Supermicro HPC/AI

Engineering Simulation, Scientific Research, Genomic Sequencing, Drug Discovery

Accelerating time to discovery for scientists, researchers, and engineers, more and more HPC workloads are augmenting machine learning algorithms and GPU accelerated parallel computing to achieve faster results. Many of the world's fastest supercomputing clusters are now taking advantage of GPUs and the power of AI.

HPC workloads typically require data-intensive simulations and analytics with massive datasets and precision requirements. GPUs such as NVIDIA's H100 provide unprecedented double-precision performance, delivering 60 teraflops per GPU, and Supermicro's highly flexible HPC platforms allow high GPU counts and CPU counts in a variety of dense form factors with rack scale integration and liquid cooling.

Systems

HGX H100 Systems

Designed for Largest AI-fused HPC Clusters

Large Workload:
4U 4-GPU or 8U 8-GPU System

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

SYS-421GU-TNXR

8U SuperBlade®

Highest Density Multi-Node Architecture for HPC, AI and Cloud Applications

Large Workload:
8U SuperBlade®

- Up to 2 (double-wide blade) or 1 (single-wide blade) H100 PCIe per blade
- 2 M.2 NVMe Drives per blade
- 2 E1.S Drives per blade
- 200G HDR InfiniBand per blade

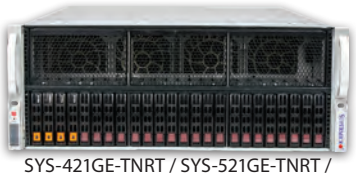
SBI-411E-1G/5G

10 GPU Systems

4U/5U 8 or 10 GPU PCIe - Maximum Performance and Flexibility

Medium Workload:
4U 10-GPU

- Up to 10 H100 PCIe
- 8 NVMe + 8 SATA drives
- 4-5 PCIe 5.0 x16 networking slots



SYS-421GE-TNRT / SYS-521GE-TNRT / AS -4125GS-TNRT

1U Grace Hopper MGX Systems

CPU+GPU Coherent Memory System for AI and HPC Applications

Medium Workload:
1U Grace Hopper MGX System

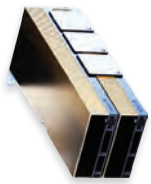
- 1 NVIDIA Grace Hopper™ SuperChip (ARM CPU and H100 with 96GB HBM3)
- 8 E1.S + 2 M.2 drives
- 480GB LPDDR5X
- 200G HDR InfiniBand

ARS-111GL-NHR

Recommended NVIDIA GPUs

H100 SXM5

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



H100 NVL

- 2 FHFL H100 GPU with NVLink Bridge
- PCIe 5.0
- 400W per GPU
- 94GB HBM3 per GPU

H100 PCIe

- FHFL
- PCIe 5.0 x16
- 350W
- 80GB HBM2e

L40S

- FHFL DW
- PCIe 4.0 x16
- 350W
- 48GB GDDR6

Accelerate HPC/AI Workloads

Engineering Simulation, Scientific Research, Genomic Sequencing, Drug Discovery

Opportunities and Challenges:

- Augmenting machine learning algorithms and GPU accelerated parallel computing to HPC workloads to achieve faster results and discoveries
- Parallel processing with massive datasets for data-intensive simulations and analytics
- Simulations requiring double precision (FP64)
- High-resolution and real-time visualization of scientific simulations and modeling

Key Technologies:

- Double-precision Tensor Cores delivering 535/268 teraFLOPs with HGX H100 SXM 8-GPU/4-GPU, or 134 teraFLOPs with H100 NVL (2 GPUs with NVLink Bridge) at FP64
- High CPU compute and high GPU compute – e.g, up to 20 CPUs and 20 GPUs in 8U
- High bandwidth GPU memory and CPU cache/integrated memory – HBM3, HBM2e
- GPU-GPU Interconnect (NVLink) and 400GbE networking for clustering, PCIe 5.0 storage
- Liquid cooling for GPUs and CPUs

Solution Stack:

- NVIDIA HPC Software Development Kit (SDK) NVIDIA CUDA
- Commercial and in-house CAE software

Use Cases:

- Manufacturing and engineering simulations (CAE, CFD, FEA, EDA)
- Bio/life sciences (genomic sequencing, molecular simulation, drug discovery)
- Scientific simulations (astrophysics, energy exploration, climate modeling, weather forecasting)

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: