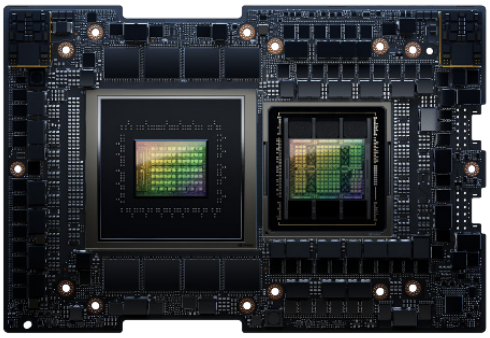
Quanta GH200

Breakthrough accelerated performance for giant-scale AI-HPC applications

- **Introducing the first gen NVIDIA® MGX™ architecture with modular infrastructure**
- **Powered by NVIDIA® Grace™ Hopper™ Superchip**
- **Coherent memory between CPU and GPU with NVLink®- C2C interconnect**
- **Optimized for memory intensive inference and HPC performance**

Quanta/QCT is the leading partner with NVIDIA in introducing the MGX architecture - an open and future compatible accelerated computing reference architecture designed to allow rapid adoption of key platform technologies including CPUs, GPUs and DPUs. The modular architecture consists of configurable bays that can house different modules to achieve desired configurations. This allows for future hardware solutions with multiple power distribution methods, cooling solutions, including hot or cold aisle configurations.



Front View

Rear View

Quanta GH200 is the first server to introduce the NVIDIA® Grace™ Hopper™ Superchip in conjunction with NVIDIA® MGX™ architecture. The Superchip combines 72 Arm Neoverse cores connected by NVLink® chip to chip high bandwidth interconnect with the Hopper™ H100 GPU to deliver a coherent memory pool that excels at accelerating AI and high performance computing applications. The modular infrastructure is designed to support multiple system configurations and reduce time to market while providing a compatible platform for future CPU, GPU and DPU solutions.

About QCT

Quanta Cloud Technology (QCT) is a global datacenter solution provider. We combine the efficiency of hyperscale hardware with infrastructure software from a diversity of industry leaders to solve next-generation datacenter design and operation challenges. QCT serves cloud service providers, telecoms and enterprises running public, hybrid and private clouds.

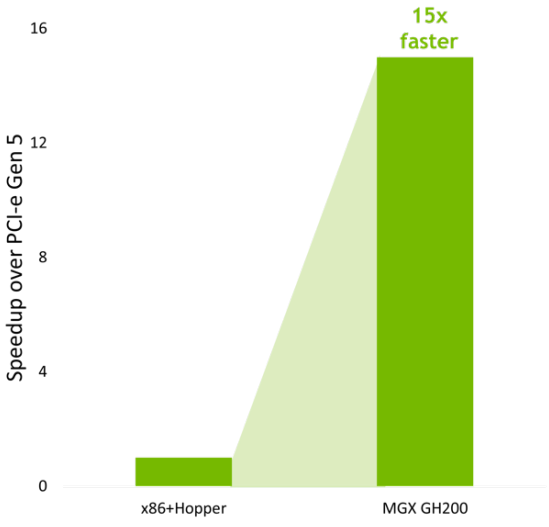
Product lines include hyper-converged and software-defined datacenter solutions as well as servers, storage, switches, integrated racks with a diverse ecosystem of hardware component and software partners. QCT designs, manufactures, integrates and services cutting edge offerings via its own global network. The parent of QCT is Quanta Computer, Inc., a Fortune Global 500 corporation.

Quanta GH200 Specifications

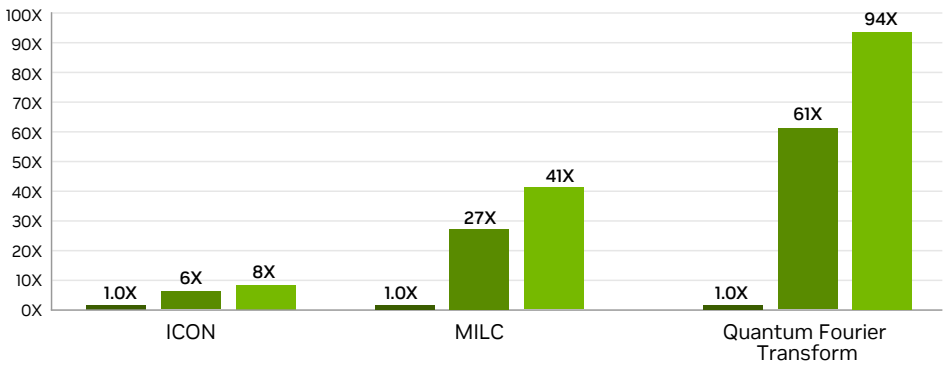
Processor	**Processor Family:** NVIDIA GH200 Grace™ Hopper™ Superchip **Processor Type:** NVIDIA Grace™ 72 Arm® Neoverse V2 cores **Max. TDP Support:** 1000W **Number of Processors:** (1) Processors **Internal Interconnect:** NVIDIA® NVLink®-C2C 900GB/s
Form Factor	2U Rackmount
Dimensions	**W x H x D (inch):** 17.24" x 3.44" x 35.43" **W x H x D (mm):** 438 x 87.5 x 900mm
Storage	960GB PCIe 1.92TB NVMe SAMSUNG
Memory	480GB LPDDR5X 96GB HBM3 GPU memory
Expansion Slot	**Default Configuration:** (3) PCIe 5.0 x16 FHFL Dual Width slots
Front I/O	Power/ID/Reset Buttons Power/ID/Status LEDs (2) USB 3.0 ports (1) VGA port
Storage Controller	Broadcom 9500-16i HBA PCIe x8
Network	Mellanox Infiniband MCX755106AS-HEAT PCIe x16 NDR Dual Port QSFP112
Power Supply	1+1 High efficiency hot-plug 2000W PSU, 80 Plus Titanium
Onboard Storage	(2) 22110/2280 PCIe M.2
Fan	(5) 6056 dual rotor fans (N+1 redundant)
Rear I/O	(1) USB 3.0 (1) Mini display port (1) ID LED (1) PWR Button/PWR LED (1) COM Port (micro USB type-B) (1) RJ45 mgmt port

