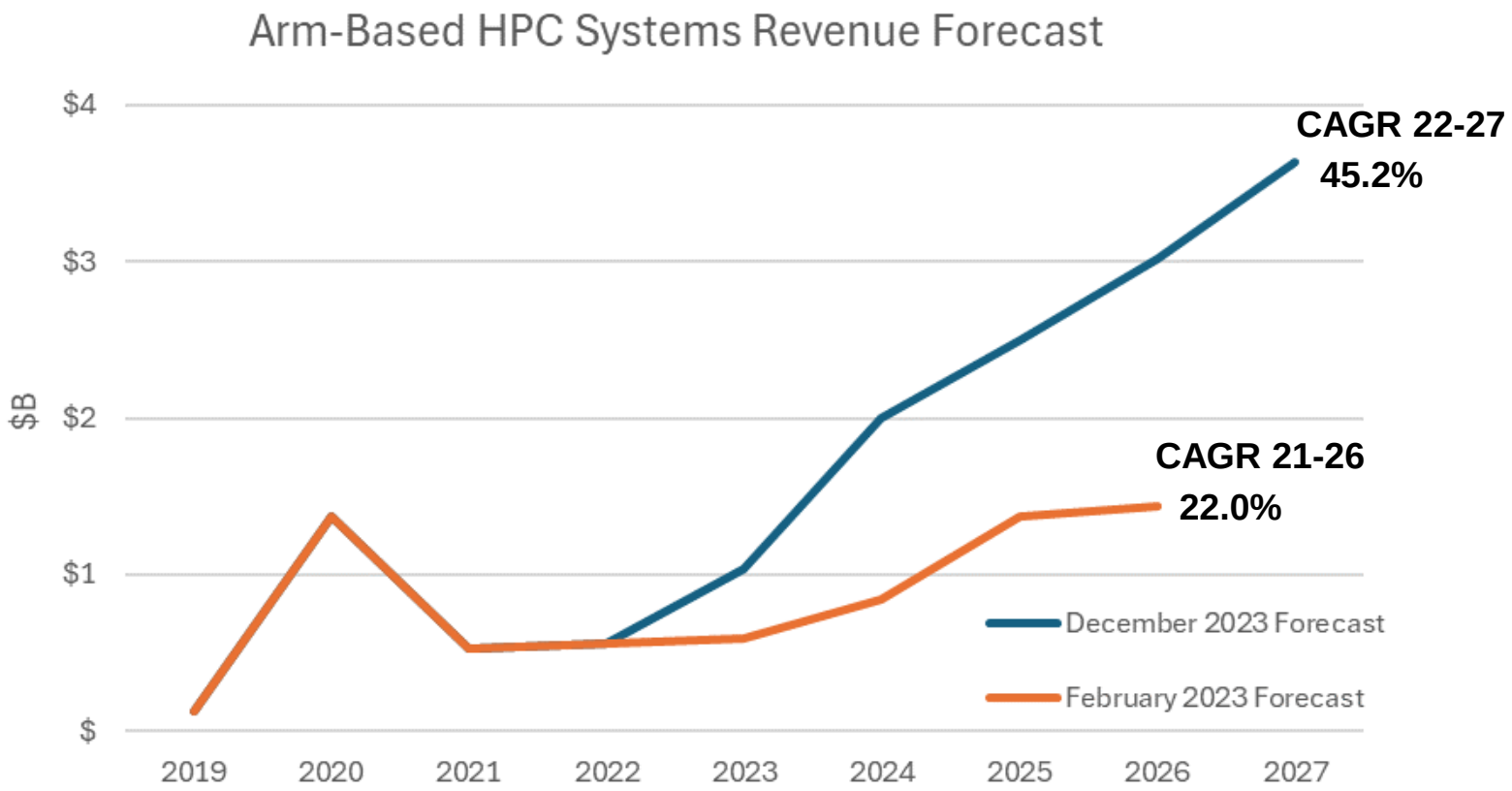
Storage Fastest Growing Broader Market Area On-Premises

Storage is expected to be the fastest growing segment, while HPC servers will provide the bulk of revenues and also drive market growth.



