HYPERION RESEARCH

White Paper

Looking Beyond Performance for HPC-AI Storage System Leadership

Sponsored by: VDURA

Mark Nossokoff and Jaclyn Ludema
October 2024

HYPERION RESEARCH OPINION

The pace of change within the HPC-AI market continues accelerating across all fronts, including the storage system. Traditional workloads such as seismic processing, life sciences, and weather analysis typically relied on checkpoint/restart mechanisms to periodically capture the state of modeling/simulation onto scratch storage to protect against system failure. Users were forgiving and could support the time to rerun failed simulations. Fast forward to today, AI training and inferencing has become prominent for recent generative AI applications as well as for the augmentation of traditional HPC modeling and simulation. Users have become much less forgiving as the cost of rerunning a training job could run into the millions of dollars and the value of the data itself is substantial, delivering unprecedented scientific and business value to researchers and companies in both traditional HPC and commercial enterprise markets.

When specifying and prioritizing requirements for new on-premises for HPC-AI systems, users almost universally place a performance-related metric at the top of their list. These metrics are typically defined as some combination of the following factors such as: raw bandwidth, throughput, and/or latency; price/performance; performance improvement target compared with existing user domain-specific application benchmarks; time to results or time to science. All elements of a system's architecture, including computing, networking, and the storage or data platform, are encompassed by the above factors.

Users, however, should be concerned with more than only performance for their solutions, especially relative to the data platform, as the amount of data required to deliver value from AI and modeling and simulation workloads is exponentially increasing. Secure, reliable, and performant storage doesn't just happen. Best-in-class storage solutions require a deep understanding of items beyond performance. Items such as reliability, availability, serviceability, usability, and installability (colloquially referred to as RASUI or "the -abilities") are equally as critical as performance. More recently, durability, or the confidence that data remains unchanged and free from corruption or loss, has emerged as an additional -ability to consider in determining leadership criteria for HPC-AI storage systems.

Understanding availability and durability is especially critical as users increasingly turn to cloud-based storage resources for appropriate workloads, where durability is well-established with some having been architected for extreme durability levels. As the cost of generating the data, and the value of the data itself, being stored far surpasses the expense of the HPC storage system, comparing availability and durability between on-premises and cloud-based resources needs to become an increased area of focus for users.

MARKET OVERVIEW

Storage represents a sizable portion of what users spend overall for their on-premises advanced HPC-AI infrastructure. The storage market is projected to grow to $9.79B in 2028, or 22.2% of the overall market. Figure 1 provides the HPC-AI broader market on-premises forecast.

TABLE 1

HPC-AI Broader Market On-premises Revenue Forecast 2020-2026

($M)	2022	2023	2024	2025	2026	2027	2028	CAGR 23-28
Server	$18,805	$20,735	$25,390	$29,559	$33,699	$37,797	$41,777	15.0%
Storage	$6,380	$6,282	$7,692	$8,745	$9,771	$10,738	$11,846	13.5%
Middleware	$1,781	$1,711	$2,026	$2,241	$2,468	$2,691	$2,968	11.6%
Applications	$5,069	$4,830	$5,684	$6,267	$6,878	$7,468	$8,240	11.3%
Service	$2,214	$2,014	$2,262	$2,411	$2,498	$2,696	$2,973	8.1%
Total Revenue	**$34,250**	**$35,573**	**$43,054**	**$49,223**	**$55,315**	**$61,390**	**$67,805**	**13.8%**

Source: Hyperion Research, 2024

THE "-ABILITIES": DEFINITIONS AND METRICS

The "-abilities" examined in this white paper are reliability, availability, and durability. The terms are related, but not interchangeable.

Reliability

Reliability refers to how often a system experiences failure. Storage systems can be designed with varying degrees of redundancy such that they are still functional while parts of the system have failed. Reliability is typically measured as Mean Time Between Failure (MTBF) in hours.

Availability

Availability refers to the ability to get to the data. Factors that impact a system's availability include redundancy and how long it may take to recover from a failure mode (Mean Time to Recover, or MTTR). In order for a system to be available, some form of redundancy must be designed into the system. There is typically some amount of performance degradation and/or loss of redundancy while a system is operating under a failure condition (e.g., a disk drive fails). Availability is typically expressed as the percentage of time it is operational in a year in terms of "9s". Table 2 provides a model that describes availability.

