



Powering the Data Lakehouse

Harnessing Supermicro and AMD Technologies to Optimize Data Lakehouse Performance for Advanced Analytics

As data volumes continue to grow exponentially, AI inference and big data analytics demand systems that can scale, while providing real-time processing capabilities.

A data lakehouse combines the horizontal scalability of data lakes with the transactional management of a data warehouse. Lakehouses decouple storage from compute, so that compute resources can be scaled on demand and queried by different tools that support open formats.

In combining the data warehouse and the data lake, the lakehouse becomes a new entity, one that includes the warehouse's transactional layer. This simplified architecture enables greater horsepower and agility for AI, BI, ML, and really, all analytics.

Supermicro offers a range of solutions built to support the multiple types of data lakehouses discussed in this paper.

AMD EPYC

SUPERMICRO AND AMD: SCALABLE SYNERGY

AMD EPYC processors and Supermicro solutions provide powerful platforms for AI and data analytics, particularly in data lakehouse implementations. AMD EPYC processors are ideal to this task, with superior performance, flexibility, security, and scalability. When coupled with Supermicro's expertise in server design and data management solutions, the results are highly efficient and effective infrastructures for modern data-intensive applications.

Data Lakehouses: Bridging Data Lakes and Warehouses

Organizations have long been faced with the challenge of balancing the strengths and limitations of both data lakes and data warehouses. These two solutions had often operated in silos, leading to inefficiencies, redundant storage, and missed opportunities for truly integrated data analytics.

Enter the data lakehouse—a hybrid architecture designed to bridge the gap between these two data giants. A data lakehouse combines the best of both worlds: it inherits the low-cost, scalable storage capabilities of data lakes while incorporating the structured data management, governance, and reliability features typically associated with data warehouses. The result? A unified platform where all data types—structured, semi-structured, and unstructured—can coexist, be governed, and accessed seamlessly.

The rise of the data lakehouse was, in many ways, a natural evolution, emerging as a direct response to the limitations of its predecessors. While data warehouses were highly effective at managing structured data for business intelligence, they struggled with unstructured data and lacked the flexibility needed to support the growing demands of modern analytics. Conversely, data lakes excelled at storing vast amounts of diverse data types, but their lack of governance and structure posed challenges for reliability and control. Organizations needed a way to harmonize the two, and that's where the data lakehouse came in—offering a solution that could handle both structured and unstructured data with the governance and scalability required for modern data workloads.

The core goals of a data lakehouse are simple but impactful. First and foremost, it serves as a single source of truth, eliminating the need for redundant data storage across multiple systems. By consolidating data from various sources into one platform, organizations can ensure consistency and accuracy. It also supports a broader range of workloads—ranging from traditional business intelligence (BI) to more advanced machine learning (ML) and artificial intelligence (AI)—without sacrificing performance or flexibility.

Two Perspectives

While some see the data lakehouse as including the data lake, others view it as a separate transactional layer built on top of lake storage. This paper follows the latter perspective: the data lake handles raw storage—be it disk-based or all-flash—while the lakehouse provides the compute and transactional features. Because a lakehouse can simply inherit the lake's storage layer, it does not require its own separate data store.

Scalability and Flexibility

By leveraging the elasticity of a data lake, organizations can store huge datasets without constant expansions or migrations. At the same time, compute and storage layers scale independently, so the infrastructure grows smoothly alongside business needs.

SUPERMICRO SOLUTIONS FOR DATA LAKEHOUSES LEVERAGING AMD EPYC PROCESSORS

Supermicro's server solutions are engineered to fully leverage the power of the 5th Gen AMD EPYC processors, delivering a robust infrastructure optimized for high-performance, scalable data lakehouses.

OPTIMIZED STORAGE

Supermicro's Petascale all-flash storage servers provide high-density NVMe storage, perfectly suited to meet the performance demands of modern data lakehouses, powered by 5th Gen AMD EPYC processors.



ASG-1115S-NE3X12R

A compact 1U Petascale storage server featuring a single socket SP5 AMD EPYC 9004 Series processor. It supports 24 DIMMs (2DPC) and offers 12 bays with 8 front hot-swap EDSFF E3.S NVMe drives and 4 fixed PCIe 5.0 x8 CXL Type 3 drive bays. Ideal for data lake environments needing robust performance and high-density storage.



