Quick Take

Effectively Managing HPC/HPDA/AI Storage Requirements: An Increasingly Critical Challenge

Mark Nossokoff and Bob Sorensen
November 2020

HYPERION RESEARCH OPINION

The growing proliferation of GPUs and related HPDA and HPC-based AI workloads is challenging the HPC storage architectures of conventional CPU-based designs and traditional HPC modeling/simulation workloads. Conventional HPC storage systems manage the well-understood needs of largely independent and segregated home directories and scratch files, keeping the system fully utilized for optimal performance. However, data-intensive HPDA and AI workloads, with much greater variety of heterogenous I/O profiles, are stressing the performance capabilities of the conventional storage systems and related architectures.

Workloads, regardless of whether they are HPC, HPDA, or AI driven, are characterized by a set of common I/O profile characteristics:

- Transfer sizes (small, large, mixed)
- Access patterns (random, sequential, mixed)
- Data type (file, block, object)
- Access frequency (hot, warm, cold)
- Performance emphasis (GBs/sec, IOs/sec, latency, mixed)

Understanding and accurately characterizing the workloads and their respective I/O profiles, including whether they will be more homogenous or heterogenous, is critical in determining the type of storage system required to optimize the performance of the system.

CURRENT SITUATION

Until recently, HPC workloads were well understood relative to their I/O profiles and associated storage requirements. Small, random I/Os were primarily driven by metadata needs, and large, sequential I/Os were required for data being transferred either to be analyzed (read) or stored (written).

As simulation and modelling datasets continue to increase in size, what were primarily independent and segregated small block, random I/O and large block sequential I/O evolved into mixed I/O profiles to address the growing metadata capacity need and faster simulation data transfer rates. Likewise, data that traditionally was characterized as active and stored on the highest performance devices, or was less-frequently accessed and stored on high capacity, slower devices, is now viewed as hot, warm or cold and dynamically moved between tiers of storage devices with diverse access patterns.

Similarly, the media type required was well defined. Fast access, low latency, high bandwidth devices with small-to-moderate capacities were required for checkpoint writes while high capacity, high bandwidth devices with less stringent latency considerations were utilized for the primary data store.

Recent Hyperion Research analysis of 194 sites representing 1,894 HPC systems suggests more than 80% of the sites employ some amount of flash-based storage and upwards of 45% of aggregate on-prem HPC capacity is on flash-based devices.

HPC WORKLOADS AND USE CASES

AI and HPDA workloads are changing the status quo and driving storage requirements beyond those of traditional HPC workloads. Traditional HPC storage for conventional modeling and simulation typically consists of project, scratch, and archive use cases. AI workflows present a different set of use cases: ingest, data preparation, training, inference, and archive. Some possess storage attributes like those of traditional HPC workloads while others drive new or more aggressive and extreme characteristics, especially relative to raw GB/s performance needs and mixed I/O profiles.

HPC and AI workloads often exhibit different I/O profiles. Traditional HPC workloads are typically segregated and based on small, random I/O for metadata and large sequential I/O for transferring modeling/simulation input data and results. In contrast, AI workloads chiefly demand a mix of large sequential and small random I/O. Metadata management for AI dataset tagging and labeling requires fast small random I/O.

Use cases also drive a variety of durability and resiliency solution needs. Archiving requires extremely cost-effective solutions without demanding performance requirements. Traditional scratch applications require high performance with the ability to offload interim results to durable storage for protection against failures. AI and HPDA solutions require a mix of storage needs for both high performance transient storage and durable, resilient storage, including a balanced intermix of large block sequential and small block random I/O profiles.

Lastly, data types drive requirements for different types of storage systems. Structured and unstructured data employ varying degrees of file, block and object access methods.

Table 1 maps HPC/HPDA/AI workloads to their respective use cases' storage system profile.

TABLE 1

Traditional HPC and AI/HPDA Workloads

Workload	Use Case	Description
Traditional HPC	Project	▪ Sometimes referred to as home directories or user files ▪ Used to capture and share final results of the modelling and simulation ▪ Mixture of bandwidth and throughput needs, utilizing hybrid flash and HDD storage
	Scratch	▪ Workspace capacity used to perform the modelling and simulation ▪ Includes metadata capacity (high throughput [IOs/sec] and flash-based) and raw data capacity for checkpoint writes to protect against system component failure during long simulation runs (high bandwidth [GB/s]) ▪ Traditionally HDD-based but now largely hybrid flash and HDDs
	Archive	▪ Long-term data retention ▪ Scalable storage without a critical latency requirement ▪ Largely nearline HDD-based systems with a growing cloud-based element ▪ Typically file or object data types
AI / HPDA	Ingest	▪ Quickly loading large amounts of data from a variety of different sources such that the data can be tagged, normalized, stored, and swiftly retrieved for subsequent analysis ▪ Very high bandwidth (GB/s) performance at scale to sustain data rates ▪ Typically object-based, using high capacity HDD-based storage and increasingly cloud-based
	Data Preparation	▪ Often referred to as data classification or data tagging, requires a balanced mix of throughput and bandwidth (hybrid flash and HDD storage systems)
	Training	▪ Utilizing Machine Learning (ML) and/or Deep Learning (DL) to build an accurate model to address respective research, design, and business needs ▪ Requires high throughput (IOs/sec) and low latency for continuous and repetitive computational analysis of the data, typically flash-based storage
	Inference	▪ Utilizing the model for experimentation and analysis to derive and deliver targeted scientific or business insights ▪ Requires high bandwidth and low latency ▪ Typically flash-based, often with a caching layer
	Archive	▪ Long-term data retention ▪ Scalable storage without a critical latency requirement ▪ Largely nearline HDD-based systems with a growing cloud-based element ▪ Typically file or object data types

Source: Hyperion Research, November 2020

FUTURE OUTLOOK

Storage systems remain a critical element in an HPC system's ability to provide optimum performance for the applications and workloads being run. As workloads evolve, decisions will need to be made whether to:

- Provision a single storage system optimized for primarily homogenous workloads
- Integrate a storage system capable of addressing the varied needs of heterogenous workloads
- Deploy multiple heterogenous storage systems, each matched to the diverse workloads expected to be run

An example of the heterogenous storage system approach is the recently announced LUMI supercomputer, targeted to go live in 2021. LUMI's three data stores will include:

- LUMI-O: CEPH – 30PB encrypted object store for sharing and staging data
- LUMI-P: Lustre – 80PB parallel file system for traditional mod/sim workload
- LUMI-F: Lustre – 7PB all-flash for extreme TB/s and IOPs capabilities

This approach can serve as a blueprint for architecting an HPC storage strategy to provide optimal performance for a wide range of diverse workloads.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com