Hyperion Research

HPC Market Update and Observations About Disaggregation/ Composable Architectures

Salishan Conference April 25, 2023

Earl Joseph

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, Associate Analyst
- Thomas Sorensen, Associate Analyst
- Cary Sudan, Principal Survey Specialist

Operations

- Jean Sorensen, COO

International Consultants

- Katsuya Nishi, Japan and Asia
- Jie Wu, China & Technology Trends

Global Accounts

- Rene Copeland, Dir. Business Development
- Mike Thorp, Sr. Global Sales Executive
- Kurt Gantrish, Sr. Account Executive

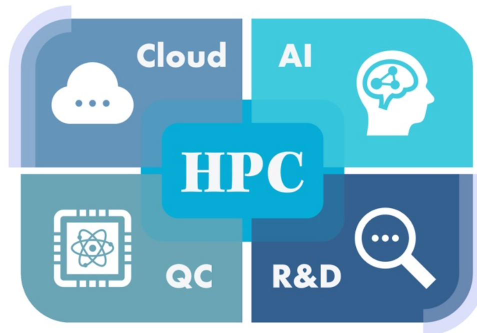
Data Collection

- Andrew Rugg, Certus Insights
- Kirsten Chapman, KC Associates

Our Research Areas

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

- Traditional HPC
- AI: ML, DL, Graph
- Cloud Computing
- Storage & Big Data
- Interconnects
- Software & Applications
- Power & Cooling
- The ROI and ROR from Using HPC
- Tracking all Processor Types & Growth Rates
- Quantum Computing
- R&D and Engineering -- All Types of High Tech
- Edge Computing
- Supply Chain Issues

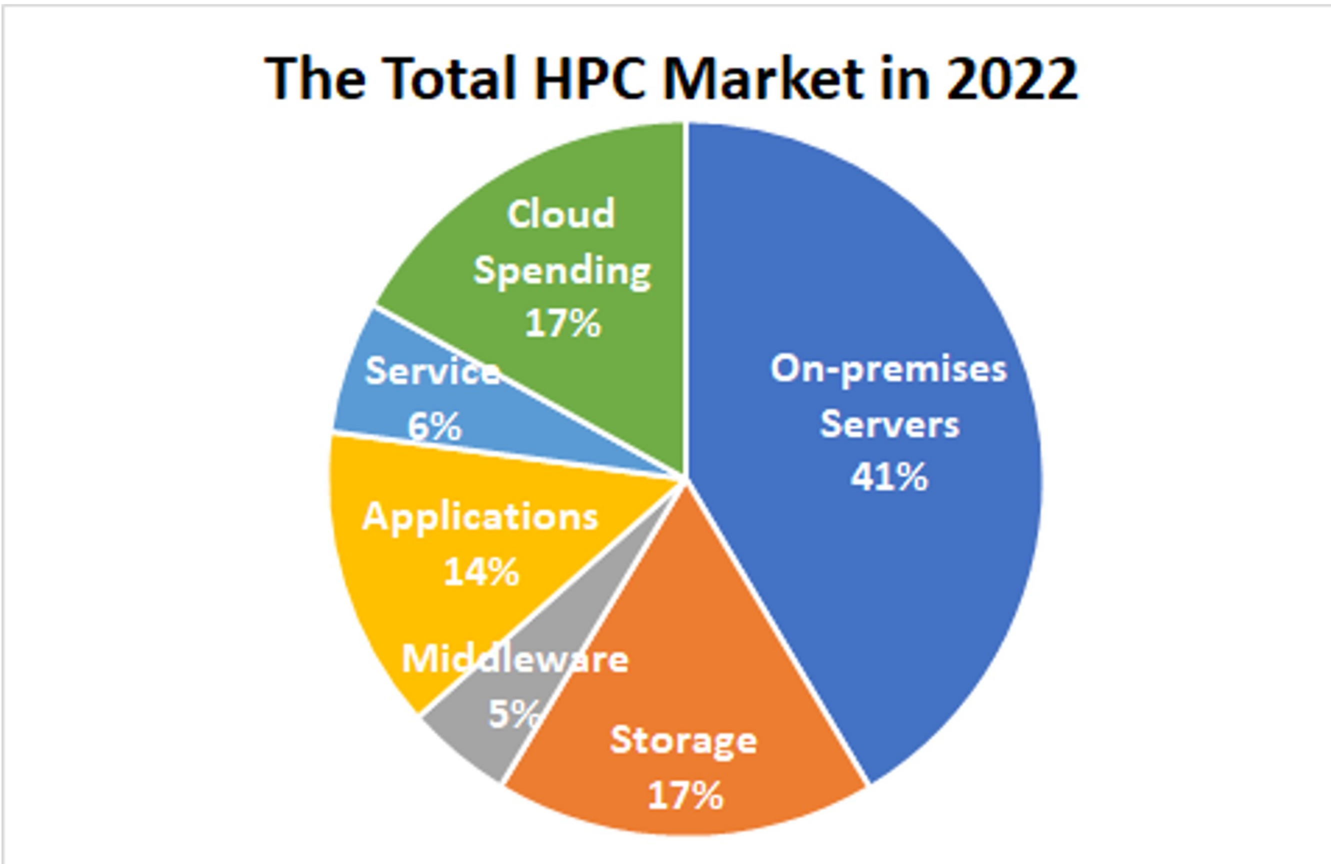


HPC Market Update

(Just Published Data)

