State of HPC Storage and Interconnects

SC23 Virtual Breakfast Briefing November 2023

Mark Nossokoff

www.HyperionResearch.com
www.hpcuserforum.com

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.
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.

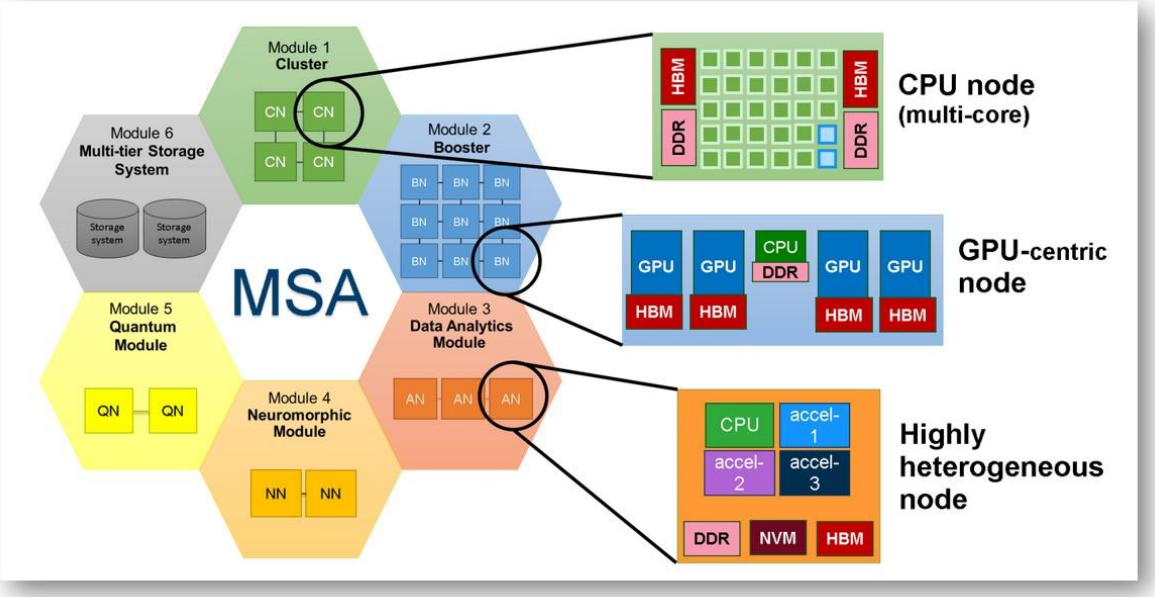
HPC System Architecture Changes

HPC system architectures will bifurcate between systems optimized for one set of applications and those designed fora myriad of applications

- **Future system designs for HPC users will factor in new requirements:**
  - New workloads, like AI and big data
  - 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
- **Heterogeneous systems will incorporate:**
  - Data intensive vs. processing intensive designs
  - Node configurations to multiple accelerators and expanded memory profiles
  - Infrastructure accelerators (e.g., DPUs), to processes from CPUs/GPUs
  - Complex storage infrastructure to address different I/O profiles
- **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

EU Plans for First Exascale System

Large HPC systems will be built with very diverse hardware building blocks



Source: https://www.fz-juelich.de/en/ias/jsc/about-us/structure/divisions/technology-division/next-gen-arch-proto/msa

- **Potential RFP benchmarks (24 in total) hinting at anticipated workloads**
  - Traditional Computing: Graph 500, HPCG, HPL, Stream
  - Accelerated: NekRS, a GPU Navier-Stokes Solver
  - Quantum
    - JUQCS,JUQCS--G
  - AI
    - AI-NLP (GPT)
    - AI-CNN (ResNet)

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
  - Shift from independent node-node and storage networks to converged 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)

