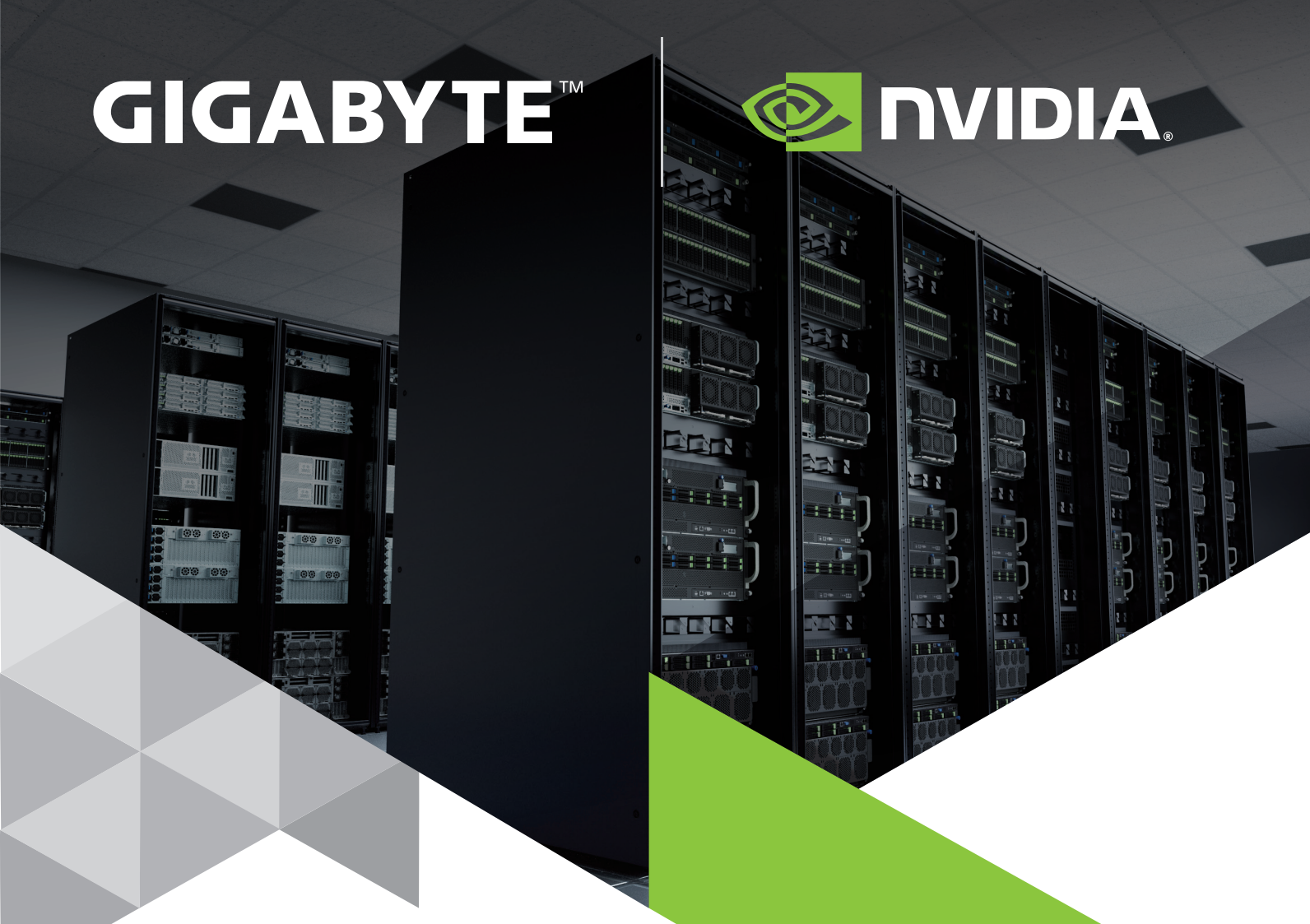


The background of the top half of the image is a photograph of a server room with rows of black server racks. The racks are filled with various electronic components, including circuit boards and fans. The lighting is dim, with some blue and green indicator lights visible. The Gigabyte logo is in the top left, and the NVIDIA logo is in the top right. A large green diagonal shape cuts across the bottom right of the image.

GIGABYTE™

 **nVIDIA®**

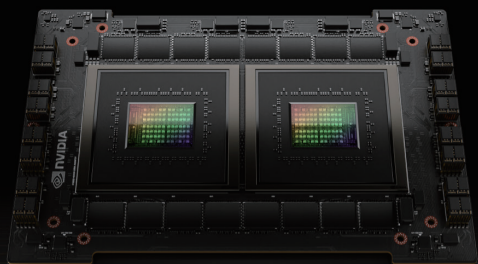
Revolutionizing Data Center Performance with **NVIDIA Grace™ CPU & GH200 Grace Hopper™ Superchip**

The breakthrough CPU accelerates data center
AI & HPC with an advanced superchip design.

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®-C2C, 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.

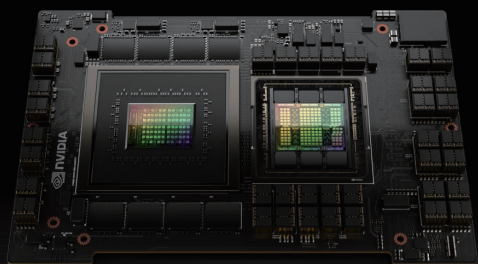


- ✓ 144 Arm Neoverse V2 Cores
- ✓ Up to 960GB CPU LPDDR5X memory with ECC
- ✓ Up to 8x PCIe Gen5 x16 links
- ✓ NVIDIA® NVLink®-C2C Technology
- ✓ NVIDIA Scalable Coherency Fabric (SCF)
- ✓ InfiniBand Networking Systems
- ✓ NVIDIA CUDA® Platform

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 Tensor Core 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 NVIDIA BlueField®-3 DPUs or SuperNIC, forming a system connected with a speed of 100GB/s for ML and HPC workloads.



- ✓ 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
- ✓ NVIDIA® 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.

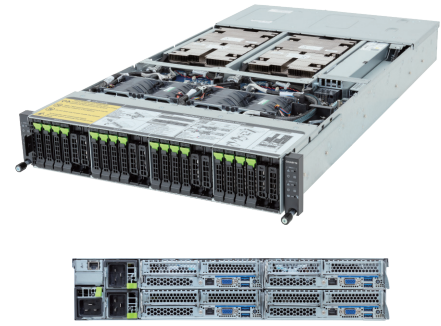
For total NVIDIA package solutions, GIGABYTE also provides a series of X-series servers, which is based on the NVIDIA MGX™ Platform with a modularized design and support for multiple add-in cards. These servers ensure compatibility between different rack standards, flexible cluster configurations, and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

GIGABYTE Servers for NVIDIA Grace™ CPU Superchip

XV23-VC0-AAJ1



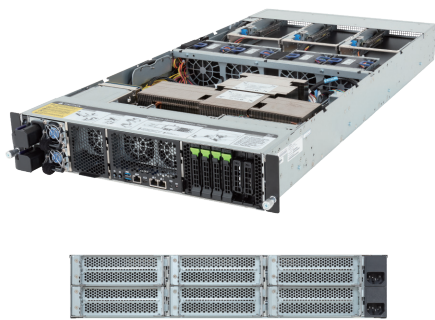
H263-V60-AAW1



Form Factor	2U MGX server system (W438 x H87 x D900 mm)	2U 4-node rear access server system (W440 x H87.5 x D850 mm)
Motherboard	MVC3-MG0	MV63-HD0
Superchip	NVIDIA Grace™ CPU Superchip: - 2 x NVIDIA Grace™ CPUs - Connected with NVIDIA® NVLink®-C2C - TDP up to 500W (CPU + memory)	NVIDIA Grace™ CPU Superchip: - 2 x NVIDIA Grace™ CPUs - Connected with NVIDIA® NVLink®-C2C - TDP up to 500W (CPU + memory)
Memory	480GB of LPDDR5X memory with ECC Memory bandwidth up to 1TB/s * Modules with 960GB of CPU memory and 768GB/s memory bandwidth are also available.	480GB of LPDDR5X memory with ECC Memory bandwidth up to 1TB/s * Modules with 960GB of CPU memory and 768GB/s memory bandwidth are also available.
Networking	1 x 10/100/1000 Mbps Management LAN	4 x 10/100/1000 Mbps Management LAN 1 x CMC Management LAN
Storage Bays	2 x 2.5" Gen5 NVMe hot-swap bays 2 x M.2 slots (PCIe 5.0 x4)	16 x 2.5" Gen5 NVMe hot-swap bays 8 x M.2 slots (PCIe 5.0 x4)
Expansion Slots	4 x FHFL PCIe Gen5 x16 slots for GPUs 2 x FHFL PCIe Gen5 x16 slots for add-in cards - Compatible with NVIDIA BlueField®-3 DPUs	8 x LP PCIe Gen5 x16 slots - Compatible with NVIDIA BlueField®-3 DPUs 4 x OCP 3.0 Gen5 x16 slots
I/O Ports	2 x USB 3.2 Gen1 1 x Mini-DP 1 x MLAN	8 x USB 3.2 Gen1 4 x VGA 4 x MLAN 1 x CMC
Security	Optional TPM2.0 kit: CTM012	Optional TPM2.0 kit: CTM012
Power Supply	2+2 redundant power supplies 2000W 80 PLUS Titanium	2+1 redundant power supplies 3000W 80 PLUS Titanium
System Management	ASPEED® AST2600 BMC GIGABYTE Management Console	ASPEED® AST2600 BMC ASPEED® AST2520 CMC GIGABYTE Management Console
Other Feature	Supports up to 4 x dual-slot Gen5 GPUs	-

GIGABYTE Servers for NVIDIA Grace™ Hopper Superchip

XH23-VG0-AAJ1



H223-V10-AAW1

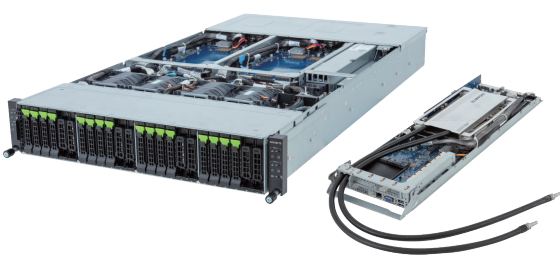
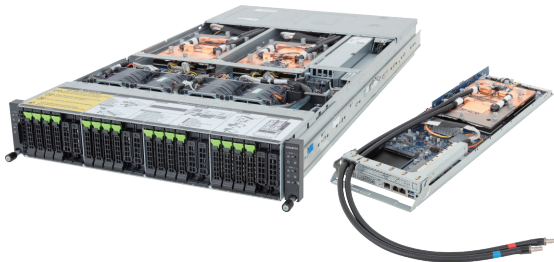


Form Factor	2U MGX server system (W438 x H87 x D900 mm)	2U 2-node rear access server system (W440 x H87.5 x D850 mm)
Motherboard	MVG3-MG0	MV13-HD0
Superchip	NVIDIA GH200 Grace Hopper Superchip: - 1 x NVIDIA Grace™ CPU - 1 x NVIDIA Hopper™ GPU - Connected with NVIDIA® NVLink®-C2C - TDP up to 1000W (CPU + GPU + memory)	NVIDIA GH200 Grace Hopper Superchip: - 1 x NVIDIA Grace™ CPU - 1 x NVIDIA Hopper™ GPU - Connected with NVIDIA® NVLink®-C2C - TDP up to 1000W (CPU + GPU + memory)
Memory	NVIDIA Grace CPU: - 480GB of LPDDR5X memory with ECC - Memory bandwidth up to 512GB/s NVIDIA Hopper GPU: - 96GB HBM3 - Memory bandwidth up to 4TB/s * Modules with 144GB of HBM3e GPU memory and 4.9TB/s memory bandwidth are also available.	NVIDIA Grace CPU: - 480GB of LPDDR5X memory with ECC - Memory bandwidth up to 512GB/s NVIDIA Hopper GPU: - 96GB HBM3 - Memory bandwidth up to 4TB/s * Modules with 144GB of HBM3e GPU memory and 4.9TB/s memory bandwidth are also available.
Networking	2 x 10Gb/s LAN (Support NCSI function) 1 x 10/100/1000 Mbps Management LAN	4 x 10Gb/s LAN (Support NCSI function) 1 x 10/100/1000 Mbps Management LAN 1 x CMC
Storage Bays	4 x 2.5" Gen5 NVMe bays 2 x M.2 slots (PCIe 5.0 x4)	8 x 2.5" Gen5 NVMe hot-swap bays 4 x M.2 slots (PCIe 5.0 x4) * Optional 4 x 2.5" NVMe via NVIDIA BlueField®-3 DPUs
Expansion Slots	3 x FHFL PCIe Gen5 x16 slots - Compatible with NVIDIA BlueField®-3 DPUs	2 x FHHL dual-slot PCIe Gen5 x16 slots 2 x FHHL single-slot PCIe Gen5 x16 slots 2 x OCP 3.0 Gen5 x16 slots
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
Security	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
System Management	ASPEED® AST2600 BMC GIGABYTE Management Console	ASPEED® AST2600 BMC ASPEED® AST2520 CMC GIGABYTE Management Console

GIGABYTE Direct Liquid Cooling Solution

H263-V11-LAW1

H263-V60-LAW1

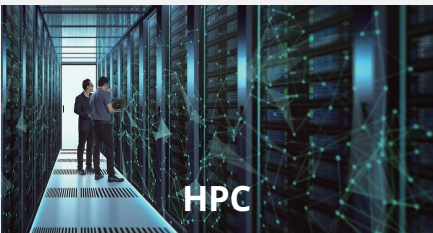


Form Factor	2U 4-node rear access server system (W440 x H87.5 x D850 mm)	2U 4-node rear access server system (W440 x H87.5 x D850 mm)
Superchip	NVIDIA GH200 Grace Hopper Superchip	NVIDIA Grace™ CPU Superchip
Memory	480GB CPU LPDDR5X ECC memory per module Up to 144GB GPU HBM3e per module	Up to 960GB LPDDR5X ECC memory per module
Networking	8 x 10GbE LAN ports (Support NCSI function) 4 x 10/100/1000 Mbps Management LAN 1 x CMC Management LAN	4 x 10/100/1000 Mbps Management LAN 1 x CMC Management LAN
Storage Bays	16 x 2.5" Gen5 NVMe hot-swap bays 8 x M.2 slots (PCIe 5.0 x4) * Optional 8 x 2.5" NVMe via NVIDIA BlueField®-3 DPUs	16 x 2.5" Gen5 NVMe hot-swap bays 8 x M.2 slots (PCIe 5.0 x4)
Expansion Slots	4 x FHHL PCIe Gen5 x16 slots 4 x OCP 3.0 Gen5 x16 slots	4 x FHHL PCIe 5.0 x16 slots - Compatible with NVIDIA BlueField®-3 DPUs 4 x OCP 3.0 Gen5 x16 slots
Power Supply	Triple 3000W 80+ Titanium redundant PSUs	2+1 3000W 80+ Titanium redundant PSUs

Applications for the NVIDIA Grace™ CPU/Hopper Superchip



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.



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.



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.



GigaComputing.com/en



Giga Computing



GigaComputing



GIGABYTEServer



GigaComputing



Designed by

**GIGA
COMPUTING**

* All intellectual property rights, including without limitation to copyright and trademark of this work and its derivative works are the property of, or are licensed to, Giga Computing Technology Co., Ltd. Any unauthorized use is strictly prohibited.

* The entire materials provided herein are for reference only. Giga Computing reserves the right to modify or revise the content at anytime without prior notice.

* 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.