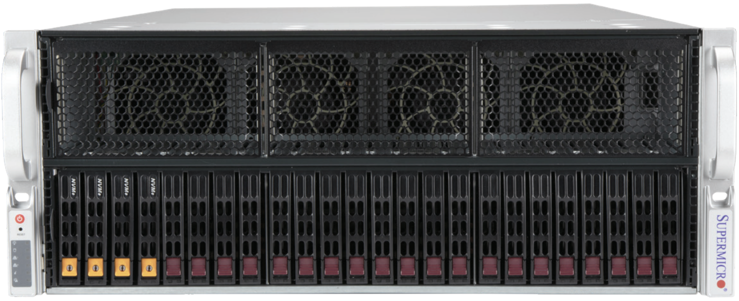
H12 4U GPU Server

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



A+ Server 4124GS-TNR+

GPU-Dense Server Optimized for High I/O Throughput

Support up to 8 GPU accelerators of your choice plus bandwidth to spare for network and disk storage:

- Up to 8 GPU accelerators directly connected to the CPUs through dedicated x16 PCI-E 4.0
- 2 additional directly connected double-width x16 slots for networking and storage access
- 2-socket design supporting 2nd or 3rd Gen AMD EPYC™ Processors
- Up to 32 DIMMs for up to 8 TB of DDR4-3200 memory
- 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 server with flexibility to host up to eight PCI-E-form-factor GPUs that best meet your needs.

Dual-Socket GPU-Optimized Server

To support GPU-accelerated workloads, we designed the server with 16 lanes of PCI-E 4.0 connecting directly to each of eight GPUs, with no latency-inducing switching or bandwidth sharing. Nothing stands in the way of your data once it is on the GPU, as the server supports AMD Infinity Fabric™ Link or NVIDIA® NVLink Bridge™ technologies for accelerated GPU-to-GPU connectivity. But that's not all. One additional x16 PCI-E 4.0 slot powers the fastest 200-Gbps InfiniBand interconnect for HPC clusters or dual 100 Gigabit Ethernet ports often used for AI and ML workloads.

The last of ten PCI-E slots connects to four hot-swap NVMe drives (default) or up to 24 hot-swap SATA3 drives with an optional RAID controller (see options on Page 2). With or without the NVMe drives, the processor's system-on-chip (SoC) also supports a pair of optional RAID-1-managed SATA drives. To hold your data close for fast access, 32 DIMM slots support up to 8 TB of main memory. Whether you access massive amounts of data on local storage or over the network, this sever puts it to work seamlessly.

The server is powered by redundant 2000W Titanium-Level power supplies and cooled by eight 11.5k RPM heavy-duty fans. This power and cooling infrastructure affords unrestricted choice of CPUs and GPU accelerators. You can choose from the fastest AMD EPYC™ CPUs, including those with AMD 3D V-Cache™ technology. You can use the fastest GPUs from the leading vendors, including the AMD Instinct™ MI100 and MI210 accelerators, and the NVIDIA® Ampere A100. If your workloads don't require top-of-the-line accelerators, you can choose from mid-range GPUs including the NVIDIA A30 and A40.

Today, the server easily powers and cools the most power-hungry GPUs, and the chassis is prepared to accept both passive and active liquid-cooled devices as needed. This makes the AS -4124GS-TNR server ready to handle your most challenging workloads now and into the future, including:

- **AI and ML workloads:** These require GPU acceleration that can harness multiple GPUs with high-speed interconnects enabling memory sharing between devices. If you need the highest inter-GPU connectivity, consider our OAM-form-factor AS-4124GQ-TNMI server.
- **HPC workloads:** These require GPU acceleration plus high-speed cluster interconnects. If a pair of RAID 1 SATA drives is sufficient for boot, the server can be equipped with a second 200-Gbps InfiniBand interface for even greater cluster and storage connectivity
- **Cloud gaming:** For these workloads, graphics rendering must be virtually instantaneous for a large number of users.

- **Molecular dynamics simulation:** Large amounts of data in three dimensions must be processed and result passed to other nodes in the cluster.

