

GIGABYTE™



NVIDIA®

Revolutionizing Data Center
Performance with

**NVIDIA Grace CPU Superchip &
GH200 Grace Hopper Superchip**

NVIDIA Grace™ CPU Superchip

Bringing a Whole New Level of Performance and Efficiency to the Modern Data Center

In response to the rapidly growing demand for high performance and low power consumption, the Arm-based NVIDIA Grace CPU excels in modern data center workloads, emphasizing massive data processing capabilities. Its extraordinary performance per watt, packaging density, and memory bandwidth sets a new standard for the industry. The NVIDIA Grace CPU Superchip represents a comprehensive upgrade from the traditional dual-processor concept. By interconnecting the two CPUs on a single module with NVIDIA® NVLink®-Chip-to-Chip (NVLink-C2C) technology, achieving an astonishing bandwidth of 900GB/s, and packing up to 960GB LPDDR5X memory into the module, the superchip can efficiently handle various workloads, especially memory-intensive applications.

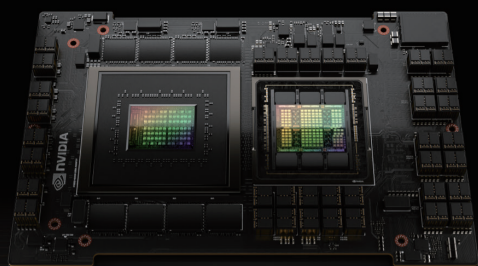


- ✓ 144 Arm Neoverse V2 Cores
- ✓ Up to 960GB CPU LPDDR5X memory with ECC
- ✓ Up to 8x PCIe Gen5 x16 links
- ✓ NVLink-C2C Technology
- ✓ NVIDIA Scalable Coherency Fabric (SCF)
- ✓ InfiniBand Networking Systems
- ✓ NVIDIA CUDA Platform

NVIDIA GH200 Grace Hopper™ Superchip

Stepping Further into the Era of AI and GPU-Accelerated HPC

Moving beyond pure CPU applications, the NVIDIA GH200 Grace Hopper Superchip is built on a combination of an NVIDIA Grace CPU and an NVIDIA H100 GPU for giant-scale AI and HPC applications. Utilizing the same NVIDIA NVLink-C2C technology, combining the heart of computing on a single superchip, forming the most powerful computational module. The coherent memory design leverages both high-speed HBM3 or HBM3e GPU memory and the large-storage LPDDR5X CPU memory. The superchip also inherits the capability of scaling out with InfiniBand networking by adopting BlueField®-3 DPUs or NICs, forming a system connected with a speed of 100GB/s for ML and HPC workloads. The upcoming GH200 NVL32 can further improve deep learning and HPC workloads by connecting up to 32 superchips through the NVLink Switch System, a system built on NVLink switches with 900GB/s bandwidth between any two superchips, making the most use of the powerful computing chips and extended GPU memory.



- ✓ 72 Arm Neoverse V2 Cores
- ✓ Up to 480GB CPU LPDDR5X memory with ECC
- ✓ 96GB HBM3 or 144GB HBM3e GPU memory
- ✓ Up to 4x PCIe Gen5 x16 links
- ✓ NVLink-C2C Technology
- ✓ InfiniBand Networking Systems
- ✓ Easy-to-Program Heterogeneous Platform
- ✓ NVIDIA CUDA Platform

Maximize Configuration Flexibility with the GIGABYTE Server Lineup

Drawing from diverse product line experience, GIGABYTE provides various options supporting the NVIDIA Grace CPU & GH200 Grace Hopper Superchips in different form factors, aiming for multiple target applications.

For the highest possible computing density, GIGABYTE designed the H263 and H223 series high-density servers to achieve the best computing power in a single rack, in either 2U 2-Node or 2U 4-Node configurations and with both air cooling and direct liquid cooling solutions.

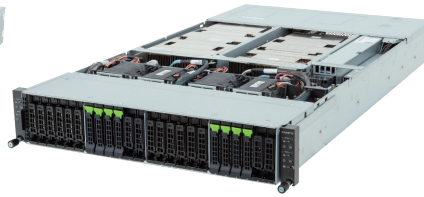
For total NVIDIA package solutions, GIGABYTE also provides a series of X-series servers, which is based on the NVIDIA MGX Platform with a modularized design and support for multiple add-in cards. These servers ensure compatibility between different rack standards, flexible cluster configurations, and compatibility with NVIDIA software and NVIDIA-defined configurations.

GIGABYTE Servers for NVIDIA Grace CPU/Hopper SuperChip

XH23-VG0-AAJ1

H223-V10-AAW1

H263-V60-AAW1

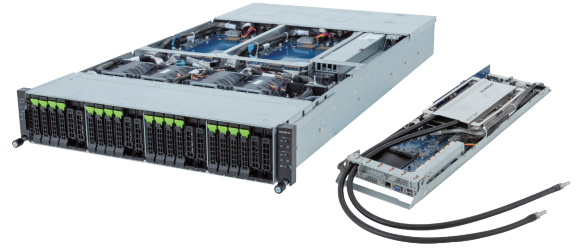
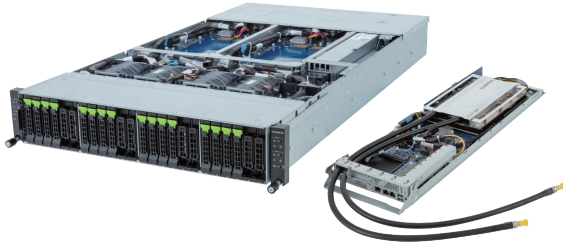


Form Factor	2U MGX server system (W438 x H88 x D900 mm)	2U 2-node rear access server system (W440 x H87.5 x D850 mm)	2U 4-node rear access server system (W440 x H87.5 x D850 mm)
Motherboard	MVG3-MG0	MV13-HD0	MV63-HD0
Superchip	NVIDIA Grace Hopper Superchip: - 1 x Grace CPU - 1 x Hopper H100 GPU - Connected with NVLink-C2C - TDP up to 1000W	NVIDIA Grace Hopper Superchip: - 1 x Grace CPU - 1 x Hopper H100 GPU - Connected with NVLink-C2C - TDP up to 1000W	NVIDIA Grace CPU Superchip: - 2 x Grace CPUs - Connected with NVLink-C2C - TDP up to 600W
Memory	Grace CPU: - Up to 480GB of LPDDR5X memory - Memory bandwidth up to 512GB/s Hopper H100 GPU: - Up to 96GB HBM3 or 144GB HBM3e - Memory bandwidth up to 4 or 4.9TB/s	Grace CPU: - Up to 480GB of LPDDR5X memory - Memory bandwidth up to 512GB/s Hopper H100 GPU: - Up to 96GB HBM3 or 144GB HBM3e - Memory bandwidth up to 4 or 4.9TB/s	Up to 960GB of LPDDR5X memory Memory bandwidth up to 1TB/s
Networking	2 x 10GbE LAN 1 x MLAN	4 x 10GbE LAN 2 x MLAN 1 x CMC	4 x MLAN 1 x CMC
Storage Bays	4 x 2.5" Gen5 NVMe hot-swap bays	8 x 2.5" Gen5 NVMe hot-swap bays	16 x 2.5" Gen5 NVMe hot-swap bays
Expansion Slots	3 x FHFL PCIe 5.0 x16 slots 2 x M.2 slots (PCIe 5.0 x4)	2 x FHFL dual-slot PCIe 5.0 x16 slots 2 x FHFL single-slot PCIe 5.0 x16 slots 2 x OCP 3.0 Gen5 x16 slots 4 x M.2 slots (PCIe 5.0 x4)	8 x LP PCIe 5.0 x16 slots 4 x OCP 3.0 Gen5 x16 slots 8 x M.2 slots (PCIe 5.0 x4)
I/O Ports	2 x USB 3.2 Gen1 1 x Mini-DP 2 x RJ45 1 x MLAN	4 x USB 3.2 Gen1 2 x Mini-DP 4 x RJ45 2 x MLAN 1 x CMC port	8 x USB 3.2 Gen1 4 x VGA 4 x MLAN 1 x CMC port
Security	Optional TPM2.0 kit: CTM012	Optional TPM2.0 kit: CTM012	Optional TPM2.0 kit: CTM012
Power Supply	Dual redundant power supplies 2000W 80 PLUS Titanium	2+1 redundant power supplies 3000W 80 PLUS Titanium	2+1 redundant power supplies 3000W 80 PLUS Titanium
System Management	Aspeed AST2600 BMC GIGABYTE Management Console	Aspeed AST2600 BMC GIGABYTE Management Console	Aspeed AST2600 BMC GIGABYTE Management Console
Other Feature	Compatible with BlueField-3 DPUs	Compatible with BlueField-3 DPUs	Compatible with BlueField-3 DPUs

