TECHBYTE
Insights Into Research Computing

ON DEMAND
systems

The Importance Of Data Storage for HPC & AI Workloads

Authored by:
Rakesh Sabharwal
Founder & CEO
On Demand Systems Pte Ltd

6 January 2025

The Importance of Data Storage for HPC and AI Workloads:

When discussing the infrastructure needed for High-Performance Computing (HPC) and Artificial Intelligence (AI) workloads, many immediately think of Graphics Processing Units (GPUs) due to their powerful computational capabilities. However, while GPUs are indeed critical, data storage plays an equally, if not more, essential role in the performance and efficiency of these workloads. Here in this document we will discuss the importance data storage.

Data Storage vs. GPU vs. Network: Understanding the Balance:

1. Data Volume and Accessibility

HPC and AI workloads often involve massive datasets that need to be accessed quickly and efficiently. Without robust data storage solutions, even the most powerful GPUs can become bottlenecked, waiting for data to process.

- ### Scalability

HPC/AI use cases have continuously increasing input data (generated by edge devices, lab equipment, experiments, manufacturing line etc.), thus data storage system scalability becomes a crucial factor in long run. Advanced storage systems can scale to accommodate growing data needs, ensuring that as datasets increase in size, the system can keep up.

- Speed

High-speed storage solutions reduce latency, ensuring data is readily available for immediate processing by GPUs.

2. Data Throughput and Bandwidth

The bandwidth between storage systems and GPUs is a crucial factor in determining overall system performance:

- High Throughput

Storage systems designed for HPC and AI must handle high throughput to feed data continuously and efficiently to GPUs.

- Network Integration

Integrating storage with high-speed networks ensures data can flow smoothly and quickly, preventing delays that can degrade performance.

AI models require extensive training before deployment. The quality of this training directly impacts the model's reliability and a key factor in effective training is access to large datasets. During the training phase, AI projects place significant demands on IT infrastructure, especially storage systems.

- **Unstructured Data:**

Large language models primarily rely on unstructured data, typically stored on file or object storage systems.

- **Structured Data:**

Financial models, in contrast, often utilize structured data, where block storage is more prevalent.

- **Hybrid Usage:**

Many AI projects leverage a combination of all three storage types, usually referred to as unified storage, depending on the data and workload requirements.

By optimizing storage solutions to match the specific needs of AI and HPC workloads, organizations can ensure efficient model training and better outcomes.

Once an AI model is trained, its demand on data storage typically decreases. In production, the system processes user or customer queries using optimized algorithms which are highly efficient. However, data storage remains essential for handling inputs and outputs-queries are submitted to the model and the model generates corresponding results.

In the operational or inference phase, AI requires high-performance I/O for optimal effectiveness. While the data volume needed is significantly smaller than during training, the system must process inputs and return outputs within milliseconds.

Key AI use cases such as cybersecurity and threat detection, IT process automation, biometric scanning for security and image recognition in manufacturing-demand rapid response times. Even in applications like generative AI chatbots designed to interact like humans, the system's speed is critical to deliver natural, seamless responses.

The Role of Data Storage in AI & HPC:

1. Preprocessing and Staging Data

Before data even reaches the GPUs for processing, it often requires significant preprocessing. Efficient storage solutions allow for this preprocessing to occur swiftly, ensuring that GPUs are utilized effectively:

- ### Data Staging

Organizing data in a way that maximizes processing efficiency is essential. Storage solutions can help by staging data near compute resources.

- ### Data Preparation

Cleaning and preparing data for analysis is a storage-intensive task that needs to be completed before GPU processing.

2. Data Management and Backup

Data storage systems are vital for managing, archiving, and backing up the massive amounts of data used in HPC and AI.

- ### Reliability:

Ensuring data is not lost and is retrievable is critical for maintaining the integrity of computational results.

- ### Data Lifecycle Management:

Storage solutions can automate the management of data throughout its lifecycle, ensuring that it is archived or deleted when no longer needed.

Storage-Centric Architecture: The Foundation for HPC & AI:

Modern HPC and AI environments are increasingly adopting storage-centric or data centric architectures, where data storage is treated as the backbone of the entire system:

- **Data Locality and Proximity:**

Storing data closer to computing resources minimizes latency and maximizes performance by reducing the time it takes for data to reach GPUs and CPUs. Enterprises also need to consider data tiering to avoid potential data gravity or related problem.

