WHITE PAPER

Accelerating AI Data Pipelines

Supermicro high-performance scale-out storage for the most demanding AI and ML applications

Organizations everywhere have recognized that their data holds value, and that has resulted in an explosion of applications that turn data into intelligence. Artificial intelligence (AI) and machine learning (ML) techniques are being applied to these datasets to create new services, reduce costs, and improve efficiency. Supermicro offers a tested and validated reference architecture designed to deliver massive amounts of data at high bandwidth and low latency to data-intensive applications, while managing data lifecycle concerns including migration and cold storage retention.

HIGHLIGHTS

The wide range of data-intensive applications have similar storage needs that the Supermicro® Scale-Out Storage Reference Architecture fulfills:

- A ***data lake*** that can contain all current and historical data, large enough so that data does not have to be discarded.
- An ***all-flash storage tier*** to cache input for data-hungry application servers and deliver high bandwidth continuously to meet demand.
- ***Specialized application servers***, ranging from high-core-count servers with AMD EPYC™ processors to GPU-dense systems for AI/ML and HPC workloads.

This white paper describes the specific needs for these three architectural elements and the Supermicro servers and network infrastructure components that can deliver data-intensive applications the balance of performance and economy that organizations require.

Introduction

Organizations in virtually every industry recognize that data holds tremendous value. Today, techniques for extracting actionable information from data continue to mature. Those who learn to harness these techniques gain a competitive edge, whether the task is drug discovery, geophysical modeling, or developing self-driving cars. As the value of transforming data into actionable insights becomes increasingly evident, the drive to collect and process data intensifies. The more granular and comprehensive the data, the more trends can be identified and used to advantage. This abundance of data enhances the training of artificial intelligence (AI) models and improves business efficiency. That's why companies are turning to Supermicro® servers with AMD EPYC™ processors to tap into the transformative potential of various data-intensive applications and workloads:

- ***Big Data Analytics:*** Retail organizations are finding that they can capitalize on trends and customer preferences by gathering and retaining data on every single transaction. Heavy industries can proactively maintain and replace equipment based on performance data. For example jet engines can be maintained more effectively when every real-time aspect of its operations are monitored, from oil consumption to combustion temperatures.
- ***Artificial Intelligence:*** The value of actual on-the-road video is incomparable when developing and training self-driving cars. When gathered and retained, footage from the cars themselves can be used to feed back into models to enable more refined driving decisions. In drug development, AI genomic techniques are enabled by more information about the molecules which can be manipulated.
- ***High-Performance Computing:*** When natural and physical processes are simulated using time series data, the more data the better. More geotechnical data helps locate the best place to identify an oil reservoir. Weather and climate predictions benefit when models analyze many fine-grained time intervals and grid squares representing the earth. Crash simulation is better when more materials data is included.
- ***Electronic Design Automation:*** As semiconductor process node sizes are reduced and electronic circuits become more complex, there is ever more data needed to simulate designs and find potential errors before taping out a new chip.
- ***Production Process Automation:*** Real-time process monitoring and management benefit from shorter data sample time intervals for mixture flows, temperatures, and pressures. Retaining data provides an audit trail and a way to debug problems if they arise in the finished product.

The reference architecture described in this document is deployed in a production environment at one of the world's largest semiconductor manufacturers for AI-based factory-floor automation. In just three weeks, the facility filled a Weka and Quantum® ActiveScale™ software installation with 25 PB of data. The production team uses AI to automatically find defects in wafers. The earlier a defect is found and the wafer is rejected, the less time and money is invested in the production and packaging process.

AMD EPYC PROCESSORS

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 processors power the most energy-efficient servers available. In addition to delivering overall better performance per watt, AMD EPYC processors make it possible to closely match CPU resources with application requirements, creating even greater efficiency.

For example, some analytic applications do not scale well to high core counts. Using high-frequency AMD EPYC processors can increase per-core performance to speed 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 and therefore unimpeded performance.

Whether you need as few as 8 or as many as 128 cores, or specialized processors, AMD EPYC processors offer the freedom to choose. All core features—including memory capacity, I/O bandwidth, and security features—are consistent within each processor family.