ASG-1115S-NE316R

A 1U all-flash storage server designed for data lakes, supporting up to 16 EDSFF E3.S NVMe drives. Powered by a single 5th Gen AMD EPYC processor with up to 192 cores, it offers high throughput and PCIe 5.0 x16 network interfaces, along with 24 DIMMs and up to 9TB of memory.

Governance and ACID Transactions

The lakehouse brings data warehouse-grade governance to the broader ecosystem. Features like schema-on-read enforcement, data cleansing, and auditing ensure data integrity. ACID transactions—covering atomicity, consistency, isolation, and durability—enable reliable updates, even in concurrent processes, making the lakehouse far more robust than a raw data lake.

Diverse Workloads

Real-time streaming, BI analytics, and ML/AI can all live on the same platform, thanks to the lakehouse's flexible design. Support for popular open-source tools (Apache Spark, Python, R) lets data scientists run advanced models directly against the same data used by BI teams, fostering unified analytics.

Data Democratization

Centralizing data under a governed platform empowers more teams to perform self-service analytics. With role-based or attribute-based access controls, employees across the organization can find the data they need, spurring collaboration and faster decision-making.

ROI

By consolidating storage into one platform and taking advantage of object storage for large data volumes, the lakehouse reduces redundancy and operational costs. Its simpler architecture also saves time previously spent maintaining multiple systems. This combination of lower overhead and flexible scalability positions the lakehouse as a future-ready solution for modern data-driven strategies.

Comprehensive Storage Solutions: Objects, Files, and HDFS

Modern data infrastructures require diverse storage approaches—object, file, and HDFS—to meet the wide-ranging demands of AI/ML, analytics, and enterprise applications. Supermicro's hardware portfolio addresses each of these storage types, providing low-latency and high-throughput access to large datasets, with flexible configurations that support different performance and capacity needs.

Storage Types

Object Storage

Ideal for unstructured data (e.g., media files, IoT logs, AI/ML training sets), object storage uses a flat address space and APIs such as S3 to manage data at scale. Each object is accompanied by unique metadata, making it straightforward to handle vast datasets in cloud-native or distributed environments. This approach is particularly effective for content repositories and large AI data lakes due to its simplicity and inherent scalability.



ASG-2115S-NE332R

A 2U Petascale storage server that provides higher storage density with 32 EDSFF E3.S NVMe slots. It features dual PCIe 5.0 x16 network interfaces, perfect for high-speed networking in all-flash data lake environments. Supports up to 192 cores and 2 PCIe 5.0 x16 slots.



AS-1125HS-TNR

A versatile 1U Hyper DP server powered by the latest AMD EPYC 9005 Series processors. It offers up to 12 NVMe, SAS, and SATA3 drives, providing a flexible platform for scale-out storage solutions in data lakes, ensuring reliable performance in both enterprise and cloud environments.



AS-1116CS-TNR

A 1U CloudDC server utilizing a single AMD EPYC 9005 Series processor, supporting up to 12 NVMe, SAS, or SATA3 drives. It is equipped with 2 PCIe 5.0 x16 slots and 2 AIOM slots, making it an ideal choice for software-defined storage in data lakehouse environments, offering scalability and flexibility.

File Storage

For applications needing a hierarchical file system, file storage (using NFS or SMB) remains essential. This structure suits legacy enterprise workloads and high-performance computing (HPC) scenarios where granular data consistency and directory-based organization are key. By mapping a networked file system, users can maintain existing workflows without major changes to application logic.

HDFS Storage (Hadoop Distributed File System)

Though its adoption for new deployments has slowed, many organizations continue to rely on HDFS for batch-oriented big data processing. It was designed for distributed computing across commodity hardware, offering strong fault tolerance and proven scalability. In traditional data lake setups, HDFS remains relevant for workloads that require parallel processing of very large datasets.

Software-Defined Storage: Flexibility and Scalability with Supermicro

Software-Defined Storage (SDS) represents a paradigm shift in data management by decoupling storage software from its underlying hardware, offering enhanced flexibility, scalability, and cost efficiency. This decoupling is a core tenet of SDS, allowing for centralized control and management, and enabling organizations to utilize standard hardware, thus reducing costs and avoiding vendor lock-in.

