

AI and GenAI deployments require a mix of performance, scale, simplicity, and productivity. Finding the right balance between resources and deployments improves AI outcomes and reduces time to value.

AI Workloads Demand Performance at Scale in File Storage

April 2024

Written by: Dave Pearson, Vice President, Worldwide Infrastructure

Introduction

Enterprises have always faced competing priorities when it comes to data infrastructure investment: maximum performance at minimum expense, extreme scale and data growth with the highest possible efficiency, and shortened time to value with more data, dependencies, and complex deployment methodologies than ever before. The advent of predictive and interpretive AI pushed these boundaries, and the recent sudden growth in generative AI (GenAI) initiatives means that balancing those priorities is a greater operational imperative for IT buyers and administrators than ever before.

In IDC's 2023 *Future of Digital Infrastructure Worldwide Sentiment Survey*, the second-most important KPI that senior leaders and board members identified was the timeliness of mission-critical data insights (as indicated by 42% of the respondents), showing just how important AI and GenAI transformation can be. However, IT and cloud operational efficiency improvements topped the list (52%), indicating that organizations still need to break down technological and operational silos, increase access to data across the enterprise to use expensive compute resources efficiently, and improve productivity among their AI and GenAI transformation teams.

The unstructured data that AI and GenAI initiatives require presents unique challenges to traditional data infrastructure, where performance was primarily the purview of block access and point solution deployment occurred on a workload-by-workload basis. The need for better storage infrastructure for unstructured and file data is clear.

AT A GLANCE

KEY TAKEAWAYS

- » Performant, scale-out file storage is key to managing the 244ZB of unstructured data that enterprises expect to generate annually by 2027.
- » AI and GenAI workloads will need to utilize hybrid multicloud deployments to leverage all enterprise data: on-premises as-a-service offerings can enable a cloud-like operating experience when combined with unified management tools across public and private clouds.
- » AI and GenAI workloads have unique demands that legacy storage systems do not meet. Performance, scale, simplicity, and efficiency can increase the productivity of AI transformation teams and initiatives.

IDC's Global DataSphere estimates that 77% of all data stored in 2023 was unstructured (e.g., imagery, video, text, and audio), incorporating everything from office productivity files to medical data to software code (see *Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2023–2027*, IDC #US50397723, June 2023). That unstructured data growth will continue unabated, as 84% of the 291ZB of data generated in 2027 will be unstructured. While text is currently the largest contributor to the volumes of new data that GenAI generates, IDC expects that 75% of the new data generated will be evenly distributed between text, imagery, and video by 2028.

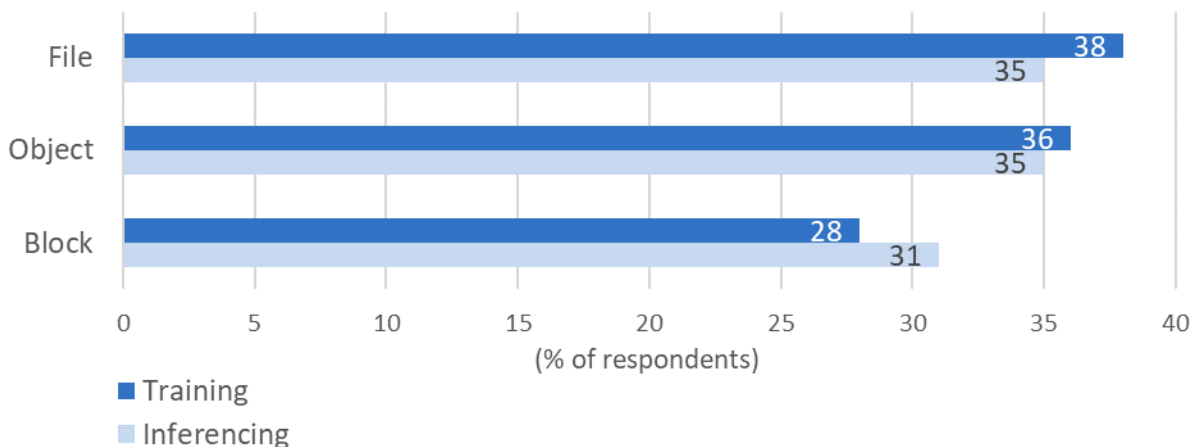
These factors are pushing organizations to face some new realities about the way they deploy their data infrastructure:

- » They require efficient, performant file storage to feed expensive, accelerated compute resources to maximize productivity and minimize time to value in their AI and GenAI initiatives.
- » They will need to manage uncertain and surging demand for capacity through the use of hybrid multicloud deployments and on-premises as-a-service offerings for data that cannot be exposed to outside systems because of security, data protection, privacy, regulatory compliance, or governance needs. About 84% of the respondents of our 2023 *IT Infrastructure for Storage and Data Management Survey* preferred a hybrid cloud or hybrid multicloud approach to data infrastructure deployments.
- » They will require application awareness, observability, and management across all their data infrastructure. About 70% of the respondents of our 2023 *IT Infrastructure for Storage and Data Management Survey* indicated that they were currently using or needed a unified multicloud management system to provide single-pane-of-glass management for disparate storage environments.

