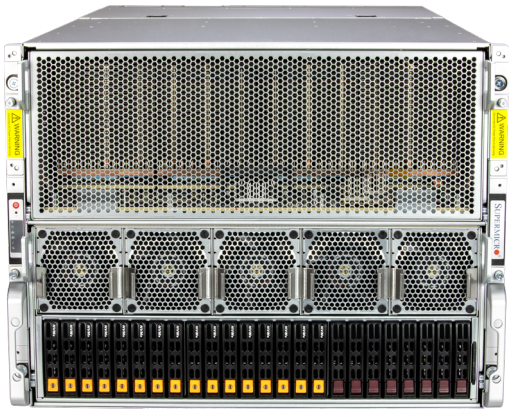
H13 8U 8-GPU System

Powered by AMD Instinct™ MI300X Accelerators



AS -8125GS-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™ MI300X accelerators

- Industry standard OCP Accelerator Module (OAM) with eight accelerators interconnected on an AMD Universal Base Board (UBB 2.0)
- Industry-leading 1.5TB HBM3 GPU memory in a single server node, over 6TB HBM3 GPU memory in an air-cooled rack
- 1:1 400G networking dedicated for each GPU designed for large scale AI and supercomputing clusters
- 2-socket design supporting 4th Gen AMD EPYC™ Processors
- Up to 24 DIMMs for up to 6 TB of DDR5-4800 memory
- Flexible PCIe 5.0 options for I/O and networking

Built on Supermicro's proven AI building-block system architecture, the new 8U 8-GPU system with AMD Instinct MI300X accelerators streamline deployment at scale for the largest AI models and reduce lead time. The 8U 8-GPU air-cooled solution is feature maximized and power optimized supporting dedicated I/O and dedicated storage per GPU, full performance GPUs, CPUs, and memory, and high-speed networking for large scale cluster deployments. These powerful GPUs enhance operations-per-second and performance-per watt at rack scale.

GPU-Optimized Server

Powered by dual AMD EPYC 9004 Series processors, up to 6 TB of DDR5-4800 memory, and 8 AMD Instinct MI300X accelerators, the server streamlines deployment at scale for the largest AI models. Within the server, 128 PCIe 5.0 lanes and up to 16 hot-swap NVMe drives provide robust I/O throughput to help accelerate data-intensive workloads. Testimony to its power, the AMD Instinct MI300X accelerator delivers a 2.6x single- and half-precision (FP16/ FP32) performance improvement over previous-generation AMD Instinct™ MI250X GPUs.

The balanced system design associates a GPU with a 1:1 networking to provide a large pool of high bandwidth memory across nodes and racks to fit today's largest language models with up to trillions of parameters, maximizing parallel computing and minimizing the training time, and inference latency. The 8U system with the MI300X OAM accelerator offers raw acceleration power of 8-GPU with AMD Infinity Fabric™ Links enabling up to 896GB/s of peak theoretical P2P I/O bandwidth on the open standard platform with industry-leading 1.5TB HBM3 GPU memory in a single server node, as well as native sparse matrix support, designed to save power, lower compute cycles and reduce memory use for AI workloads. Each server features dual socket AMD EPYC™ 9004 series processors with up to 256 cores in total. At rack scale, over 1000 CPU cores, 24TB of DDR5 memory, 6.144TB of HBM3 memory, and 9728 Compute Units are available for the most challenging AI environments.

Fast Time to Value with the AMD ROCm Platform

Whether you are deploying AI or HPC applications, AMD ROCm™ software opens doors to new levels of freedom. With mature drivers, compilers, and optimized libraries supporting AMD Instinct accelerators, ROCm is open and ready to deploy. Proven in some of the world's largest supercomputers, ROCm software provides support for leading programming languages and frameworks for AI, including PyTorch, TensorFlow, ONYX-RT, Triton, and JAX.

Open Management

Our approach to management enables you to deliver the scale your organization requires. Supermicro® SuperCloud Composer with open-source Redfish® compliant 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, our accessible Redfish-compliant APIs provide access to higher-level tools and scripting languages. More traditional management approaches, including IPMI 2.0, are available as well. Regardless of your data center needs, our open management APIs and tools are ready to support you.

H13 Generation	AS -8125GS-TNMR2
Form Factor	• 8U rackmount
Processor Support	• Dual SP5 sockets for AMD EPYC™ 9004 Series processors (two CPUs required)[1] • Support for CPUs with AMD 3D V-Cache™ technology • Up to 128 cores and up to 400W cTDP[1] per processor (up to 256 cores per server)
Memory Slots & Capacity	• 12-channel DDR5 memory support • 24 DIMM slots for up to 6 TB ECC DDR5-4800 RDIMM
On-Board Devices	• System on Chip • Hardware Root of Trust • IPMI 2.0 with virtual-media-over-LAN and KVM-over-LAN support • ASPEED AST2600 BMC graphics
GPU Support	• AMD Instinct MI300X Platform with 8 MI300x 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 • Optional 4 PCIe 5.0 x4 NVMe U.2 drives[2] • 2x M.2 NVMe 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