New Directions in HPC Storage and Interconnects

April 2023

www.HyperionResearch.com
www.hpcuserforum.com

Mark Nossokoff

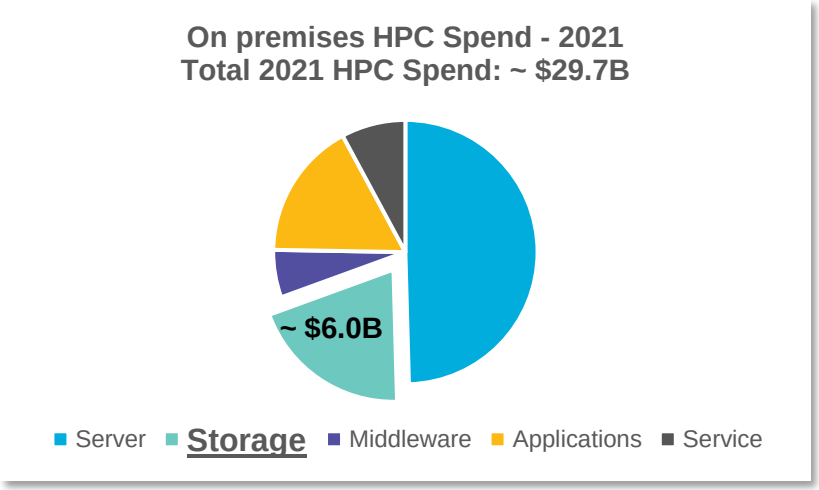
Contents

- **Storage Market Overview**
  - On-premises
  - Cloud
  - Prediction - Storage and Interconnects: A New Architectural Focal Point
- **“Recent” Happenings**
  - DPUs
  - Interconnects
  - Standards and protocols
- **Composability**
  - What is it?
  - Why now?
  - Is it for everyone?

© Hyperion Research 2023

HPC Storage Growth Continues

Demand increasing across all sectors and verticals



Source: Hyperion Research, 2022

- **Storage historically the highest growth HPC element**
- **Storage represents ~ 20% of on-premises HPC spending and growing**
- **Almost half of sites surveyed expect their storage budgets to increase more than 5%**

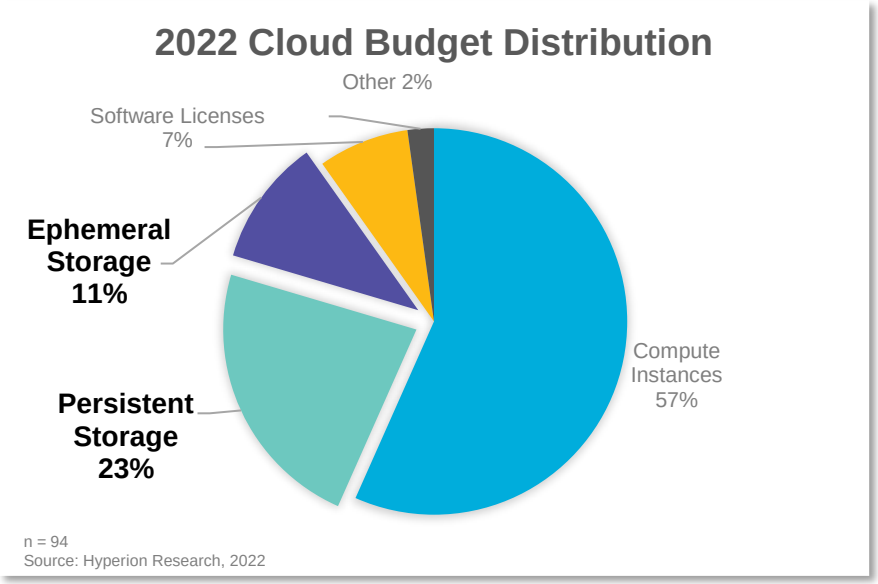
Area ($M)	2021	2022	2023	2024	2025	2026	CAGR 21-'26
Server	$14,748	$16,077	$17,738	$19,565	$19,495	$20,481	6.8%
Add-on Storage	**$5,971**	**$6,677**	**$7,457**	**$8,388**	**$8,491**	**$9,027**	**8.6%**
Middleware	$1,729	$1,863	$2,030	$2,212	$2,172	$2,268	5.6%
Applications	$4,948	$5,302	$5,731	$6,195	$6,089	$6,326	5.0%
Service	$2,266	$2,316	$2,389	$2,468	$2,336	$2,296	0.3%
Total Revenue	$29,662	$32,236	$35,345	$38,828	$38,584	$40,398	6.4%

