WHITE PAPER

Object Storage in Enterprise AI

As enterprise AI adoption accelerates, object storage has transformed from a simple archival solution into a high-performance backbone for modern data infrastructure. With innovations like RDMA acceleration, it can deliver the low-latency, high-throughput access required by demanding AI workloads.

In this environment, object storage can address many key data management challenges to optimize enterprise data infrastructure, manage surging data volumes, and accelerate AI initiatives. Supermicro's portfolio includes several high-performance storage platforms specifically engineered for AI workloads on object storage, delivering the high throughput and low latency required for both inference and training.

If organizations aren't careful, massive amounts of data can result in high energy use. Efficiency is key for data-intensive applications, and AMD EPYC™ server CPUs power the most energy-efficient servers available.

In addition to delivering overall better performance per watt, AMD EPYC server CPUs make it possible to closely match CPU resources to application requirements, resulting in even greater efficiency. For example, some analytic applications do not scale well to high core counts. Using high-frequency AMD EPYC server CPUs can increase per-core performance, speeding these applications without the burden of carrying additional cores not essential to the mission. Some technical computing applications operate best when processors are equipped with large L3 caches.

AMD EPYC processors with AMD 3D V-Cache™ technology free CPUs to process data with fewer cache misses, thereby enabling unimpeded performance.

The State of AI Storage Today

Data fuels AI. To extract insights, businesses must collect, process, train on, and be retrained on petabytes of data. This data needs to be stored and readily accessible to powerful compute engines throughout the process. As organizations drive AI innovation, their data storage infrastructure must evolve to keep up.

AI needs high-performance, highly scalable storage. Object storage can now meet these requirements and sits at the heart of modern AI strategies.

The AI Workflow

AI data workflows follow a relatively straightforward, widely understood process. The key to implementing them, however, is to ensure that the business requirements that address the problem always come first. These requirements guide decisions across the technology stack and data sources at every step of the workflow, as well as which datasets and AI models are best suited to an organization's needs.

An AI workflow is made up of four general stages:

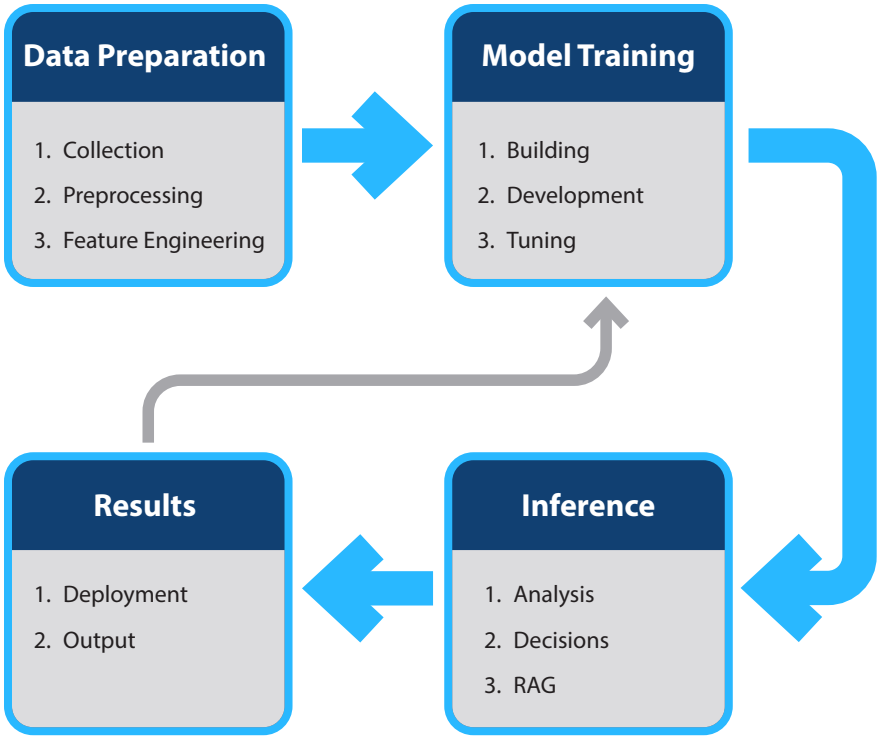

Figure 1: *Steps in an AI Workflow*

SUPERMICRO PETASCALE STORAGE SERVERS

ASG-1115S-NE316R

SSG-121E-NES24R

ASG-2115S-NE332R

Supermicro Petascale storage servers support all-flash data lakes and feature high-density, high-performance Enterprise Datacenter Standard Form Factor (EDSFF) NVMe drives for future-forward compatibility and growth. Our 16- and 32-drive servers, powered by a single 4th Gen AMD EPYC™ server CPU with up to 128 cores for high throughput, are populated with EDSFF E3.S drives, and our 24-drive system is a 1U server with ESDFF E1.S drives for high density. Each server supports PCIe 5.0 x16 network interfaces for connectivity with the fastest interfaces available, such as the NVIDIA ConnectX®-7 card with 400 Gb/s InfiniBand.

Data preparation

The first step is to prepare the data for the AI model.

1. **Data collection:** Data is collected from various sources, including siloed enterprise applications, databases, business documents, sensor readings, images, videos, or log data. Data will mainly be unstructured (documents, videos, etc.) with some structured data from applications and databases.

2. **Preprocessing:** After being extracted into a staging area, the raw data is transformed to meet the model's requirements. Data often must be normalized, debiased, or otherwise filtered. For example, image data may need to be cropped to the same scale or rotated to the same orientation, etc. Data at edge locations may be transformed in place before being sent back to the central location. This process, called ETL (extract, transform, and load), is powered by CPUs.

3. **Feature engineering:** Raw data is refined into a format the model can use, enhancing its performance and accuracy. This process involves creating more meaningful features by combining variables, transforming non-numeric data into numerical values, and selecting only the most relevant features to focus the model's efforts.

The data preparation steps can be time-consuming if the enterprise has not previously cataloged data sources and implemented a data lake.

Raw data is often spread across clouds, data centers, edge locations, and archival systems. To conserve bandwidth, data from remote locations is sometimes cleaned before being sent back to a central storage location. Ultimately, the data is staged for subsequent training, using massive, scalable storage systems. This is where object storage systems shine: centralizing large amounts of diverse data, scaling effortlessly to the size needed, and connecting them to CPU-based systems for processing.

How much capacity will you need? This depends on factors such as data type (e.g., text, images, videos) and the algorithm chosen for the business use case. These provide a starting point for estimating data capacity. Higher model accuracy may require more data, and new algorithms can impact capacity needs. Some organizations assign storage quotas to researchers and adjust them regularly to meet evolving requirements.

SIMPLY DOUBLE SYSTEMS

The Supermicro Simply Double systems double the number of hard-disk drives most 2U systems can support, providing for up to 24 with the M.2 (standard 3.5") form factor. Also powered by AMD EPYC™ server CPUs, with your choice of core count to suit the application, Supermicro SimplyDouble efficiently and compactly supports object storage systems for data lakehouses and enterprise AI.

Model Training

Once the data is prepared, the process of building and training the AI model begins.

1. **Model building:** An AI model is a serialized file containing parameters (weights) developed via machine learning, packaged with source code to define its architecture. In most enterprise use cases, data scientists select one or more foundational models that meet the business needs for their AI workflow. Foundational model types include large language models (LLMs), multimodal models, generative adversarial networks (GANs), and computer vision models. Data scientists can also train their own model—for instance, when the goal is to identify patterns specific to the organization's industry sector or area of inquiry using the organization's own proprietary data.

2. **Model development:** AI foundational models are trained on massive amounts of data to generate outputs across a wide range of inputs (e.g., prompts).[1] After a foundational model is selected, data scientists can further train it on the data they have collected so that the analysis and decisions align with business requirements.

3. **Hyperparameter tuning:** These models' inference behavior can also be fine-tuned to meet the business requirements before deployment. The model's configuration settings, known as hyperparameters, are adjusted so that the model's outputs align with those most productive for the business needs and context.

