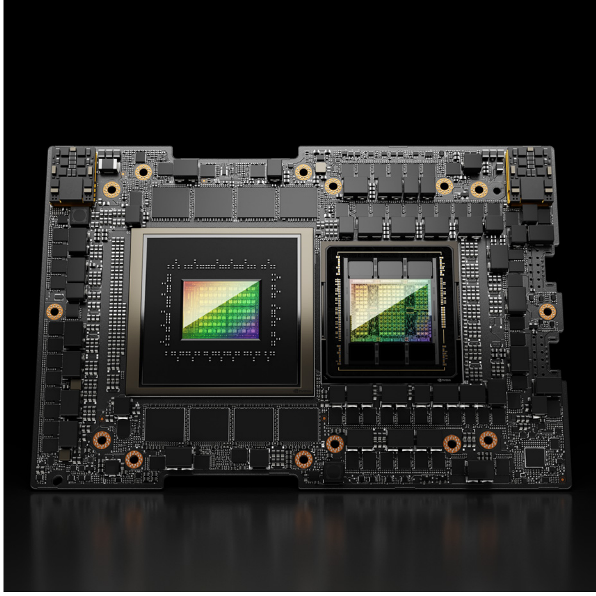
Datasheet

NVIDIA GH200 Grace Hopper Superchip

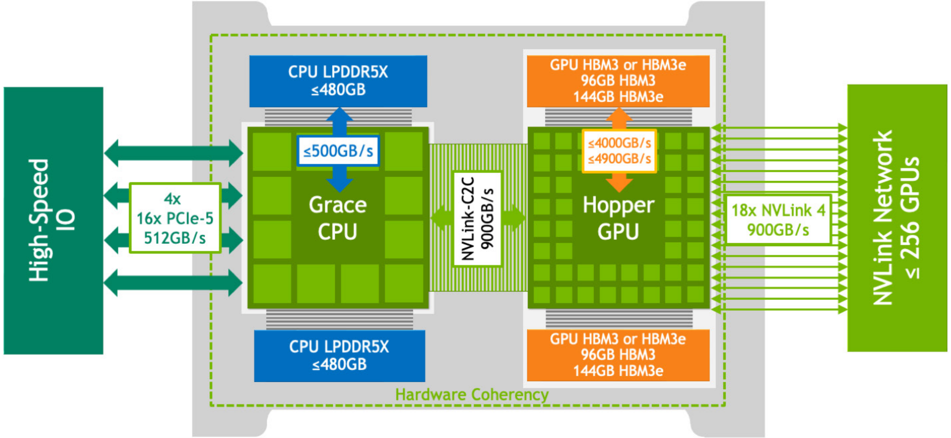
The breakthrough processor for large-scale AI and high-performance computing (HPC) applications.



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.



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

Power and Efficiency With the Grace CPU

The NVIDIA Grace CPU delivers 2X the performance per watt of conventional x86 platforms. The Grace CPU was designed for high single-threaded performance, high-memory bandwidth, and outstanding data-movement capabilities. The NVIDIA Grace CPU combines 72 Neoverse V2 Armv9 cores with up to 480GB of server-class LPDDR5X memory with ECC. The wide memory subsystem delivers up to 500GB/s of bandwidth at one-fifth the power of traditional DDR memory at similar cost.

Performance and Speed With the Hopper H100 GPU

The H100 Tensor Core GPU is NVIDIA's ninth-generation data center GPU, and it delivers an order-of-magnitude performance leap for large-scale AI and HPC over the prior-generation NVIDIA A100 Tensor Core GPU. The NVIDIA H100 based on the new Hopper GPU architecture features multiple innovations:

- New fourth-generation Tensor Cores perform faster matrix computations than ever before on an even broader array of AI and HPC tasks.
- A new Transformer Engine enables H100 to deliver up to 9X faster AI training and up to 30X faster AI inference compared to the prior GPU generation.
- Secure Multi-Instance GPU (MIG) partitions the GPU into isolated, right-size instances to maximize quality of service (QoS) for smaller workloads.

