

H13 4U GPU Systems

Flexible, High-Density GPU Server Family for AI, ML, and HPC



A+ Server 4125GS-TNRT



A+ Server 4125GS-TNRT1 (Coming Soon)



A+ Server 4125GS-TNRT2 (Coming Soon)

Three GPU-Dense Servers Optimized for Compute-Intensive Workloads

Support up to 10 GPU accelerators of your choice plus bandwidth to spare for networking and disk storage:

- PCIe 5.0 connectivity throughout including direct-connected, single-root, and dual-root options, plus CXL 1.1+ support
- Supports the latest GPUs, including AMD Instinct™ MI210 and NVIDIA® H100
- 2-socket design supporting AMD EPYC™ 9004 Series Processors
- Up to 6TB DDR5-4800MHz in 24 DIMMs
- Flexible PCIe slot options for I/O and networking
- Titanium-Level efficiency power supplies

If your goal is to power artificial intelligence (AI), machine learning (ML) or high performance computing (HPC) workloads, look no further than our 4U GPU systems with flexibility for up to 10 PCIe-form-factor accelerators.

Dual-Socket GPU-Optimized Server

To support AI, ML, and HPC workloads, we designed a family of servers with 16 lanes of PCIe 5.0 connectivity to up to ten GPUs, with support for the fastest accelerators from AMD and NVIDIA. To accelerate GPU-to-GPU connectivity, the family of servers support both AMD Infinity Fabric™ Link and NVIDIA® NVLink Bridge™ technologies. We have optimized these systems with a range of options that enable you to choose the right balance of computation, acceleration, I/O, and local storage to best suit your workload needs:

- **AS -4125GS-TNRT Server:** features a dual-root PCIe architecture that directly connects each of 8 GPUs to CPUs with 16 lanes of connectivity so that nothing stands in the way of the flow of data to the accelerators. This server is ideal for AI and machine-learning workloads that are very I/O intensive and that need a balance of CPU and GPU performance. Direct connectivity is also provided to two 16-lane PCIe 5.0 slots and the server includes support for up to 4 NVMe and 2 SATA drives.
- **AS -4125GS-TNRT1 Server:** uses a single-root architecture that connects 10 GPU accelerators to a

single CPU through a PLX PCIe 5.0 switch. This server is tailored for deep learning applications where most of the computation takes place in the GPU. It supports up to 8 NVMe drives and 2 SATA drives.

- **AS -4125GS-TNRT2 Server:** is based on a dual-root configuration that connects up to five GPU accelerators to each CPU through a PLX switch. This server provides a balance between CPU and GPU capacity and is ideal for HPC applications (such as molecular dynamics simulation) that demand intensive computation from both components. This server boasts the capability to 8 NVMe drives and 2 SATA drives: up to 2 SATA and up to 24 NVMe.

Designed with PCIe 5.0 Connectivity Throughout

The H13 family of 4U GPU servers is designed with PCIe 5.0 connectivity throughout, helping to speed the flow of data within the server and also to provide high network and cluster interconnect connectivity for scale-out applications. PCIe and AIOM / OCP 3.0 interfaces can support 100-Gbps InfiniBand and 100 Gigabit Ethernet connectivity today, with the bandwidth to support 400 Gbps interfaces as they become available.

Each server in the family supports up to 6TB of main memory. They are powered by redundant 2000W Titanium-Level power supplies and cooled by eight 11.5k RPM heavy-duty fans. This

power and cooling infrastructure supports the fastest CPUs and GPU accelerators with either air or active or passive liquid cooling.

Made Possible by AMD EPYC™ 9004 Series Processors

Our H13 servers are made possible by AMD EPYC™ 9004 Series Processors, with up to 96 cores per CPU and 192 cores per server. You can choose the number of cores, cache size, and clock frequency appropriate for your application and the rest of the features are included at no cost.



