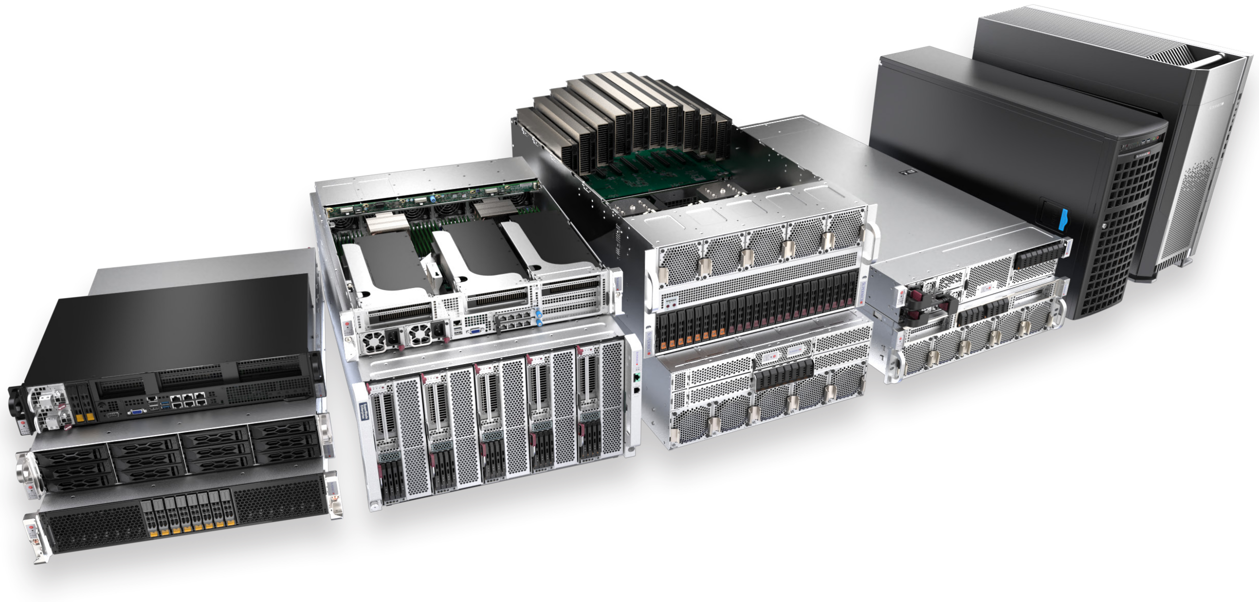
DATASHEET

Supermicro PCIe GPU Systems

Broad range of systems for LLM and Gen AI inference & fine-tuning, agentic AI, visualization, graphics & rendering, and virtualization



Featuring latest-generation NVIDIA RTX PRO™ 6000 Blackwell Server Edition, H200 NVL, and H100 NVL PCIe GPUs

- Broad portfolio of 100+ systems optimized for PCIe GPUs
- Acceleration from data center to edge with a wide range of form factors
- Wide-ranging workload support to adapt to almost any application, including virtualized and cloud environments with NVIDIA Multi-Instance GPU (MIG)
- Open architectures based on the industry-standard PCIe interconnect and optimized for air-cooled environments

Generative and Agentic AI

LLM Inference and Fine-Tuning

Rendering and 3D Graphics

Virtualization

Supermicro offers a full range of systems supporting industry-standard form factor GPUs, delivering flexible yet powerful acceleration of AI and graphics workloads from the data center to the edge. These include NVIDIA Certified systems which guarantee compatibility and support for NVIDIA AI Enterprise software to simplify the process of developing and deploying production AI.

Designed for maximum flexibility, Supermicro systems can be adapted to a wide range of applications including AI inference and fine-tuning, 3D rendering, media encoding, and virtualization with support for the latest generation of NVIDIA PCIe GPUs including NVIDIA RT PRO 6000 Blackwell Server Edition and H200 NVL. Systems can also support prior-generation GPUs including L40S and L4 for thermal and space-constrained environments.

Cloud and virtualization workloads will also benefit from new support for Multi-Instance GPU (MIG) on the NVIDIA RTX PRO 6000. Supermicro's thermally-optimized architectures maximize performance in air-cooled environments and are also designed to support NVIDIA SuperNICs such as BlueField®-3 and ConnectX®-7 for the best infrastructure scaling and GPU clustering with NVIDIA Quantum-2 InfiniBand or Spectrum™-X Ethernet.

Featured Products

5U GPU-Optimized

Up to 10 GPUs

RTX PRO™ 6000 | H200 NVL

MGX™ Systems

Up to 8 GPUs

RTX PRO™ 6000 | H200 NVL

Edge-Optimized

Up to 10 GPUs in 3U

RTX PRO™ 6000 | H100 NVL

SuperBlade®

Up to 120 GPUs per rack

RTX PRO™ 6000 | H200 NVL

Rackmount Workstation

Up to 4 GPUs

RTX PRO™ 6000 | L40S

Workstation

Up to 4 professional-grade GPUs

RTX PRO™ 6000 Blackwell Workstation Edition | L40S

4U GPU-Optimized

Up to 10 GPUs

H100 NVL | L40S

Rackmount

Up to 4 double-width or 8 single-width GPUs

H100 NVL | L40S

Multi-Processor

2 GPUs in 2U or up to 12 GPUs in 6U

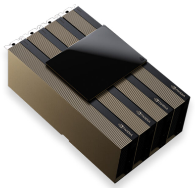
H100 NVL | L40S

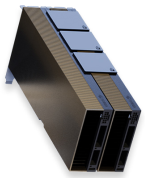
Edge

Up to 1 double-width GPU in a compact form factor

RTX PRO™ 6000 | L40S

NVIDIA PCIe GPU Specification Comparison





	NVIDIA RTX PRO™ 6000	NVIDIA H200 NVL	NVIDIA H100 NVL
Best For	**Generative AI, Graphics, Video**	**Large AI Model Inference & Fine-Tuning, Scientific Research, HPC**	**AI Inference & Fine-Tuning, HPC**
GPU Architecture	NVIDIA Blackwell	NVIDIA Hopper™	NVIDIA Hopper™
GPU-GPU Interconnect	PCIe 5.0 x16	4-way or 2-way NVIDIA NVLink™ at 900GB/s	2-way NVIDIA NVLink™ at 600GB/s
GPU Memory	96GB GDDR7	141GB HBM3e	94GB HBM3
GPU Memory Bandwidth	1.6 TB/s	4.8 TB/s	3.9 TB/s
MIG Instances	Up to 4 @ 24GB each	Up to 7 @ 16.5GB each	Up to 7 @ 12GB each
Media Engines	4 NVENC 4 NVDEC	7 NVDEC 7 NVJPEG	7 NVDEC 7 NVJPEG
Power (per GPU)	400W-600W (configurable)	Up to 600W (configurable)	350-400W (configurable)
Form Factor	2-slot FHFL	2x 2-slot FHFL 4x 2-slot FHFL	2x 2-slot FHFL

Visit http://www.supermicro.com/pcie-gpu or scan the QR code to visit the Supermicro PCIe GPU solutions web page: