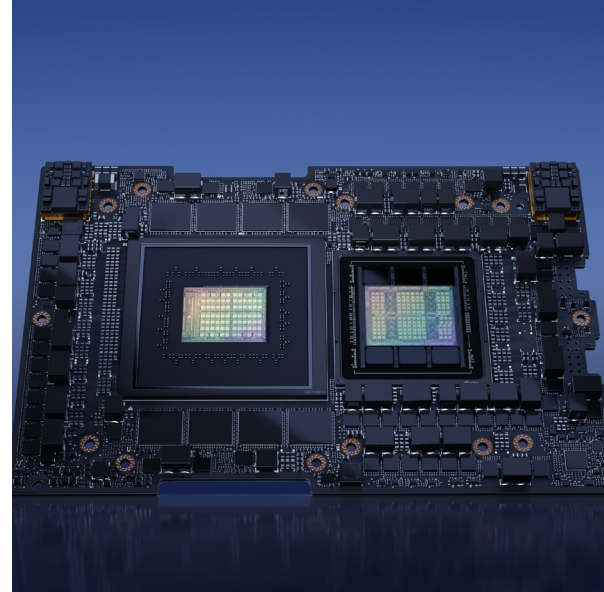




# NVIDIA GH200 Grace Hopper Superchip

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



## The World's Most Versatile Computing Platform

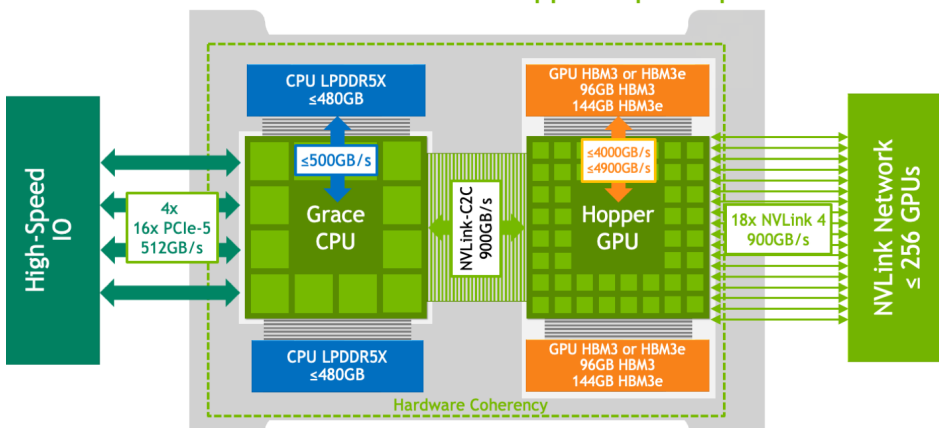
The NVIDIA Grace Hopper™ architecture brings together the groundbreaking performance of the NVIDIA Hopper™ GPU with the versatility of the NVIDIA Grace™ CPU in a single superchip, connected with the high-bandwidth, memory-coherent NVIDIA® NVLink® Chip-2-Chip (C2C) interconnect.

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 enables applications to oversubscribe the GPU's memory and directly utilize NVIDIA Grace CPU's memory at high bandwidth. With up to 480GB of LPDDR5X CPU memory per GH200 Grace Hopper Superchip, the GPU has direct access to 7X more fast memory than HMB3 or almost 8X more fast memory with HBM3e. GH200 can be easily deployed in standard servers to run a variety of inference, data analytics, and other compute and memory-intensive workloads. GH200 can also be combined with the NVIDIA NVLink Switch System, with all GPU threads running on up to 256 NVLink-connected GPUs and able to access up to 144 terabytes (TB) of memory at high bandwidth.

## Key Features

- > 72-core NVIDIA Grace CPU
- > NVIDIA H100 Tensor Core GPU
- > Up to 480GB of LPDDR5X memory with error-correction code (ECC)
- > Supports 96GB of HBM3 or 144GB of HBM3e
- > Up to 624GB of fast-access memory
- > NVLink-C2C: 900GB/s of coherent memory

### NVIDIA GH200 Grace Hopper Superchip



## Power and Efficiency With the Grace CPU

The NVIDIA Grace CPU delivers 2X the performance per watt of conventional x86-64 platforms and is the world's fastest Arm® data center CPU. 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. This design strikes the optimal balance of bandwidth, energy efficiency, capacity, and cost. Compared to an eight-channel DDR5 design, the Grace CPU LPDDR5X memory subsystem provides up to 53 percent more bandwidth at one-eighth the power per gigabyte per second.