Storage Challenges

All data-intensive workloads put high demands on storage systems. Those that retain every piece of data that might ever be used again require virtually limitless storage capacity. Workloads that feed data into analytic models or machine-learning training need storage with the highest performance possible. These characteristics are difficult to reconcile, as high-capacity storage is too slow to sustain the bandwidth that applications need, while high-performance storage is too expensive to use for all of an organization's data. The answer is to create a tiered storage architecture.

For high-capacity storage, organizations need a very large, cost-effective object store that can act as a data lake. The data lake needs to be large enough to retain all of the raw, historical data that could be reused, and cost-effective enough that old data doesn't have to be discarded to make room for new data. It's impossible to know ahead of time which historical data might be needed again, so a data lake must have the capacity for everything. For AI applications, historical training data is needed for new models as they are introduced. For big data applications, a history of scientific measurements or business transactions are what enrich the value of the data analyses conducted. Although the performance demands on the data lake are not high, the software that manages it must be tuned to manage data lifecycles, in particular protecting it from intentional or inadvertent modification.

A very high-performance tier is used to support active data. This tier needs to be able to cache subsets of the data lake and provide the data to the application servers for continuous operation, without stalling. Performance is the most important characteristic for this tier, because the costs of performance mismatches can be significant. If millions are invested in the latest GPUs for accelerated model training or inferencing, then any time these resources are idle is an investment not realized. Stalling and waiting for data can also affect time to value: if AI models are used to inspect manufacturing processes, then any delay in feedback impacts process improvement and lengthens the time to cull-out defective products. The earlier this information is received and decisions are made, the lower the cost to the manufacturing process.

Supermicro offers a broad server product line with AMD EPYC processors. Many of these servers are tuned to deliver high-performance, all-flash storage that is ideal for active data, and high-capacity systems replete with spinning disks for high-capacity, low-cost storage. The more you can deploy servers that are well matched to their workloads, the more highly utilized their resources are, and the best return on investment can be achieved. Supermicro's engineering organization is focused on addressing the challenges of data-intensive computing for many years, and developed a general scale-out storage architecture that serves the range of application needs with cost-effective solutions.

Reference Architecture Overview

Supermicro designed a scale-out storage reference architecture to address the needs described above. The architecture is flexible and scalable, able to adapt to a wide range of storage-intensive application needs and grow as data sets expand. It consists of three tiers:

- ***All-flash tier:*** This tier stores active data (data that needs to be accessed and stored the fastest.) Typically this amounts to 10 to 20 percent of an organization's overall data, and is essentially a cache for the application tier.
- ***Object tier:*** This tier provides long-term, capacity-optimized storage for data that is not needed on an immediate basis, typically 80 to 90 percent of overall data. While the all-flash tier acts as a data cache, the object storage tier acts as a data lake.
- ***Application tier:*** This tier includes the systems that run workloads and consume data. The example deployment described in the next section uses Supermicro 8-GPU servers with AMD EPYC processors to support machine learning training.

The server architecture is designed with modularity in mind, allowing different tiers to be scaled according to specific customer needs. The storage tiers can be adjusted to balance performance and capacity, and various Supermicro servers can be utilized to meet cost requirements. The network architecture is built on 400 Gb/s InfiniBand® fabrics, supporting protocol features such as Remote DMA (RDMA) and NVMe over Fabrics (NVMeOF).

The software architecture is also designed with modularity in mind. This allows for deployment across a variety of products for the object-storage tier, catering to the preferences of different organizations. For instance, educational institutions and governmental labs often lean towards open-source software, while business organizations typically favor commercial software with well-defined support models.