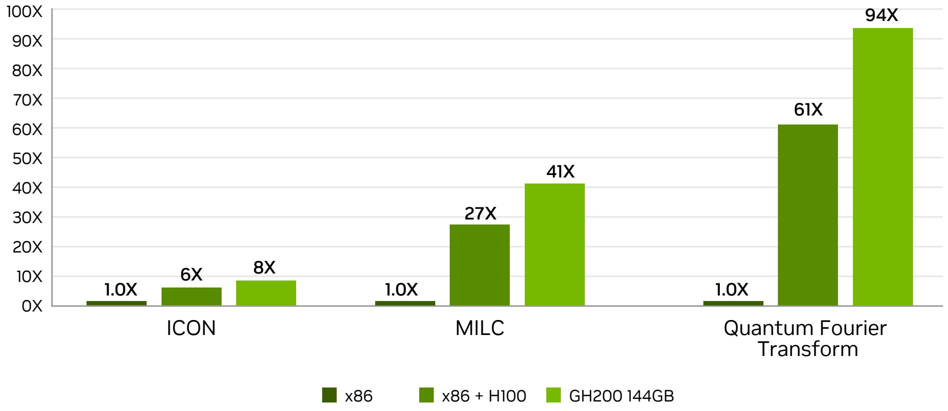
The Power of Coherent Memory

NVLink-C2C memory coherency increases developer productivity, performance, and the amount of GPU-accessible memory. CPU and GPU threads can concurrently and transparently access both CPU and GPU resident memory, allowing developers to focus on algorithms instead of explicit memory management. Memory coherency lets developers only transfer the data they need and not migrate entire pages to and from the GPU. It also provides lightweight synchronization primitives across GPU and CPU threads by enabling native atomics from both the CPU and GPU. Fourth-generation NVLink allows accessing peer memory with direct loads, stores, and atomic operations, so accelerated applications can solve larger problems more easily than ever.

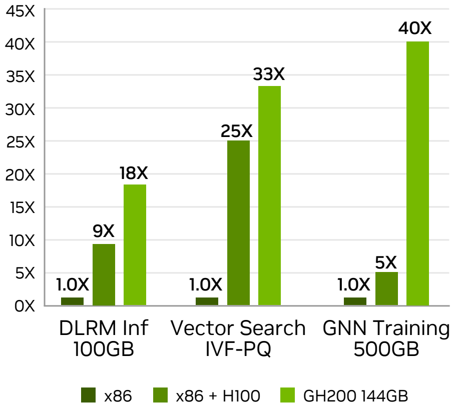
Class-Leading Performance for HPC and AI Workloads

The GH200 Grace Hopper Superchip is the first true heterogeneous accelerated platform for HPC and AI workloads. It accelerates any application with the strengths of both GPUs and CPUs while providing the simplest and most productive heterogeneous programming model to date, enabling scientists and engineers to focus on solving the world's most important problems. For AI inference workloads, GH200 Grace Hopper Superchips combine with NVIDIA networking technologies to provide the best TCO for scale-out solutions, letting customers take on larger datasets, more complex models, and new workloads using up to 624GB of fast-access memory. The NVIDIA GH200 also comes in a GH200 NVL2 configuration with two Grace Hopper Superchips fully connected by NVLink to deliver 288GB of HBM3e and 1.2TB of fast memory for both compute- and memory-intensive workloads.

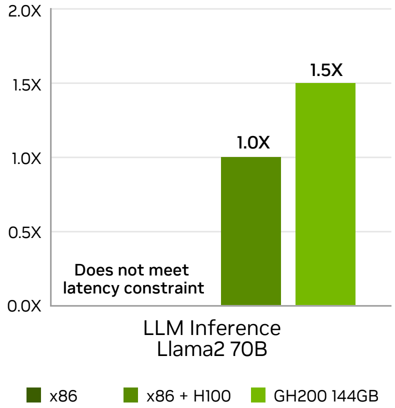
GH200 HPC Performance



GH200 AI Performance



GH200 LLM Performance



* Comparison: 2S Xeon Platinum 8480+, Xeon Platinum 8480+ and H100 Tensor Core GPU, and NVIDIA GH200 144GB: ICON (QUBICC 191 levels 80km radiation), MILC (APEX Medium), Quantum Fourrier Transform, DLRM 100GB inference, vector search (batch size = 10,000 | queries = 10,000 over 85M vectors IVF-PQ) GNN (GraphSage OGB-100M papers dataset), Llama2 70B (batch size = 64 (GH200), 96 (2x H100s) | precision = FP8 | TensorRT-LLM - throughput per GPU).

Results subject to change.

Full NVIDIA Platform Support

The NVIDIA GH200 Grace Hopper Superchip extends the existing large and diverse ecosystem of 64-bit Arm processors. The very same containers, application binaries, and operating systems that run on other Arm products run on Grace Hopper without modification—only faster. And for customers who wish to leverage and build upon NVIDIA's software expertise, the NVIDIA Grace Hopper Superchip is supported by the full NVIDIA software stack, including the NVIDIA HPC and NVIDIA AI platforms.

Product Specifications

Feature	GH200	GH200 NVL2
CPU core count	72 Arm Neoverse V2 cores	144 Arm Neoverse V2 cores
L1 cache	64KB i-cache + 64KB d-cache	
L2 cache	1MB per core	
L3 cache	114MB	228MB
Base frequency | all-core single instruction, multiple data (SIMD) frequency	3.1GHz | 3.0GHz	
LPDDR5X size	480GB 120GB, 240GB	960GB 240GB, 480GB
Memory bandwidth	Up to 384GB/s Up to 512GB/s	Up to 768GB/s Up to 1024GB/s
PCIe links	Up to 4x PCIe x16 (Gen5)	Up to 8x PCIe x16 (Gen5)

Feature	GH200	GH200 NVL2
FP64	34 teraFLOPS	68 teraFLOPS
FP64 Tensor Core	67 teraFLOPS	134 teraFLOPS
FP32	67 teraFLOPS	134 teraFLOPS
TF32 Tensor Core	989 teraFLOPS* | 494 teraFLOPS	1,979 teraFLOPS* | 990 teraFLOPS
BFLOAT16 Tensor Core	1,979 teraFLOPS* | 990 teraFLOPS	3,958 teraFLOPS* | 1,979 teraFLOPS
FP16 Tensor Core	1,979 teraFLOPS* | 990 teraFLOPS	3,958 teraFLOPS* | 1,979 teraFLOPS
FP8 Tensor Core	3,958 teraFLOPS* | 1,979 teraFLOPS	7,916 teraFLOPS* | 3,958 teraFLOPS
INT8 Tensor Core	3,958 TOPS* | 1,979 TOPS	7,916 teraFLOPS* | 3,958 teraFLOPS
High-bandwidth memory (HBM) size	96GB HBM3 | 144GB HBM3e	Up to 288GB HBM3e
Memory bandwidth	Up to 4TB/s | Up to 4.9TB/s	Up to 9.8TB/s
NVIDIA NVLink-C2C CPU-to-GPU bandwidth	900GB/s	1800GB/s
Power	Configurable 450 to 1000W (Memory + CPU + GPU)	Configurable 900W to 2000W (Memory + CPU + GPU)
Thermal solution	Air-cooled or liquid-cooled	

* With sparsity

Ready to Get Started?

To learn more about the NVIDIA Grace Hopper Superchip, visit nvidia.com/grace-hopper-superchip/

To download the Grace Hopper architecture whitepaper, visit resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

© 2025 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, ConnectX, DGX, HGX, Hopper, NGC, NVIDIA-Certified Systems, 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. 3773000. MAR25.