Benefits

In IDC's September 2023 *AI View 2023 Survey*, respondents indicated that file, object, and block storage were all critical to their AI infrastructure when it came to data access, but file storage was slightly more prevalent in training activities and just as important as object storage for inferencing (see Figure 1).

FIGURE 1: **Importance of File, Object, and Block Storage for AI Infrastructure**



n = 400

Source: IDC's *AI View 2023 Survey*, September 2023

A performant file storage system with the scalability to support the variable and growing demands of AI and GenAI workloads can ensure the optimal utilization of expensive compute resources at all stages of AI/GenAI by removing storage performance bottlenecks. All flash offerings with NVMe can provide the throughput and latency characteristics necessary for many stages of AI and GenAI workflows within a single storage tier, thus consolidating and optimizing the data pipeline for these highly demanding activities. This improves the productivity of even more expensive resources — such as data scientists and engineers within the enterprise — and it can optimize time to insights and accelerate AI transformation.

Unified file storage management, monitoring, and protection can enable developers, data engineers, and data scientists to extract value from unstructured file data in a global namespace from a variety of sources and deployment modalities, including on premises, edge, and cloud via a simplified management console that extends capabilities and file services across the enterprise. Increasing productivity and reducing time to value can maximize ROI on AI and GenAI transformation activities by enabling collaborative efforts between individuals, teams, and business units.

Legacy architectures and technical debt are part of the complexity problems that organizations face and are holding back digital transformation. By simplifying administration and addressing the AI transformation skills gap and lack of resourcing, it is possible to enhance administrative productivity while data life-cycle management, data protection, security, compliance, and governance are taken care of through data services, not by time-constrained resources.

Increased ROI, decreased TCO, and greater sustainability are achievable through the efficient use of denser, higher-capacity resources and appropriate scaling and data reduction capabilities, leading to improved operational efficiency, better power and cooling metrics, and improved ESG outcomes for stakeholders throughout the organization.

Considering HPE GreenLake for File Storage

Hewlett Packard Enterprise (HPE) has provided datacenter infrastructure services for almost a decade as an independent company and, prior to 2015, as part of the Hewlett-Packard Company. With \$29.1 billion in annual revenue in 2023 (up 2% from 2022), HPE is a named vendor in IDC's Enterprise Storage Systems Tracker as a top provider of external storage systems as well as internal (server-based) storage solutions. HPE's leading storage systems product brand for 2023 was the HPE Alletra Storage product family, which is bolstered by the HPE SimpliVity, HPE Primera, and HPE Nimble Storage product lines, among others.

In November 2017, HPE launched the HPE GreenLake edge-to-cloud platform, offering customers on-premises infrastructure solutions akin to a public cloud's on-demand or as-a-service model. Since its inception, HPE has significantly broadened its HPE GreenLake portfolio, now comprising over 50 services. These options include features for seamlessly managing hybrid multicloud environments, spanning on-premises, edge, and public and private cloud infrastructures.

HPE GreenLake for File Storage was launched in April 2023 on a new hardware platform, HPE Alletra Storage MP, in conjunction with VAST Data's software. This novel combination of modular, all-flash, NVMe-enabled hardware with The VAST Disaggregated Shared Everything (DASE) architecture for file storage enables performant, exabyte-scale storage that avoids the typical trade-offs associated with solutions that scale compute and storage resources and capacity in tandem.

In March 2024, this offering was refreshed with a higher-performance version, increasing compute and storage performance, density, and capacity. The company now provides higher compute and storage density options along with the standard configurations to deliver greater flexibility for customers at different stages of their AI journey. Some customers may have outgrown the capacity or capabilities of earlier offerings and require increased enterprise performance at AI scale for AI and GenAI workloads, as well as machine learning (ML), high-performance computing (HPC), big data, life sciences, financial analytics, media and entertainment, data lakes, and fast-access archive deployments.

The centralized management, monitoring, and protection capabilities of the HPE GreenLake platform are designed to provide a simplified, intuitive management experience for hybrid multicloud environments. The simplified setup with a global namespace is meant to ease collaboration and simplify deployments, as is automatic device discovery, onboarding, and configuration. HPE GreenLake has a suite of rich data services that enhance the productivity of data engineers and data scientists by removing much of the management overhead associated with legacy architectures, namely life-cycle management, backup, and disaster recovery as well as nondisruptive upgrades and proactive maintenance managed through the single-pane-of-glass HPE GreenLake console.