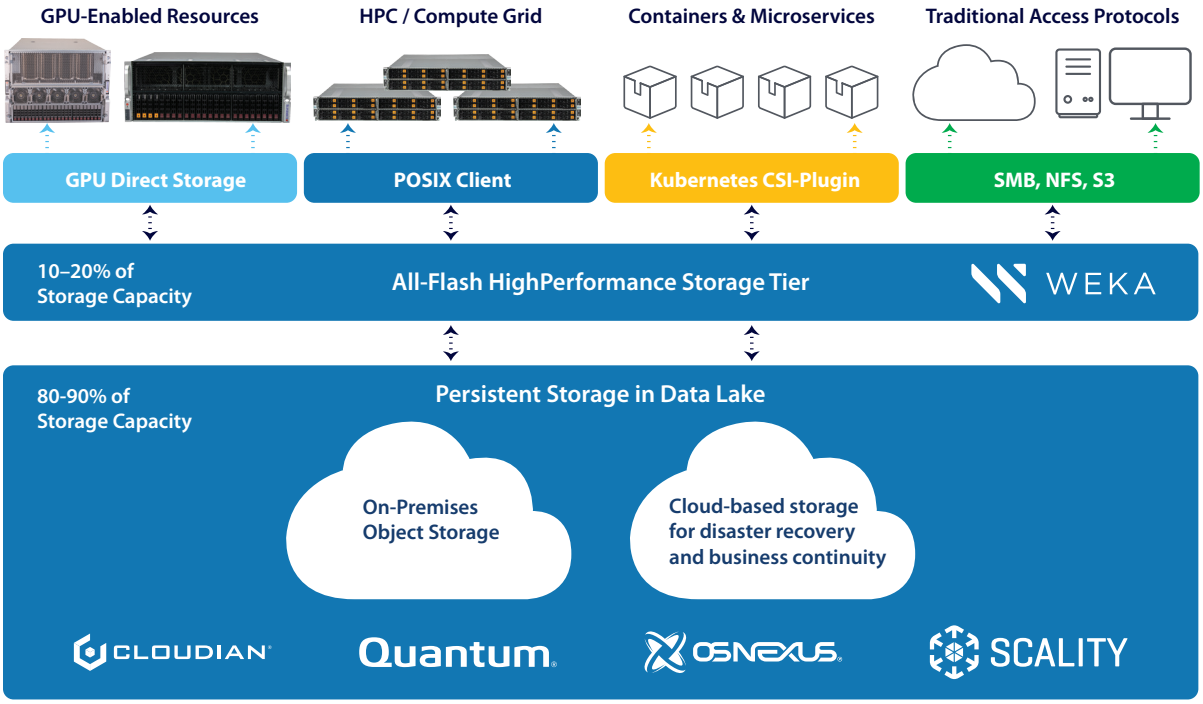


Figure 1. Reference architecture logical structure

OBJECT TIER SOFTWARE

There are many choices for object-storage tier software that interoperates with the Supermicro architecture, including commercial and open-source options.

- ***Quantum ActiveScale*** provides S3-compatible object storage that is durable, secure, and supports archiving and long-term retention of cold data.

- ***OSNexus*** is a software-defined storage solution that supports scale-out object storage solutions and provides a single pane of glass management system across data centers and sites worldwide. It is deployed using a variety of Supermicro hybrid, all-flash and all disk storage servers.

- ***Cloudian*** is an enterprise-class, S3-compatible object storage system that provides a combination of a data lake and data warehouse dubbed a "Data Lakehouse." It provides immutable storage, safe from hacker encryption, and it is built to support data sovereignty regulations.

- ***Scality*** is enterprise-class software-defined storage with distributed file and S3 object storage with the capability to tier into the cloud. It provides hacker-resilient storage to help protect data from ransomware attacks.

All-Flash Tier

The core of the solution, the all-flash tier, holds the data that is actively being used by applications. Typically 10 to 20 percent of overall data, this tier uses all-flash storage servers to store data on a distributed, scale-out file system to a range of clients. It also is capable of coordinating data tiering to the object-storage tier and into the cloud. Using servers optimized for storage—instead of more costly purpose-built appliances—enables linear scaling, where bandwidth rises along with the capacity to store more data as additional servers are added to the cluster. In addition to basic filesystem support, the software supporting this tier must provide resiliency to failures so that a server can fail and not impact service delivery or result in a loss of data.

Storage clusters in this tier need to be interconnected with the highest-bandwidth networking available, either 400 GbE or 400 Gb/s InfiniBand, in order to maintain the lowest latency possible.

