SOLUTION BRIEF

Qumulo Cloud Data Platform

Qumulo Cloud Data Platform

The rapid adoption of AI is driving a critical need for seamless integration with a high-performance storage platform capable of handling the massive volumes of unstructured data that AI applications demand. Legacy storage solutions can struggle to provide the scalability, speed, and integration necessary for real-time data processing to support ongoing training, RAG-enabled applications, and cloud-based AI service integrations.

For years, Qumulo has supported HPC customers running AI workloads like image recognition, drug discovery, driver assistance, and LLM training using thousands of parallel connections. As evidenced through Qumulo's record-breaking AI benchmark, Qumulo has proven itself the most cost-effective, highest-performing data platform for AI.

Cloud-Based AI

Organizations seeking to run AI in the cloud are compelled to construct solutions that transfer data from low-cost object storage to the high-cost local file caches connected to GPU-powered applications. A two-tier system using file caches also means that high cost GPUs are idle up to 40% of the time, waiting to load the data from object storage into the file cache. Qumulo's cloud architecture (Azure Native Qumulo and Cloud Native Qumulo for AWS) delivers intelligent cache management for the object store, executing parallelized, prefetched reads that prestage hot data to feed GPU-hosted AI applications and minimize idle cycles.

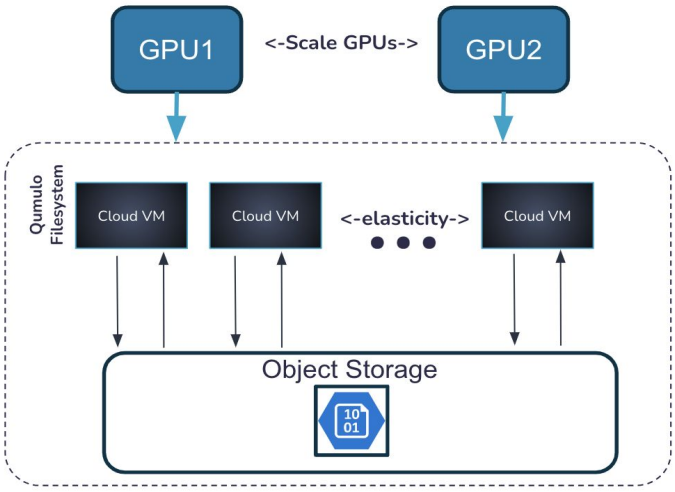

Fig.1

Industry's Fastest AI Benchmark

The AI_Image benchmark from SpecStorage provides a robust measure of performance in production AI workloads by simulating common I/O patterns seen in frameworks like TensorFlow and PyTorch, utilizing widely available datasets. In this benchmark, Qumulo demonstrated exceptional responsiveness, maintaining an overall response time of just 0.84ms, even as the workload scaled from zero to 700 synthetic AI jobs. The price of running this benchmark was only $400 – a proof point of Qumulo's extreme value and disruptive pricing. This performance highlights Qumulo's ability to handle high-demand AI workflows, ensuring rapid data access and minimizing latency for intensive machine learning tasks. Read about it HERE.

Qumulo accelerates GPU-side performance by *eliminating load times* between the object layer and the file system, achieving both sub-millisecond latency and high throughput, even as AI demand increases [Fig.1].

The net result is up to 40% higher performance for total-time-to-train.

And since Qumulo allows you to pay only for what you use, including throughput, IOPS and capacity, charges halt the moment the AI jobs subside. This is how Qumulo is able to achieve industry-leading performance at an industry-disruptive price point.

Qumulo has strategically developed integrations with Azure Copilot and Amazon Q for Business, enabling customers to create custom connectors that effortlessly retrieve and index common file types such as PDFs, text files, and spreadsheets from Qumulo instances on Azure and AWS. Through the use of natural language queries via interfaces like Microsoft Copilot, these integrations significantly enhance efficiency by automating or eliminating manual, repetitive tasks, freeing staff to focus on higher-value work, accelerating productivity and innovation.

Use Case: Image Recognition

A leading South Korean company, specializing in the production of printed circuit boards (PCBs) and semiconductor packaging substrates, has transformed its manufacturing process by integrating AI-based image recognition technology. This advanced AI solution analyzes images of PCBs in real-time during production, identifying defects or inconsistencies that might be overlooked by traditional inspection methods. By automating quality control through AI, the company significantly improves product reliability and reduces human error, ensuring the highest standards of precision and efficiency in PCB manufacturing.

Use Case: Advanced Driver Assistance Systems

A leading German company has successfully integrated AI into its Advanced Driver Assistance Systems (ADAS) to enhance critical safety features such as lane-keeping assistance, emergency braking, and traffic sign recognition. Leveraging Qumulo's file data platform over NFS, the company's AI processes real-time data from vehicle sensors and cameras, enabling more accurate and reliable vehicle responses. By utilizing AI-driven insights, the company delivers superior safety performance in their vehicles, ensuring rapid and precise reactions to dynamic road conditions, ultimately increasing the safety of drivers and passengers.

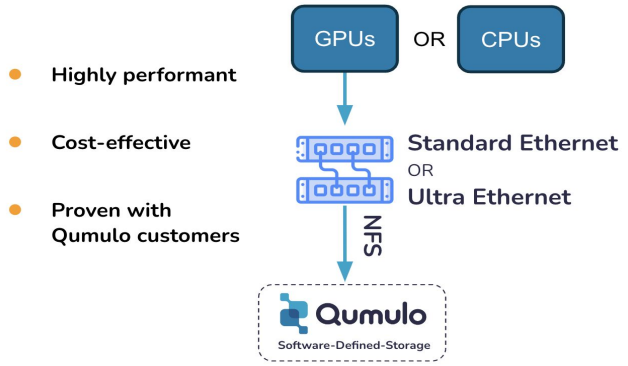

Fig.2

On Premises AI

Organizations who invest in costly NVIDIA Superpods to run AI workloads on-premises may put their overall return on investment at risk. Qumulo enables organizations to efficiently scale AI operations using existing infrastructure, eliminating the need for expensive, specialized hardware. Customers who leverage Qumulo's file data platform over NFS have successfully run thousands of AI jobs in parallel using standard Ethernet environments.

As the first storage vendor to join the Ultra Ethernet Consortium, Qumulo is also pioneering efforts to simplify the integration of compute, AI processing, and storage through a standards-based approach. With Qumulo, customers can run thousands of HPC-based parallel AI jobs over NFS [Fig.2], achieving industry-leading performance and scalability, all while reducing costs associated with specialized hardware investments.

Cloud Native, Cloud Ready

Qumulo's AI-ready data platform is purpose-built to meet the demands of modern AI applications, offering seamless integration, unparalleled performance, and cost-effective scalability for managing massive volumes of unstructured data. Qumulo's integrations with Azure Copilot and Amazon Q for Business enhance operational efficiency by automating data retrieval and enriching AI-driven queries, positioning Qumulo as the optimal choice for AI-driven organizations seeking to maximize performance and value.

On-Premises HPC AI Training

"Our UC San Diego customers require dedicated networks for storage interconnect, based on a standard ethernet infrastructure at up to 200 Gbps of provisioned performance. The Data Science Machine Learning Platform (DSMLP) runs on Qumulo, where thousands of students execute performance-sensitive AI workloads concurrently. These students require optimal configurations to ensure efficiency when running GPUs over thousands of NFS connections."

~Brian Balderston, Director of Infrastructure at the San Diego Supercomputer Center

 | info@qumulo.com