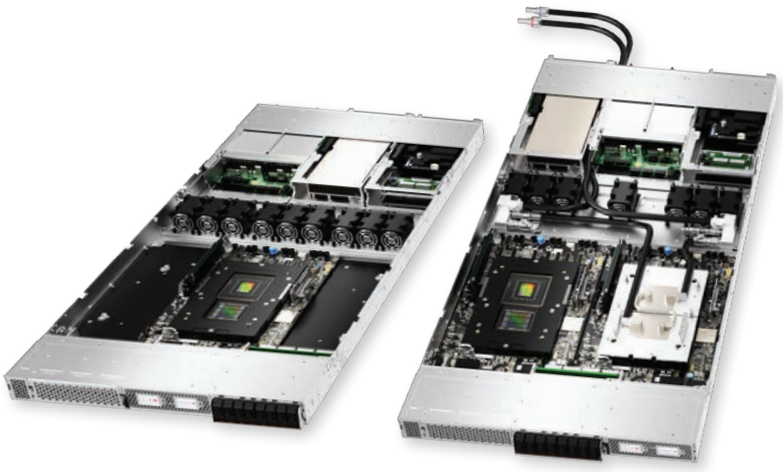
Supermicro NVIDIA MGX™ Systems

1U NVIDIA GH200 Grace Hopper™ Superchip Systems



Maximum AI Performance Per-Rack-Unit

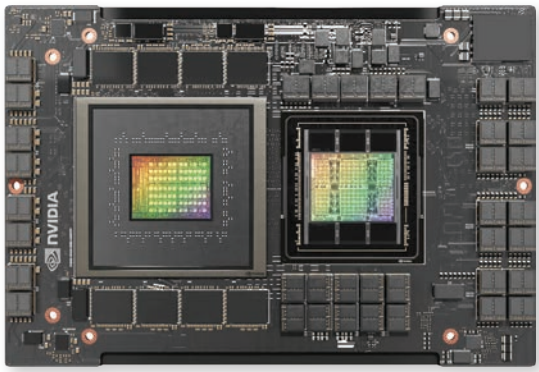
Supermicro NVIDIA MGX™ Systems are purpose-built for modern AI and accelerated computing. This modular platform enables new possibilities in system design and bleeding-edge technologies, including support for NVIDIA GH200 Grace Hopper™ Superchip which combines the power of an NVIDIA H100 GPU and NVIDIA Grace CPU on a single chip. In a mere 1- rack unit form factor, Supermicro NVIDIA MGX™ Systems can be equipped with up to 2 Grace Hopper Superchips and deliver the highest accelerated computing density in this compact form factor. 1U Grace Hopper Superchip Systems are ready to run any application in the NVIDIA AI software stack.

Grace Hopper Superchip

H100 GPU+ Grace CPU on one Superchip

The Grace Hopper Superchip addresses a key bottleneck in training and inference of AI models: access to plenty of high-bandwidth memory. NVLink® Chip-2-Chip(NVLink-C2C) provides a coherent CPU-GPU link that is 7x faster than PCIe 5.0.

Utilize a coherent pool of the 96GB of HBM3 for GPU and 480GB of LPDDR5X for CPU totaling 576GB of memory to accelerate AI and HPC applications. CPU and GPU threads can now concurrently and transparently access both CPU and GPU resident memory, allowing developers focus on algorithms instead of explicit memory management.



MGX: A Modern System Architecture

Optimized for thermals, compatibility, and flexibility.

MGX is a modular and flexible platform with support for the leading GPUs, CPUs, and DPUs of today and the future. The Supermicro MGX systems also support both air-cooling and liquid-cooling in the same compact chassis to enable highly dense compute featuring GPUs with 400W+ TDP while providing high energy efficiency. The modular bays on both sides facilitate constructing finely-tuned systems, such as providing flexible PCIe 5.0 slots supporting NVIDIA BlueField®-3 or NVIDIA ConnectX®-7 networking for the next generation of HPC supercomputing clusters and AI factories.

Cooling + Efficiency + Power Delivery

Increased Operations-Per-Second. Decreased OPEX.

Due to its mechanical design and component selection, Supermicro MGX™ Systems optimize cooling, efficiency, and power delivery without sacrifice.

Supermicro's proven Direct-to-Chip liquid cooling solutions can reduce OPEX by more than 40%. Up to 2x 2700W redundant Titanium Level power supplies deliver ample power to handle up to the 2000W power requirements of dual Grace Hopper Superchips, with headroom to spare.

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

Accelerate NVIDIA MGX Systems

1U NVIDIA GH200 Grace Hopper™ Superchip Systems

Construct new solutions for accelerated infrastructures enabling scientists and engineers to focus on solving the world's most important problems with larger datasets, more complex models, and new generative AI workloads. Within the same 1U chassis, Supermicro's dual NVIDIA GH200 Grace Hopper Superchip systems deliver the highest level of performance for any application on the CUDA platform with substantial speedups for AI workloads with high memory requirements. In addition to hosting up to 2 onboard H100 GPUs in 1U form factor, its modular bays enable full-size PCIe expansions for present and future of accelerated computing components, high-speed scale-out and clustering.

	ARS-111GL-NHR	ARS-111GL-NHR-LCC	ARS-111GL-DNHR-LCC
Form Factor	1U system with single NVIDIA GH200 Grace Hopper Superchip	1U system with single NVIDIA GH200 Grace Hopper Superchip	Dual node 1U system with single NVIDIA GH200 Grace Hopper Superchip per node
CPU	72-core Grace Arm Neoverse V2 CPU + H100 Tensor Core GPU in a single chip	72-core Grace Arm Neoverse V2 CPU + H100 Tensor Core GPU in a single chip	72-core Grace Arm Neoverse V2 CPU + H100 Tensor Core GPU in a single chip per node
GPU	NVIDIA H100 Tensor Core GPU with 96GB of HBM3 or 144GB of HBM3e (coming soon)	NVIDIA H100 Tensor Core GPU with 96GB of HBM3 or 144GB of HBM3e (coming soon)	NVIDIA H100 Tensor Core GPU with 96GB of HBM3 or 144GB of HBM3e (coming soon) per node
Memory	Onboard memory: CPU: 480GB integrated LPDDR5X with ECC	Onboard memory: CPU: 480GB integrated LPDDR5X with ECC	Onboard memory: CPU: 480GB integrated LPDDR5X with ECC per node
Drive	8x hot-swap E1.S drives and 2x M.2 NVMe drives	8x hot-swap E1.S drives and 2x M.2 NVMe drives	4x hot-swap E1.S drives and 2x M.2 NVMe drives per node
Networking	3x PCIe 5.0 x16 slots supporting NVIDIA BlueField-3 or ConnectX-7	3x PCIe 5.0 x16 slots supporting NVIDIA BlueField-3 or ConnectX-7	2x PCIe 5.0 x16 slots supporting NVIDIA BlueField-3 or ConnectX-7
Interconnect	NVLink-C2C with 900GB/s for CPU-GPU interconnect	NVLink-C2C with 900GB/s for CPU-GPU interconnect	NVLink-C2C with 900GB/s for CPU-GPU interconnect
Cooling	Air-cooling	Liquid-cooling	Liquid-cooling
Power	2x 2000W Redundant Titanium Level power supplies	2x 2000W Redundant Titanium Level power supplies	2x 2700W Redundant Titanium Level power supplies

Go to www.supermicro.com/mgx or scan the QR code to learn more.

© 2023 Copyright 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.