- **Parallel File Systems:**

High-performance file systems like BeeGFS, Lustre, GPFS and others allow simultaneous access to large datasets, enabling faster and more efficient processing than traditional enterprise storage could manage alone.

- **Emerging Storage Technologies:**

Innovations in storage, such as NVMe, storage class memory (SCM), and flash-based systems, significantly enhance throughput and speed, meeting the intensive demands of HPC and AI better than legacy storage solutions.

Data Storage for Managing the Entire Data Lifecycle:

HPC and AI workloads generate data that needs careful management throughout its lifecycle. Efficient data storage systems help in:

- ## Data Governance:

With increasing data compliance requirements, advanced storage solutions support traceability, version control, and regulatory compliance, essential for sectors like healthcare, finance, and scientific research.

- ## AI Training and Iterative Workflows:

AI model training often requires multiple passes over vast datasets. Optimized storage ensures that iterative training processes run without bottlenecks, maximizing GPU usage and reducing training time.

Enabling Edge Computing with Robust Storage:

As HPC and AI applications expand to the edge, where data is generated in real-time (e.g., IoT sensors, autonomous vehicles), storage solutions must handle data locally and efficiently:

- ## On-Edge Data Processing:

Storage systems at the edge can preprocess and filter data before it's transmitted to centralized HPC or AI environments, reducing data transfer times and processing costs.

- ## Data Aggregation & Synchronization:

Robust edge storage systems help manage data aggregation and synchronization, so data from diverse sources can be efficiently integrated into the primary HPC and AI workflows.

Case Studies: Industry Applications of Storage-Driven AI & HPC:

- ## Healthcare:

In genomics and medical imaging, rapid data retrieval from storage is crucial for real-time diagnostics and research.

- ## Finance:

For applications like real-time fraud detection, data storage systems must quickly process and retrieve data to enable immediate analysis.

- ## Scientific Research:

Large-scale projects, such as climate modeling, rely on scalable, high-performance storage to manage petabytes of data over extended research periods.

Future Trends in Data Storage for HPC & AI:

Looking ahead, data storage will continue to play an expanding role in HPC and AI with developments such as:

- **AI-Driven Data Storage Management:**

AI is being integrated into storage solutions to predict and manage data needs dynamically, helping allocate resources in real-time for maximum efficiency.

- **Hybrid and Multi-Cloud Storage:**

With cloud computing growing, data storage solutions that seamlessly integrate with cloud platforms will offer cost-effective, scalable options for HPC and AI workloads.

Conclusion:

While GPUs have transformed HPC and AI by delivering unparalleled computational power, the role of data storage remains equally critical. A well-designed storage system enhances the efficiency, speed, and scalability of HPC and AI workloads, unlocking the full potential of GPU performance. By prioritizing storage infrastructure as the foundation, we can create robust, efficient systems capable of addressing modern computational challenges.

AI storage solutions are purpose-built to manage the massive data volumes generated, processed, and stored by AI/ML workloads. These solutions enable fast, reliable, and efficient data handling, ensuring optimal application performance.

In contrast, traditional storage systems often fall short in meeting the intensive data demands of AI/ML operations. They lack the speed, capacity, and scalability required to handle rapid data influxes, leading to bottlenecks and reduced application performance. Investing in specialized AI storage is essential for overcoming these challenges and achieving peak system efficiency.

About Us:

On Demand Systems Pte Ltd provides Secure High Performance Computing (HPC) & Artificial Intelligence (AI) infrastructure solutions. We provide end-to-end services from conceptualization, design and deployment to management.

Our HPC/AI solutions are complemented by On Demand PFS & On Demand ObjectStor, data storage solutions specifically designed & engineered to optimize performance for HPC & AI Workloads. On Demand PFS delivers high-throughput parallel file systems, while On Demand ObjectStor ensures scalable & efficient object storage for massive datasets. These solutions empower enterprises to handle data-intensive workloads seamlessly & efficiently.

In addition, the HPC/AI solutions and services are complemented with identity and access management software and cloud-based Identity-as-a-Service, enabling enterprises to securely manage identities and control access across computer networks and cloud computing environments.