Arm-Based HPC Systems Revenues Expected to More than Triple by 2027

The Arm-based on-premises HPC server market forecast has been dramatically increased since last year due to increased optimism about Arm HPC adoption and expanding scale



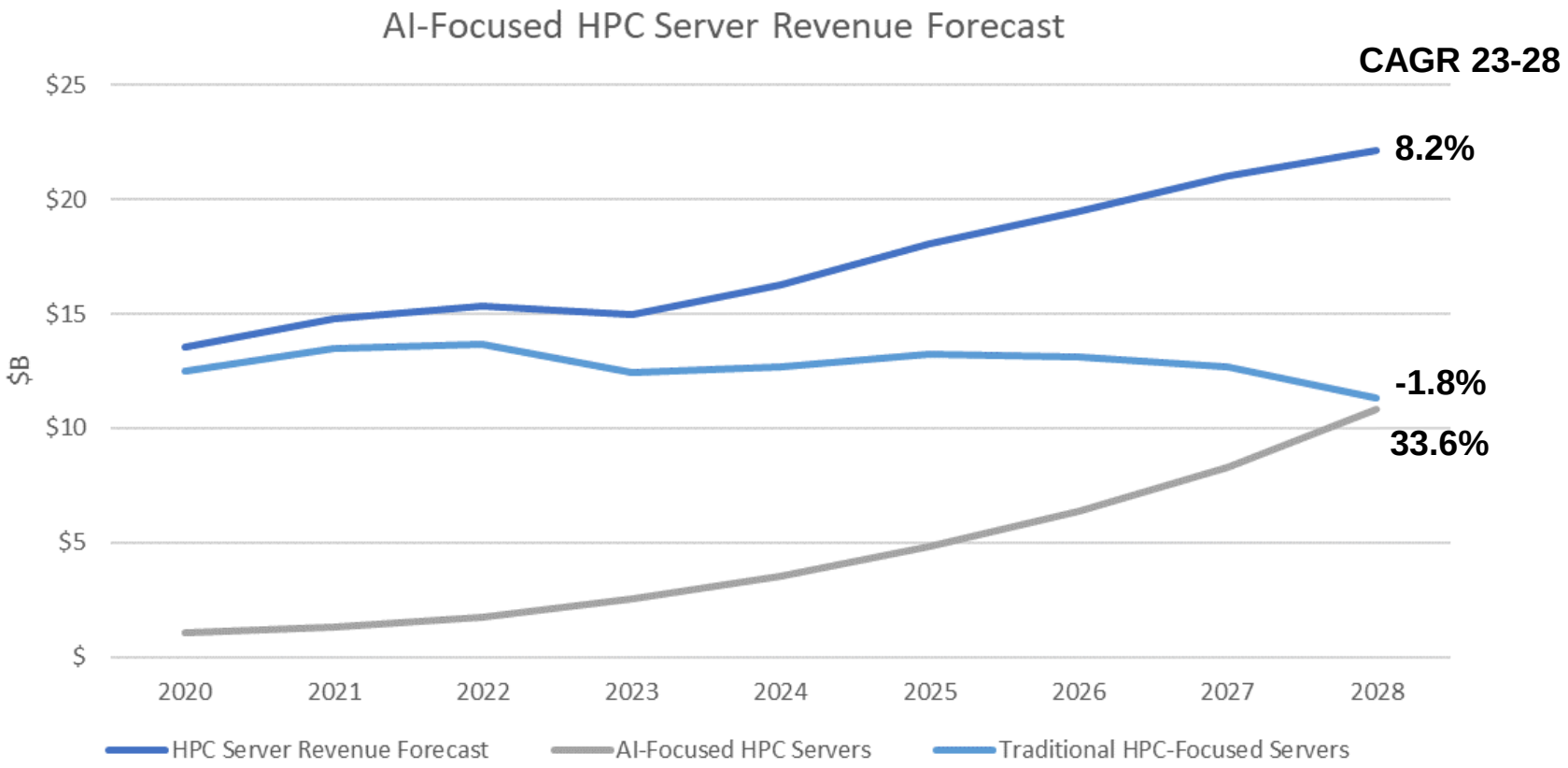
Arm-Based HPC Systems Revenue to Double in 2024