A fast, scalable storage system is vital for model training and inference. It needs to be quick to load large datasets into GPU memory. Models may outgrow their storage space as they are used. The models produce data in the form of checkpoints, logs, and decisions. It is critical to be able to scale capacity.

The data fabric that connects storage devices to GPU memory includes PCIe buses, network cards, DPU (data processing units), and other components that sit between the storage and the GPU. Storage systems for AI should support this broad range of connectivity to meet the growing variety of AI needs.

Another storage consideration is managing checkpoints. These are similar to snapshots, allowing data scientists to roll back to a point in time if the training results begin to diverge from the targeted outputs, and to save intermediate state results in case of failures that require restarting the training process.

Inference

Once trained, the model can operate on new data it has not seen before, generate relevant follow-on data, make predictions, and provide recommendations or decisions.

1. **Analysis:** This is the form of AI most commonly used today, generative AI, in which a prompt yields likely conclusions, answers, or content. The prompt does not have to be in language form; it could, for example, be an incomplete DNA sequence, with the generated output predicting what DNA should fill the gaps.

2. **Decisions:** Based on its analysis, the model produces either directives for other applications (AI or non-AI) to follow or recommendations for human reviewers to act on, whether via a workflow or a more informal process.

3. **RAG (retrieval-augmented generation):** Inferencing that uses RAG adds content retrieved on the fly to the prompt, via search or another method. This content could be proprietary corporate documents formatted and stored in vector databases, straightforward public search results, or almost anything in between. The additional data sharpens the model's context, enabling it to produce more relevant and useful outputs. RAG inference requires a data lake of enterprise-specific data, which can be used to improve the GenAI output to include company-specific proprietary data.

The value of the AI solution is delivered at this stage, so the ability to scale out capacity and performance is more critical than ever. Storage must be able to store and deliver data for inference.

Results

The final phase focuses on delivering the model's output and ensuring sustained performance.

1. **Deployment:** The fully trained and optimized AI model is integrated into a real-world application or system, such as a customer support chatbot or a predictive maintenance system.

2. **Output:** The workflow delivers results, such as an automated response to a customer, a machine failure prediction, or a data-driven insight. The system should be continuously monitored and provide feedback to refine the model further and improve the workflow over time.

Storage Requirements Across the AI Lifecycle

The AI workflow is a multi-phase process, and each phase interacts with data storage systems. From preparing raw data to delivering actionable insights, the storage system must adapt to varying requirements for performance, scalability, and efficiency. The table below outlines the key goals and storage needs for each stage of the AI lifecycle.

AI Workflow Stage	Stage Goal	Storage Requirements	Estimated Capacity
Data Preparation	Prep data so it suits the AI model	• Efficiently handle large files and small files. • Robust metadata management. • High I/O performance to accelerate data cleaning and feature engineering.	Massive GBs to PBs
Model Training	Develop, train, and tune the model	• Provide training data to keep GPUs utilized using high-performance, scalable storage with low latency. • Save and restore checkpoints.	Massive GBs to PBs
Inference	Provide reliable predictions or recommendations from input data prompts	• Provide rapid, concurrent access to models and data. • Ensure minimal latency for predictions and insights.	Dependent on use case, user adoption MBs to TBs
Results	Integrate into applications and support refinement and responsiveness to new conditions	• Provide scalability for massive volumes of results and logs. • Ensure data durability and availability. • Provide low-latency, high-throughput performance for real-time applications.	Dependent on use case, user adoption MBs to TBs

Table 1: *Storage requirements for each AI phase*

The Challenges of Modernizing Data Architectures For AI

As AI unlocks the value of business data, choosing the right storage solution is a critical yet multifaceted decision. The demands of AI workflow stages sometimes conflict: data collection requires cost-effective, high-capacity storage, while training and inference need high-performance solutions to maximize GPU efficiency. High-capacity storage is typically too slow for the intense demands of model training, while high-performance storage is too expensive to house an entire organization's data.

Object storage-based data lakes and lakehouses offer an effective solution for this dilemma.

