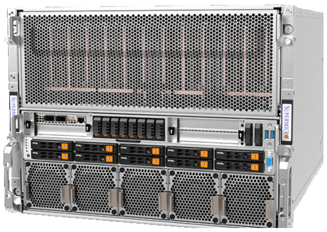
DATASHEET

H14 8-GPU Systems

Next-Generation Large-Scale AI Training Platforms



AS -8126GS-TNMR

AS -4126GS-NMR-LCC

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

Proven high-performance, 8-GPU system design with AMD Instinct™ MI350X and MI355X accelerators:

- Industry-standard OCP accelerator module (OAM) with 8 GPUs interconnected on an AMD universal base board (UBB 2.0)
- Industry-leading 2.3 TB of HBM3E GPU memory in a single server node
- 400-Gbps networking dedicated to each GPU for large-scale AI clusters
- 2-socket design supports 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 systems built with our proven AI building-block system and powered by 5th Gen AMD EPYC™ processors and AMD Instinct™ MI350X and MI355X GPU accelerators.

AMD Instinct-Accelerated Servers

The 8U system hosts the AMD Instinct MI350X Platform with GPUs built on the 4th Gen AMD CDNA™ architecture to deliver exceptional AI inference, training, and HPC workload performance with massive 288 GB HBM3E memory and 8TB/s bandwidth. The platform offers a seamless upgrade from the AMD Instinct MI325X Platform with integration eased by the AMD ROCm™ platform that delivers optimized performance for existing code on day zero.

The 4U system hosts the liquid-cooled AMD Instinct MI355X platform for even greater density and AI performance. With the GPU clocks able to run faster due to the system's liquid cooling, higher performance is virtually guaranteed.

New to the AMD Instinct MI300 Series is support for FP6 and FP4 datatypes to speed AI inference optimized for use with pruned models quantized to low-precision variables, and enhanced FP16 and FP8 processing to deliver exceptional performance for advanced Generative AI models—pushing the boundaries of AI acceleration.

Built with eight MI350X accelerators, the industry-standard-based universal baseboard (UBB 2.0) hosts an aggregate 2.3 TB of HBM3 memory supported by 8 TB/s memory bandwidth to help contain and quickly process the most demanding AI models. Each accelerator on the platform connects to the other seven with 160 GB/s AMD Infinity Fabric™ Link technology for an aggregate 1120 GB/s bandwidth. Each accelerator can connect to the host system through 16 lanes of PCIe 5.0 bandwidth (128 GB/s).

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.

Both systems are 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 x8 hot-swap NVMe drives dedicated to each GPU in the server for a total of one or two drives per GPU, depending on the server.

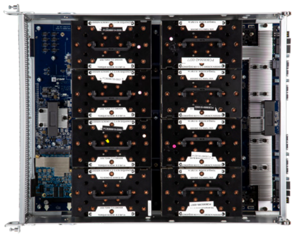
The AMD EPYC CPU's system-on-chip (SoC) design supports built-in functions including IPMI-based management, on-board M.2 drive, and built-in SATA controllers for two drives. The SoC-oriented design reduces the number of external chip sets, helping to reduce complexity and power consumption. Titanium-Level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

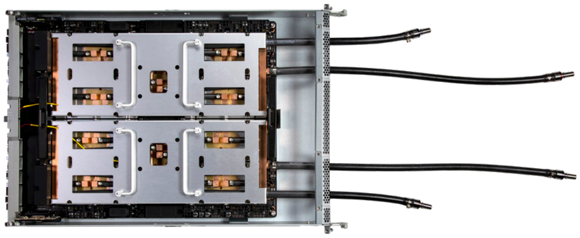
Open and Optimized AI Software Stacks

Built on AMD's commitment to open-source innovation, AMD Instinct MI350 Series GPUs are seamlessly integrated with the next-generation AMD ROCm™ software stack—the industry's premier open alternative for AI and HPC. The ROCm platform supports all major AI and HPC frameworks, inference engines, and model-serving systems including PyTorch, TensorFlow, JAX, ONNX Runtime, Kokkos, Raja, SGLang, Triton, vLLM, and more—enabling effortless model deployment with minimal code changes and maximum flexibility.

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-TNMR	AS -4126GS-NMR-LCC
Form Factor	• 8U rackmount	• 4U rackmount
Processor Support	• Dual SP5 sockets for AMD EPYC™ 9005 Series processors up to 500W and up to 192 cores per CPU (two CPUs required)[1]	• Dual SP5 sockets for AMD EPYC™ 9005 Series processors up to 500W and up to 192 cores per CPU (two CPUs required)[1]
Memory Slots & Capacity	• 12-channel DDR5 memory support • 24 DIMM slots for up to 9 TB ECC DDR5-6000 RDIMM	• 12-channel DDR5 memory support • 24 DIMM slots for up to 6 TB ECC DDR5-6400 RDIMM
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	• 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 MI350X Platform with 8 MI350X OAM GPUs	• AMD Instinct MI355X Platform with 8 MI355X GPUs with liquid cooling
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	• 8 PCIe 5.0 x16 low-profile slots • 2 PCIe 5.0 x16 full-height full-length slots
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]	• 8 front-panel hot-swap PCIe 5.0 x4 NVMe U.2 drives • 2 M.2 NVMe/SATxA boot drives
I/O Ports	• 1 RJ45 Dedicated IPMI LAN port • 2 USB 3.0 Ports (rear) • 1 VGA Connector • 1 TPM Header	• 1 RJ45 Dedicated IPMI LAN port • 2 USB 3.0 Type -A ports (rear) • 1 VGA Connector • 1 TPM header
BIOS	• AMI 64 MB SPI flash EEPROM	• AMI 64 MB SPI flash EEPROM
System Management	• Built-in server management tool (IPMI 2.0, KVM/media over LAN) with dedicated LAN port; SuperCloud Composer®; Supermicro Server Manager (SSM); Supermicro Update Manager (SUM); Super Diagnostics Offline (SDO); Supermicro Thin-Agent Service (TAS); SuperServer Automation Assistant (SAA); and Redfish APIs	
System Cooling	• Dual-zone cooling optimized for performance and operational costs with 5 front and 5 rear counter-rotating fans with optimal speed control	• 5x 6 cm heavy-duty fans with optimal fan speed control • Direct-to-chip (D2C) cold plate
Power Supplies	• 6x 5250W redundant (3+3) Titanium-Level (96%) power supplies	• 4x 6600W redundant (3+3) Titanium-Level (96%) 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