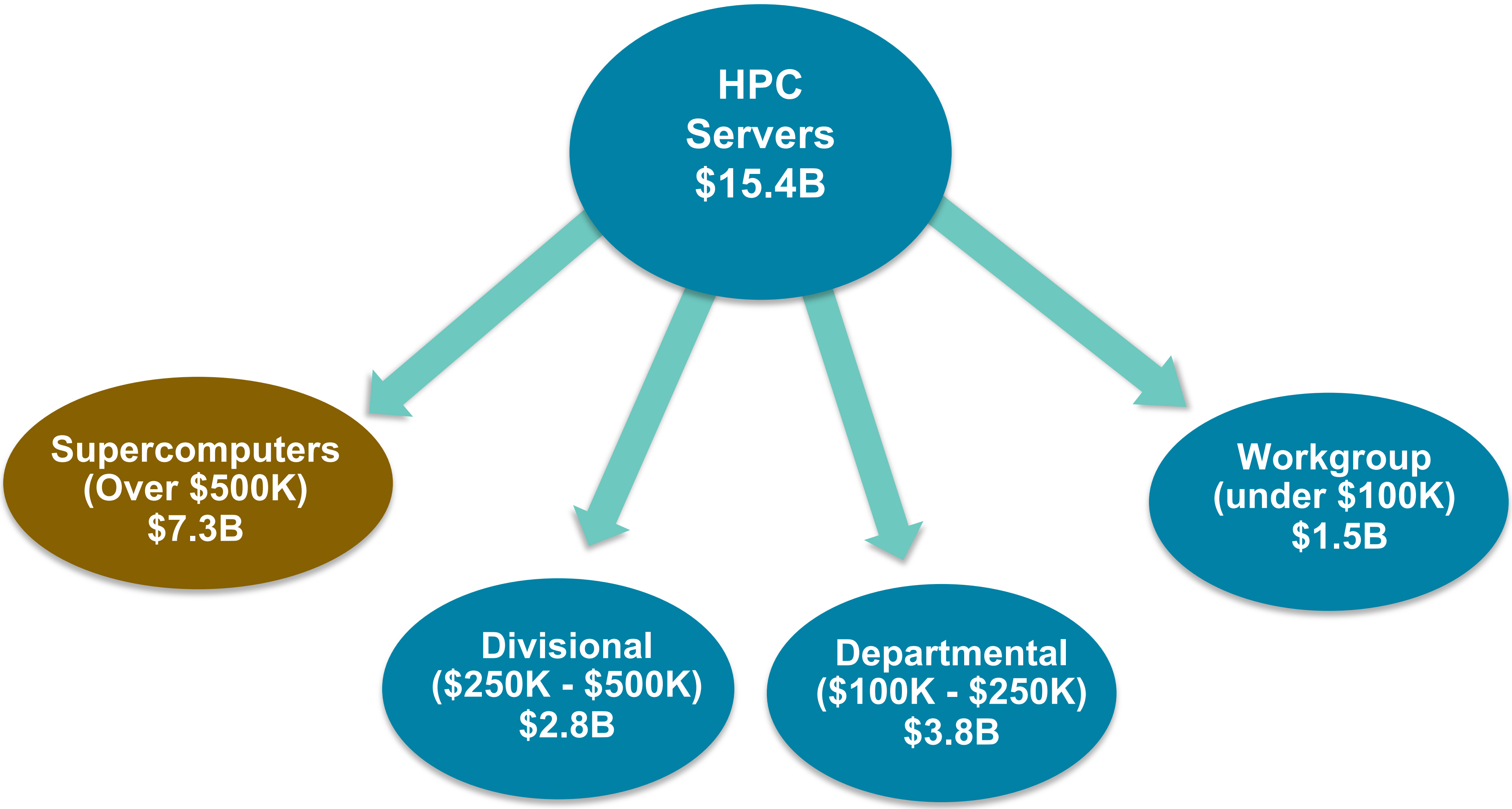
The Overall HPC Market in 2022

Looking at the overall HPC market, including servers, cloud usage, storage, software and repair services = $37.3 billion USD



The 2022 Worldwide On-Prem HPC Server Market: $15.4 Billion (up 4.3%)

2023 is projected to be around $17 Billion



2022 WW HPC On-Prem Market by Vendor and Sector ($ Millions)

HPC On-premises Server Market ($M)	
Vendor	**2022**
HPE	$5,137
Dell Technologies	$3,575
Lenovo	$1,201
Inspur	$1,073
Sugon	$603
IBM	$505
Atos	$480
Fujitsu	$230
NEC	$207
Penguin	$442
Other	$1,988
Total	**$15,441**

Source: Hyperion Research, 2023

HPC On-premises Server Market ($M)	
Sector/Vertical	**2022**
Bio-Sciences	$1,449
CAE	$1,768
Chemical Engineering	$173
DCC & Distribution	$826
Economics/Financial	$757
EDA / IT / ISV	$873
Geosciences	$998
Mechanical Design	$57
Defense	$1,602
Government Lab	$3,342
University/Academic	$2,677
Weather	$700
Other	$221
Total	**$15,441**

Source: Hyperion Research, 2023

The HPC Market Should Grow in 2023

Several exascale systems should be accepted in 2023
AI and cloud spending are growing quickly

- **2023 is forecasted to reach an all-time high of around US $17 billion in on-prem HPC servers with US $33 billion in total on-premises HPC spending**
- **But there are a number of issues:**
  - The overall economy is putting pressure on many buyers
  - Covid and the resulting supply chain issues have been a major concern for 3 years, and are expected to continue to be a problem
  - The lower end of the on-premises market continues to struggle
- **Growth drivers include:**
  - 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
  - New technological developments in AI, processors, etc. are providing many new areas for users to advance their research and engineering
  - Cloud computing is becoming more useful to a larger set of HPC workloads

5-Year On-Prem HPC Server Forecast

6.8% yearly average growth over the next 5 years

5 Year On-premises Server Forecast							
	2021	**2022**	**2023**	**2024**	**2025**	**2026**	**CAGR 21-26**
Supercomputer	$6,971	$7,288	$8,083	$8,926	$9,359	$10,094	7.7%
Divisional	$2,783	$2,804	$3,103	$3,512	$3,680	$3,930	7.1%
Departmental	$3,614	$3,828	$4,047	$4,456	$4,584	$4,889	6.2%
Workgroup	$1,412	$1,520	$1,483	$1,575	$1,593	$1,663	3.3%
Total	$14,781	$15,441	$16,715	$18,468	$19,216	$20,576	6.8%
Source: Hyperion Research, 2023							

The Broader Market ($ Millions)

2022 total HPC spending reached $37B
2026 is projected to exceed $52B

The Broader HPC Market	
	2022
On-premises servers	$15,441
Storage	$6,408
Middleware	$1,790
Applications	$5,092
Service	$2,224
Total On-premises	$30,956
Cloud Spending	$6,304
Total	**$37,260**

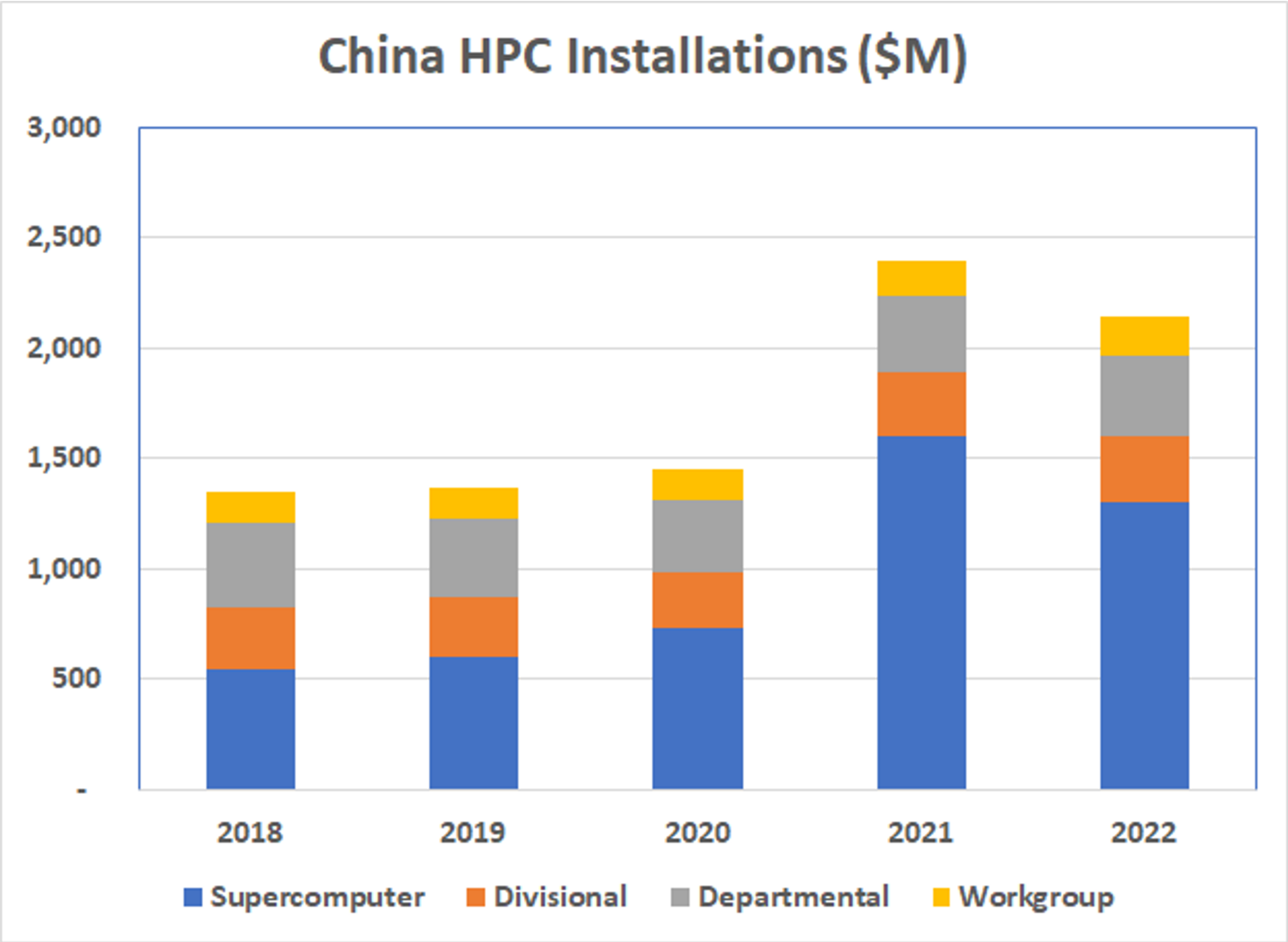
Source: Hyperion Research, 2023

The Broader HPC Market	
	2026
On-premises servers	$20,576
Storage	$9,068
Middleware	$2,281
Applications	$6,349
Service	$2,308
Total On-premises	$40,582
Cloud Spending	$11,613
Total	**$52,195**

Source: Hyperion Research, 2023

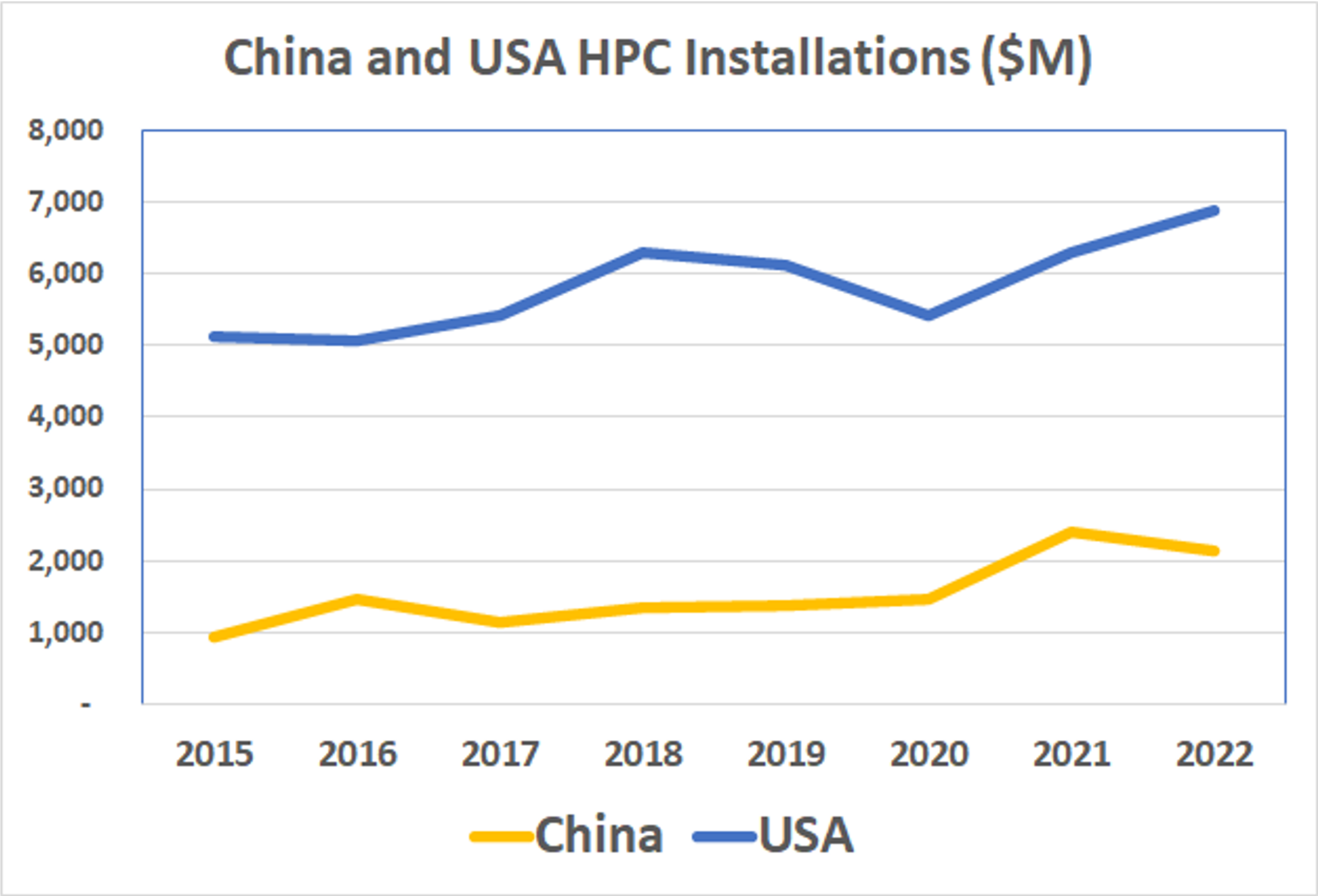
The HPC Market in China

We assume that China now has 3 operational exascale systems
Two in 2021 and one in 2022



The HPC Market in China

China's growth has slowed from the previous decade of high growth



HPC Market Predictions

Hyperion Research's 2023 Predictions

1. **Strong growth in the leadership-class segment will support modest growth across the global on-premises HPC market.**
2. **The advanced computing sector and its associated supply chain will become increasingly driven by national and regional government policies that stress domestic capabilities.**
3. **Sustainability and energy efficiency considerations will become a dominant factor in many procurements.**
4. **Cloud utilization will shift towards production workloads leading to initial erosion of on-premises spending in low end of the market.**
5. **2023 will be the year of AI regulation.**

Hyperion Research's 2023 Predictions

6. AI will become more pervasive in production tier deployments due to users' higher confidence in its abilities and ease of use.
7. HPC system architectures will bifurcate between systems optimized for one set of applications and those designed to address many.
8. Divergent requirements of traditional and modern workloads will move architectural focal points from compute to interconnects and storage systems.
9. Interest in edge computing for HPC will rise in 2023, especially in the industry sector, but spending will be muted.
10. Growth at many HPC sites will be stunted due to the continued difficulty in acquiring and retaining talent.

High Growth Areas

The Exascale Market (System Acceptances)

Over 30 systems and over $10 billion in value

Exascale and Near-Exascale Leadership Systems (2020 to 2027)							
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 (~half in 2022 and half in 23)	--	4	$1.3B
2023	1 exascale system ~$350M	1 or 2 pre-exascale systems ~$150M each	1 near-exascale system ~$150M	1 exascale system ~$600M	--	4-5	$1.3B - $1.4B
2024	1 exascale system ~$350M	1 exascale ~$350M, plus 1 exascale (or pre) system ~$200M	?	1 exascale system ~$600M	1 pre-exascale system ~$200M	5	~$1.7B
2025	1 or 2 exascale system ~$300M each	1 or 2 exascale systems ~$350M each	1 exascale system ~$200M	1 or 2 exascale systems ~$350M each	1 near-exascale system ~$150M	5-8	$1.4B - $2.4B
2026	1 or 2 exascale system ~$300M each	1 or 2 exascale systems ~$325M each	?	1 or 2 exascale systems ~$350M each	1 or 2 exascale systems ~$150M each	4-8	$1.1B - $2.3B
2027	1 or 2 exascale systems ~$250M each	1 or 2 exascale systems ~$300M	?	1 or 2 exascale systems ~$300M each	1 or 2 exascale systems ~$150M each	4-8	$1.0B - $2.0B
Total	**8-11**	**9-13**	**3+**	**7-10**	**4-6**	**31-43**	**$10B - $13B**

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

Source: Hyperion Research, March 2023

94.3% of Sites Have Accelerators in Their Largest System Today

Up from 82.7% in 2021

In Mid 2021

How many co-processors or accelerators are in your largest HPC technical server?

	Responses	Percent
None	23	17.3%
Less than 32	28	21.1%
32 to less than 64	18	13.5%
64 to less than 100	19	14.3%
100 to less than 500	18	13.5%
500 to less than 1,000	11	8.3%
1,000 to less than 5,000	10	7.5%
5,000 to less than 10,000	4	3.0%
10,000 or more	2	1.5%
n = 133		
Source: Hyperion Research, 2021		

In Late 2022

Largest System Accelerator Count

Q: How many compute-oriented accelerators/co-processors are in your largest on-premises HPC technical server?

		Overall Percent
None		5.7%
Less than 32		24.4%
32 to less than 64		15.3%
64 to less than 100		12.5%
100 to less than 500		13.1%
500 to less than 1,000		7.4%
1,000 to less than 5,000		7.4%
5,000 to less than 10,000		2.8%
10,000 to less than 50,000		2.3%
50,000 to less than 100,000		4.0%
100,000 to less than 250,000		3.4%
250,000 to less than 500,000		0.6%
750,000 to less than 1,000,000		0.6%
1,000,000 to less than 5,000,000		0.6%
n = 176; 104; 20; 52		
Source: Hyperion Research, 2023		

