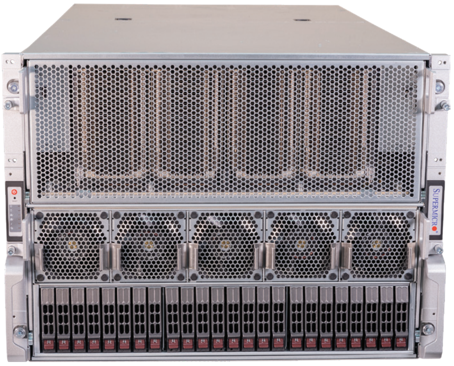
H13 8U GPU Systems

Next-Generation Machine Learning Platform



A+ Server 8125GS-TNHR

Propel ML Workloads with 4th Gen AMD EPYC™ Processors and NVIDIA® HGX H100 8-GPU Modules

Optimized for machine learning through massive I/O capacity:

- End-to-end nonblocking I/O to keep GPUs hydrated with data
- 2-socket design supporting 4th Gen AMD EPYC™ Processors
- Up to 24 DIMMs for up to 6 TB of DDR5-4800 memory
- Flexible PCI-E 5.0 options for I/O and networking
- Dual-zone cooling optimized for performance and lower operating costs
- Titanium-Level efficiency power supplies

Imagine your data lake cascading directly into GPU memory at an aggregate rate of 3.2 terabits per second, hydrating the new engine of AI factories, the NVIDIA HGX H100 8-GPU module. If your challenges include feeding machine-learning models with massive amounts of data, and processing that data with the most advanced accelerators available today, the Supermicro AS -8125GS-TNHR server is designed to meet your challenges.

End-to-End Nonblocking I/O

The NVIDIA HGX H100 8-GPU module can transform training exercises that once took weeks into ones that take only days. Keeping utilization high—and total cost of ownership low—depends on satisfying the eight GPUs' voracious appetite for data.

Whether data is supplied from main memory, or from a network-based data lake, the AS -8125GS-TNHR is designed to move massive amounts from its source to GPU memory with low-latency, nonblocking PCI-E 5.0 connectivity. Each of the system's GPUs is closely coupled with a PCI-E 5.0 switch that provides 16 lanes of connectivity to one of the system's AMD EPYC 9004 Series processors, 4 lanes to an optional NVMe drive, and 16 lanes to a dedicated PCI-E slot that can support up to 400 Gpbs of network connectivity. Remote DMA (RDMA) can set up high-speed transfers from network-based storage through the network interface card, and direct to GPU memory with no intervening CPU cycles necessary. Once data is loaded into one accelerator's memory, it can share with its peers through the NVIDIA NVSwitch™ that interconnects all eight GPUs on the 8-GPU module. This helps to speed complex models that pipeline data through multiple GPUs. With 900 GBps of bandwidth between any two GPUs, your data moves without bottlenecks.

GPU-Optimized Server

The entire server architecture is built to support the power, cooling, and I/O demands of machine learning training, inferencing, and other data-intensive analytics operations. Powered by two AMD EPYC 9004 Series processors and enabled with up to 6 TB of DDR5-4800 memory, the interconnections on this server are based on PCI-E 5.0 technology, doubling the bandwidth of prior generations. Eight slots support tightly-coupled NICs while up to 4 additional slots can be used for general I/O purposes. Similarly, 8 NVMe drives are tightly coupled to individual accelerators, while up to 8 optional drives can provide additional storage.

Titanium-Level power supplies with up to 96% efficiency keep the GPUs accelerating your machine-learning software while dual-zone cooling with 10 counter-rotating fans keep the accelerators within their thermal envelopes.

© 2023 Copyright Super Micro Computer, Inc. All rights reserved

Made Possible by 4th Gen AMD EPYC Processors

This server is made possible by 4th Gen AMD EPYC processors, with up to 128 cores per CPU and 256 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 supports massive I/O capacity, with 160 lanes of PCI-E 5.0 connectivity in this 2-socket server. The system-on-chip (SoC) design supports built-in functions including IPMI-based management, on-board M.2 drive, and built-in SATA controllers for two drives. 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 -8125GS-TNHR
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[†] 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 • IMPI 2.0 with virtual-media-over-LAN and KVM-over-LAN support • ASPEED AST2600 BMC graphics
GPU Support[3]	• NVIDIA HGX H100 8 GPU SXM5 board
Expansion Slots	• 8 PCI-E 5.0 x16 low-profile slots connected to GPU via PCI-E switch • 2 PCI-E 5.0 x16 full-height half-length slots • Optional 2 PCI-E 5.0 x16 full-height half-length slots via expansion kit
Storage	• Up to 8 PCI-E 5.0 x4 U.2 NVMe drives connected to GPU via PCI-E switch • 4 PCI-E 5.0 x4 NVMe U.2 drives • Optional 4 PCI-E 5.0 x4 NVMe U.2 drives[2] • 1 M.2 NVMe/SATA 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

[3] GPU support is limited to specific conditions

© 2023 Super Micro Computer, Inc. Specifications subject to change without notice. All other brands and names are the property of their respective owners. All logos, brand names, campaign statements and product images contained herein are copyrighted and may not be reprinted and/or reproduced, in whole or in part, without express written permission by Supermicro Corporate Marketing.