GIGABYTE Direct Liquid Cooling Solution

H263-V11-LAW1

H263-V60-LAW1



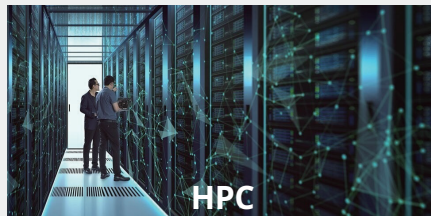
Form Factor	2U 4-Node (W440 x H87.5 x D850 mm)	2U 4-Node (W440 x H87.5 x D850 mm)
Superchip	NVIDIA Grace Hopper Superchip	NVIDIA Grace CPU Superchip
Memory	Up to 480GB CPU LPDDR5X ECC memory Up to 96GB HBM3 or 144GB HBM3e per module	Up to 960GB LPDDR5X ECC memory
Networking	8 x 10GbE LAN, 4 x MLAN, 1 x CMC	4 x MLAN, 1 x CMC
Storage Bays	16 x 2.5" Gen5 NVMe hot-swap bays	16 x 2.5" Gen5 NVMe hot-swap bays
Expansion Slots	4 x FHHL PCIe 5.0 x16 slots - Compatible with NVIDIA BlueField-3 DPUs 4 x OCP 3.0 Gen5 x16 slots 8 x M.2 slots (PCIe 5.0 x4)	4 x FHHL PCIe 5.0 x16 slots - Compatible with NVIDIA BlueField-3 DPUs 4 x OCP 3.0 Gen5 x16 slots 8 x M.2 slots (PCIe 5.0 x4)
Power Supply	Triple 3000W 80 PLUS Titanium PSUs	2+1 3000W 80 PLUS Titanium PSUs

Applications for the NVIDIA Grace CPU/Hopper SuperChip



AI

With the fast-growing adoption of AI, either for training massive language models or real-time responsive inference, the Arm-based NVIDIA Superchips benefit from the seamless communication between CPUs and GPUs and lower power consumptions on CPUs. Together with high-bandwidth chip-to-chip connections and coherent memory design for large AI model computations, these superchips fulfill the computational needs in modern AI applications.



HPC

As HPC developed throughout the years, the applications have gradually moved from traditional x86 platforms to the more power efficient Arm-based platform. Based on the existing Arm ecosystem, a range of HPC applications can be transferred to the new platform with ease. Along with packaged high-bandwidth low-power memory, the superchips deliver outstanding performance with low power consumption and encourage diverse platform choices for those seeking new solutions.



Cloud Computing

Benefiting from much higher core density and better core scalability, the NVIDIA Grace CPU Superchip offers outstanding performance, power efficiency, and system scalability in an era where the cloud has become an essential part of our daily lives. Providing low-latency and high scalability solutions to adapt to increasing needs in private and public cloud computing services.



Learn more about GIGABYTE server, visit <https://www.gigacomputing.com>

* All specifications are subject to change without notice. Please visit our website for the latest information.
* NVIDIA, the NVIDIA logo, Grace, Grace Hopper, Hopper, and NVLink are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated.

© 2024 Giga Computing Technology Co., Ltd. All rights reserved.

Designed by