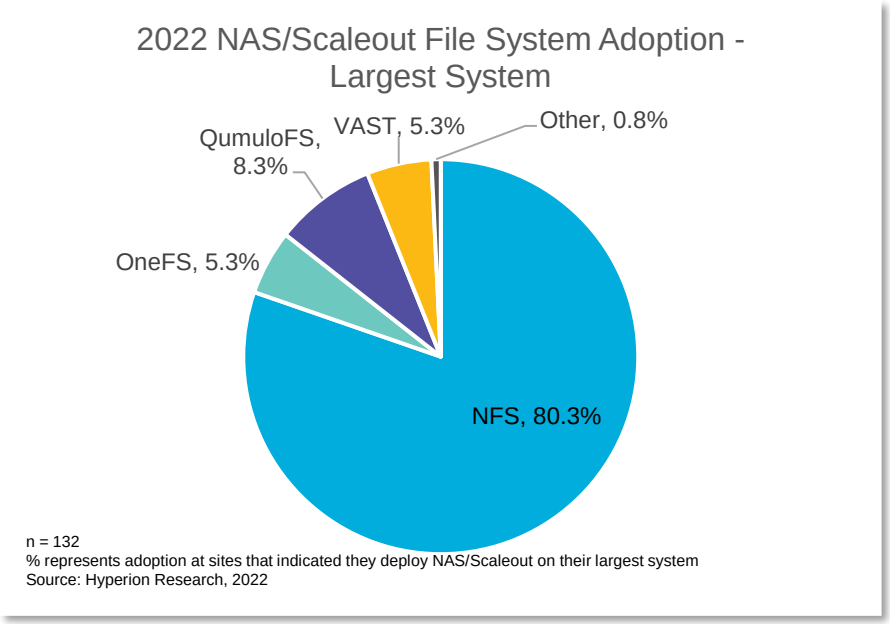
Large Implications for File Systems

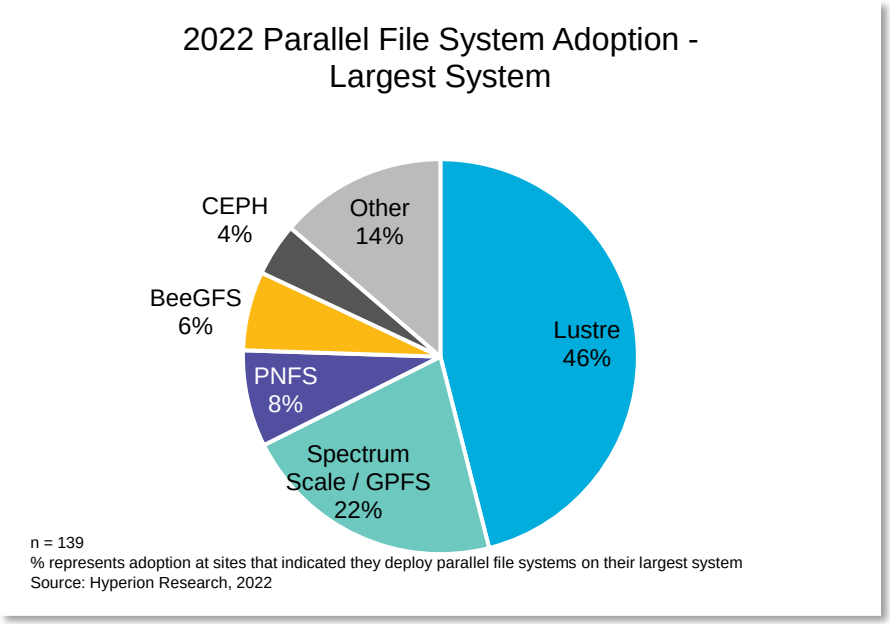
Lustre, Spectrum Scale, NFS currently dominate

- **Systems typically require multiple file systems to address the variety of I/O profiles**
  - Compute-intensive vs. data-intensive
  - IO profiles
  - Access methods
  - Access frequency
  - Locality
  - Enforced consistency
- **CSPs providing support for parallel file systems**
  - AWS: FSx for Lustre
  - Google: Parallelstore (based on DAOS)
  - Microsoft Azure: Managed Lustre Service
- **Advancements are occurring in each file systems but not equally across each area**
- **In search of the elusive global parallel file system**
  - Captive storage system investments
  - Independent start-ups

File System Preferences – Largest System

Large % of sites adopt both NAS/scaleout and parallel file systems on their largest systems





- ***Note: Data is from 2022 Global Site Survey. 2023 Global Site Survey will launch in December.***

Interconnect Snapshot Since ISC23

- **Heightened Regional Focus**
  - EuroHPC call for *Innovation Action in Low Latency and High Bandwidth Interconnects*
- **Many moving parts within the interconnect ecosystem**
  - NVIDIA IB, Ethernet, DPU,
  - Broadcom Jericho3 AI switch and PCIe Gen6&7 roadmap
  - Rockport Networks re-branding as Cerio and releases new PCIe platform
  - Cornelis Networks continuing OPA roadmap investment
  - CSPs and hyperscalars making internal investments
  - HPE Slingshot maturing (Frontier, Aurora)
  - Strong ecosystem and broad emergence of products for CXL
- **Ethernet gaining momentum for increased adoption in HPC/AI**
  - Evolution driven by AI impacts to networks
  - Ultra Ethernet Consortium
  - ETH Zurich research
- **Chiplet interconnects**
  - UCIe
  - OCP BoW
  - Start-up innovations
- **Further investment in and deployment of optical I/O**
  - Ayar Labs funding
  - Lightmatter funding
  - Lightelligence products and solutions
  - Google Apollo OCS (Optical Circuit Switching)

EuroHPC call for *Innovation Action in Low Latency and High Bandwidth Interconnects**

Support the R&I technology development of innovative and competitive European HPC inter-node interconnect technology

- **Call open from August 1, 2023 – January 31, 2024**
- **Proposals should outline ability to:**
  - Develop a roadmap for European scalable inter-node interconnects targeting HPC exascale and post-exascale systems. The roadmap should take into account the EuroHPC supported work in this area such as the components being developed in the EuroHPC RED-SEA project as well as in the area of processors and accelerators.
  - Develop the inter-node interconnect hardware addressing design, development, testing and tape-out as well as integration in test-beds. The work should foster synergies with the EuroHPC supported work in the area of processors and accelerators.
  - Develop the software, installation, configuration and management tools for the developed interconnect, driven by the needs of relevant HPC workflows and application requirements.
  - Address issues like high bandwidth, low latency, power efficiency, virtualisation, scalability, reliability, security, etc.

* Source: https://eurohpc-ju.europa.eu/innovation-action-low-latency-and-high-bandwidth-interconnects_en

Report Card

High marks for both predictions

- **Taking the following together…**
  - Jupiter's architecture
  - Industry and ecosystem happenings since ISC23
  - Identification of interconnect technology of key strategic importance to EuroHPC

- **…supports pretty high grades for both the system architecture prediction and the storage and interconnect prediction.**
- **Looking towards 2024**
  - Continued and expanded storage and interconnect research
  - A new slate of predictions in January

HYPERION RESEARCH

Questions?

mnossokoff@hyperionres.com