Made Possible by AMD EPYC Processors

The AS -4124GS-TNR server is made possible by the features of AMD EPYC Processors:

- **Massive I/O capacity** is supported by AMD EPYC CPUs to achieve 160 lanes of PCI-E 4.0 connectivity. In this server design, 80 lanes per processor are dedicated to the expansion slots, each one supporting 16 lanes of direct CPU connectivity. There are no intervening controllers, no bandwidth sharing, and no PCIe switches that could slow the flow of information.
- **System-on-chip design** that supports the server's built-in functions including support for dual Gigabit Ethernet ports, USB and KVM functions, and even support for a pair of RAID 1 SATA drives that can be used for system boot. The SoC-oriented design eliminates chip sets, helping to reduce complexity and power consumption.
- **High performance** powered by up to 64 cores per processor and up to 768 MB of L3 cache (in processors with AMD 3D VCache technology). Choose from main-line CPUs having from 8 to 64 cores each, high-frequency processors with excellent per-core performance, and AMD EPYC 7003 Series CPUs with AMD 3D V-Cache™ technology.

All of these AMD EPYC features are consistent across the product line, meaning that you can match the processor to your workload without concern for whether or not a particular feature is supported.

Open Management

Our approach to management enables you to deliver the scale your organization requires. Supermicro® SuperCloud Composer with open-source Redfish® compatibility 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 API provides 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's philosophy, our open management APIs and tools are ready to support you.

H12 Generation	AS -4124GS-TNR Server	AS -4124GS-TNR+ Server
Form Factor	• 4U rackmount	
Processor Support	• Dual SP3 socket for AMD EPYC™ 7002 or 7003 Series processors (two CPUs required) • Up to 64 cores and up to 280W TDP[†] per processor (up to 128 cores per server)	
Memory Slots & Capacity	• 32 DIMM slots for Registered ECC DDR4 3200-MHz SDRAM	
On-Board Devices	• System on Chip • Marvell 9230 RAID controller for optional pair of SATA drives • Intel i350 2-port Gigabit Ethernet controller • IMPI 2.0 with virtual-media-over-LAN and KVM-over-LAN support • ASPEED AST2600 BMC graphics	
Expansion Slots	• Each of 10 PCI-E 4.0 x16 slots have direct CPU connectivity, 5 slots are connected to each CPU • 8 slots for GPU accelerators • 1 slot for network connectivity including 200-Gbps InfiniBand and 2x 100 Gigabit Ethernet • 1 flexible slot for local disk connectivity with several options: – 4 hot-swap NVMe drives (default, uses all 16 lanes); or – 2 hot-swap NVMe drives (frees 8 lanes for customer use); or – 0 drives (frees 16 lanes for customer use)	
GPU Support	– NVIDIA A100, A4000, A5000, A6000, A4500, A2, A10, A16, A30, A40, V100, T4 Quadro RTX, AMD Radeon, AMD Instinct MI100, MI210 – 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 24 Hot-swap 2.5" drives using PCIe slot as described above: – 2x or 4x 2.5" NVMe drives (4x is the default configuration); or – Up to 24 HDDs with optional PCI-E RAID controller • 2x 2.5" hot-swap SATA drives with optional RAID 1 via onboard controller	
	• NVMe drives: PCI-E 3.0	• NVMe drives: PCI-E 4.0
I/O Ports	• 2 RJ45 Gigabit Ethernet ports • 1 RJ45 Dedicated IPMI LAN port • 2 USB 3.0 ports (rear) • 1 VGA Connector • 1 COM port (header)	
BIOS	• AMI 32Mb SPI Flash ROM	
System Management	• Supermicro SuperCloud Composer • IPMI 2.0 • KVM with dedicated LAN • SSM, SPM, SUM • SuperDoctor® 5 • Watchdog	
System Cooling	• 8x 11.5K RPM heavy-duty fans	
Power Supply	• Redundant 2000W Titanium-Level power supplies with PMBus	

[†] Certain high TDP CPUs may be supported only under specific conditions. Please contact Supermicro Technical Support for additional information about specialized system optimization.