11. Arm-based processors will rise sharply in adoption, with expected revenues for Arm-based HPC systems doubling compared with the previous year

- **This reflects growth in many areas of the Arm ecosystem:**
  - More HPC users are starting to adopt Arm processors into their systems
  - Existing Arm users are increasing the number, size, and complexity of Arm-based HPC systems
  - There is an ongoing expansion of available Arm-based products and vendors
    - E.g., The burgeoning interest in Arm-based superchips incorporating GPUs (e.g., NVIDIA's Grace Hopper)
  - The resulting overall rising cost of Arm-based offerings
  - Pricing inelasticity with respect to Arm-based GPUs has allowed system prices to rise without a normally expected corresponding drop in demand

On-Premises AI-Focused HPC Servers Experiencing High Growth

The AI-focused subset of HPC servers is expected to quadruple between 2023 and 2028



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

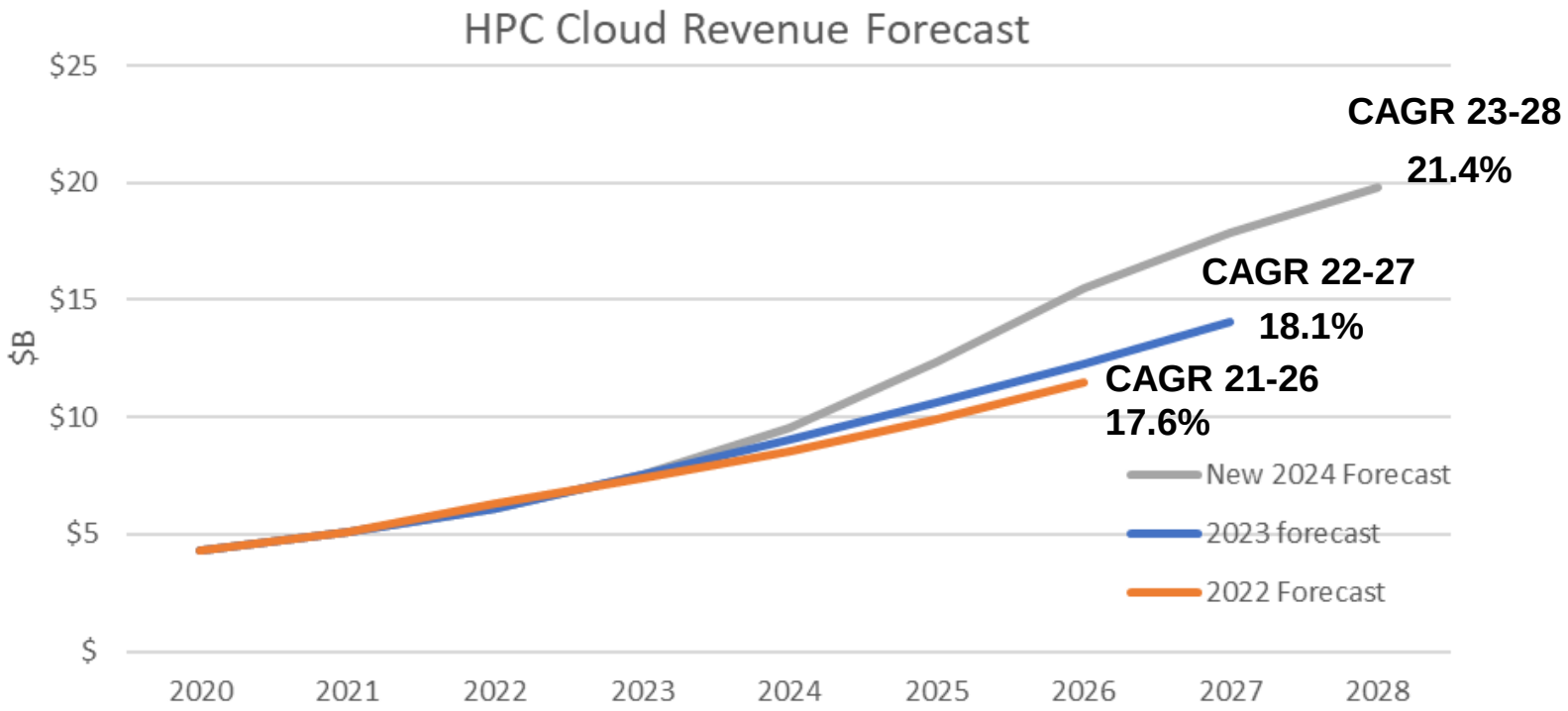
AI Training and Inferencing to Accelerate Utilization of Cloud HPC Resource

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

- **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

HPC Cloud Revenue Expected to Experience Even Higher Growth

The HPC cloud revenue forecast has been increased substantially over last year's predictions due to strong market drivers



HYPERION RESEARCH

Questions?

mriddle@hyperionres.com

Considerations for Cloud

ISC24 Market Update Briefing
May 2024

Mark Nossokoff

www.HyperionResearch.com
www.hpcuserforum.com

Not Your Father's HPC… …Now Add AI

An Almost Untenable Increase in Complexity

HPC + AI

Research

Engineering

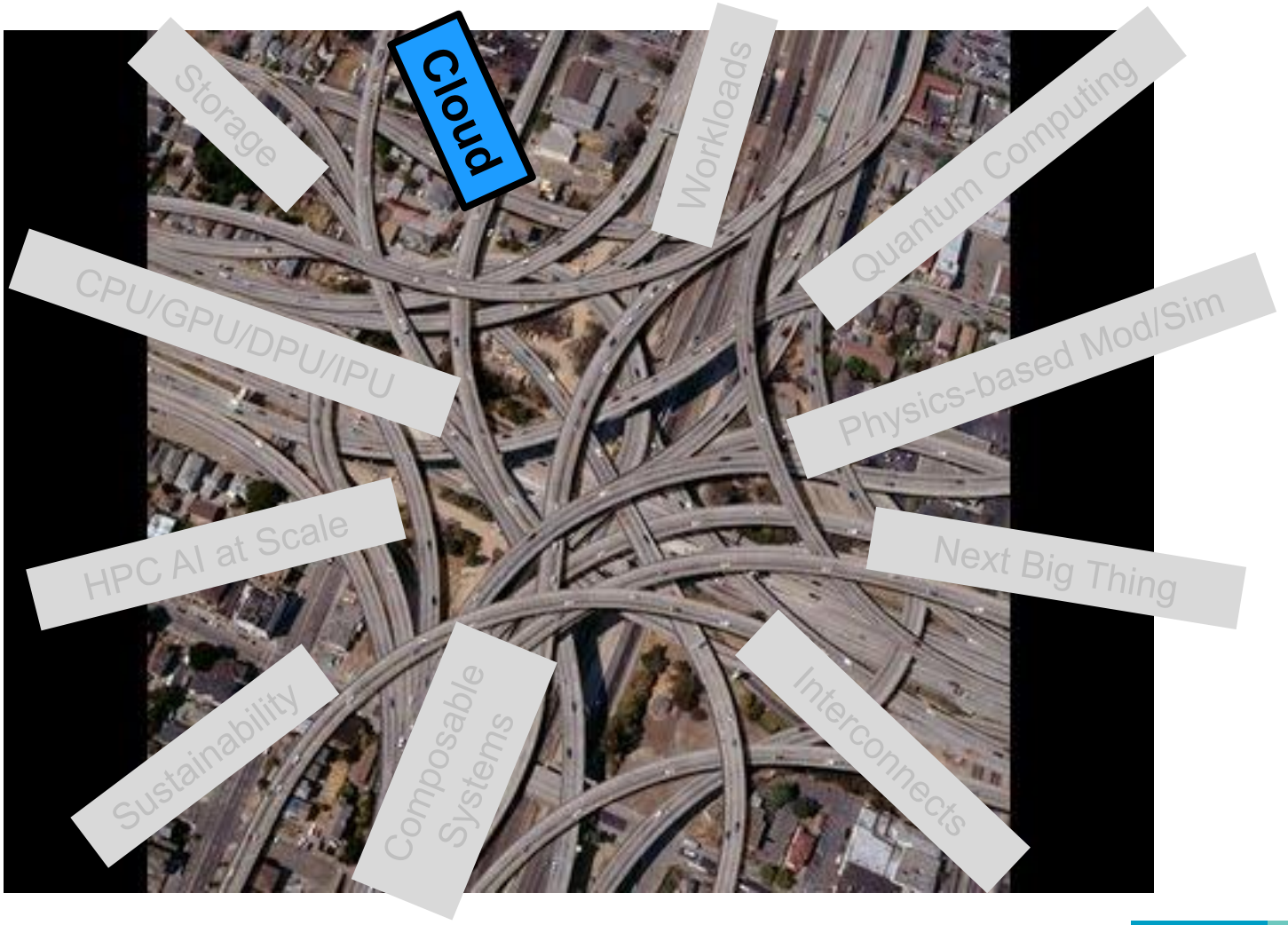
Scientific Discovery

Business Analytics

LLMs

SLMs

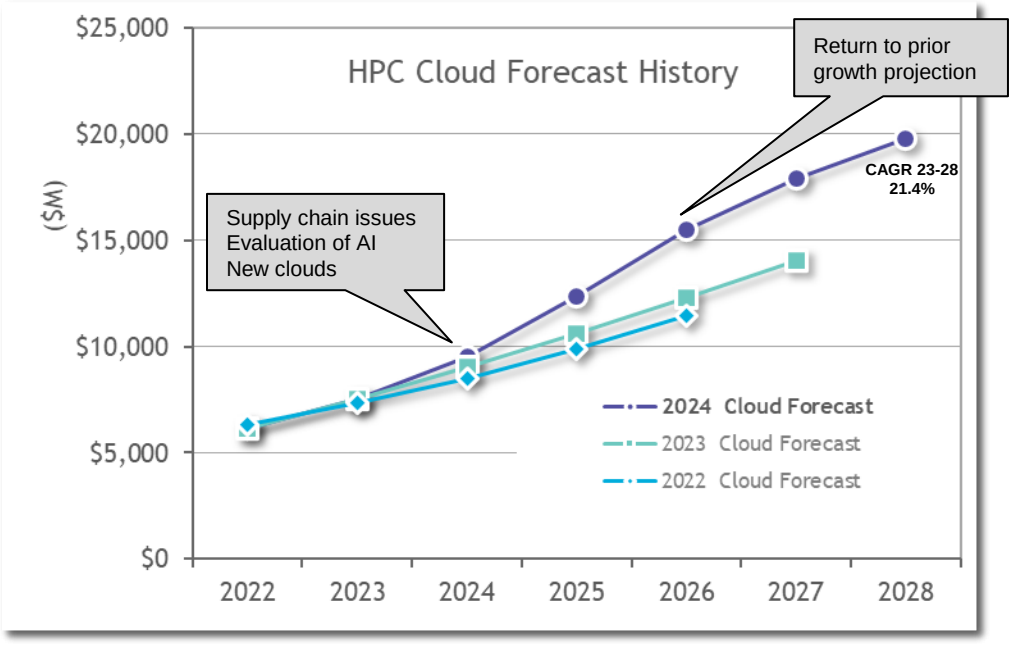
Multimodal GenAI



HPC/AI Cloud Forecast

Cloud revenue expected to approach $20B by 2028

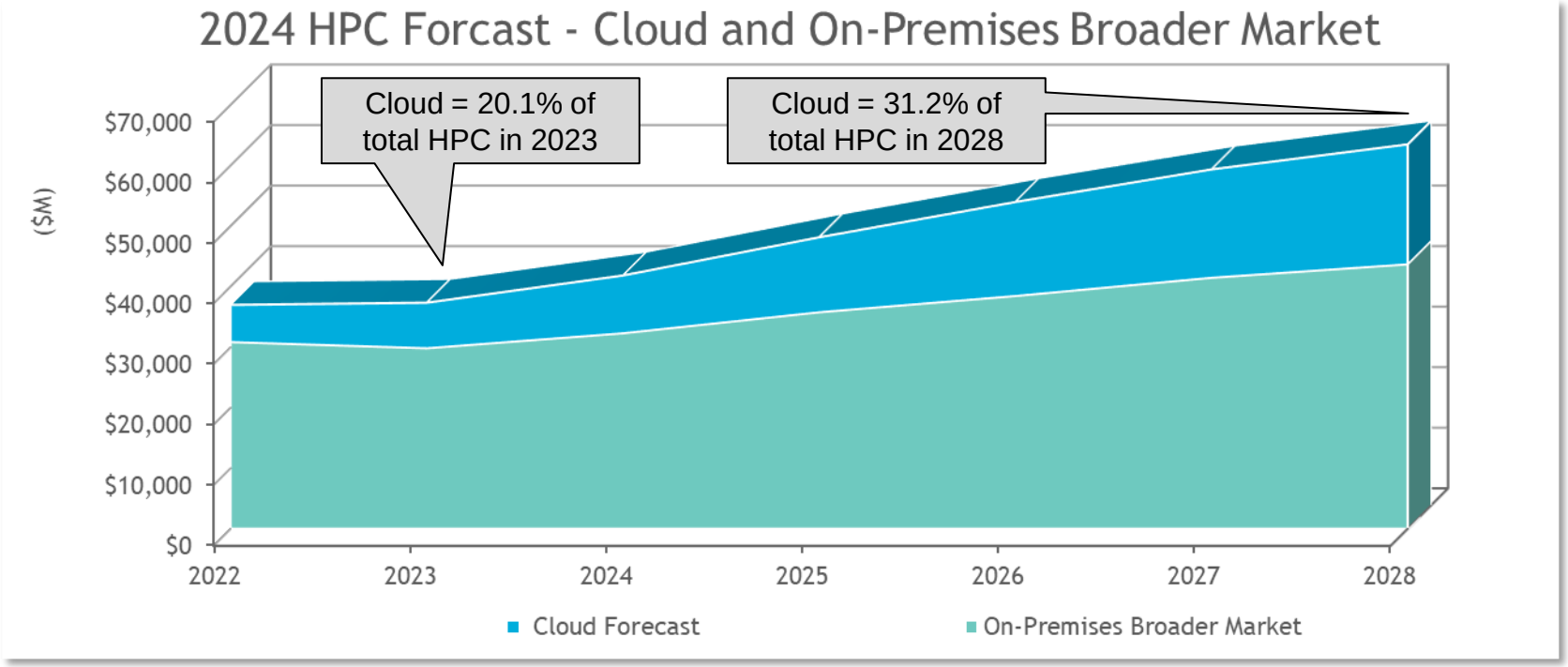