TABLE 2

Representation of Availability

# of Nines	Annual % Uptime	Minute Online per Year	Minutes Offline per Year
1	90%	473,040	52,560 (36.5 days)
2	99%	520,344	5,256 (3.7 days)
3	99.9%	525,074	525.6 (8.8 hours)
4	99.99%	525,547	52.6 (< 1 hour)
5	99.999%	525,595	5.3
6	99.9999%	525,599	0.5

Source: Hyperion Research, 2024

Durability

Durability refers to the data existing on the storage media as it was written when the application reads the data. In other words, there has been no data corruption or data loss. Durability is typically measured as a probability for data loss (data loss incurred by a solution across a configuration of systems within the solution) or mean time to data loss (MTTDL). Table 3 describes a model for understanding durability.

TABLE 3

Representation of Durability

# of Nines	Percentage	Data Loss within a Defined System
1	90%	1 object* with data loss within a configuration of 10 objects
2	99%	1 object with data loss within a configuration of 100 objects
3	99.9%	1 object with data loss within a configuration of 1,000 objects
4	99.99%	1 object with data loss within a configuration of 10,000 objects
5	99.999%	1 object with data loss within a configuration of 100,000 objects
6	99.9999%	1 object with data loss within a configuration of 1,000,000 objects
7	99.99999%	1 object with data loss within a configuration of 10,000,000 objects
8	99.999999%	1 object with data loss within a configuration of 100,000,000 objects
9	99.9999999%	1 object with data loss within a configuration of 1,000,000,000 objects
10	99.99999999%	1 object with data loss within a configuration of 10,000,000,000 objects
11	99.999999999%	1 object with data loss within a configuration of 100,000,000,000 objects

Note: *An object is defined as the lowest common denominator unit that can be lost that will contribute to data loss, depending on the level and type of redundancy and protection. Within a storage system, for example, an object could be a file or a complete storage system.

Source: Hyperion Research, 2024

Taken together, how a system is architected to address reliability, availability, and durability determines how resilient the system is. Table 4 summarizes definitions and metrics for these -abilities.

TABLE 4

"-ability" Definitions and Metrics

Term	Definition	Typical Metrics
Reliability	The probability that a storage system will function correctly without failure during a specific period.	Mean Time Between Failure (MTBF)
Availability	The ability to get to the data. The percentage of time a storage system is operational and accessible for use. It is often expressed as a percentage of uptime, such as 99.999% (five nines)	% of uptime, often expressed a "9s" (e.g., 99.999%, or 5 9s)
Durability	The ability of the data to last. The ability of a storage system to preserve data without loss or corruption over time.	Data durability percentage (e.g., 99.99995%) Data loss probability (e.g., 0.00005% chance of loss) Mean Time to Data Loss (MTTDL)

Source: Hyperion Research, 2024

Note that durability is required for a system to achieve availability, while a system can be durable and not be available. Consider a safe with a combination lock. The safe may be indestructible (e.g., durable) and the contents are safe from flood and fire and will be there when the safe is opened. If, however, the combination is lost or forgotten, the contents of the safe are not available. And conversely, the combination may be known, but if the safe is not water-tight, the contents may be destroyed when it is opened if there has been a flood. The safe must be durable in order for it to be available.

THE IMPORTANCE OF AVAILABILITY AND DURABILITY

Using the broader market spending breakout as a proxy for initial purchase costs for an on-premises system, servers and storage comprise approximately 50% and 20%, respectively, of the initial purchase expense. Systems that don't provide adequate availability and durability risk sitting idle should any failures cause downtime.

When considering the costs of today's leadership machines, particularly those that employ thousands of GPUs for accelerated modern AI workloads, it's imperative that the computing nodes be able to access and retrieve the required data. Should the storage system be unable to provide the data when it's needed, the expensive GPUs sit underutilized. Making proper investments to achieve high levels of availability and durability in the "20%" storage element costs is needed to ensure high utilization of the "50%" server element costs, as referenced above.

Efficiency costs are equally, if not more, critical than equipment expense to consider relative to system downtime. Engineers performing complex physics-based simulations may sit idle or need to re-run jobs if their systems go down. Time to science and discovery is extended should scientists' and researchers' systems be impacted, not to mention the integrity of their results could be compromised in the event of any data loss or corruption. Consider some specific examples from select verticals which heavily employ HPC-AI infrastructure:

- **Seismic processing:** The integrity of data directly impacts the accuracy of subsurface imaging and resource exploration. Any data corruption can lead to significant misinterpretations and costly errors.
- **Life sciences**: Research and clinical trials rely on precise data to drive drug discovery and patient outcomes. Here, data availability is paramount; any disruption could hinder vital research processes or delay critical advancements in healthcare.
- **Weather analysis**: Accurate and timely data is essential for predicting and responding to environmental changes. The ability to access reliable data quickly can significantly affect forecasting accuracy and disaster response efforts.

Business impact can be even more significant should systems not have enough availability or durability. Data center operations that support hundreds or thousands of businesses would stand to lose substantial revenue or long-term business contracts should systems go down and/or data be lost.

ARCHITECTING FOR THE "-ABILITIES"

Anatomy of an HPC-AI Storage System

Storage systems incorporate a number of architectural elements that impact their performance and "-ability" characteristics, including:

- **Systems that house the controllers (RAID) and physical devices (HDDs, SSDs**) that respectively provide the storage services (replication, snapshots, redundancy) and storage media that manage, store and maintain the data
- **Expansion storage enclosures** to provide additional storage media that scales out from a storage server
- **File systems and servers** dedicated to running the file system inclusive of primary storage, metadata storage and archive storage
- **Storage interconnect switches and cabling** that provide the connectivity between HPC-AI compute servers and storage servers, and between storage servers and enclosures

Redundancy

Most storage systems, whether they be on-premises or cloud-based storage resources, employ some form of redundancy by distributing the data across multiple storage devices and applying a striping and parity algorithm. RAID and erasure coding are the predominant methods of doing this.

RAID

RAID combines multiple disk drive components into a logical unit for the purposes of data redundancy and distributes the data and parity calculated from the data across multiple storage devices. In the event of a device failure, the data can be reconstructed from the parity information and be made available to the application. The level of protection and performance under failure and reconstruction is determined by the number of physical devices defined for the system and the allocation of data capacity and parity capacity desired. Tradeoffs can be made between the level of durability and availability required by the application and the budget constraints for the storage infrastructure.

RAID has been widely used in on-premises enterprise storage systems, providing a reliable method for data protection and performance enhancement in traditional disk-based storage environments.

Erasure Coding

Erasure coding is a method of data protection in which data is broken into fragments, expanded, and encoded with redundant data pieces, then stored across a set of different locations or storage media. The goal of erasure coding is to allow for data recovery even when some fragments are lost or corrupted. Unlike traditional RAID, which typically uses simple parity or mirroring, erasure coding employs more sophisticated mathematical algorithms to achieve higher storage efficiency while maintaining or improving data durability.

Erasure coding is particularly well-suited for distributed storage systems and cloud environments, where data is spread across multiple nodes or even geographic locations. Erasure coding has gained popularity in recent years, especially in object storage systems and large-scale cloud storage platforms.

Table 5 compares the redundancy methods against several criteria.

TABLE 5

Comparison of RAID and Erasure Coding Redundancy Methods

Criteria	RAID	Erasure Coding
Flexibility	Moderate (limited to predefined RAID levels)	High (customizable protection schemes)
Scalability	Limited (typically bound to a single array or node)	High (can scale across multiple nodes or locations)
Performance	Good (for small-scale systems; can be a bottleneck in large systems)	Moderate to high (depends on implementation)
Cost	Moderate (may require specialty hardware for optimal performance)	Low to moderate (can be implemented on commodity hardware)
Risk of data loss	Low to moderate (depending on RAID level)	Very low (especially in distributed systems)
Durability	Good (protects against drive failures)	Excellent (can protect against media, node, or site failures)
Resource Utilization	Moderate (fixed overhead based on RAID levels)	Efficient (customizable overhead)

Source: Hyperion Research, 2024

In addition to RAID and erasure coding techniques, redundancy can also be architected into the controller and interconnect elements of the storage system. Dual controllers with failover capabilities protect against a controller failure and can provide multiple paths to the storage devices.