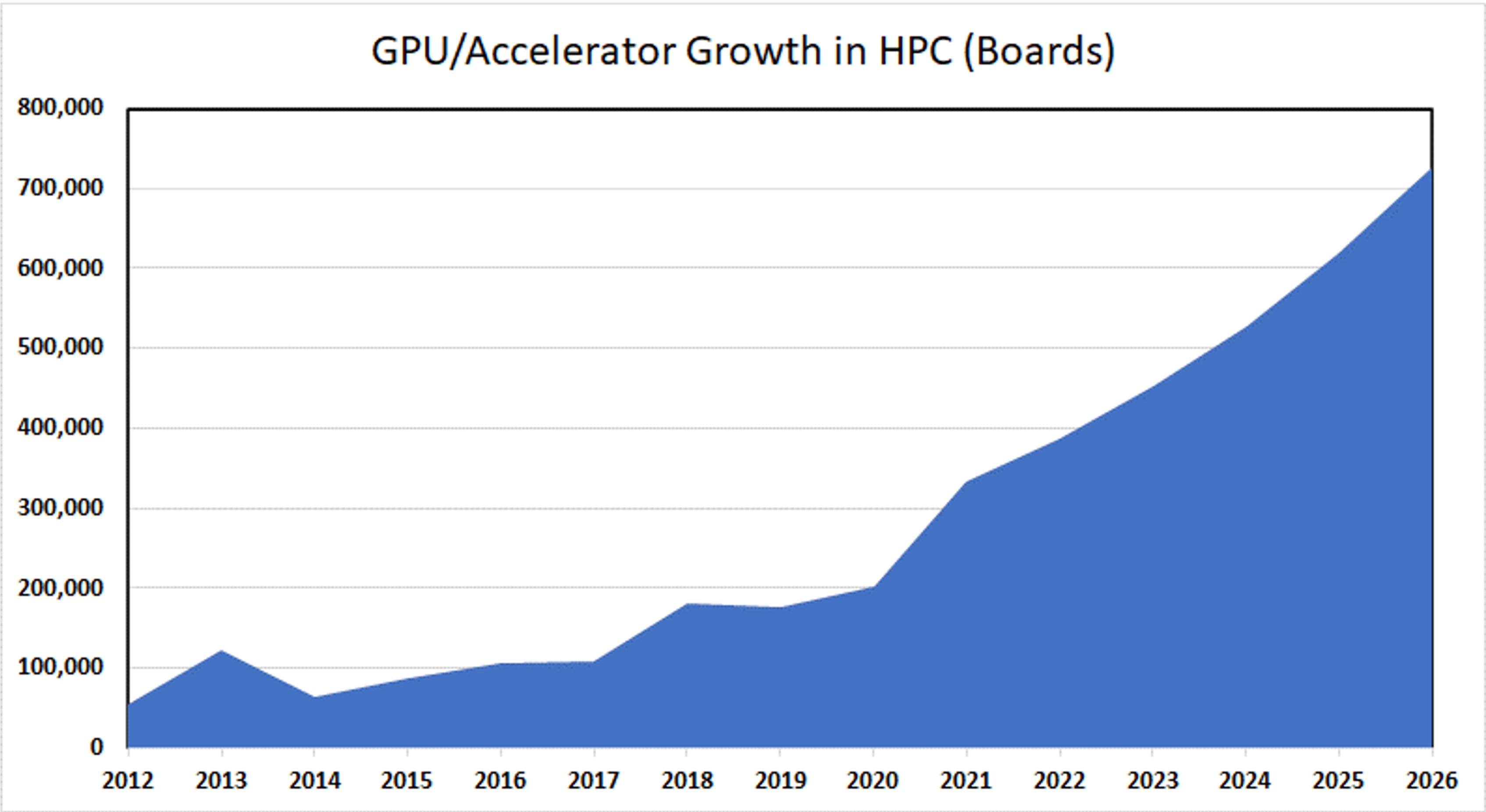
Accelerator Plans for Next Purchases

From our recent end-user MCS study

Planned Processing Elements by Sector				
Q: In the next 12 – 18 months, which of these processing elements do you expect will be incorporated into your HPC/AI/HPDA compute resources? Select all that apply:				
	Overall Percent	**Industry Percent**	**Government Percent**	**Academia Percent**
GPUs	74.0%	67.9%	85.0%	82.7%
TPUs (tensor processing units)	24.3%	27.5%	25.0%	17.3%
FPGAs	22.7%	28.4%	15.0%	13.5%
Single-purpose AI processors	11.0%	12.8%	5.0%	9.6%
ASICs	8.3%	11.9%	0.0%	3.8%
Neuromorphic processors	7.7%	9.2%	10.0%	3.8%
eASICs	2.2%	3.7%	0.0%	0.0%
Other	2.8%	2.8%	0.0%	3.8%
None	5.5%	7.3%	5.0%	1.9%
n = 181; 109; 20; 52				
Source: Hyperion Research, 2023				

GPU/Accelerator Forecast

Anticipated high growth for accelerators over next 5 years

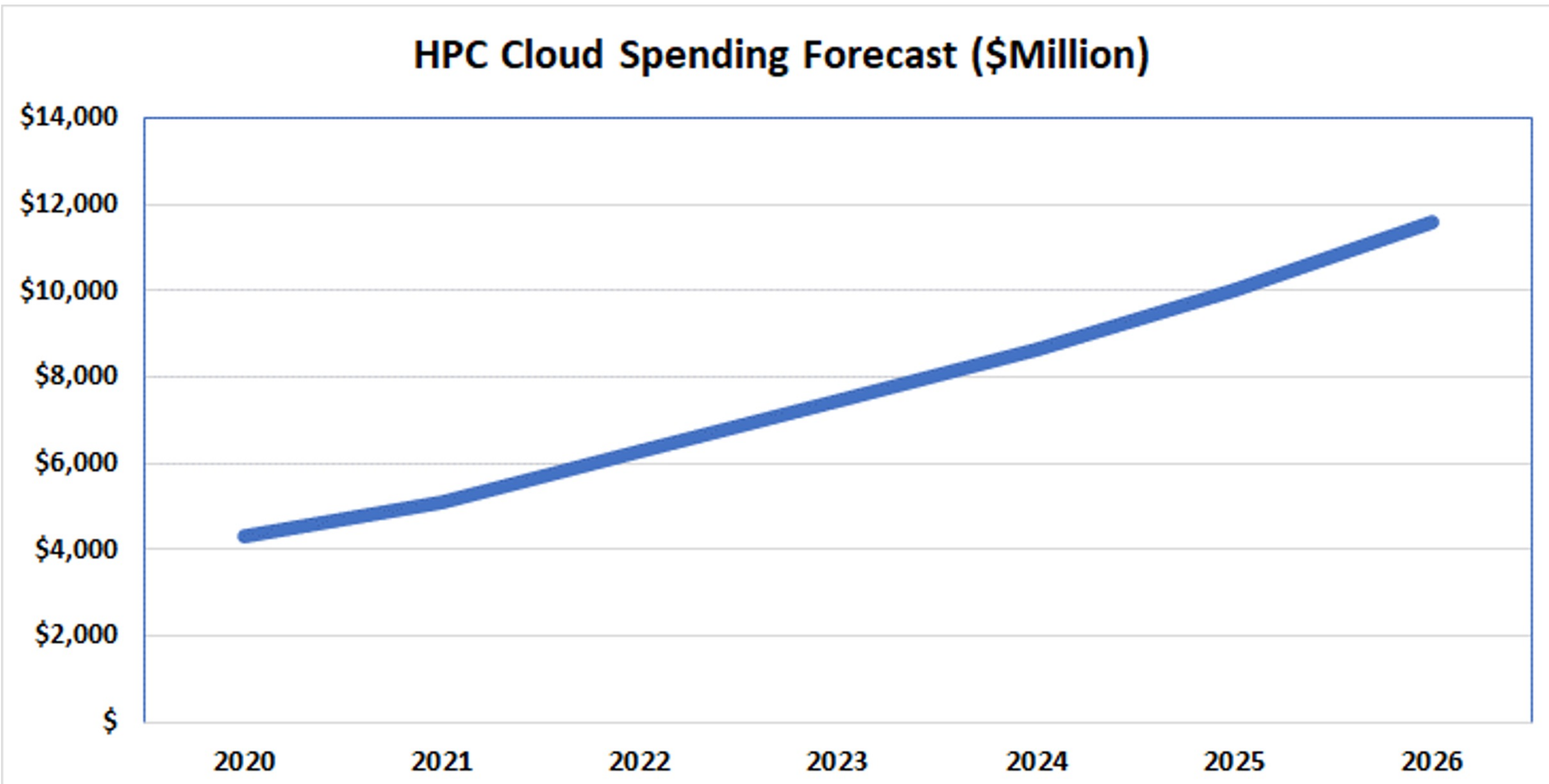


HPC Cloud Usage Forecast

17.9% growth over the next 5 years

HPC Cloud Spending ($ Million)								
	2020	2021	2022	2023	2024	2025	2026	CAGR 21 to 26
HPC Cloud Spending	$4,300	$5,100	$6,304	$7,472	$8,630	$10,011	$11,613	17.9%

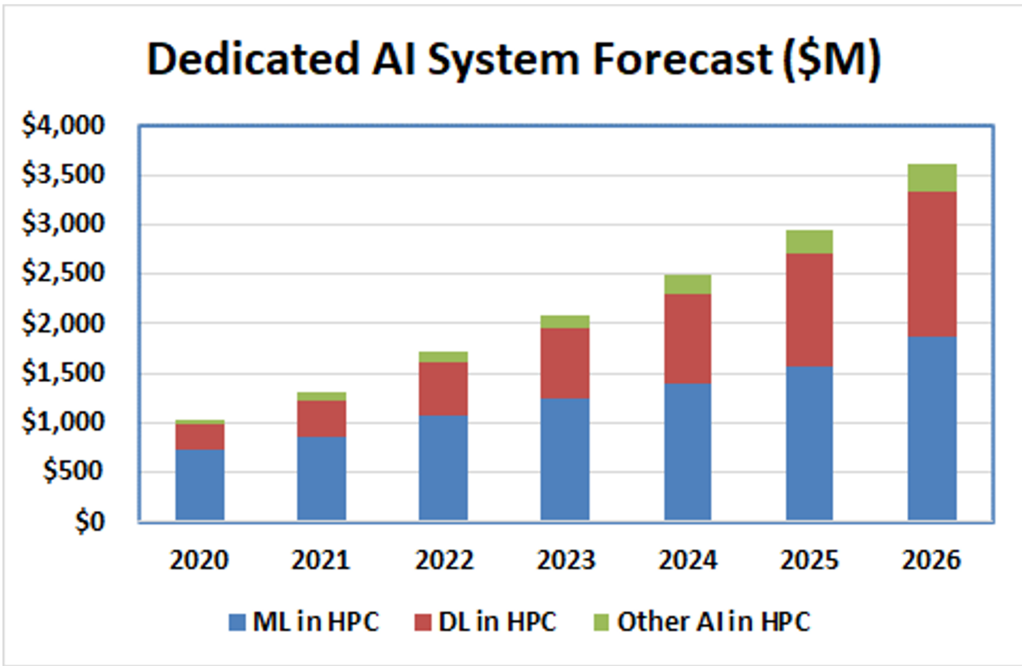
Source: Hyperion Research, 2023



AI Forecast

22.7% 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**	**CAGR 21-26**
ML in HPC	$719	$861	$1,081	$1,243	$1,391	$1,568	$1,859	16.6%
DL in HPC	$263	$364	$532	$708	$919	$1,147	$1,468	32.2%
Other AI in HPC	$57	$75	$104	$132	$173	$226	$292	31.3%
Total AI Server Revenue	$1,039	$1,300	$1,718	$2,083	$2,484	$2,941	$3,619	22.7%
Source: Hyperion Research, 2023								



Observations About Architectures Changes

HPC System Architecture Changes

7. *HPC system architectures will bifurcate between systems optimized for one set of applications and those designed to address a myriad of applications*

- **Future sites will move to have a broader mix of supercomputers**
  - Most will still have a large central system
  - But will invest more in differently focused secondary systems
- **Future system designs for HPC users will have to factor in new requirements:**
  - New workloads, like AI and big data
  - A push for faster time to solution
  - New areas of research
  - New anticipated scale of data and computation
- **Major system decisions will split between:**
  - Support of a much larger and diverse set of building blocks
  - Single, heterogeneous system to address wide set of applications
  - Multiple, smaller systems with specific applications in mind
  - Public clouds for specific sets of applications

HPC System Architecture Changes

7. HPC system architectures will bifurcate between systems optimized for one set of applications and those designed to address a myriad of applications

- **Heterogeneous systems will incorporate:**
  - Data intensive vs. processing intensive designs
  - A variety of node configurations to incorporate accelerators and expanded memory profiles
  - Infrastructure accelerators, like DPUs, to offload some processes from processors/accelerators
  - Complex storage infrastructure to address different I/O profiles of workloads
- **Smaller systems will be designed to target applications like AI, Big Data, or traditional modeling/simulation**
  - AI systems will most likely have more accelerated nodes
  - This scenario requires data centers to be knowledgeable of the requirements of novel and established applications

Storage and Interconnects: A New Architectural Focal Point

8. The divergent requirements of traditional HPC modeling/simulation and AI workloads will move HPC architectural focal points from compute to system interconnects and storage systems.

- **Internode system interconnects will be critical for performance and scalability of composable system elements**
  - InfiniBand and Ethernet dominance is expected to continue
  - Other and new technologies are gaining some traction
  - Trade-offs will be made between converged storage and MPI fabric and independent node-storage and node-node networks
- **Intranode interconnects such as CXL are emerging to address composable memory**
- **Storage architectures are evolving to address broad challenges across the entire ecosystem**
  - 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)
  - Enforced consistency (strict POSIX vs. relaxed POSIX)

From a Recent Hyperion Research Study on Interconnects

- **Optical I/O was rated as the technology that has the highest potential to improve HPC architectures over the next 2-6 years**
  - By both users and vendors
- **In-memory computing was identified as the second highest impact technology area**
- **75% of respondents felt that there is a strong need for disaggregation of system resources to enable workload-driven composable infrastructure**
  - Also ranking high as a high-impact technology are physical interface standards for chiplets (e.g. UCIe) that can enable standardized connection between the host SoC and in-package optical I/O chiplet
- **Predominant system issues for future architectures: system scale-out, lack of system composability, and network throughput**

What is Disaggregation and Composability?

