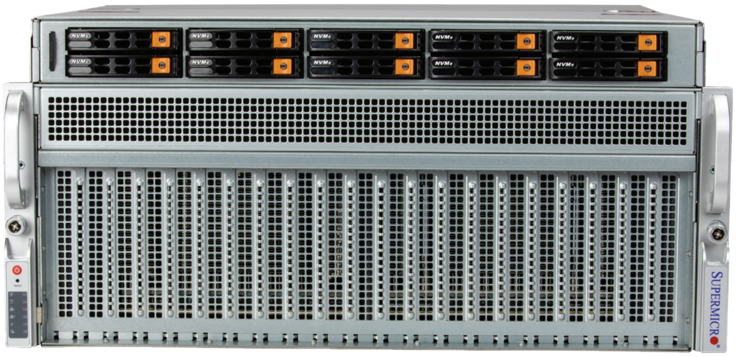
H12 Universal GPU

Open Standards-Based Server Design for Architectural Flexibility



A+ Server 4124GQ-TNMI with optional 1U expansion module

High-Density Servers for the Most Demanding HPC and AI Workloads

Supports 4 OAM-form-factor GPUs for extreme workload acceleration:

- Modular design for flexibility and future-proof operation
- Optimized storage, networking, power, and cooling for high-performance GPU accelerators
- 2-socket design supporting 3rd Gen AMD EPYC™ processors
- Up to 32 DIMMs for up to 8 TB of DDR4-3200 memory
- 10 hot-swap NVMe/SAS/SATA drive bays
- 8 PCI-E 4.0 x16 expansion slots
- 2 additional PCIe or AIOM slots with optional expansion module
- Titanium-Level efficiency power supplies

One of the industry's most advanced and flexible GPU servers, the A+ Server 4124GQ-TNMI is designed to deliver maximum acceleration power for large-scale machine learning and HPC workloads. This modular, open-standards-based platform supports the family of GPU accelerators selected for the exascale Frontier supercomputer: the AMD Instinct™ MI200 Series.

Designed for Open Standard GPU Accelerators

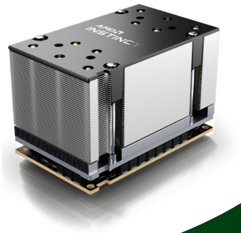
The AS -4124GQ-TNMI is part of a family of servers with a consistent architecture to support a range of GPU accelerator form factors including Open Compute Project (OCP) Accelerator Modules (OAM). The server is powered by two 3rd Gen AMD EPYC™ processors and supports four OAM-form-factor AMD Instinct™ MI250 GPUs with integrated AMD Infinity Fabric™ Link technology that provides a total of up to 600 GB/s of aggregate bandwidth between modules.

To speed communication between the CPU and GPUs, we designed the server with 32 lanes of PCI-E 4.0 bandwidth to each of the four GPUs. The server supports 10 hot-swap 2.5" NVMe U.2 or SAS/SATA drives. Intracluster communication is very important in HPC workloads, so eight x16 low-profile PCIe slots are provided for networking and cluster interconnects. The server is powered by four 3000W Titanium-Level power supplies and cooled by five 11.5K RPM fans. Whether you access massive amounts of data on local storage or over the network, this sever puts your data to work seamlessly.

The server's thermal design gives it headroom to support GPU accelerators well into the future. The 4U version of the server supports modules (including the MI250) with up to 560W TDP. An optional 1U expansion module increases airflow so the server can accommodate up to 700W per GPU. With the expansion module, you also gain two open-standard AIOM 3.0 or PCI-E 4.0 x16 slots.

Acceleration for the Exascale Era

The AMD Instinct™ MI200 series accelerators are AMD's newest data center GPUs, designed to power discoveries in Exascale systems, enabling scientists to tackle the most pressing challenges in both artificial intelligence and high-performance computing. The Instinct MI250 makes a quantum leap by supporting HPC codes with up to 90.5 TFLOPs of FP64 matrix operations, and supporting AI endeavors with up to 362.1 TFLOPs of FP16 capacity. The accelerators house two GPU chips in each package and support up to



128 GB of high-speed HBM2e memory. The AMD ROCm™ platform provides the most popular HPC and AI platforms and libraries so that you can compile for ROCm and run on the GPU accelerator of your choice.

Power for HPC and AI Deep Learning

This is the server you need to overcome your most difficult HPC and AI/ML challenges, including:

- **Deep-learning workloads:** These applications use process massive amounts of data to train machine-learning models in areas such as video detection and life sciences, drug discovery, autonomous driving, and robotics.
- **High-performance computing:** These workloads use proprietary and open-source platforms to model geophysical systems, molecular dynamics, physics, viruses, computational chemistry and climate sciences.

The Power of AMD EPYC Processors

Depending on your workload, the CPU may have an important role to play in providing highly parallel compute power to speed time to results. Third-generation AMD EPYC processors are 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.

The server's massive I/O requirements are handled easily by AMD EPYC processors that provide 128 lanes of PCI-E 4.0 connectivity for server designs. Clock synchronization between the CPU and memory clocks speeds transfer of data throughout the system.

AMD EPYC features such as these 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 -4124GQ-TNMI Server
Form Factor	• 4U rackmount • 5U rackmount (with optional expansion module)
Processor Support	• Dual SP3 socket for AMD EPYC™ 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 up to 8 TB registered ECC DDR4 3200-MHz RDIMM/LRDIMM
GPU Support	• 4 AMD Instinct MI250 GPU accelerators with integrated AMD Infinity Fabric™ Link for GPU-to-GPU connectivity • Power and cooling support for 560W per GPU
On-Board Devices	• System on chip • IMPI 2.0 with virtual-media-over-LAN and KVM-over-LAN support • ASPEED AST2600 BMC graphics
Expansion Slots	• 8 PCI-E 4.0 x16 slots via PLX • Optional 1U expansion module provides 2 additional PCI-E slots or AIOM module slots via CPU
Storage	• 10 hot-swap 2.5" U.2 NVMe/SATA/SAS hybrid drive bays
I/O Ports	• 1 RJ45 Dedicated IPMI LAN port • 2 USB 3.1 Gen 1 ports (front) • 1 VGA connector (front) • 1 COM port (header)
BIOS	• AMI 32Mb SPI Flash ROM
System Management	• Supermicro SuperCloud Composer • Integrated IPMI 2.0 plus KVM with dedicated LAN • Supermicro Server Manager (SSM) • Supermicro Update Manager (SUM) • Supermicro SuperDoctor® 5 • Watch Dog
System Cooling	• 5x 11.5K RPM hot-swap heavy-duty fans
Power Supply	• 4x 3000W Titanium-Level power supplies

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