Data lakes: a foundation for AI workflows

Data lakes, built on object storage, emerged as a solution to the growing need for centralized repositories capable of storing vast amounts of structured, semi-structured, and unstructured data. Modern data lakes now support flexible access via file protocols (NFS, SMB), object storage (S3), and legacy systems (HDFS).

While data lakes excel at supporting the data preparation and results analysis stages of AI workflows, they fall short in meeting the high-performance demands of model training and inference.

Data lakehouses: bridging the gap

The data lakehouse model combines the flexibility and low-cost storage of data lakes with the robust management and structuring capabilities of data warehouses. By adding a metadata management layer to the object store, lakehouses enable better data organization, governance, and access.

Modern data lakehouse features include:

- **ACID (atomicity, consistency, isolation, durability) transaction support:** to provide data integrity during concurrent read/write operations.
- **Schema enforcement or scheme validation tools:** maintain data quality and prevent inconsistencies by rejecting writes that don't match the schema.
- **High-performance data access:** to meet the demands of AI model training and inference, ensuring GPUs and accelerators operate at full capacity.

By efficiently managing and streamlining access to large-scale datasets, data lakehouses address the performance bottlenecks inherent in traditional data lakes, making them a compelling choice for AI-ready architectures.

The evolving nature of AI data

Modern AI workloads are reshaping storage requirements at the file level. Unlike traditional systems optimized for large files, AI and analytics workloads often involve processing millions, or even billions, of small objects, such as images, audio clips, or text documents.

This shift introduces several challenges:

- **High metadata overhead:** Small files generate significant metadata, straining traditional storage systems.
- **Intensive I/O operations:** Inefficient handling of small objects can create bottlenecks during data preparation and model training, slowing down AI pipelines.
- **Data growth and diversity:** AI models increasingly rely on a mix of structured, unstructured, and semi-structured data, requiring architectures that can scale and adapt to manage this variety.

The "overcloudification" of AI

While public cloud platforms offer unparalleled flexibility and scalability for AI workloads, they also introduce significant challenges:

- **Data transfer costs:** Moving large datasets between on-premises systems and the cloud can result in spiraling expenses.
- **Security and compliance risks:** Regulated industries often face restrictions on moving sensitive or proprietary data to the cloud.
- **Performance trade-offs:** The model training and inference stages of an AI workflow require optimized infrastructure that delivers low-latency, high-bandwidth access to data.
- **Data gravity:** Once data is stored in the cloud, it becomes difficult and costly to move, leading to vendor lock-in.

Organizations must carefully evaluate the trade-offs between cloud-based and on-premises solutions, particularly when sensitive or proprietary data is involved.

Balancing performance and cost

Achieving consistently high IOPS (input/output operations per second) without overspending is a critical challenge in designing AI-ready architectures. High-capacity storage may be affordable, but it lacks the speed required for AI workloads. However, high-performance storage is fast but expensive.

A well-architected solution must balance scalability, flexibility, and performance to support AI initiatives effectively.

The Latest Innovations in Object Storage Benefit AI

Disaggregated architectures

Disaggregated object storage architectures decouple compute and storage resources, allowing organizations to scale each independently. This decoupling enables the ability to add additional compute power, such as GPUs, when required to process AI workloads. It also enables expanding storage capacity as data volume increases, a classic strength of object storage.

By eliminating the rigid, server-bound constraints of traditional systems, disaggregated architectures organize storage into flexible pools accessible to any server over high-speed networks. This approach acts as a central data hub, supporting multiple stages of the AI workflow, from training clusters to inference engines.

Key benefits include:

- **Eliminating silos:** Shared data pools reduce redundant copies and simplify workflows.
- **Unified storage layer:** Balances high-capacity, low-cost needs with high-performance throughput demands.
- **Rapid experimentation:** Data science teams can quickly set up environments, test models, and access datasets without waiting for IT provisioning, accelerating development cycles.

RDMA acceleration

In high-performance computing, every microsecond matters. Remote Direct Memory Access (RDMA) accelerates data transport by allowing one computer's processor to access another's memory directly, bypassing the operating system and CPU. Compared with traditional TCP/IP-based communication, RDMA reduces software overhead, resulting in significantly lower latency and higher throughput.