- **Global HPC/AI buyers around the world continue shifting portions of their on-premises budgets to spending in the cloud**
- **Increasing growth rate for 2024, returning to prior projected growth in 2026**
- **Primary growth drivers:**
  - Supply chain issues
  - Access to latest GPU technology
  -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

The cloud market approaching 1/3 of total market 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	$30,813	$29,791	$32,265	$35,735	$38,420	$41,433	$43,656	7.9%
Total	$36,945	$37,308	$41,805	$48,112	$53,939	$59,326	$63,460	11.2%

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
- **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
    - Potentially detrimental performance impact in counties where they get it wrong
  - What is the correct metric to key 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?

What's next in clouds?

Dynamic environment anticipated 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 welcome questions, comments, and suggestions

Please contact us at:
mnossokoff@hyperionres.com

Conclusions

ISC24

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

www.HyperionResearch.com
www.hpcuserforum.com

Conclusions

- **2023 was a flat growth year for HPC overall**
- **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 strong 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.
- **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/AI Expertise Shortage

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

Two major trends:

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

Join Us For Our 2024 HPC User Forums And Hyperion Research Briefings (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**

HYPERION RESEARCH

Thanks For Joining Us Today! And Have A Great ISC24!

We welcome questions, comments and suggestions

Please contact us at:
info@hyperionres.com