The AMD EPYC™ 9004 series Processors supports massive I/O capacity, with up to 160 lanes of PCIe 5.0 connectivity in our two-socket systems. The system-on-chip (SoC) design supports built-in functions such as Gigabit Ethernet ports, USB and KVM functions, and even support for M.2 drives that can be used for system boot. The SoC-oriented design reduces the number of external chip sets, helping to reduce complexity and power consumption.

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.



H13 Generation	AS -4125GS-TNRT Server	AS -4125GS-TNRT1 Server (Coming Soon)	AS -4125GS-TNRT2 Server (Coming Soon)
Form Factor	<ul style="list-style-type: none"> • 4U rackmount 		
Processor Support	<ul style="list-style-type: none"> • Dual SP5 socket for AMD EPYC™ 9004 Series processors (two CPUs required) • Up to 96 cores and cTDP up to 400W¹ per processor (up to 192 cores per server) 		
Memory Slots & Capacity	<ul style="list-style-type: none"> • Up to 6TB 3DS ECC DDR5-4800MHz RDIMM in 24 DIMMs 		
On-Board Devices	<ul style="list-style-type: none"> • System on Chip • Hardware Root of Trust • IMPI 2.0 with virtual-media-over-LAN and KVM-over-LAN support • ASPEED AST2600 BMC graphics 		
PCIe 5.0 Topology	<ul style="list-style-type: none"> • Dual root, direct connected • 4 GPUs and one PCIe x16 slot directly connected to each CPU 	<ul style="list-style-type: none"> • Single root • 10 GPUs and 1 PCIe x16 slot connected to one CPU via two PLX switches 	<ul style="list-style-type: none"> • Dual root • 5 GPUs and 2 PCIe x16 slots, connected to each CPU via PLX, • 4 PCIe Gen 5 NVMe drives directly connected to each CPU
Expansion Slots³	<ul style="list-style-type: none"> • 8 PCIe 5.0 x16 slots for double-width GPU accelerators • 1 PCIe 5.0 x16 or 2 x8 slots 	<ul style="list-style-type: none"> • Up to 10 PCIe 5.0 x16 slots for double-width GPU accelerators • 1 PCIe 5.0 x16 slot • 1 x16 slot 	<ul style="list-style-type: none"> • Up to 10 PCIe 5.0 x16 slots for double-width GPU accelerators • 1 PCIe 5.0 x16 slot • 1 x16 slot
GPU Support⁴	<ul style="list-style-type: none"> • NVIDIA® A100, H100, AMD Instinct™ MI200 Series • Supports both air and active and passive water-cooled GPUs • Optional NVIDIA NVLink™ Bridge, AMD Infinity Fabric™ Link for GPU-to-GPU connectivity 		
Storage	<ul style="list-style-type: none"> • Up to 4 hot-swap 2.5" NVMe drives² • 2x 2.5" hot-swap SATA drives² • 1 M.2 NVMe boot drive 	<ul style="list-style-type: none"> • Up to 8 hot-swap 2.5" NVMe drives² • 2x 2.5" hot-swap SATA drives² • 1 M.2 NVMe boot drive 	<ul style="list-style-type: none"> • Up to 8 hot-swap 2.5" NVMe drives² • 2x 2.5" hot-swap SATA drives² • 1 M.2 NVMe boot drive
I/O Ports	<ul style="list-style-type: none"> • 2 RJ45 Gigabit Ethernet ports • 1 RJ45 Dedicated IPMI LAN port • 2 USB 3.0 Ports (rear) • 1 VGA Connector • 1 TPM 2.0 header 		
BIOS	<ul style="list-style-type: none"> • AMI Code Base 256 Mb (32 MB) SPI EEPROM 		
System Management	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 8x 11.5K RPM heavy-duty fans 		
Power Supply	<ul style="list-style-type: none"> • 4x 2000W Titanium-Level power supplies with PMBus 		

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
 3. Specifications subject to change
 4. GPU support is limited to specific conditions