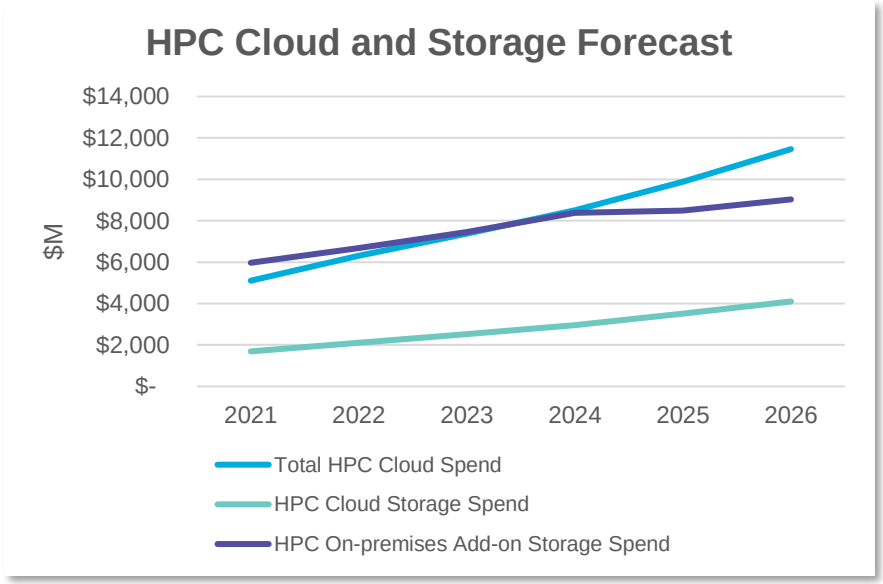
Source: Hyperion Research, 2021

© Hyperion Research 2023

HPC Storage and the Cloud

Cloud adoption for storage remains strong and growing





- **Storage ~ 1/3 total HPC spending in the cloud**
- **Spending on persistent, durable storage 2x greater than ephemeral, temporal storage**

- **~ $1.7B cloud storage spend in 2021**
- **Cloud storage growth ~ 2.3x on-premises storage growth**
- **Total cloud spending projected to overtake on-premises storage spending in 2024**

© Hyperion Research 2023

Storage and Interconnects: A New Architectural Focal Point

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)

© Hyperion Research 2023

“Recent” Happenings

- **Consolidation of and interest in DPUs**
  - AMD acquisition of Pensando (old news)
  - Microsoft acquisition of Fungible assets (recent news)
  - NVIDIA Networking Bluefield 3 (on-going investment)
  - Intel IPU
- **Interconnect evolutions**
  - InfiniBand, Ethernet, and OmniPath evolution (line rates, features)
  - Vendor augmentations and innovations
    - HPE Slingshot
    - Rockport
  - Captive CSP investments
    - AWS: EFA
    - Google: Aguila, Apollo
    - Alibaba: collaboration with Broadcom
- **Disaggregation and Composability**

© Hyperion Research 2023

Why now for composability?