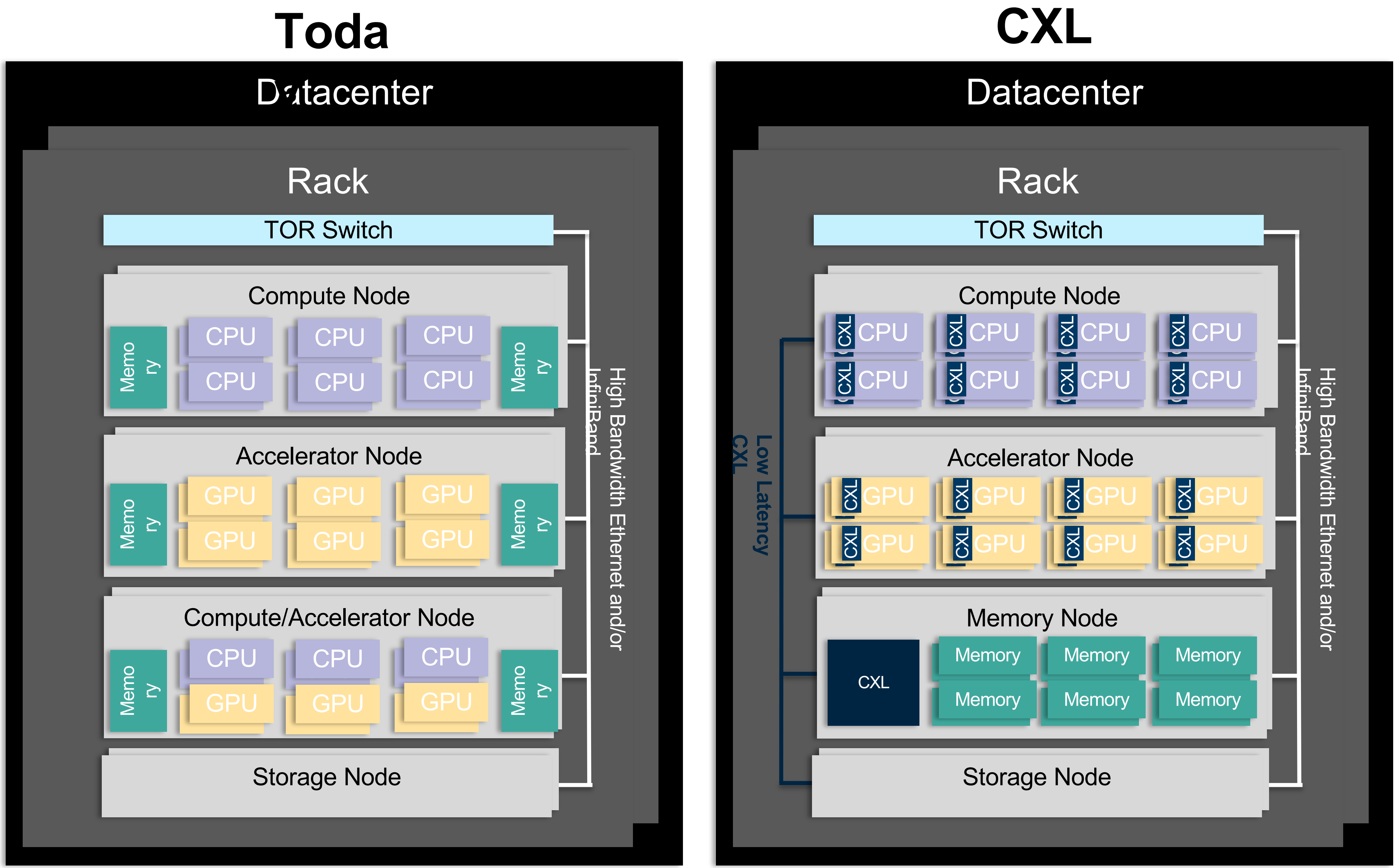
What is Disaggregation and Composability?

- **Disaggregation**
  - An architectural paradigm that moves system elements typically integrated together □ to their own respective element-specific subsystems
    - And are networked together to create a system
  - System elements primarily include CPUs, GPUs and memory
- **Composability**
  - Dynamic allocation and provisioning of system resources based on the requirements of <u>individual</u> jobs and workloads
  - System elements are reserved independently of each other based on each specific job
  - Leverages emerging innovations such as Compute Express Link (CXL), and advancements in existing interconnect standards

Workflow in Today’s Architectures

- **Typical steps:**
  - Users submits a job to the scheduler
  - Scheduler determines resource requirements and identifies when all will be available
  - Resources (e.g., compute, memory) are locked in fixed, aggregate amounts for the duration of the job
  - Resources are released for the next job once the current job completes
- **Challenges**
  - Resources are allocated in fixed amounts based on the extent of the most critical element, oftentimes leaving other resources idle
  - Jobs remain in the queue that could otherwise run on the idle resources
- **Proposed solution: disaggregation and composable systems**

HPC Architectures – Before and After CXL



Why Now for Composability?

- **Protocols have been developed or are on the horizon to <u>support interoperability across different vendors' components</u> that are being composed**
  - Compute Express Link (CXL): Cache-coherent interconnect for processors, memory expansion and accelerators
- **<u>Interconnects are becoming performant enough for many applications to support disaggregation</u> of multiple system resources (e.g., memory, CPUs, GPUs) to overcome increased latencies introduced by composability**
- **<u>Job types and component requirements are changing frequently</u>**

HYPERION RESEARCH

Major Benefits Of Disaggregation/ Composable Architectures

Major Promised Benefits

Lower costs, higher utilization, faster job turnaround times

1. **Lower costs for purchasing future systems**
   - Don't need to buy as much hardware
   - Simplified system expansion and reduced system costs via modular resource-specific nodes
2. **Reduced que wait times**
   - Fit more jobs within a given pool of resources
   - Improved system utilization by more fully leveraging expensive on-premises assets
   - Accelerated time to completion for workloads that would otherwise be sitting idle in a queue
3. **Improved system utilization**
   - More work completed per unit of time

The intent is to implement it in a way that doesn't require major software or application changes

Promised Benefits

Lower costs & wait times

1. **Save costs by needing fewer parts/devices, e.g., one can buy fewer GPUs and, in some cases, fewer CPUs to provide the same amount of work**
   - Or one can buy a more capable system for a given budget
   - Its easier to add additional parts, e.g., more GPUs, different accelerators, different types of CPUs, into an existing system or center
2. **Jobs can run faster, by more quickly getting a more optimal mix of parts for each application job run**
   - Each job only locks in the resources needed at a device level (CPUs, GPUs, storage, …) and not a whole node
   - Unused resources are freed up more quickly

Promised Benefits

Better utilization

3. **Increased overall system utilization & number of jobs ran per unit of time**
   - Both increase system efficiency and lower costs
   - Allows moving away from always having a fixed ratio of CPUs and GPUs assigned to a job
     - Each job can get the specific mix of hardware and only use what's needed for only as long as needed
   - More quickly free up resources/devices
   - Reduce the amount of idle hardware
   - More easily mix and match hardware with each job run

Note: System utilization can be 85% to 95% while components can be at 10% to 20% -- the promise is to greatly improve utilization at the device level, in addition to at the system level

Other Observations

System expansion can be done more easily

- **Systems can be expanded by adding "only" the specific resources needed (more memory, or more GPUs, or more CPUs, etc.)**
- **It can improve reliability & maintainability**
    - If a GPU breaks, it is directly taken out of the pool
    - Additional hardware can more easily be added to a system:
        - e.g., if workloads starts to need more of a specific GPU or CPU, those can be added as an additional rack
        - New types of hardware can be more easily added to an existing system
- **It can help improve sustainability – by better use of fewer resources**

