Supermicro Large Scale AI Training

Large Language Models, Generative AI Training, Autonomous Driving, Robotics

Large-Scale AI training demands cutting-edge technologies to maximize parallel computing power of GPUs to handle billions if not trillions of AI model parameters to be trained with massive datasets that are exponentially growing. Leverage NVIDIA's HGX™ H100 SXM 8-GPU/4-GPU and the fastest NVLink™ & NVSwitch™ GPU-GPU interconnects with up to 900GB/s bandwidth, and fastest 1:1 networking to each GPU for node clustering, these systems are optimized to train large language models from scratch in the shortest amount of time. Completing the stack with all-flash NVMe for a faster AI data pipeline, we provide fully integrated racks with liquid cooling options to ensure fast deployment and a smooth AI training experience.

Systems

AI Rack Solutions

Multi-Architecture Flexibility with Future-Proof Open-Standards-Based Design for POD, and SuperPOD with Liquid Cooling

Extra Large Workload:
Liquid Cooled AI Rack Solutions

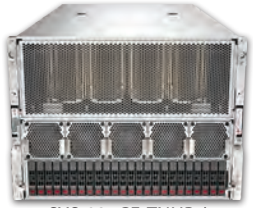
- NVIDIA HGX H100 SXM 8-GPU
- Up to 80 kW/Rack

HGX H100 Systems

Multi-Architecture Flexibility with Future-Proof Open-Standards-Based Design

Large Workload:
8U 8-GPU System

- NVIDIA HGX H100 SXM 8-GPU
- 16 U.2 NVMe Drives
- 8 PCIe 5.0 x16 networking slots



SYS-821GE-TNHR /
AS -8125GS-TNHR

Medium Workload:
4U 4-GPU

- NVIDIA HGX H100 SXM 4-GPU
- 6 U.2 NVMe Drives
- 8 PCIe 5.0 x16 networking slots

SYS-421GU-TNXR

Petabyte Scale Storage

High Throughput and High-Capacity Storage for AI Data Pipeline

Petabyte Scale NVMe Flash:

1U 24-Bay E1.S
SSG-121E-NES24R

1U 16-Bay E3.S
SSG-121E-NE316R /
ASG-1115SNE316R

2U 24/32-Bay E3.S
SSG-221E-NE324R /
ASG-2115S-NE332R

Petabyte Scale HDD:

4U 60/90-Bay
Top-Loading

SSG-640SP-E1CR60 /
SSG-640SP-E1CR90

Recommended NVIDIA GPUs

HGX H100 SXM5 4-GPU or 8-GPU

- H100 SXM5 board with 4-GPU or 8-GPU
- NVLink & NVSwitch Fabric
- PCIe 5.0
- 700W per GPU
- 80GB HBM3 per GPU

Accelerate Large Scale AI Training Workloads

Large Language Models, Generative AI Training, Autonomous Driving, Robotics

Opportunities and Challenges:

- Pool of 10,000+ GPUs and GPU memory to fit large AI models to maximize parallel computing and minimize training time
- Training with massive amount of data with continuous growth of data size (e.g. over 1 trillion tokens)
- Serve AI models (inference) to millions of concurrent users
- High performance everything: GPUs, memory, storage, and network fabric

Key Technologies:

- NVIDIA HGX H100 SXM 8-GPU/4-GPU with 900GB/s NVLink interconnect
- Dedicated, lots of high performance, high bandwidth GPU memory - HBM3, HBM2e
- 400GbE networking (Ethernet or InfiniBand), PCIe 5.0 storage for fast AI data pipe
- NVIDIA GPUDirect RDMA and Storage to keep feeding data to GPUs with minimum latency
- Liquid cooling for GPUs and CPUs

Solution Stack:

- DL Frameworks: TensorFlow, PyTorch
- Transformers: BERT, GPT, Vision Transformer
- NVIDIA AI Enterprise Frameworks (NVIDIA Nemo, Metropolis, Riva, Morpheus, Merlin)
- NVIDIA Base Command (infrastructure software libraries, workload orchestration, cluster management)
- High performance storage (NVMe) for training cache
- Scale-out storage for raw data (data lake)

Use Cases:

- Large Language Models (LLMs)
- Autonomous Driving Training
- Recommender Systems

GPU Acceleration for Complete Range of Workloads

Go to www.supermicro.com/ai or scan the QR code to download the AI Workload Solution Brochure: