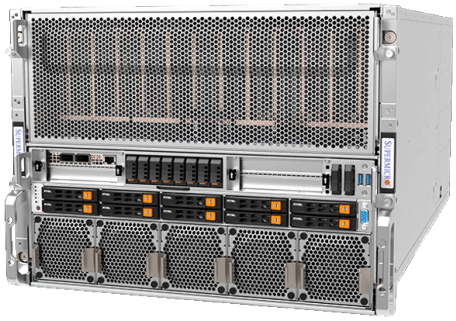
H14 8-GPU System

Next-Generation Large-Scale AI Training Platform



AS -8126GS-TNMR2

Streamline Deployment at Scale for the Largest AI and Large-Language Models

Proven 8U high-performance fabric 8-GPU system design with AMD Instinct™ MI325X accelerators:

- Industry-standard OCP accelerator module (OAM) with 8 GPUs interconnected on an AMD universal base board (UBB 2.0)
- Industry-leading 2 TB of HBM3E memory in a single server node
- 400-Gbps networking dedicated to each GPU for large-scale AI clusters
- 2-socket design supports 4th and 5th Gen AMD EPYC™ Processors
- Up to 24 DIMMs for up to 9 TB of DDR5-6000 memory (with 5th Gen AMD EPYC processors)
- Flexible PCIe 5.0 options for I/O and networking
- Titanium-Level efficiency power supplies

When artificial intelligence (AI) workloads can tap into massive computational power, scientists and researchers can solve the unsolvable. Supermicro unleashes the power of large-scale infrastructure with a server built with our proven AI building-block system and powered by 5th Gen AMD EPYC™ processors and AMD Instinct™ MI325X GPU accelerators.

AMD Instinct-Accelerated Server

The 8U server hosts the AMD Instinct MI325X Platform, an industry-standard-based universal baseboard (UBB 2.0) model with 8 AMD Instinct MI325X accelerators and a total of 2 TB of HBM3 memory to help process the most demanding AI models. The Instinct MI325X boasts an improved HBM memory system with a 33% increase in HBM3 memory compared to the prior-generation MI300X, and 6 TB/s memory bandwidth. It is designed to hold a one-trillion parameter model in its memory. Along with a 25% improvement in FP8 throughput, it brings delivered teraFLOPS to 1565, speeding AI inference and model fine tuning. Native sparsity support helps save power, use fewer compute cycles, and reduce memory use. Each accelerator on the UBB platform connects to the other seven with 128 GB/s AMD Infinity Fabric™ Link technology for an aggregate 896 GB/s capacity. Each accelerator can connect to the host through 16 lanes of PCIe 5.0 bandwidth, and the AS -8126GS-TNMR2 is optimized for I/O throughput. Together, these features give the capacity to propel the most challenging AI workloads and large-language models.

Balanced System Design

You can achieve faster time to results when accelerators can consume the data they need—when they need it. AMD EPYC 9005 Series processors provide up to 192 cores per CPU and up to 9 TB of memory for the parallelism you need to manage data before and/or after processing by the GPU. For tasks requiring fast per-core speed with less parallelism, the 64-core, frequency-optimized, EPYC 9575F is AI-optimized to deliver exceptional performance per core and per thread.

This AI building-block server is designed to provide each accelerator with x16 connectivity to a dedicated 400-Gbps networking device and to the host CPU—so whether data is arriving from main memory or a networked-based data lake, it can transfer directly to accelerator memory. When buffering is needed, each GPU is switched to two x8 hot-swap NVMe drive slots for a total of 16 drives dedicated to GPUs per server.

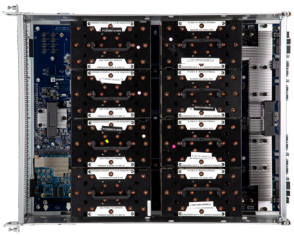
The AMD EPYC CPU's system-on-chip (SoC) design supports built-in functions including IPMI-based management, on-board M.2 drive, and built-in SATA controllers for two drives. The SoC-oriented design reduces the number of external chip sets, helping to reduce complexity and power consumption. Titanium-Level power supplies keep the GPUs accelerating your workloads while dual-zone cooling with 10 counter-rotating fans keep the accelerators within their thermal envelopes.

Fast Time to Value with the AMD ROCm Platform

Whatever the source of your AI platforms, AMD ROCm™ software opens doors to new levels of freedom. With support for open frameworks like PyTorch and TensorFlow, ROCm simplifies AI model migration and deployment, optimizing hardware efficiency with minimal code changes. Through strategic partnerships with AI leaders such as OpenAI, PyTorch, Hugging Face, and Databricks, the ROCm ecosystem delivers high-performance, out-of-the-box AI solutions, empowering enterprises to meet their goals with seamless integration and robust partner support.

Open Management

Regardless of your data center's management approach, our open management APIs and tools are ready to support you. In addition to a dedicated IPMI port, and a Web IPMI interface, Supermicro® SuperCloud Composer software helps you configure, maintain, and monitor all of your systems using single-pane-of-glass management. If your DevOps teams prefer to use their own tools, industry-standard Redfish® APIs provide access to higher-level tools and scripting languages.



H14 Generation	AS -8126GS-TNMR2
Form Factor	• 8U rackmount
Processor Support	• Dual SP5 sockets for AMD EPYC™ 9004 Series processors up to 400W or AMD EPYC 9005 Series processors up to 500W (two CPUs required)[1] • Up to 128 cores (EPYC 9004 Series) or 192 cores (EPYC 9005 Series) per CPU
Memory Slots & Capacity	• 12-channel DDR5 memory support • 24 DIMM slots for up to 6 TB ECC DDR5-4800 RDIMM (with EPYC 9004 Series) • 24 DIMM slots for up to 9 TB ECC DDR5-6000 RDIMM (with EPYC 9004 Series)
On-Board Devices	• System on Chip • Hardware Root of Trust • IMPI 2.0 with virtual-media-over-LAN and KVM-over-LAN support • ASPEED AST2600 BMC graphics
GPU Support	• AMD Instinct MI325X Platform with 8 MI325X OAM GPUs
Expansion Slots	• 8 PCIe 5.0 x16 low-profile slots connected to GPU via PCIe switch • 2 PCIe 5.0 x16 full-height full-length slots • Optional 2 PCIe 5.0 x16 slots via expansion kit
Storage	• 12 PCIe 5.0 x4 NVMe U.2 drives • 4 PCIe 5.0 x4 NVMe U.2 drives (optional)[2] • 1 M.2 NVMe/SATxA boot drive • 2 hot-swap 2.5" SATA drives[2]
I/O Ports	• 1 RJ45 Dedicated IPMI LAN port • 2 USB 3.0 Ports (rear) • 1 VGA Connector
BIOS	• AMI Code Base 256 Mb (32 MB) SPI EEPROM
System Management	• Built-in server management tool (IPMI 2.0, KVM/media over LAN) with dedicated LAN port • Redfish APIs • Supermicro SuperCloud Composer • Supermicro Server Manager (SSM) and Supermicro Update Manager (SUM)
System Cooling	• Dual-zone cooling optimized for performance and operational costs with 5 front and 5 rear counter-rotating fans with optimal speed control
Power Supplies	• 6x or 8x 3000W N+N redundant Titanium-Level power supplies

1. Certain CPUs with high TDP (320W and higher) air-cooled support is limited to specific conditions. Please contact Supermicro Technical Support for additional information about specialized system optimization
2. Optional parts are required for NVMe/SAS/SATA configurations