RDMA is particularly valuable for AI workloads, where massive datasets are constantly moved between storage and GPU servers. Benefits include:

- **Maximized GPU utilization:** Faster data access minimizes idle time, shortening training cycles.
- **Efficient handling of small files:** RDMA reduces per-operation overhead, enabling the system to process millions of small objects efficiently, critical for data preparation, checkpointing, and inference.
- **Support for disaggregated architectures:** RDMA means storage nodes can deliver data at the speeds required for demanding AI workflows.

Object Storage Features for Enterprise AI

Object storage organizes data as discrete objects with metadata and unique identifiers, making it ideal for managing massive datasets and diverse file types. Modern AI development relies on seamless integration with data access protocols, APIs, and support frameworks, all of which are elements of many object storage systems today.

- **Ecosystem integration:** Robust API support, compatibility with data management platforms, and support for access protocols including POSIX (portable operating system interface), NFS (network file system), SMB (server message block), and CSI (container service interface).
- **Support S3-compatible APIs:** These have become the standard for cloud-native applications and AI frameworks, enabling direct dataset loading into training scripts.
- **Provide native integration for AI frameworks:** Use standard libraries and S3 endpoints to enable PyTorch, TensorFlow, and CUDA, reducing the need for custom connectors and enabling data scientists to focus on model development over pipeline management.
- **Enable end-to-end workflows:** A well-integrated system offers smooth transitions from data preparation to model training and inference.

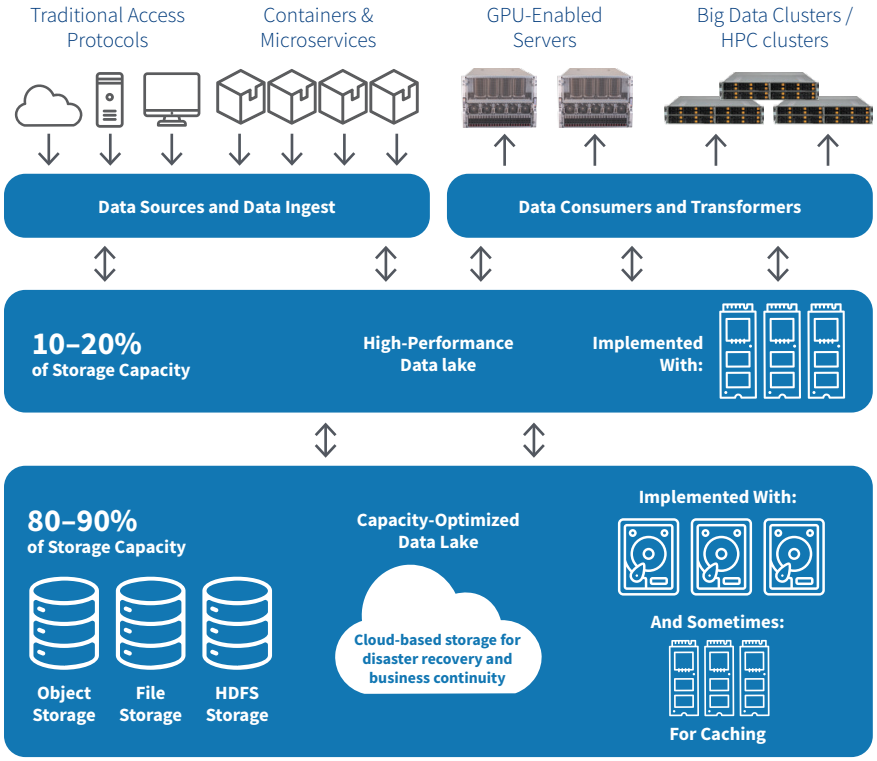

Figure 2: *Tiered Data Lake Architecture for AI Workflows*

Modern, Well-Integrated Object Storage Arrays

At the heart of AI-ready architecture is physical storage. Modern object storage arrays are designed to meet the challenges of data-intensive workloads by combining:

- **High-density, cost-effective capacity:** For storing large datasets.
- **High-performance flash:** To handle hot data and high throughput demands, either through tiering or via all-flash arrays.
- **Disaggregated scalability:** Compute and storage scale independently, aligning with business needs.

These arrays eliminate silos, reduce complexity, and provide the flexibility needed to support evolving AI workflows.

Use Case: Buttressing the AI Infrastructure Strategy with Object Storage

The following example scenario illustrates how an organization might use an object storage strategy to prepare its infrastructure for future AI initiatives while addressing current challenges.

The Challenge

An IT infrastructure team, led by a forward-thinking decision-maker, recognizes the need to modernize its data architecture. While their current high-performance storage systems are sufficient for existing siloed analytics workloads, they lack the scalability, flexibility, and cost-efficiency required for future, more complex AI initiatives.

Anticipating exponential growth in data and the eventual deployment of diverse AI models, the team seeks a durable, scalable solution that avoids a disruptive "rip-and-replace" overhaul. Their goal is to build a future-proof infrastructure that evolves with the company's AI ambitions.

The Solution

The organization adopted a modern, disaggregated object storage solution as the foundation of its long-term data strategy. To optimize performance for demanding AI workloads, the architecture incorporates RDMA acceleration, enabling high data throughput with minimal latency.

Greenfield vs Brownfield Integration

The team deployed object storage as a new centralized data lake for ingesting and archiving structured and unstructured data. This greenfield approach is non-disruptive and delivers scalable, cost-effective capacity immediately. However, taking a brownfield approach, where an existing data lake is migrated or adapted to a data lake built on object storage, is frequently desirable as well.

Scalable foundation for AI pipelines

As data science teams start experimenting with AI, the object storage system serves as a single source of truth. Its S3-compatible API allows seamless integration with AI frameworks such as TensorFlow and PyTorch, enabling direct access to data for model training and experimentation. This eliminates the need for time-consuming data copying, accelerating development cycles.

Cost-effective and flexible scaling

The disaggregated architecture allows independent scaling of storage and compute resources. As data volumes grow, the organization can add storage nodes without over-provisioning compute. When ready to deploy AI models at scale, they can expand GPU clusters with confidence that the storage backend will meet data throughput demands.

Outcome

By implementing object storage, the organization establishes a scalable, cost-effective infrastructure that supports its long-term AI strategy. This future-proof solution aligns with industry trends and ensures compatibility with emerging tools and innovations.

The enterprise is now positioned to innovate at its own pace, confident that its data architecture can handle increasingly sophisticated AI workloads without incurring prohibitive costs or requiring disruptive changes. This strategic investment enables the organization to manage data at scale and accelerate its transformation into an AI-driven enterprise.

Conclusion: The Significance of Object Storage in Enterprise AI

To successfully navigate the complexities of enterprise AI, organizations must ensure their data architecture is prepared for future demands. By leveraging object storage-based data lakes and lakehouses, enterprises can maintain control over sensitive data, ensure compliance with regulatory requirements, and avoid the performance trade-offs and costs associated with cloud-based solutions.

Modern object storage offers the flexibility, performance, and cost-efficiency necessary to support the full AI lifecycle, from data preparation to model training and inference. Disaggregated architectures and newer capabilities of object storage power scalable foundations that adapt to data growth and prevent performance bottlenecks.

By implementing these strategic infrastructure choices, enterprises can fully realize AI's transformative potential for sustained success.

Why Supermicro?

As AI continues to reshape industries, the demands on data storage systems grow exponentially. From data preparation to inference and results delivery, each stage of the AI workflow requires storage solutions that balance scalability, performance, and cost-efficiency.

Supermicro's high-performance storage servers, designed specifically for AI workloads, provide the low-latency, high-throughput capabilities needed to keep GPUs fully utilized during training and inference. With innovations such as RDMA acceleration and disaggregated architectures, Supermicro enables organizations to build scalable, future-proof infrastructures that support the entire AI lifecycle, providing seamless performance and adaptability for evolving AI demands.