Failure Domains

Considerations need to be made for how broadly a failure within a part of the storage system will impact the entire storage system or cluster. Sometimes referred to as the "blast radius", the smaller the failure domain, the least impact to the system. Availability and durability can be achieved regardless of the size of the failure domain, with corresponding cost and performance tradeoffs.

Performance Under Failure and Rebuild/Reconstruction

Highly available systems are operational while a failure exists within the system. System performance may be sustained or de-graded while the failure condition exists, depending on the RAID or erasure coding design as some of the system performance may diverted to returning it to an optimal state. Techniques can be employed to ensure degraded performance does not dip below a specified threshold of required sustained performance.

The longer it takes to return to an optimal state, the higher the chances of an additional failure to impact the system's availability, depending on how much redundancy has been implemented. Additional performance could be designed into the system, or overprovisioned, to accommodate faster reconstruction times and minimize a system's window of vulnerability to a failure that would take it offline.

BEST PRACTICES WHEN CONSIDERING HPC-AI STORAGE AVAILABILITY AND DURABILITY

One size does not fit all when it comes determining how much availability and durability are "enough" for a given workload or user need. Users and vendors alike can greatly influence optimizing the system's TCO relative to availability and durability.

Users

First and foremost, users need to understand their risk profile for downtime. Several areas to understand include:

- How much downtime can the business afford?
- How much business is lost if the data is corrupted or not available?
- Are other resources (e.g., personnel, data products or services) sitting idle if the data is not available for them to run their simulations or experiments?
- Can the budget support the expense required to attain the desired levels of availability and durability?
- What is the minimum performance degradation that can be tolerated if the system is operating with a failure condition?
- How long is acceptable for the system to be operating with degraded performance?

Once the above parameters are established, users should include specific requirements for availability and durability in the RFIs, RFPs, and RFQs they provide to vendors in assessing sourcing options. This applies both to infrastructure being acquired on-premises, as well as determining whether cloud-based resources are appropriate.

Vendors

Recognizing that users will have varying degrees of availability and durability requirements, vendors should provide:

- Flexibility for users to determine related tradeoffs (e.g., performance, cost)
- Design and configuration guidelines for users to achieve desired availability and durability goals
- Telemetry for users to monitor the system's availability and durability

Additionally, vendors should continue to invest and innovate in this area. While performance capabilities tend to receive the lion's share of investment, it is essential for systems to not only be fast but also to ensure that data is readily available and intact. If the data necessary for engineering simulations or training AI models is inaccessible or corrupt, then high performance becomes irrelevant. A balanced approach that prioritizes business-required speed and data availability/durability will lead to a more effective solution.

FUTURE OUTLOOK

Storage is a critical element of leading HPC-AI system architectures. While performance capabilities of the storage platform are important, they are not the only factors that define leadership for HPC-AI storage solutions. Availability and durability are also essential elements, and their importance is growing in the assessment of HPC-AI storage leadership.

Data quality and reliability are becoming increasingly important in the age of AI, particularly relative to training AI models. Corrupt data can lead to inaccurate models and producing unreliable inferencing. No data (e.g., the storage system is down and unable to provide any data for training) causes expensive resources to sit idle and impacts delivery time for new models and updates to existing models with current information.

Users cannot tolerate any data loss or corruption in any form. Durability will continue to be an absolute de facto table stakes requirement for any storage system to be considered for today's HPC and AI workloads. Availability should also be non-negotiable in most cases, although some users do have varying degrees of tolerance. The more availability a system can provide, the less impact there will be to user business, engineering simulation performance and results, and scientific research capabilities.

Ultimately, users and vendors who fully understand and balance the "-abilities" including the newly emergent durability, alongside performance and budget considerations may realize greater returns on their investments. By prioritizing both data integrity and availability, organizations can enhance their HPC workloads across all workloads, ensuring robust performance and impactful outcomes.

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). Hyperion Research provides 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

Copyright Notice