## 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.

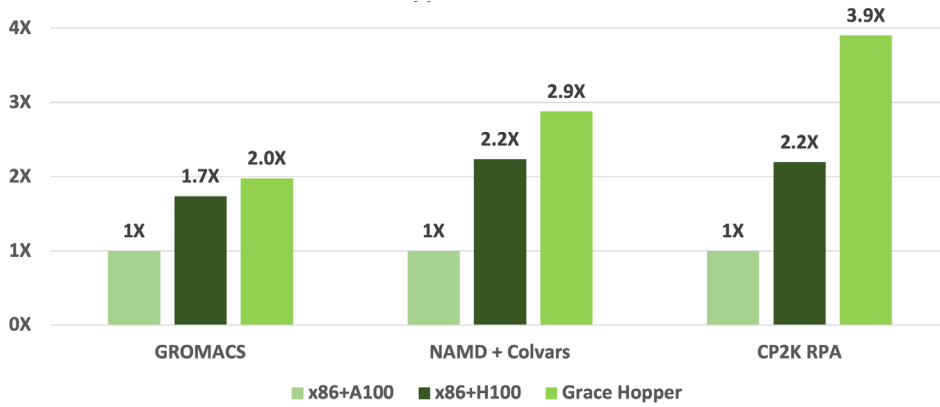
## The Power of Coherent Memory

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.

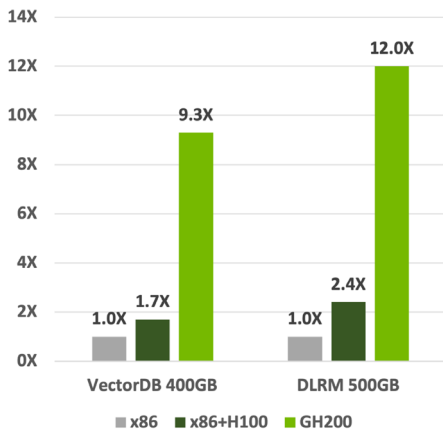
## Class-Leading Performance for HPC and AI Workloads

The GH200 Grace Hopper Superchip is the first true heterogeneous accelerated platform for HPC 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. For AI training, up to 256 NVLink-connected GPUs can access up to 144TB of memory at high bandwidth for large language model (LLM) or recommender system training.

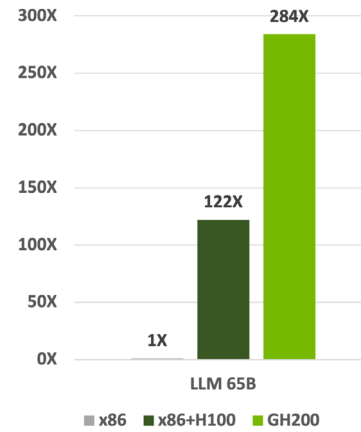
## Grace Hopper HPC Performance



## Relative AI Performance



## Relative AI Performance



## 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, NVIDIA AI, and NVIDIA Omniverse™ platforms.

## Product Specifications

Grace CPU	Feature
<b>CPU core count</b>	72 Arm Neoverse V2 cores
<b>L1 cache</b>	64KB i-cache + 64KB d-cache
<b>L2 cache</b>	1MB per core
<b>L3 cache</b>	117MB
<b>LPDDR5X size</b>	Up to 480GB
<b>Memory bandwidth</b>	Up to 512GB/s
<b>PCIe links</b>	Up to 4x PCIe x16 (Gen5)

Hopper H100 GPU	Feature
<b>FP64</b>	34 teraFLOPS
<b>FP64 Tensor Core</b>	67 teraFLOPS
<b>FP32</b>	67 teraFLOPS
<b>TF32 Tensor Core</b>	989 teraFLOPS*   494 teraFLOPS
<b>BFLOAT16 Tensor Core</b>	1,979 teraFLOPS*   990 teraFLOPS
<b>FP16 Tensor Core</b>	1,979 teraFLOPS*   990 teraFLOPS
<b>FP8 Tensor Core</b>	3,958 teraFLOPS*   1,979 teraFLOPS
<b>INT8 Tensor Core</b>	3,958 TOPS*   1,979 TOPS
<b>High-bandwidth memory (HBM) size</b>	Up to 96GB   144GB HBM3e
<b>Memory bandwidth</b>	Up to 4TB/s   Up to 4.9TB/s HBM3e
<b>NVIDIA NVLink-C2C CPU-to-GPU bandwidth</b>	900 GB/s bidirectional
<b>Module thermal design power (TDP)</b>	Programmable from 450W to 1000W (CPU + GPU + memory)
<b>Form factor</b>	Superchip module
<b>Thermal solution</b>	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/](https://nvidia.com/grace-hopper-superchip/)

To download the Grace Hopper architecture whitepaper, visit [resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper](https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper)

© 2023 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, Grace, Grace Hopper, Hopper, NVLink, and Omniverse 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. 2902113. AUG 23.