Choosing Supermicro simplifies the purchase process. With our storage and AI software partners, you can purchase a complete solution from Supermicro. This includes sizing and server selection, installation, rack-level delivery, and ongoing software and hardware support with a single point of contact.

Our partnership with AMD further enhances performance and efficiency, delivering innovative solutions for AI and storage workloads.

Supermicro's partnerships for software-defined storage provide the complete foundation for your AI, data lake, and data lakehouse use cases.

Cloudian Hyperstore is an on-premises, S3-compatible data lake platform, particularly suited for AI/ML applications. HyperStore can scale to exabyte levels without disruption to operations.	DDN offers the EXAScaler parallel file system to support a wide range of data-intensive workflows in a hyperconverged model. It supports POSIX, NFS, SMB, HDFS, and S3 access to storage. Its Infinia product supports file and object stores.	EDB Postgres AI solves common enterprise challenges by extending enterprise-grade Postgres with native AI vector processing, an analytics lakehouse, and a unified platform for observability and hybrid data management.
Hammerspace Global Data Platform unifies unstructured data across edge, data centers, and clouds. It provides extreme parallel performance for AI, GPUs, and high-speed data analytics, and orchestrates data to any location and supports a wide range of protocols including NFS, pNFS, SMB, S3 and CLI.	IBM Storage Ceph is a software-defined scale-out enterprise storage platform that provides block, file, and object capabilities. Its best in class S3 capabilities support the largest and most intensive data applications.	MinIO Enterprise Object Store is an ultra high-performance object store that is used to deliver against AI/ML, analytics and archival workloads - all from a single platform. Software-defined, MinIO is ideal for a broad class of SuperMicro servers.
OSNexus QuantaStor is a scale-out shared-storage solution that delivers multi-tenant, S3-compatible, object storage with unique dynamic tiering features to optimally place objects for maximum performance and cost efficiency.	Quantum ActiveScale merges flash, disk, and optionally, tape libraries, to build data lakes, storage clouds, and computing clusters at any scale with outstanding performance, efficiency, availability, and durability at up to 80% lower cost than alternative solutions.	Qumulo is a single platform for all unstructured data, no matter where that data resides. Qumulo is built for managing geographically distributed file and object data on-premises, at the edge, in the core, and in the cloud.
Scality solves organizations' biggest data storage challenges — security, performance, and cost. The world's most discerning companies trust Scality so they can grow faster and execute AI data-driven ideas quicker — while increasing efficiency and avoiding lock-in.	VAST Data enables data-intensive enterprises to effortlessly capture, catalog, refine, and enrich data through an API-driven data pipeline, providing real-time insights for applications like agentic AI. VAST empowers organizations to challenge conventional thinking and unlock transformative possibilities through its innovative data platform.	WEKA delivers a software-defined data platform that helps get your data out of silos and islands and into streaming data pipelines to fuel information workloads like AI and HPC.

Table 2: *Supermicro software-defined storage partners supporting data lakes*

For more information:

- Storage Servers Solutions For Enterprise Architectures | Supermicro
- AI Storage Solutions

[1] https://www.supermicro.com/en/glossary/foundation-model

Supermicro (NASDAQ: SMCI) is a global leader in Application-Optimized Total IT Solutions. Founded and operating in San Jose, California, Supermicro is committed to delivering first-to-market innovation for Enterprise, Cloud, AI, and 5G Telco/Edge IT Infrastructure. We are a Total IT Solutions manufacturer with server, AI, storage, IoT, switch systems, software, and support services. Supermicro's motherboard, power, and chassis design expertise further enable our development and production, enabling next-generation innovation from cloud to edge for our global customers. Our products are designed and manufactured in-house (in the US, Taiwan, and the Netherlands), leveraging global operations for scale and efficiency and optimized to improve TCO and reduce environmental impact (Green Computing). The award-winning portfolio of Server Building Block Solutions® allows customers to optimize for their exact workload and application by selecting from a broad family of systems built from our flexible and reusable building blocks that support a comprehensive set of form factors, processors, memory, GPUs, storage, networking, power, and cooling solutions (air-conditioned, free air cooling, or liquid cooling).