By leveraging SDS, businesses can quickly adapt to changing storage needs without the constraints of traditional hardware limitations. This agility leads to significant cost savings through optimized resource utilization and simplified management. SDS allows for more frequent hardware refreshes and the integration of new technologies, including the latest CPUs, GPUs, flash drives, and disk drives.

Key Features of SDS:

- **Policy-Based Automation:** SDS platforms offer policy-driven management, enabling administrators to define rules for data placement, replication, and access. This automation simplifies storage provisioning and ensures compliance with organizational policies. Policies can be created to automatically load more data into data lakes, manage data migration from hot to cold storage tiers, and ensure data is stored according to compliance requirements. This simplifies data management, reduces human error risks, and frees IT resources for strategic initiatives.
- **Vendor-Neutral Integration:** The decoupled nature of SDS ensures compatibility across diverse hardware and software ecosystems, providing organizations the freedom to choose components that best fit their needs without being tied to a single vendor. This is important because it allows rapid integration of new hardware components, letting customers take advantage of the newest technologies. This also includes adherence to industry standards and open protocols, such as the Open Compute Project (OCP) specifications, which provide benefits like higher density and improved airflow. By supporting open protocols like the S3 API, SDS solutions can easily integrate with various tools and applications.

HIGH-PERFORMANCE COMPUTING

Supermicro's servers with 5th Gen AMD EPYC processors ensure high-speed data access for faster processing and reduced latency. Specialized systems are designed with high-frequency cores, large L3 caches, and 3D V-Cache technology to enhance per-core performance.



AS-2126FT-HE-LCC

A liquid-cooled 2U FlexTwin™ server with four hot-pluggable nodes, each supporting dual AMD EPYC 9005 Series processors with up to 500W of power. It offers up to 24 DIMMs, with support for 9TB DDR5-6000 and flexible networking, ideal for high-performance computing (HPC) and AI workloads.



SBS-820H-4114S

A SuperBlade server system that offers a single AMD EPYC 7003/7002 Series processor per node, with up to 64 cores, two hot-plug NVMe/SATA3 drive bays, and 8 DIMM slots. This system is designed for HPC environments that require both high computational power and efficient storage.

Supermicro validates all SDS software with specific configurations and media for compatibility.

- **Innovations:** Supermicro is committed to delivering robust SDS solutions by collaborating with various software vendors and incorporating the latest hardware technologies. Supermicro's approach to SDS is rooted in its building block architecture, which allows quick delivery of new server technologies using common components in different form factors or application-optimized servers. This enables customers to rapidly adopt new technologies and integrate new flash and disk drives as they become available.

Such flexibility future-proofs your data lakehouse strategy: the infrastructure can evolve alongside data growth and changing business demands. Whether integrating additional drives for capacity, introducing faster networking (e.g., 400 Gb/s InfiniBand), or deploying updated AMD EPYC processor generations, Supermicro's SDS-ready platforms ensure your environment remains agile and cost-effective.

AI Infrastructure

The data lakehouse can store data in its native format—whether structured, semi-structured, or unstructured—giving AI/ML practitioners the flexibility to work with diverse datasets without imposing a rigid schema. This approach allows for seamless experimentation and faster analytics.

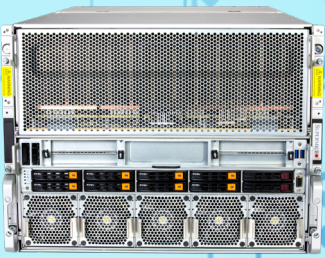
One of the most critical aspects of AI infrastructure is ensuring GPUs have low-latency, high-speed access to data. To meet this need, Supermicro provides high-performance storage options, including Petascale all-flash servers with NVMe drives—delivering the bandwidth required for demanding AI/ML workloads. At the same time, Supermicro offers high-capacity disk-based servers for more cost-effective handling of large datasets via a tiered storage approach (flash for active data, HDD for archival).

Supermicro's latest Petascale solutions employ E3.S (7.5mm) Gen 5 NVMe drives, supporting up to 61TB of throughput in 1U and 2U form factors. Meanwhile, PCIe 5.0 x16 network interfaces enable rapid data transfer between storage and compute nodes, ensuring GPUs stay well-fed and avoiding idle cycles. Finally, having a single platform for different data types eliminates silos. Lakehouse environments that use columnar or otherwise optimized formats can accelerate analytical queries over raw data, boosting efficiency compared to traditional data lakes.

AI Pipelines for LLMs and RAG

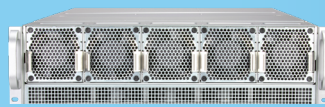
Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) applications present significant computational and storage challenges due to their complexity and scale. Training these models involves processing vast datasets, which requires high-performance computing resources and efficient data management strategies.

In particular, LLMs demand substantial computational power and memory capacity for training, often involving trillions of parameters and necessitating the use of GPU-accelerated systems. RAG methods enhance LLMs by retrieving relevant information from external databases, which necessitates rapid data



AS-8126GS-TNMR

A new 8U GPU server system featuring dual AMD EPYC 9005 Series processors, supporting up to 8 OAM GPU accelerator cards. It is optimized for AI/deep learning training, HPC, and industrial automation, offering significant computational power for demanding workloads.



AS-5126GS-TNRT2

A 5U GPU server with dual AMD EPYC 9005 Series processors and up to 10 PCIe GPUs via a PLX switch. It supports AI, deep learning, 3D rendering, and high-performance computing, offering a flexible architecture tailored to specific workloads requiring substantial GPU computing power.

MODULAR SCALABILITY

Organizations can select from modular designs: 1U, 2U, multi-node, and blade servers, all compatible with 5th Gen AMD EPYC processors and customizable for specific requirements.

access and low-latency storage solutions. The ability to quickly retrieve and integrate information is crucial for generating context-aware responses in real-time. These applications push the limits of existing infrastructure and require solutions that can deliver both high performance and high scalability.

Supermicro's Optimized Workflows

Data Ingestion

- It supports the ingestion of structured, semi-structured, and unstructured data in its native format, accommodating the wide variety of data types used in LLM and RAG applications.
- High-capacity disk-based storage servers are used for cost-effective storage of large datasets, while Petascale all-flash servers utilizing NVMe drives provide low-latency access for active data. This hybrid approach ensures that both capacity and performance needs are met.
- High-speed networking interfaces, such as PCIe 5.0 x16, are used to ensure rapid data transfer between storage and compute nodes, offering 400 Gb/s InfiniBand bandwidth.

Training

- For training purposes, Supermicro offers server systems enabling AI developments and delivery to run small and large AI models.
- Supermicro's GPU solutions, such as the SuperCluster systems, are optimized for LLM training, deep learning, and high-volume inference, ensuring rapid data retrieval and processing.

Inference

- In inference scenarios, low-latency storage is crucial for real-time **retrieval-augmented generation (RAG)** applications, where rapid data access directly impacts the quality and speed of responses.
- Supermicro's **Petascale all-flash storage servers with NVMe drives** are optimized to deliver the low-latency, high-bandwidth access required for inference workloads. These systems are designed to handle the most demanding AI inference tasks, including large language models and real-time decision-making processes.
- These systems support the latest **E3.S (7.5mm) Gen 5 NVMe drives**, ensuring high throughput and low latency in a compact form factor, ideal for applications that need to process large amounts of data in minimal time.
- To minimize bottlenecks, Supermicro systems are built with **high-performance NVMe storage** and **PCIe Gen 5** connectivity, providing fast and efficient data flow, enabling real-time analysis and response generation without delays.
- **Flexible deployment options** are available, tailored to various organizational needs. Supermicro offers configurations such as **single-socket or dual-socket servers, rackmount or blade servers, and all-flash storage solutions**. These options can be customized with **AMD EPYC processors** and **AMD Instinct GPUs** for an optimal balance of computational power and storage throughput. Additionally, **cloud-ready architectures** allow for seamless integration into existing cloud environments, whether public, private, or hybrid, providing scalability and flexibility for growing workloads.



ASG-2115S-NE332R

A 2U Petascale all-flash storage server that offers 32 EDSFF E3.S NVMe slots. Powered by a single 5th Gen AMD EPYC processor, it combines compute and storage capabilities for scalable data lakehouse architectures.



AS-1125HS-TNR

A 1U Hyper server powered by the AMD EPYC 9005 Series processors, featuring up to 12 NVMe, SAS, or SATA drives and 3 PCIe 5.0 x16 slots. Its compact form factor enables dense, scalable deployments in data lakehouse environments.



AS-1116CS-TNR

A 1U CloudDC server, compliant with OCP specifications, featuring a single AMD EPYC 9005 Series processor. It supports up to 12 NVMe, SAS, or SATA drives, with 2 PCIe 5.0 x16 slots and 2 AIOM slots, ideal for software-defined storage and scalable data lakehouse environments.

ENERGY EFFICIENCY

The 5th Gen AMD EPYC processors deliver excellent performance while minimizing energy consumption. Supermicro's commitment to energy-efficient systems is demonstrated through the Supermicro H12 Ultra Servers, powered by AMD EPYC 9005 processors, which achieve significant improvements in energy efficiency for data-intensive applications.

Supermicro supports multiple open-source and commercial frameworks geared toward LLMs and RAG, ensuring broad compatibility. This includes standard APIs (e.g., **S3**) for unified data access. By combining **powerful compute**, **scalable storage**, and **high-speed networking**, Supermicro delivers an **end-to-end solution** for the unique demands of advanced AI pipelines. Whether ingesting multi-petabyte datasets or serving real-time inference requests, organizations benefit from energy-efficient, high-performance infrastructure designed to handle the challenges of next-generation AI applications.

Key Supermicro Systems

It's important to note that while data lakes focus primarily on storage, a lakehouse adds compute and transactional layers. Consequently, Supermicro's storage lineup remains central to data-lake implementations, yet the company also provides compute-oriented products that power the lakehouse's higher-layer operations.

- **GPU Integration**

Supermicro offers a broad spectrum of GPU-accelerated platforms supporting **AMD EPYC** CPUs alongside **AMD Instinct** or **NVIDIA** GPUs. This flexibility addresses varying workload requirements, from deep learning to data analytics, helping organizations scale AI pipelines seamlessly.

- **High-Performance Options**

For workloads requiring **high throughput** and **low latency**, Supermicro's **Petascale** all-flash servers use NVMe drives, making them ideal for active data in real-time analytics, AI/ML training, or large-scale processing scenarios.

- **Advanced Networking**

Many of Supermicro's servers support **high-speed networking** options (e.g., 200GbE, 400 Gb/s InfiniBand), which are crucial for parallel processing and data-intensive tasks in lakehouse environments.

- **Multi-Node Systems**

The **Flex Twin**, **Big Twin**, and **Grand Twin** platforms deliver multi-node compact form factors. These systems excel in **cloud-native** or **virtualized** infrastructures, offering high-density computing, advanced networking, and I/O flexibility for high availability.

- **Tailored Solutions**

From **all-flash arrays** to **high-capacity disk-based** servers, Supermicro maintains a diverse portfolio to match any workload's performance or capacity demands. Customers can deploy these solutions for **object**, **file**, or **HDFS** storage, forming a flexible foundation for data lakehouses.

- **Tiered Storage**

Supermicro supports **tiered storage** that combines fast flash with cost-effective disk systems. This architecture ensures high-performance media for active data, while colder information resides on lower-cost storage. Organizations can also implement hot and cold tiers **within** an object store, distinct from the specialized tiered approach used in AI training pipelines.

- **High-Capacity Options**

To address **large-scale storage** needs, top-loading servers with up to **90 HDDs** or "SimplyDouble" 2U systems effectively double typical capacity. These configurations suit object-storage or archival data, yet still integrate seamlessly into the broader lakehouse.



AS-1024US-TRT

A 1U Ultra server with 4 SATA bays (optional NVMe/SAS) and dual AMD EPYC 7003 Series processors. It offers 32 DIMMs and 4 PCIe 4.0 x16 slots, ideal for virtualization, cloud computing, and high-end enterprise servers that require energy-efficient performance.

SYNERGISTIC ADVANTAGES

The combination of AMD EPYC processors and Supermicro servers allows for optimized performance in AI inferencing, model training, and big data analytics. The EPYC processors provide the computational power, while Supermicro's platforms ensure high-speed access to data. And built-in security of AMD EPYC processors, enhanced by Supermicro's focus on secure solutions, provides robust protection for sensitive data. The synergistic relationship between AMD and Supermicro allows organizations to seamlessly scale their infrastructure to a full range of AI workloads.

