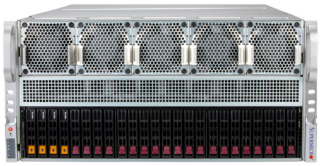
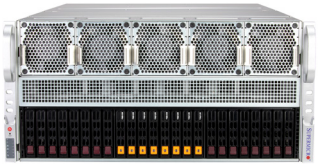
H14 5U PCIe GPU Systems

Flexible High-Density GPU Systems for AI and HPC



A+ Server 5126GS-TNRT



A+ Server 5126GS-TNRT2

GPU-Dense Servers Optimized for Compute-Intensive Workloads

Support up to 10 PCIe GPU accelerators, plus bandwidth to spare for networking and disk storage:

- Dual-socket systems supporting 4th and 5th Gen AMD EPYC™ processors
- Dual-root PCIe 5.0 topology including direct-connected and switched GPU connectivity
- Supports PCIe GPUs up to 600W including AMD Instinct™ and NVIDIA® HGX
- 24 DIMMs for up to 9 TB of DDR5-6000 memory
- Up to 8 NVMe hot-swap front-panel drives; up to 10 bays total
- Flexible PCIe and OCP 3.0 slot options for I/O and networking
- Up to 8 redundant Titanium-Level efficiency power supplies

If your goal is for maximum acceleration of artificial intelligence (AI), machine learning (ML) or high performance computing (HPC) workloads, look no further than Supermicro's 5U GPU systems with flexibility for up to 10 PCIe-form-factor accelerators.

1- and 2-Socket GPU-Optimized Server

To support AI, ML, and HPC workloads, we designed a family of servers powered by one or two AMD EPYC 9004 or 9005 Series CPUs with up to 160 lanes of PCIe 5.0 connectivity to up to ten AMD or NVIDIA GPUs. 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 -5126GS-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 for an additional 16-lane PCIe 5.0 and one x16 OCP 3.0 AIOM slot. The server includes support for up to 4 NVMe and 2 SATA hot-swap, front-panel-accessible drives.
- **AS -5126GS-TNRT2 Server:** uses a dual-root PCIe topology 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 CPU and GPUs and high-speed interconnects within clusters. For I/O connectivity, the server supports up to 3 x16 PCIe 5.0 and one x16 OCP 3.0 AOIOM slot. This server supports up to 8 NVMe and 2 SATA hot-swap, front-panel-accessible drives.

Designed with PCIe 5.0 Connectivity Throughout

Supermicro's H14 family of 5U 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 Open Compute Project (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 24 DDR5-6000 DIMMs for up to 9 TB of main memory. They are powered by up to eight redundant 2700 Titanium-Level power supplies and cooled by 10 heavy-duty fans. This power and cooling infrastructure supports

CPUs up to 500W and GPU up to 600W accelerators with either air or active or passive liquid cooling.

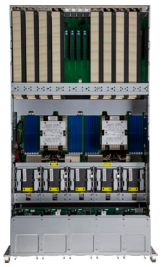
Compatible with 4th and 5th Gen AMD EPYC Processors

Our H14 servers are designed with SP5 processor sockets that support AMD EPYC 9004/9005 Series processors, with up to 128 cores per CPU (EPYC 9004 Series) or up to 192 cores per CPU (EPYC 9005 Series) for up to 256 or 384 cores per server (respectively). The EPYC 9005 Series features a set of AI-optimized processors with up to 64 cores running at high frequencies to speed single-threaded performance that is often needed to quickly prepare and marshal data to the GPUs. For in-memory AI inference, the EPYC 9005 Series now supports full 512-bit data paths to accelerate vector processing with AVX-512 instructions.

AMD EPYC processors support 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.



H14 Generation	AS -5126GS-TNRT Server	AS -5126GS-TNRT2 Server
Form Factor	• 5U rackmount	
SP5 CPU Sockets	• 2	• 2
Processor Support	• Up to 2 AMD EPYC™ 9004 or 9005 Series processors • Up to 293 cores and up to 500W TDP[1] per processor (up to 384 cores per server)	
Memory Slots & Capacity	• 12-channel DDR5 memory support • 24 DIMM slots for up to 9 TB ECC DDR5-6000 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	
PCIe 5.0 Topology	• Dual root • 4 GPUs up to 600W directly connected to each CPU	• Dual root • 5 GPUs up to 600W connected to each CPU via PLX switches
Expansion Slots[3]	• 8 PCIe 5.0 x16 FHFL slots for double-width GPU accelerators • 1 PCIe 5.0 x16 LP or 2 x8 FHFL slots • 1 x16 OCP 3.0 AIOM slot	• Up to 10 PCIe 5.0 x16 FHFL slots for double-width GPU accelerators • 3 PCIe 5.0 x16 slots • 1 x16 OCP 3.0 AIOM slot
GPU Support[4]	• AMD Instinct™ and NVIDIA®GPU accelerators • Supports both air and active and passive water-cooled GPUs • Optional NVIDIA NVLink™ Bridge, AMD Infinity Fabric™ Link for GPU-to-GPU connectivity	
Storage	• Up to 4 hot-swap 2.5" NVMe drives[2] • Up to 2 hot-swap 2.5" SATA drives[2] • 1 M.2 NVMe boot drive	• Up to 8 hot-swap 2.5" NVMe drives[2] • Up to 2 hot-swap 2.5" SATA drives[2] • 1 M.2 NVMe boot drive
I/O Ports	• 2 RJ45 Gigabit Ethernet ports (or does the photo show an AIOM card installed?) • 1 RJ45 Dedicated IPMI LAN port • 2 USB 3.0 Ports (rear) • 1 VGA Connector • 1 TPM 2.0 header	
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	• 10 hot-swap heavy-duty fans	
Power Supply	• Up to 8 redundant 2700W Titanium-Level power supplies with PMBus	

1. Certain CPUs with high TDP 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