Omdia by informa techtarget

OCTOBER 2025

Google Cloud Managed Lustre

Tony Palmer, Practice Director, Technical Validation, Omdia

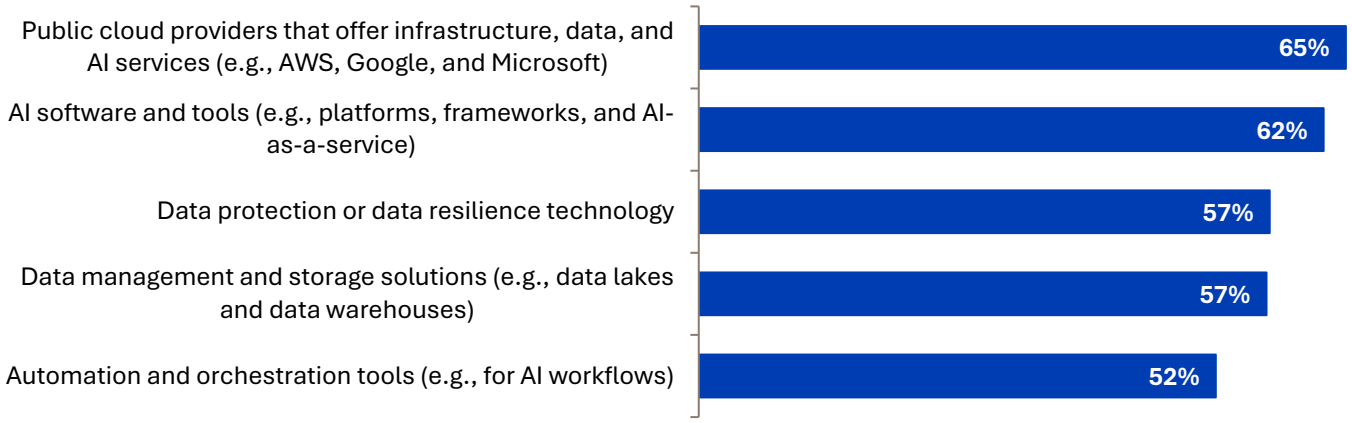
AI Infrastructure Challenges

AI is profoundly affecting nearly every facet of how businesses operate today. Storage infrastructure is facing heightening requirements for massive data storage capacity to support AI projects while keeping the environment secure and meeting performance demands. This First Look shows how Google Cloud Managed Lustre satisfies these requirements, decreasing training time, improving inference performance, and driving business value.

Organizations strongly believe in the potential of AI to drive transformational value. In fact, 84% of organizations surveyed by Enterprise Strategy Group said AI is critical to their future strategy. Most recognize the need for significant infrastructure investments, considering that 87% reported that AI adoption will drive or is driving substantial data growth, while 70% stated that storage-related challenges present a significant barrier to AI success.[1] In another Enterprise Strategy Group research survey, respondents identified public cloud as a leading area of AI investment compared with other technologies (Figure 1).[2]

Figure 1. Cloud Services Are Targeted for AI Infrastructure Investments

Which of the following areas are part of your organization's IT or cloud infrastructure investment plan to support its AI initiatives? (Percent of respondents, N=317, multiple responses accepted)



Source: Omdia

These issues highlight the key role that the storage environment is set to play in driving enterprise AI success.

[1] Source: Enterprise Strategy Group Research Report, *The Critical Role of Storage in Building an Enterprise AI Infrastructure*, September 2025. All Enterprise Strategy Group research references in this Technical First Look are from this report unless otherwise noted.
[2] Source: Enterprise Strategy Group Research Report, *IT Transformed: Inside the Convergence of Hybrid Cloud and AI* , July 2025.

Google Cloud Managed Lustre for AI/ML Workloads

Managed Lustre, a parallel file system based on DDN's EXAScaler software and offered as a fully Google-managed service, provides multi petabyte zonal persistent storage, 1 TB/sec throughput, and sub-millisecond latency for reads. The Managed CSI driver for Google Kubernetes Engine (GKE) simplifies operations and allows tens of thousands of GKE clusters to access the same Lustre instance. Managed Lustre technology and capabilities make it well suited to address key storage requirements for AI workloads. Because parallel file system solutions are so common in enterprise on-premises deployments, the transition to Google Cloud should be reasonably straightforward.

Enterprises' total storage corpus—all their training data—consumes petabytes. A majority of respondent enterprises (55%) have 1 PB or more of installed capacity across all locations, with almost one in five (19%) having 10 PB or more. Because AI workload requirements are often not known in advance, the AI data pipeline requires data to be curated for training and fine-tuning a model, and this is just one of the ways where Managed Lustre and Cloud Storage integration is valuable. Organizations can easily import or export data from Managed Lustre to Cloud Storage, simplifying management of the overall lifecycle of the data in their AI pipelines.

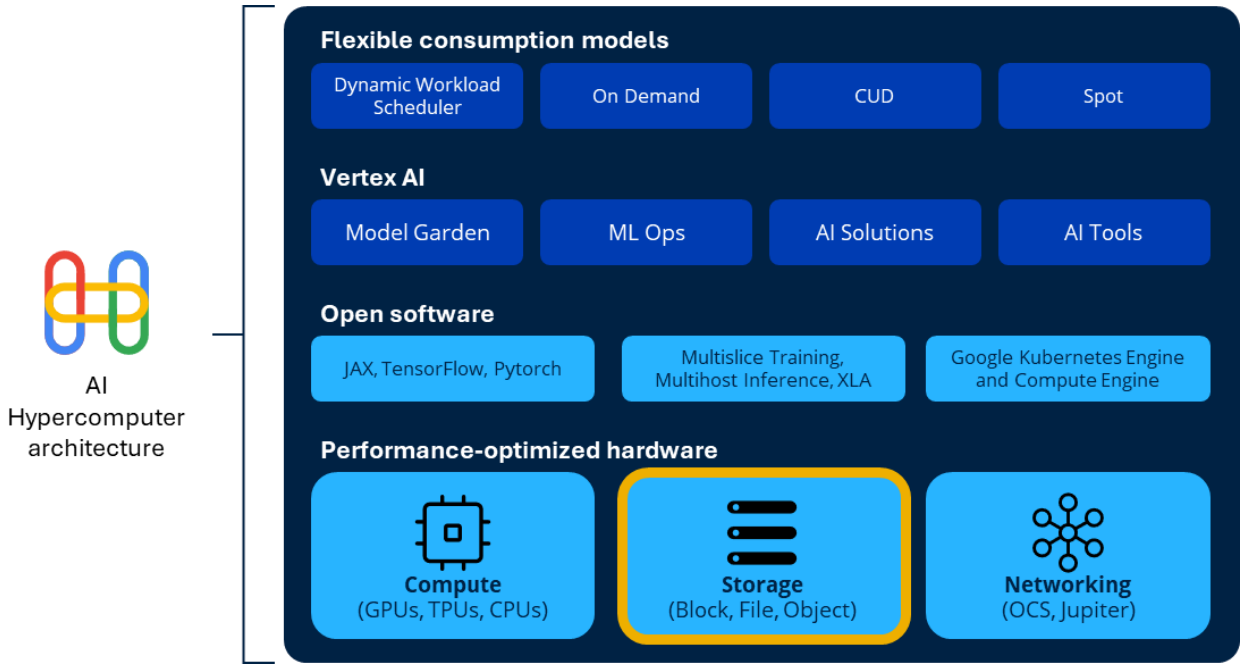
Saving and restoring checkpoints depends on the model size (number of parameters), which typically dictates the size of the Checkpoint, most commonly gigabytes to terabytes. Additionally, the time to load models for inference or serving depends on the model size (typically gigabytes to terabytes), the period of updates, and the number of hosts. Managed Lustre is well suited for long context serving like key-value (KV) cache and virtual Large Language Model (vLLM) with GPU integration. KV cache is a technique that helps speed up repetitive calculations by remembering important information from previous steps. Rather than recomputing everything, the model reuses what it has already calculated, making text generation faster and more efficient. vLLM helps solve resource requirements of AI models by improving memory use and processing speed, making the models faster and cheaper to run.

First Look

Managed Lustre is a fully managed, high-throughput, low-latency parallel file system with high concurrency. Managed Lustre provides highly scalable persistent storage—in both performance and capacity—that is built on an architecture optimized for AI/ML and high-performance computing workloads.

Managed Lustre is a core component of Google Cloud's AI Hypercomputer, a system of technologies encompassing hardware, software, and consumption models. AI Hypercomputer builds on a foundation of performance-optimized hardware across compute, storage, and networking delivered as ultra-scale data center infrastructure. Storage is a foundational element that underpins the ability to deliver services throughout the stack.

Figure 2. Google Cloud's AI Hypercomputer



Source: Google Cloud and Omdia

Google Cloud's Open Software layer enables extensive and optimized support for leading ML frameworks, including JAX, TensorFlow, and Pytorch, with the ability to scale AI workloads across physically disparate hardware deployments with multi-slice training and multi-host inference.

Google Kubernetes Engine integration for orchestration helps developers accelerate deployments and maximize the utilization of the underlying hardware.

Highly flexible consumption models, including Dynamic Workload Scheduler, designed specifically for AI workloads, make it simple and cost effective to consume these services.

Vertex AI, Google Cloud's fully managed, unified AI development platform for building and using generative AI, runs on top of the AI Hypercomputer architecture.

Google Cloud Managed Lustre Delivers AI-optimized Infrastructure

This Omdia Technical First Look examines how Managed Lustre optimizes infrastructure for AI across training, checkpointing, and delivery.

Large-scale deep learning models depend on extensive datasets for training. Managed Lustre provides distributed data access that accelerates training processes. This approach supports improved model accuracy and facilitates complex AI project execution. The system's scalable architecture maintains consistent performance as data volumes increase, delivering massive parallelism with millions of IOPS and up to 1 TB/sec of read throughput, accelerating data pipelines for AI with sub-millisecond latency and the ability to scale from terabytes to petabytes, while delivering simplicity of deployment and scaling to minimize

storage-related performance limitations. It's important to note that Google Cloud offers multiple performance and capacity tiers for Managed Lustre.

Managed Lustre also shines for checkpoints (writes) across multiple parallel jobs. Managed Lustre's ability to facilitate more frequent checkpoints complements Google Cloud Storage, which is why it is often used in conjunction for long-term archiving. This ensures organizations can maximize GPU utilization and ensure rapid recovery.

Maximizing GPU/TPU utilization for serving and inference should be a top priority, as model load times can vary by storage solution, node count, and model size. Managed Lustre significantly accelerates AI inference and model serving by providing high-throughput, low-latency read and write access to massive datasets. Managed Lustre distributes data across multiple storage nodes, enabling concurrent access by numerous GPUs. This parallel access eliminates bottlenecks that occur with traditional file systems, enabling AI models to rapidly ingest and process the vast amounts of data required for training, inference, and serving.

High per-VM throughput (>20 GB/sec) and aggregate cluster throughput reduce model load time and can be used for any number of serving VMs.

Conclusion

AI is fundamentally transforming business operations across all sectors. Organizations now require storage systems that can handle enormous data volumes for AI initiatives, while simultaneously maintaining robust security and delivering the high-performance capabilities these projects demand.

Organizations recognize AI's transformational potential, with 84% considering it critical to their future strategy, according to Enterprise Strategy Group research. However, infrastructure challenges persist: 87% reported actual or expected AI-driven data growth, and 70% cited storage issues as significant barriers to AI success. Research also indicated that public cloud leads AI technology investments.[3]

Managed Lustre delivers massive parallelism with millions of IOPS and up to 1 TB/sec of read throughput and accelerates data pipelines for AI with <1ms latency, which can have a massive impact on AI training. Managed Lustre's massive parallelism shines for checkpoints (writes) across multiple parallel jobs as well, ensuring organizations can maximize GPU utilization and ensure rapid recovery. It is also a cost-effective and performant way to serve models, with high per-VM throughput (>20 GB/sec) and aggregate cluster throughput, reducing model load time for any number of serving VMs. Delivered as a fully managed service, Managed Lustre removes uncertainty and reduces risk by eliminating the need for configuration and maintenance by customers.

Omdia validated that Google Cloud Managed Lustre delivers a high-performance, fully managed parallel file system optimized for AI applications. With multi-petabyte scale and up to 1 TB/sec throughput at sub-millisecond latency, Managed Lustre facilitates the efficient migration of demanding AI workloads to the cloud with uncompromising performance.

[3] Ibid.

Get in touch: www.omdia.com askananalyst@omdia.com