Object Tier

The all-flash tier acts as a cache for the most actively used data in the organization. With most data-intensive businesses, however, data can't just be deleted when it is no longer frequently used. It still has value, and it needs to be retained on storage that is optimized for capacity and durability instead of performance, which reduces the overhead for the 80 to 90 percent of data that belongs in this category. The servers to support this mission have a high density of disk drives, and they have comparatively low bandwidth and IOPS requirements for handling data transfers in and out of the object-storage cluster. Similarly, the requirements for networking are relaxed, and 100 GbE is an economical choice that balances cost and performance for object-storage needs.

Ideally, software to manage this tier needs to provide a searchable, protected content repository for unstructured data where it participates in the global name space provided by the all-flash tier. It needs to be focused on protecting data from loss due to hardware failures and also from threats including ransomware or even accidental deletion. It needs to manage storage tiering, so it can migrate data to cloud-based storage when directed to by data lifecycle policies.

The software needs to provide many of the same capabilities as the all-flash tier, but needs to focus on storage lifecycle management. It needs to scale nondisruptively and continue to operate without data loss through the failure of any of the tier's components. But it also must protect and preserve data through erasure coding data protection, versioning, remote replication, object locking, threat mitigation, and implementation of lifecycle data retention and migration policies.

WEKA

WEKA DATA PLATFORM SOFTWARE

The all-flash tier is supported by the Weka Data Platform. This distributed parallel file system connects to the object store. It supports clusters from eight to thousands of nodes, and uses validated server configurations that offer several advantages:

- ***Scalable performance*** that drives low-latency interconnects like 100 GbE and 400 Gb/s InfiniBand.
- ***Protocol support*** for GPUDirect transfers as well as standard file system protocols, including NFS, SMB, and S3.
- ***Global namespace*** that can include stores that are in the Weka cluster and also reside on object or cloud-based storage.
- ***Integrated tiering*** that moves older or infrequently used data to object storage on premises or in the cloud.
- ***Enterprise storage features*** including instant, space-efficient snapshots, end-to-end data encryption, and immutable data for data governance.

Application Tier

The application tier, where data-intensive workloads reside, directly connects to the all-flash tier. This connection is facilitated through a 400 Gb/s InfiniBand link, ensuring high performance and low latency.

- For workloads that are GPU accelerated, the all-flash tier supports NVIDIA® GPUDirect™ protocols to transfer data directly from storage to GPU memory.
- For HPC applications using the high core counts in AMD EPYC processors, standard file protocols are supported.
- For data-intensive applications deployed using cloud-native approaches including containers and microservices, the Kubernetes Container Storage Interface (CSI) is often the preferred access protocol.
- For access from ordinary workstations, file system sharing protocols are used, including Server Message Block (SMB), Network File System (NFS) and Amazon S3.

The application tier uses 400 Gb/s InfiniBand to access storage in the all-flash tier. A separate data center Ethernet network is used for in-band management of the data-processing applications.

Reference Architecture at Work

A wide range of customers use the Supermicro Storage Architecture for applications including machine learning for autonomous driving, biomedical research, and social media databases. Nowhere are the demands more intensive than in machine learning. Efficiently training models on image, video, and language data requires an uninterrupted stream of data to feed arrays of GPU accelerators. Supermicro offers a diverse set of servers that can populate the storage architecture to meet a range of capacity and performance requirements, as well GPU-dense servers to support the most demanding client applications.

Supermicro works with organizations to carefully size and configure the architecture and infrastructure for their use case. The reference architecture described below serves as an example that is tested, validated, and deployed for machine learning training. The solution can scale up and down depending on customer needs, and can be constructed with one or two storage tiers.

- ***The all-flash tier*** is the heart of the storage architecture, with high performance requirements satisfied with our latest storage servers running the Weka Data Platform.
- ***The object tier*** provides optional object storage. Within the reference architecture, the object tier is managed by Quantum ActiveScale Object Storage Software. This scalable, 'always-on', long-term data repository delivers a cost-effective storage solution with extreme data durability, accessibility, and security.
- ***The application tier*** is built to run demanding machine-learning workloads where most of the work is accomplished in GPU accelerators.

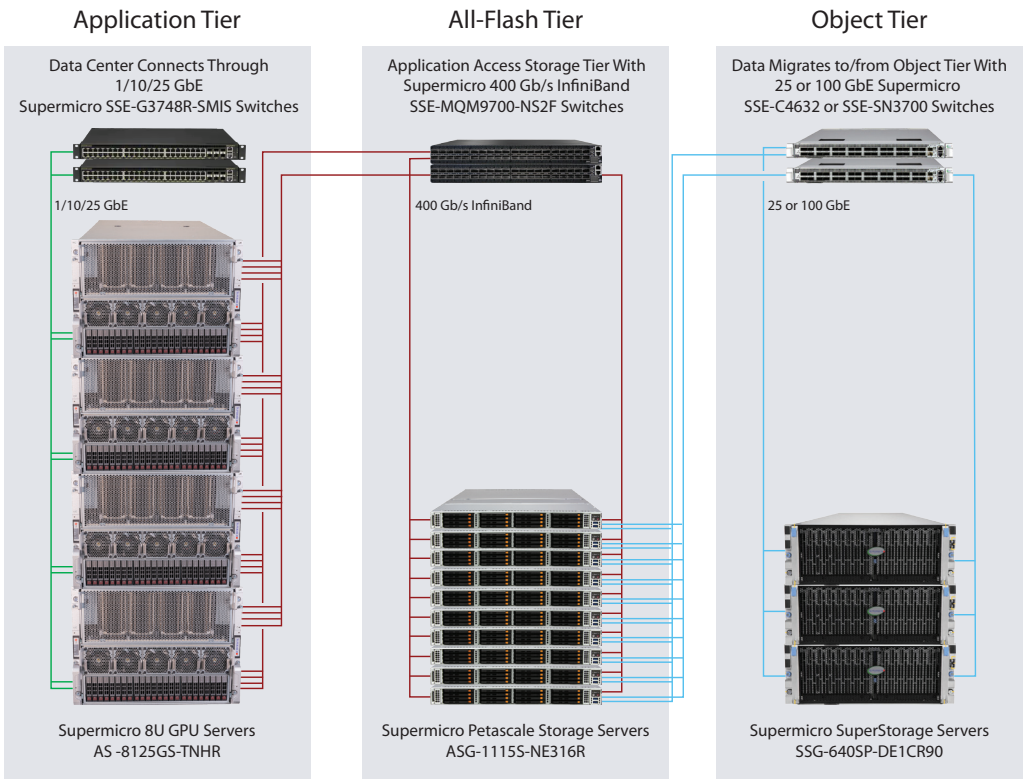


Figure 2. Example implementation of the Supermicro Storage Architecture

ALL-FLASH TIER SERVER OPTIONS

Supermicro ASG-1115S-NE316R

The 1U Petascale storage server used in the storage architecture opens the door to the future with up to 16 EDSFF E3.S NVMe slots for the fastest and most dense storage solutions available. Powered by a single 4th Gen AMD EPYC processor with up to 128 cores, the server supports two PCIe 5.0 x16 network interfaces plus two OCP 3.0 SFF-compliant AIOM slots.

Supermicro ASG-2115S-NE332R

This 2U Petascale storage server doubles the capacity of the 1U offering to 32 EDSFF E3.S NVMe slots for solutions that require a balance of storage and I/O capacity. It supports two PCIe 5.0 NVIDIA ConnectX®-7 cards for a total of 800 Gb/s of InfiniBand bandwidth.

Supermicro AS -1115CS-TNR

This 1U Supermicro CloudDC rack server is powered by a single 4th Gen AMD EPYC processor and it supports up to 10 U.2 form factor 2.5” NVMe drives. It supports two PCIe 4.0 NVIDIA ConnectX-7 cards.