Accelerated Time to Job Completion

- **Jobs can run faster, by more quickly getting a more optimal mix of parts for each application job run**
  - Each job only locks in the resources needed at a device level (CPUs, GPUs, storage, …) and not a whole node
  - Unused resources are freed up more quickly
- **Queue can be significantly reduced**

What's Holding Back Adoption

There Are Some Critical Open Questions/Concerns

- **Disaggregation requires high levels of intra-resource communication**
  - Including stringent requirements for ultra-low latency and ultra-high transmission bandwidth
- **Some key questions:**
  - When, where, and to what extent does disaggregation make sense for HPC systems?
  - Will CXL, a cache-coherent interconnect for data centers, be deployed widely in HPC?
  - Will large-scale supercomputers be disaggregated beyond rack-scale?
  - Should we disaggregate main memory?
  - What is the future of optical I/O, and how fast will it be adopted?

Concerns

Unclear performance and costs

1. **Will it actually perform well – given that resources are farther away?**
2. **Does it actually work at scale?**
   - And for hard HPC jobs?
3. **How much more will this actually cost?**
   - It's adding an additional network and additional software
   - What will be the true savings vs. additional costs?
4. **How hard will this be to support?**
   - Data centers and HPC systems are already very complex and getting more complex with each new generation
   - How much technical support will be needed?

Composability Challenges

- **Resource impacts**
  - How much (if at all) will application codes need to change to support composability?
  - Will it increase or reduce support requirements?
- **Performance impacts**
  - How much latency must be added to manage, provision, monitor, and re-claim system resources between jobs?
  - Will increased physical distance also add latency?
  - How far can it scale?
- **Usage and operational impacts**
  - How to determine workloads most suitable to composability?

Composability Usage and Operational Considerations

HPC Usage & Operational Considerations	Conditions Amenable to Composable System	Examples
Utilization	• Overall low system utilization • Resource bottlenecks that lock up idle system resources for long periods of time • Mismatched resource allocation	• 24-hour wall clock system utilization < 50% • Long storage access delays that idle CPUs • Low count GPU jobs running on high GPU count nodes
Scale	• Small-to-medium scale systems • Jobs with minimal data dependencies and minimal interprocessor communication	• Single processor or single node jobs • Financial risk modeling, drug discovery, big data analysis, imaging
Performance	• Specific requirements for one or more system resource	• AI applications running GPU intensive jobs, CPU intensive CFD models
Workload	• Short-to-medium run times • Small to medium size jobs • Cloud-friendly	• R&D software development runs test codes, test bed or devops programs • Highly parallel, limited data, modular codes
Talent	• Sites with limited on-site HPC expertise • Sites running new workloads that aren't dependent on legacy codes	• New compute facilities, academic sites, start-ups facilities

Vendor Adoption/Intertest -- Today

- **Broad ecosystem emerging, particularly with CXL**
  - 15-member Board of Directors
  - 70 contributors
  - 130 adopters
- **Innovations occurring alongside CXL**
  - Existing interconnects and protocols
  - Augmentation
- **HPC community is represented at all levels:**
  - Compute
  - Systems
  - Interface
  - Networking
  - Memory
  - Software
  - Storage

Composability Ecosystem

Broad industry representation today

System Element	Contributors
Standards and Consortia	Compute Express Link (CXL)*, Ethernet Technology Consortium, InfiniBand Trade Association (IBTA)
Compute	AMD, ARM, Intel, NVIDIA, SiPearl
Systems	Dell, HPE, Huawei, IBM
Interface	Broadcom, IntelliProp, Marvell,
Networking	Ayar Labs, Cornelis, GigaIO, NVIDIA, Rockport
Memory	Micron, Rambus, Samsung, SK Hynix
Software	Google, Liqid, MemVerge, Microsoft
Storage	Seagate, Western Digital

Industry Adoption – What's Needed?

- **CXL 3.0 for full cache coherency (or alternative solution <u>support by multiple vendors</u>)**
- **Availability of interoperable components <u>from multiple vendors</u> for each element in the solution stack**
  - CXL interfaces
  - CPU
  - GPU
  - Memory
  - Interconnect
- **<u>Application proof points</u> to confirm:**
  - Performance on HPC applications
  - Supported scale
  - Breadth of workload applicability

Upcoming Research

HYPERION RESEARCH

Special Report

Perspectives on Composable Systems and HPC/AI Architectures

Mark Nossokoff, Bob Sorensen, and Earl Joseph

April 2023

HYPERION RESEARCH OPINION

Traditional HPC architectures have been designed to address either homogenous workloads (such as physics-based modeling and simulation) with similar, and perhaps more important, fixed, compute, memory, and I/O requirements or, more recently, heterogenous workloads with a diverse range of compute, memory, and I/O requirements. Most HPC data center planners and operators, however, don't have the luxury of focusing on one main type of workload; they typically must support a large number of HPC users and their associated workloads sporting a wide range of compute, memory, and I/O profiles. Ensuing architectures typically, then, consist of a fixed set of resources, resulting in an underutilized system with expensive elements sitting idle a costly and unacceptable amount of time. One approach being explored to increase system utilization by exposing resources that would otherwise sit idle to appropriately matched jobs waiting in a queue is via composable systems.

Closing Thoughts on Composability

- **Driving factors:**
  - It works well in the non-HPC world
  - Complex, heterogenous modern workloads will continue to stress existing system architectures
    - Incorporating new or different amounts of CPUs, accelerators, memory will be needed more often than full system upgrades
  - Storage, interconnects, and data management will grow in importance for future architectures
  - Increasing interdependence between complexities of new workloads (e.g., AI), access to resources at scale (e.g., cloud), and user demands for accelerating time to results
  - A reduction in available HPC talent

- **These are creating an opportunity for composability**

Overall Conclusions

- **2022 was a soft growth year with a 4.3% increase**
  - 2023 is expected to be a moderate growth year
    - Exascale systems will drive growth in 2023 & 2024
    - GPUs, cloud, AI/ML/DL & big data are high growth areas
- **New technologies are showing up large numbers:**
  - Processors, AI hardware & software, memories, new storage approaches, Quantum, etc.
  - Composability may fit well in certain applications
- **The cloud has become a viable option for many HPC workloads**
  - HPC in the cloud is lifting the sector writ large
- **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**

A Concern: HPC 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**
  - Increasing HPC systems per site
  - 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
  - And 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

We Welcome Questions, Comments And Suggestions

Please contact us at:

info@hyperionres.com