DATASHEET

Vultr Cloud GPU Accelerated by NVIDIA GH200 Grace Hopper™ Superchip

The NVIDIA GH200 Grace Hopper™ Superchip's breakthrough CPU-GPU design enables the era of large-scale AI and high-performance computing (HPC) applications.

VULTR.COM

Vultr Cloud GPU Accelerated by NVIDIA GH200 Grace Hopper™ Superchip

Introduction

The NVIDIA GH200 Grace Hopper™ Superchip 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.

Why it's important right now

GPU acceleration has made the revolution in AI possible, increasing AI performance over one million times over the last decade. AI is now being used across every industry, with GPUs accelerating GenAI, deep learning recommendations, and scientific discoveries. In a quest for accuracy and generalizability, models and datasets have been increasing in size and complexity.

These massive models are pushing the limits of today's systems. Continuing to scale these models requires addressing shifting bottlenecks. This needs fast access to a large pool of memory and a tight coupling of the CPU and GPU.

The NVIDIA GH200 Grace Hopper™ Superchip accelerates large-scale AI and HPC workloads with the strengths of both GPUs and CPUs, enabling scientists and engineers to focus on solving the world's most important problems.

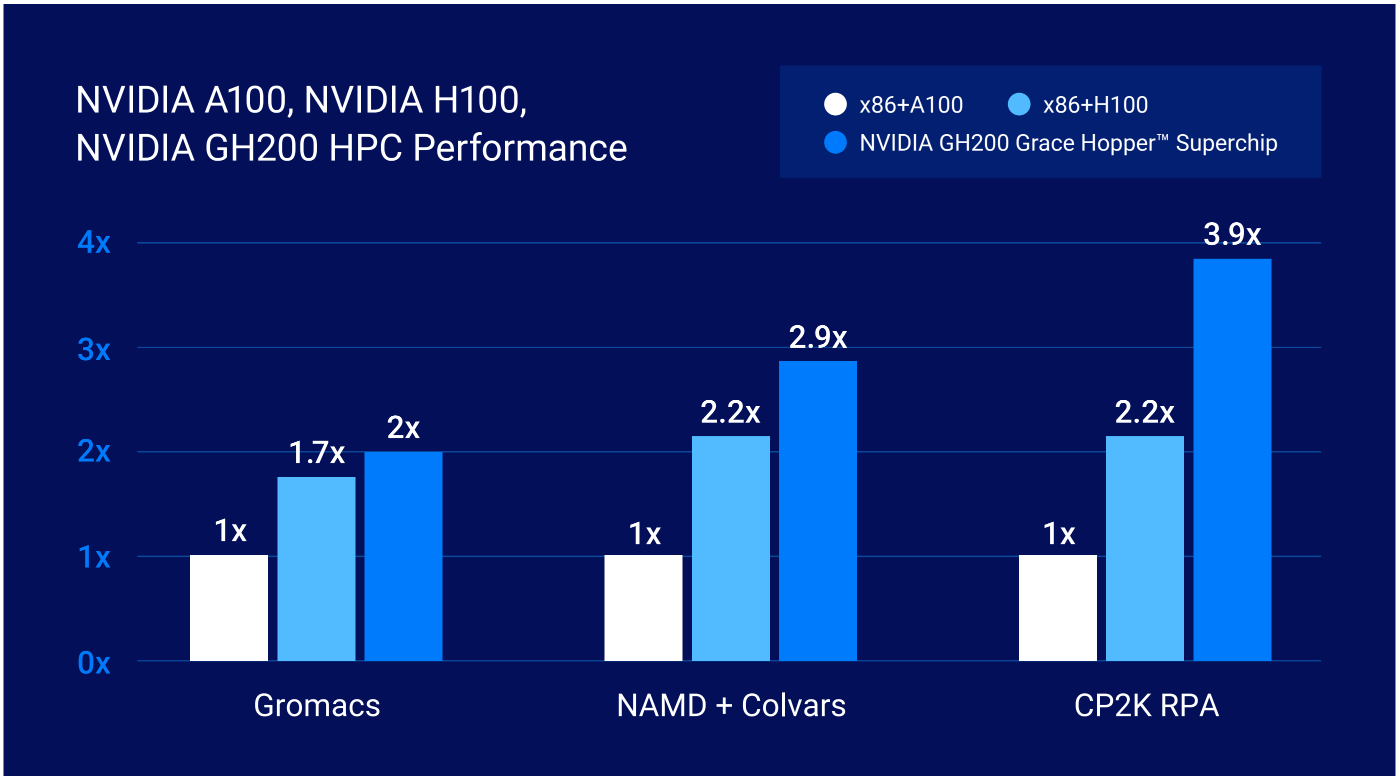
Why it's important right now

AI inference

As AI training models have gotten dramatically larger, so have the resulting AI Inference models. Transformers are among the most influential AI model architectures today and are shaping the direction for future R&D in AI. Invented first as a tool for natural language processing (NLP), they're now used for almost any AI task, including computer vision, automatic speech recognition, classification of molecule structures, and processing of financial data. For large transformer-based models, such as ChatGPT, NVIDIA GH200 Grace Hopper™ Superchip delivers 4x more inference performance compared to the prior generation.

High-performance computing (HPC)

HPC is a fundamental pillar of modern science. To unlock next-generation discoveries, scientists look to simulation to better understand complex molecules for drug discovery, physics for potential new sources of energy, and atmospheric data to better predict and prepare for extreme weather patterns. HPC is evolving toward a fusion of both AI and simulation, and a tight integration between the CPU and GPU is critical to delivering the performance needed to advance science.


NVIDIA A100, NVIDIA H100, NVIDIA GH200 HPC Performance

NVIDIA Grace™ CPU Specifications

GPU Memory	72 Arm Neoverse V2 cores
L1 cache	64KB i-cache + 64KB d-cache per core
L2 cache	1MB per core
L3 cache	117MB
LPDDR5X size	Up to 480GB
Memory bandwidth	Up to 512GB/s
PCIe links	Up to 4x PCIe x16 (Gen 5)
NVIDIA NVLink-C2C Chip-to-Chip bandwidth	900 GB/s bidirectional

NVIDIA Hopper™ H100 GPU Specifications

FP64	34 teraFLOPS
FP64 Tensor Core	67 teraFLOPS
FP32	67 teraFLOPS
TF32 Tensor Core	989 teraFLOPS* | 494 teraFLOPS
BFLOAT16 Tensor Core	1,979 teraFLOPS* | 990 teraFLOPS
FP16 Tensor Core	1,979 teraFLOPS* | 990 teraFLOPS
FP8 Tensor Core	3,958 teraFLOPS* | 1,979 teraFLOPS
INT8 Tensor Core	3,958 TOPS* | 1,979 TOPS
High-bandwidth memory (HBM) size	Up to 96GB
Memory bandwidth	Up to 4TB/s
NVIDIA NVLink-C2C Chip-to-Chip bandwidth	900 GB/s bidirectional

Learn more about Vultr Cloud GPU accelerated by NVIDIA GH200 Grace Hopper Superchip™

Contact us at vultr.com to get started. →