All-Flash Tier

The Weka Data Platform is enterprise-grade software that supports a parallel, scale-out distributed file system. Its most important feature for AI/ML training is its support of the NVIDIA GPUDirect feature that enables data to transfer using remote DMA directly from the storage server’s NVMe drives to GPU memory. The zero-copy resource manager enables transfers with no caching or copying, reducing latency and speeding throughput. The platform supports a range of other clients, including standard POSIX clients, Kubernetes containers via the Container Services Interface (CSI), as well as standard network file protocols including Server Message Block (SMB), Network File System (NFS), and Amazon S3. Weka software features erasure coding for data protection and resiliency, and data is distributed across the cluster to mazimize throughput. The software can support clusters from 8 to 1024 nodes, and with storage capacities into the hundreds of petabytes range.

Supermicro Petascale Storage Systems are ideal servers for the all-flash tier. Powered by a single AMD EPYC 9004 Series processor, Weka’s recommendation for 32 cores per storage server is easily met with a single AMD EPYC CPU. The Supermicro 1U ASG-1115S-NE316R is configured with 16 15.36 TB hot-swap Enterprise Data Center Standard Form Factor (EDSFF) E3.S NVMe drives for up to 245.76 TB of storage per server. The EDSFF form factor is optimized for NVMe media power and cooling requirements, unlike legacy 2.5" SFF M.2 drives. Each drive is directly connected with four lanes of PCIe 5.0 bandwidth. In this example, ten servers provide 2.4 PB of raw storage, for an effective capacity of 1.5 PB when storage system overhead is factored in. The operating system and Weka software are hosted on two internal M.2 NVMe boot drives.

To support the voracious appetite of machine learning models for data, the capacity and performance of the servers’ NVMe storage must be matched with appropriate network connectivity. Each of the 1U servers is equipped with two NVIDIA ConnectX-7 SmartNICs, each with a single InfiniBand NDR 400 Gb/s port. These occupy each server’s PCIe 5.0 x16 expansion slots. Two independent data fabrics are supported by a pair of Mellanox NDR InfiniBand switches (available as part SSE-MQM9700-NS2F), each with 64 nonblocking ports.

Note the I/O balance established so far in this example architecture. The ten storage servers have a total of 20 InfiniBand links capable of 400 Gb/s each. The four GPU servers have a total of 32 InfiniBand links. If the storage servers were to serve 400 Gb/s of storage continuously, each GPU server would have access to 250 Gb/s of storage bandwidth on each of its links. But storage utilization in AI model training is more bursty than continuous, so this architecture can support multiple bursts of 400 Gb/s on the InfiniBand links when data is needed.

Object Tier

Not all data needs to be in the fastest, all-flash tier. Indeed, when a model is trained with a large amount of image or video data, or a large amount of big data is analyzed, it can be retained on a cost-effective storage capacity tier until it is needed again. The Weka Data Platform supports migrating data to cloud storage or to a second storage tier, and it interoperates with a number of scale-out distributed object-based storage platforms that can be used to support this

Quantum

QUANTUM ACTIVESCALE

The object tier is supported by Quantum ActiveScale Object Storage software. This durable object storage repository connects to the all-flash tier and supports active archiving and long-term retention of cold data. The software provides:

- ***Unlimited scale*** with support for billions of objects and exabytes of capacity. Objects are placed across resources for performance at scale.
- ***Automatic scaling*** of capacity and performance, with newly added nodes immediately joining the cluster and servicing requests and tasks.
- ***High performance*** through highly parallelized software that load balances data and operations across cluster nodes.
- ***Always available data access*** with rolling system upgrades and the capability to tolerate component and site failures.
- ***Data durability and security*** with support for advanced erasure coding, end-to-end encryption, object locking, versioning, monitoring, and repair of both active and cold data.

tier. For very large training data sets used in LLMs that cannot fit into the all-flash performance tier, the training set is stored on the capacity tier and migrated to the performance tier as needed by the Weka file system.

The reference architecture tested for AI/ML workloads uses Quantum ActiveScale Object Storage, a platform that provides durable, secure, S3-compatible storage. It can scale nondisruptively, distributing objects across nodes for optimal performance without the need to manually rebalance. Data is always available due to rolling upgrades that only take one node out of use at a time, and the stateless S3 protocol used to access storage. ActiveScale delivers data durability and security with advanced erasure coding, versioning, end-to-end encryption, object locking, and ongoing monitoring and repair.