Operating Environment	Operating temperature: 5°C to 35°C (41°F to 95°F) Non-operating temperature: -40°C to 70°C Operating relative humidity: 20% to 85%RH Non-operating relative humidity: 10% to 95%RH
TPM	TPM 2.0 SPI module (optional)

NVIDIA GH200 NVL32 is 8 time faster for recommender training



GH200 HPC Performance



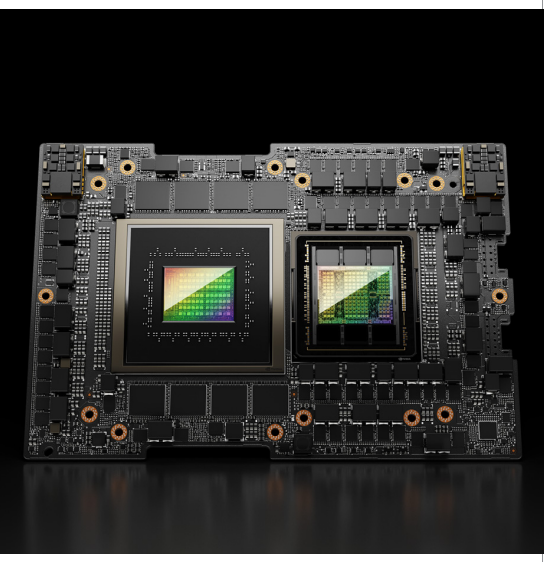
NVIDIA

Quanta GH200

Ideal for strong-scaling giant AI LLM and Recommender Workloads

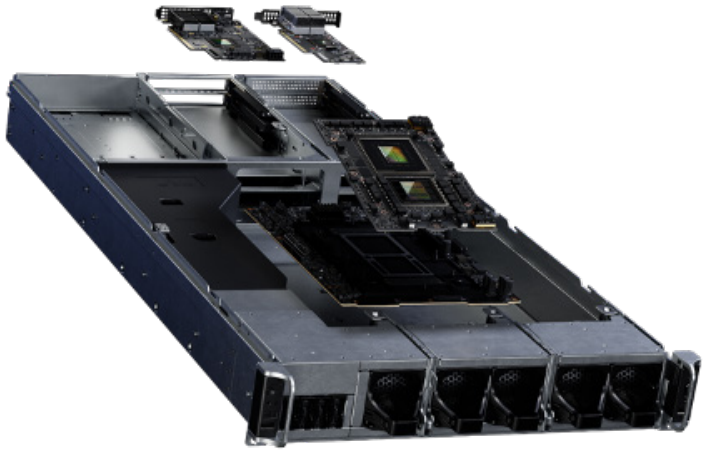
Key Features

- NVIDIA Hopper GPU
- Supports up to 96GB of HBM3 or 144GB of HBM3e
- 72-core NVIDIA Grace CPU
- Up to 480GB of LPDDR5X memory with error-correction code (ECC)
- Up to 624GB of fast-access memory
- NVLink-C2C: 900GB/s of coherent memory



NVIDIA GH200 and InfiniBand or Ethernet

Ideal for Scale-out Machine Learning and HPC Workloads



NVIDIA GH200 Grace Hopper Superchip

The breakthrough processor for large-scale AI and high-performance computing (HPC) applications.

The World's Most Versatile Computing Platform

The NVIDIA Grace Hopper™ architecture brings together the groundbreaking performance of the NVIDIA Hopper™ GPU with the versatility of the NVIDIA Grace™ CPU in a single superchip, connected with the high-bandwidth, memory-coherent NVIDIA® NVLink® Chip-2-Chip (C2C) interconnect.

NVIDIA NVLink-C2C is a memory-coherent, high-bandwidth, and low-latency interconnect for superchips. The heart of the GH200 Grace Hopper Superchip, it delivers up to 900 gigabytes per second (GB/s) of total bandwidth, which is 7X higher than PCIe Gen5 lanes commonly used in accelerated systems. NVLink-C2C memory coherency increases developer productivity, performance, and the amount of GPU-accessible memory. GH200 can be easily deployed in standard servers to run a variety of inference, data analytics, and other compute- and memory-intensive workloads.