- **Active Data Utilization**

Modern data lakes increasingly focus on **real-time analytics** rather than passive archiving. Supermicro's solutions accommodate this shift by delivering the necessary throughput and compute: combining **NVMe** for high-speed reads/writes with **GPU-accelerated** servers. This synergy fuels near-instant analytics and AI processing across a unified platform.

Additional Design Considerations for Data Lakehouses

Data Volume and Types

Lakehouses must handle a range of data—transactional, machine-generated, multimedia—while scaling to future demands. Supermicro designs ensure robust throughput and flexible capacity, enabling the ingestion and analysis of virtually **all** data types in one governed environment.

Three-Tier Architecture

Supermicro's **three-tier** blueprint optimizes storage and compute access by splitting data into distinct layers—providing the **balance** of performance and cost essential to large data lakehouse deployments.

- **All-Flash Tier**

Typically accounting for ~10-20% of data, this tier handles the most **latency-sensitive** and frequently accessed information. **Supermicro Petascale** all-flash servers with AMD EPYC processors are ideal here, offering low-latency, high-bandwidth access to power immediate analytics or AI/ML requests.

- **Object Tier**

This capacity layer (~80-90% of data) stores historical or less-active information. By tapping high-capacity solutions, organizations can achieve cost-effective scale, turning this tier into a persistent data lake of large, often multi-petabyte repositories.

- **Application Tier**

This layer comprises servers running high-level workloads—like BI dashboards, AI training, or HPC computations—and connects directly to the all-flash tier for real-time data access. Supermicro's multi-GPU or HPC-optimized servers (e.g., 8-GPU AMD EPYC systems) excel at data-intensive tasks, forming the compute backbone of the lakehouse.

For more information

- [Data lakehouses with Supermicro](#)
- [Data lakes with Supermicro](#)
- [AI deep learning solutions](#)
- [AI storage solutions](#)
- [Storage solutions](#)
- [Software-defined storage](#)

Supermicro software-defined storage partners

Supermicro collaborates with a variety of software-defined storage vendors. These collaborations ensure that Supermicro’s hardware is optimized for popular SDS platforms, offering customers various options for their specific needs.

Table 1. Partners supporting data lakehouses

- Enterprise DB offers data lakehouse software, services, and support for financial services, government, media & communications, and information technology organizations.
- VAST Data enables data-intensive enterprises to capture, catalog, and refine data via an API-driven pipeline for real-time insights.












	
<p><u>Enterprise DB</u> offers data lakehouse software, services, and support for financial services, government, media & communications, and information technology organizations.</p>	<p><u>VAST Data</u> enables data-intensive enterprises to capture, catalog, and refine data via an API-driven pipeline for real-time insights.</p>

Table 2. Partners supporting data lakes

		
<p>Cloudian HyperStore is ideal for hybrid data lakehouses, with military-grade security and S3 API compatibility for seamless integration and bimodal data access for efficient data management across file-based and object-based data.</p>	<p>VAST Data provides a comprehensive software infrastructure to manage AI data. Their collaboration with Supermicro delivers a unified hyperscale data platform suitable for large-scale AI deployments with storage solutions starting at 1.4PB and scaling to exabyte levels.</p>	<p>IBM Storage Ceph provides block, file, and object capabilities with best-in-class S3 support for large data applications.</p>
		
<p>MinIO AI Store is an ultra-high-performance object store used for AI/ML, analytics, and archival workloads from a single platform.</p>	<p>OSNexus QuantaStor is a scale-out shared-storage solution delivering multi-tenant, S3-compatible object storage with dynamic tiering.</p>	<p>Quantum ActiveScale provides durable object storage seamlessly integrates with Supermicro's scale-out architecture and offers S3-compatible object storage and data lifecycle management and a six-node cluster configuration for cost-effective storage and retrieval.</p>
		
<p>Qumulo is a single platform for managing geographically distributed file and object data across on-premises, edge, core, and cloud.</p>	<p>Scality's object storage solutions offer security, performance, and cost, helping organizations manage data and avoid vendor lock-in.</p>	<p>WEKA delivers a software-defined data platform that helps get data into streaming pipelines for AI and HPC workloads.</p>