Sustainability and operational efficiency are key to achieving the appropriate ROI for AI and GenAI activities. Independent scaling of performance and capacity by the flexible, separate addition of controller or capacity nodes avoids the underutilization of key resources in legacy systems that require upgrades to scale in lockstep. Increasing the density of storage resources through capacity-optimized flash media, along with data reduction technologies, can reduce the footprint, carbon emissions, and power and cooling requirements of storage systems in constrained environments already struggling to support power-hungry, GPU-accelerated, compute resources. HPE leverages the Similarity data reduction algorithm available with HPE GreenLake for File Storage to provide fine-grained data reduction capabilities across the global namespace. This approach combines the most positive aspects of compression and deduplication without adding significant overhead to storage systems. HPE GreenLake for File Storage includes support for Infiniband, NVIDIA GPUDirect, and remote direct memory access to optimize and maximize GPU utilization, lowering time to insight and increasing ROI for critical AI and GenAI initiatives.

Challenges

AI and GenAI explorers often begin their transformation journey with cloud AI services, which comprise much of today's market. However, as organizations mature, they come to understand that much of their most critical data must remain under their control and not "out in the wild." Operating in this hybrid cloud or multicloud mode can increase complexity in these environments unless vendors can provide frictionless workload and data mobility as well as observability and management across the organization. Providing technology, services, and skilled resources directly or through partner ecosystems to accelerate maturity among AI and GenAI users is a requirement for the successful implementation of these high-value initiatives.

HPE is not the only company with as-a-service offerings in the storage space, and they have considerable competition from providers targeting AI and GenAI workloads through a variety of hardware and software technologies and deployments. Amplifying the HPE GreenLake ecosystem's value proposition and the importance of performance,

HPE GreenLake for File Storage offers customers a workload-appropriate performant file storage solution with the flexibility to opt for traditional storage procurement and licensing models or embrace the pay-per-use approach facilitated by GreenLake.

simplicity, and efficiency at AI scale will be a key challenge. Many customers in the early stages of AI exploration are seeking simple cloud services or full-stack "in a box" solutions because of their lack of skilled resources, so HPE will need to demonstrate that HPE GreenLake for File Storage can accelerate their journey along the AI maturity curve and provide greater value over the life cycle of these initiatives.

Conclusion

Unstructured data growth remains a concern for enterprises at every stage of digital and AI transformation. Coupled with the need for timely insights from advanced analytics, AI, and GenAI, the growth of on-demand, scale-out storage architectures is expected to continue unabated for at least the next five years. Enterprises, which previously gravitated toward public cloud solutions for rapid deployment and scalable storage capabilities, now find themselves with the need to keep certain data within the relative safety of their own datacenter infrastructure and require on-premises alternatives aimed at simplifying the acquisition, implementation, operation, and expansion of unstructured data storage.

HPE GreenLake for File Storage is deployed on a unified, high-performance hardware platform that is purpose-built to accommodate data-intensive workloads. With increases to performance and capacity, HPE offers customers a data-intensive workload-appropriate performant file storage solution with the flexibility to opt for traditional storage procurement and licensing models or embrace the pay-per-use approach that HPE GreenLake facilitates, providing cloud services' operational and management paradigms. The intersection of enterprises' need for higher performance and capacity for data-intensive AI workloads and control over their most critical and valuable data creates an opportunity for scale-out performant file storage. Continued interest in AI, GenAI, ML, and HPC initiatives should provide vendors with such offerings opportunities to grow in these emerging workloads.

The intersection of enterprises' need for higher performance and capacity for data-intensive AI workloads and control over their most critical and valuable data creates an opportunity for scale-out performant file storage.

About the Analyst



Dave Pearson, Research Vice President, Worldwide Infrastructure Research

Dave Pearson is research vice president within IDC's Worldwide Infrastructure Research organization and global research lead for IDC's Storage and Converged Systems practice. Dave and his team provide global insights on storage; integrated, hyperconverged, and composable infrastructure technology trends; vendor strategies; and market adoption. It includes storage for performance-intensive computing use cases such as high-performance computing, artificial intelligence, and analytics.

MESSAGE FROM THE SPONSOR

HPE GreenLake for File Storage delivers enterprise-grade, scale-out file storage to accelerate rapidly growing, data-intensive AI workloads. This cloud service offers enterprise performance, simplicity, and enhanced efficiency, all at AI scale. HPE GreenLake for File Storage unlocks more value from all your data with faster time to insights and discovery for competitive advantage; empowers data scientists and LOB application owners with increased productivity and focus on discovery and innovation; and achieves high sustainability with capacity density, datacenter footprint, power, and cooling efficiencies for reduced carbon footprint.

Learn how to accelerate your AI journey with the industry's most comprehensive file storage solution to store, manage, and protect data across hybrid cloud while achieving the required performance, simplicity, and enhanced efficiency for AI scale.

[Watch the demo video.](#)

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2024 IDC. Reproduction without written permission is completely forbidden.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com