Advancements occurring to address workload complexities

- **Heterogenous modern workloads exacerbating stranded resources and costly under utilization**
- **Protocols have been developed or are on the horizon to support interoperability across different vendors' components that are being composed.**
  - Compute Express Link (CXL): Cache-coherent interconnect for processors, memory expansion and accelerators.
- **Interconnect features and performance have evolved to support disaggregation of multiple system resources (e.g., memory, CPUs, GPUs)**

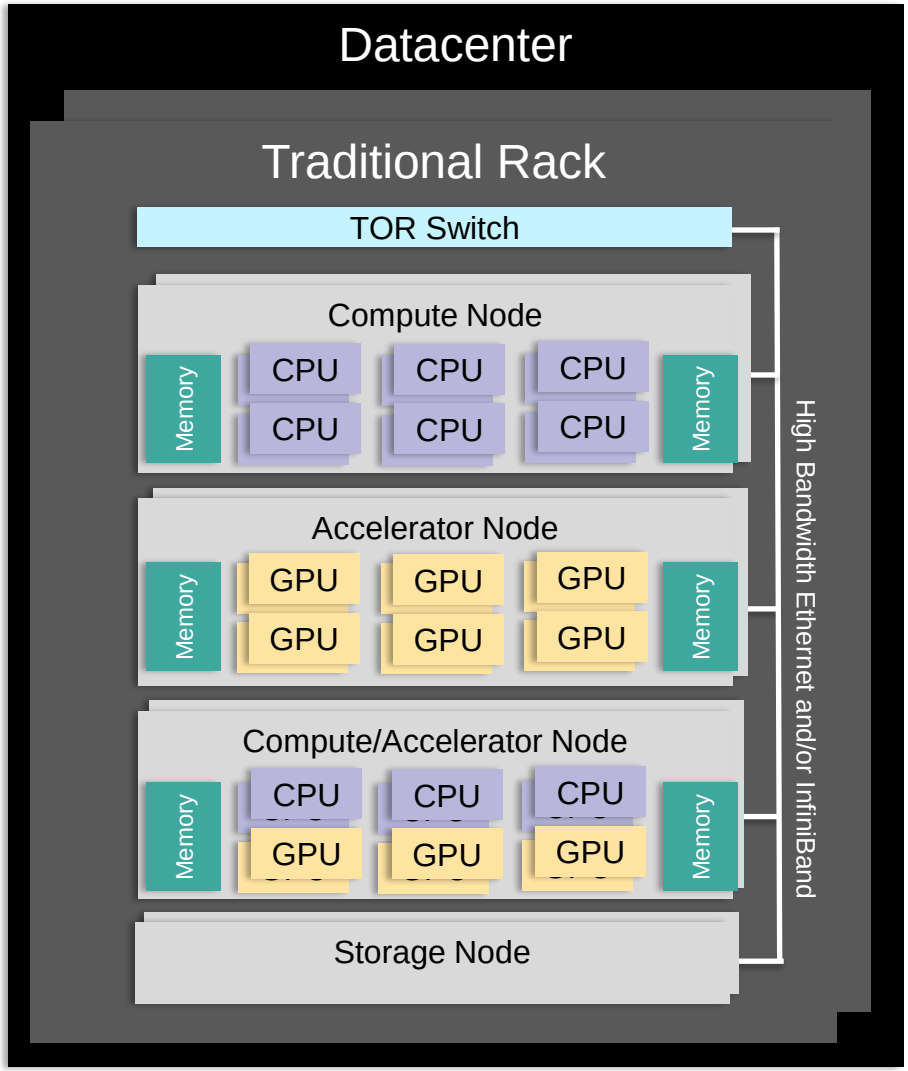
© Hyperion Research 2023

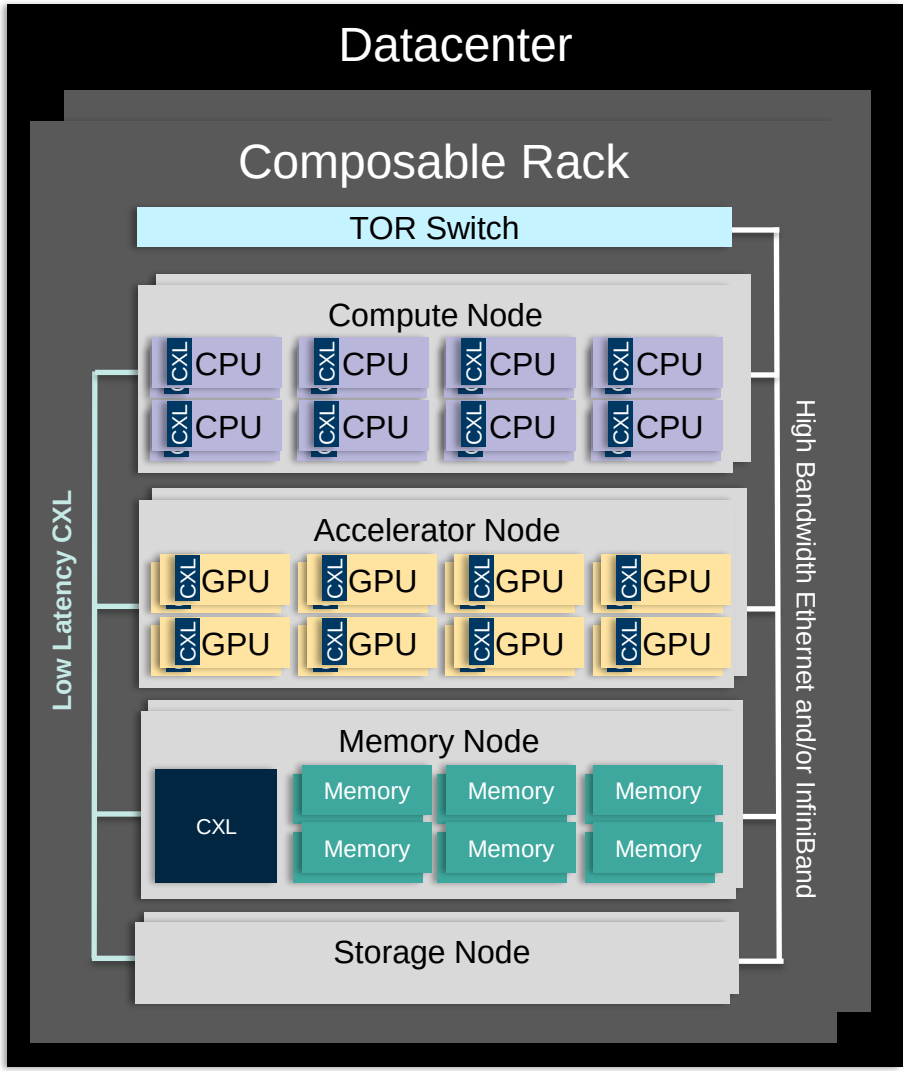
What are Disaggregation and Composability?

- **Disaggregation**
  - An architectural paradigm that moves system elements typically integrated together in a scalable turnkey node moves them into their own respective element-specific subsystems to then be networked together to create a complete solution
  - System elements include CPUs, GPUs, and memory
- **Composability**
  - Dynamic allocation and provisioning of system resources based on the requirements of individual 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

© Hyperion Research 2023

Traditional and Composable Architectures





© Hyperion Research 2023

Composability Benefits

- **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
- **Simplified system expansion and reduced system costs** via modular resource-specific nodes

© Hyperion Research 2023

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 might 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?
- **Cost impacts (e.g., an additional network)**
- **Usage and operational impacts**
  - How to determine workloads most suitable to composability?

© Hyperion Research 2023

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

© Hyperion Research 2023

Composability Ecosystem

Broad industry representation

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

© Hyperion Research 2023

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.

© Hyperion Research 2023

Closing Thoughts

- **Complex, heterogenous modern workloads will continue to stress existing system architectures**
- **Storage, interconnects, and data management grow in importance for future architectures**
- **Increasing interdependence between complexities of new workloads (e.g., AI, quantum), access to resources at scale (e.g., cloud), and user demands for accelerating time to results**

© Hyperion Research 2023

HYPERION RESEARCH

Questions?

We welcome questions, comments, and suggestions

Please contact us at:
mnossokoff@hyperionres.com

© Hyperion Research 2023