The Quantum ActiveScale software is supported on a a six-node cluster of three dual-node Supermicro SuperStorage servers, each with two internal CPU nodes. Three Supermicro SSG-640SP-DE1CR90 servers comprise this tier, each configured with 90 22 TB 3.5" SAS hard disk drives, with the operating system and ActiveScale software hosted on the two 2.5" or U.2 SSDs per node. The total raw storage in the cluster is 5.94 PB, with 4.59 PB usable for object storage after overhead.

The throughput requirements for the object tier are much lower than with the all-flash tier. Therefore, the three Supermicro SuperStorage servers are connected to a pair of redundant Supermicro SSE-C4632 or SSE-SN3700 25/100 GbE switches. This pair of 100 Gb/s networks connects to the all-flash tier for fast data migration between tiers.

Application Tier

In this example, the application tier is populated with four 8U GPU systems. The AS -8125GS-TNHR is a next-generation machine-learning platform that is designed to propel ML workloads with 4th Gen AMD EPYC processors and industry-leading NVIDIA HGX H100 8-GPU modules. It supports up to 3.2 terabits per second of I/O, hydrating the GPU accelerators with massive amounts of data. This 2-socket server is configured with 6 TB of memory, Titanium Level (96% efficiency) power supplies, and 10 counter-rotating fans with dual-zone cooling to keep the high-power GPU accelerators within thermal operating ranges.

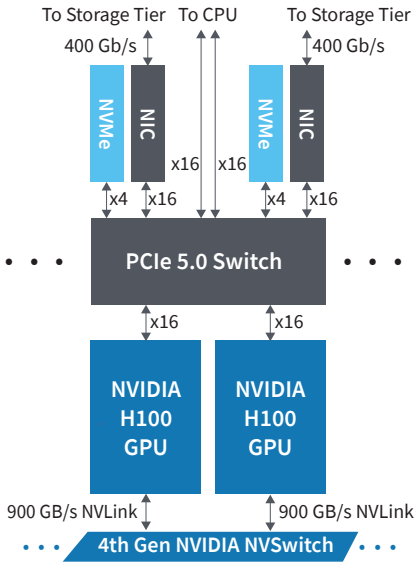
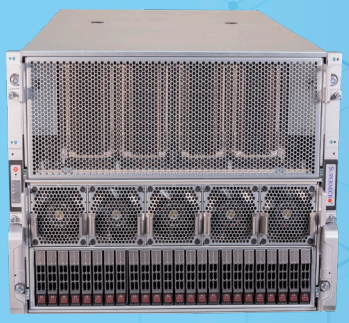


Figure 3. Each pair of GPUs has a dedicated switch with two NICs and two NVMe drive slots attached to them.

GPU SERVER OPTIONS



Supermicro AS -8125GS-TNHR

With eight NVIDIA HGX H100 8-GPUs , the processing capacity of this server is unparalleled, delivering the most GPU acceleration in the Supermicro product line.

Supermicro AS -4125GS-TNRT, TNRT1, and TNRT2

In a 4U form factor, these 4th Gen AMD EPYC processor-powered servers support up to 10 PCIe-form-factor GPU accelerators, giving customers a broad choice of vendors and GPUs. With three models that vary in PCIe architecture, you can choose the server that best addresses the balance of CPU and GPU computing power your workloads require. These servers each offer a single PCIe x8 slot for network connectivity plus a x16 OCP 3.0 AIOM slot.

Whether accessing data from main memory or from the storage tier, the server is designed to move data with low-latency, nonblocking PCIe 5.0 connectivity. This is Supermicro's most specialized GPU server, with data optimized through a set of PCIe switches that can directly connect GPU, I/O, and storage devices with no CPU intervention.

Each of the four pairs of GPUs is connected to a PCIe 5.0 switch that provides x16 connectivity to the GPU, to the NIC, and to the CPU. It also connects two of the server's NVMe drive slots with four lanes of bandwidth. This provides for unimpeded, nonblocking I/O between the storage tier and the GPU.

Once a remote DMA operation is set up, high-speed transfers can move data directly to GPU memory with no intervening CPU cycles or buffering. The same can be said for data that may be cached on one of the pair of NVMe drives on the same PCIe switch. Input or output can be stored close to the GPU.

After data is loaded into an accelerator's memory, it can share with its peers through the NVIDIA NVSwitch™ that interconnects all eight GPUs on the 8-GPU module. This helps to speed complex models that pipeline data through multiple GPUs. With 900 GB/s of NVIDIA NVLink bandwidth between any two GPUs, data moves without bottlenecks.

The application tier is connected to the storage tier through eight PCIe NVIDIA ConnectX-7 interfaces over 400 Gb/s InfiniBand.

Management

The Supermicro storage architecture is designed to work with data center management approaches, with open management APIs and integrated tools supporting deployments. In addition to dedicated IPMI ports and Web IPMI interfaces on all servers, Supermicro's 8U GPU and all-flash storage servers can be managed through Supermicro® SuperCloud Composer. This software helps IT staff configure, maintain, and monitor systems using single-pane-of-glass management interface. If operational teams prefer to use existing management tools, industry-standard Redfish® APIs provide access to higher-level tools and scripting languages.

Delivering Business Advantage

Choosing to support data-intensive applications with the scale-out storage architecture from Supermicro is a strategic business decision that can reap benefits for years to come.

Maintain Data On Premises

The first decision is to maintain data on premises with absolute control over where data is stored. Next is identifying an infrastructure design with sufficient bandwidth and low latency to meet application requirements. Business continuity is also a factor. Every organization is susceptible to downtime and data loss from natural disasters, human error, and cyberattacks. Including data replication practices in disaster recovery plans can help reduce downtime and maintain businesses continuity. There are several data replication options, such

as onsite replication and replication to the cloud or remote sites. It is often the most cost-efficient approach, since the price of cloud-based storage and the bandwidth to access it can be prohibitive.

Choose the Supermicro Scale-Out Storage Architecture

Next comes the decision to choose the Supermicro scale-out reference architecture. Supermicro is known for offering a diverse set of servers optimized for every workload. So whether the application tier needs the highest density of GPU accelerators to speed machine learning or the highest density of AMD EPYC processor cores, Supermicro offers servers that are optimized to meet specific needs. Choosing Supermicro delivers flexibility. As needs change, organizations can use the same storage architecture to power different data-driven applications, knowing that a range of choices is always available for applications.

Configure the Storage Tiers for Workloads

The scale-out storage reference architecture is designed with high bandwidth and low latency in mind, and tested and validated for optimum performance. Supermicro tuned the Petascale server configurations to deliver optimal data flow over NVIDIA ConnectX-7 SmartNICs and equipped the GPU servers in the application tier with matching interfaces for unimpeded data flow. With Supermicro's unique building block approach, capacity, performance, and cost can be balanced to deliver a solution that meets application needs within budget constraints. For example, organizations can tap into high-performing Supermicro 1U and 2U Petascale servers in the all-flash tier or choose traditional U.2 NVMe storage for an economical solution. Similary, Supermicro SuperStorage servers provide four different object-storage platforms, offering maximum capacity and performance in minimal space for cost-effective long-term data storage protection.

Make Applications Perform

This reference architecture, combined with Supermicro servers, is designed to maximize application performance. Because each organization has a unique set of requirements, Supermicro's engineering team is prepared to help size, design, and implement an optimized version of the scale-out storage architecture that meets performance and capacity demands and helps ensure businesses stay on the cutting edge.

Get Started

The Supermicro Scale-Out Storage Architecture provides the foundation for transforming data into actionable insights. When deployed with Supermicro servers with AMD EPYC processors, this architecture can help organizations collect and process data for better business outcomes. To learn more, visit supermicro.com and the links below or contact us at www.supermicro.com/en/contact.

- Supermicro rack servers, GPU servers, and storage servers
- Weka Data Platform software
- Quantum ActiveScale Object Storage software
- NVIDIA ConnectX-7 adapters

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 enables 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).