Accelerate Everything

Supermicro Enterprise AI Inference & Training

Generative AI Inference, AI-enabled Services/Applications, Chatbots, Recommender System, Business Automation

The rise of generative AI has been recognized as the next frontier for various industries, from tech to banking and media. The race to adopt AI has begun as a source to breed innovation, significantly boost productivity, streamline operations, make data-driven decisions, and improve customer experience.

Whether it is AI-assisted applications and business models, intelligent human-like chatbots for customer service, or AI to co-pilot code generation and content creation, enterprises can leverage open frameworks, libraries, pre-trained AI models, and fine-tune them for unique use cases with their own dataset. As the enterprise adopts AI infrastructure, Supermicro's variety of GPU optimized systems provide open modular architecture, vendor flexibility, and easy deployment and upgrade paths for rapidly evolving technologies.

Systems

10 GPU Systems

4U/5U 8 or 10 GPU PCIe — Highly Flexible Architecture

Extra Large Workload: 8 -10 GPU (PCIe)

- 8 NVIDIA H100 NVL or 10 H100 PCIe
- 8 NVMe and 8 SATA Drives
- 32 DIMMs DDR5-4800

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

6U SuperBlade®

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

Large Workload: 6U SuperBlade®

- 2 NVIDIA H100 PCIe
- 2 U.2 NVMe Drives
- 3 M.2 NVMe Drives
- 2 E1.S Drives
- 2x25GbE LOM

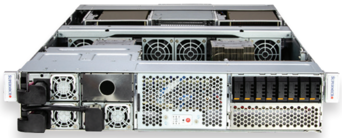
SBI-611E-5T2N

2U MGX System

Modular Building Block Platform Supporting Today's and future GPUs, CPUs, and DPUs

Medium Workload: 2U MGX System

- 4 NVIDIA H100 PCIe or NVL
- 8 E1.S + 2 M.2 drives
- 16 DIMMs DDR5-4800



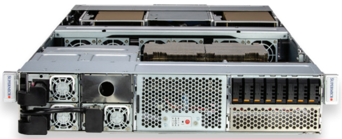
SYS-221GE-NR

2U Grace MGX System

Modular Building Block Platform with Energy-efficient Grace CPU Superchip

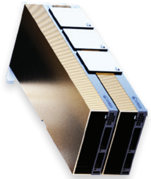
Medium Workload: 2U Grace MGX System

- 4 NVIDIA H100 PCIe, NVL, or L40S
- 8 E1.S + 2 M.2 drives
- 960GB LPDDR5X



ARS-221GL-NR

Recommended NVIDIA GPUs



H100 NVL

- 2 FHFL H100 GPU with NVLink Bridge (4x faster than PCIe)
- PCIe 5.0
- 400W per GPU
- 94GB HBM3 per GPU

H100 PCIE

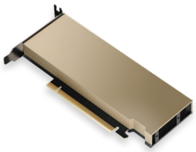
- FHFL DW
- PCIe 5.0 x16
- 350W
- 80GB HBM2e

L40S

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

L40

- FHFL DW
- PCIe 4.0 x16
- 300W
- 48GB GDDR6



L4

- HHHL SW
- PCIe 4.0 x16
- 72W
- 24GB GDDR6

Accelerate Enterprise AI Inference & Training Workloads

Generative AI Inference, AI-enabled Services/Applications, Chatbots, Recommender System, Business Automation

Opportunities and Challenges:

- AI adoption across industries to boost productivity, streamline operations, make data-driven decisions, and improve customer experience
- Open architecture, vendor flexibility, fast/easy deployment for rapidly evolving technologies
- High computational and resource costs, cloud vs. on-prem
- Utilization of frameworks, pre-trained models, open-source AI models with fine-tuning

Key Technologies:

- Flexible, modular, highly configurable rackmount servers with different form factors to balance compute, storage, networking, and cost for various enterprise AI workload needs for today and the future
- PCIe 5.0 supported platforms for future proofing – GPUs, storage, networking
- FP8 and FP16 support to boost performance with less resources and cost
- Intel, AMD, ARM CPU options
- NVIDIA Certified with NVIDA AI Enterprise and NGC catalog to fully leverage pre-trained models and optimized libraries and toolset

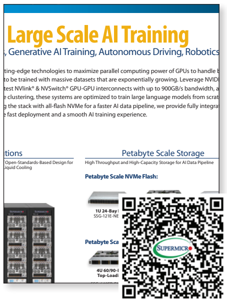
Solution Stack:

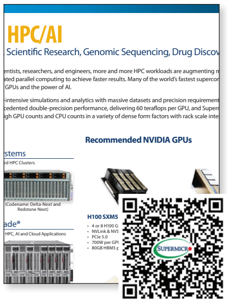
- NVIDIA AI Enterprise software
- NVIDIA NGC™ catalog: containers, pre-trained models
- RedHat OpenShift, VMWare

Use Cases:

- Content creation (image, audio, video, writing)
- AI-enabled office applications and services
- Enterprise business process automation

GPU Acceleration for Complete Range of Workloads


Large Scale AI Training


HPC/AI

Enterprise AI Inference & Training

Visualization & Design

Content Delivery & Virtualization

AI Edge

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

SUPERMICRO