VULTR

DATASHEET

Vultr Cloud GPU: Accelerated by NVIDIA HGX™ B200

Propelling a new era of accelerating computing and generative AI, integrating NVIDIA Blackwell GPUs with high-speed interconnect to accelerate AI performance at scale

VULTR.COM

VULTR

Vultr Cloud GPU: Accelerated by NVIDIA HGX™ B200

Propelling a new era of accelerating computing and generative AI, integrating NVIDIA Blackwell GPUs with high-speed interconnect to accelerate AI performance at scale

Introduction

The NVIDIA HGX™ B200 is built to deliver a new era of accelerated computing and generative AI. Accelerated by NVIDIA Blackwell GPUs with high-speed interconnect, the NVIDIA HGX™ B200 delivers remarkable performance for the most demanding AI training, AI inference, and HPC workloads. Combined with Vultr's global infrastructure, featuring 32 global data center regions reaching 90% of the world's population within 2-40ms, the NVIDIA HGX™ B200 is available to deliver computing power with efficiency and scale.

Why it's important right now

As the adoption of generative AI continues apace, the need for advanced computing power remains unsatiated, with demand increasing as AI models continue to migrate from training to inference. Newer, larger large language models (LLMs) require a new paradigm to deliver AI's transformation efficiently and cost-effectively. Vultr Bare Metal and Vultr Cloud GPU, accelerated by NVIDIA HGX™ B200, address this demand by providing scalable, high-performance AI resources globally, enabling organizations to support the computational needs of LLMs without the complexity of managing infrastructure.

As sustainability concerns mount, the need to accommodate growing parameters and computational complexity contrasts with the need to reduce energy usage. The NVIDIA HGX™ B200 delivers both outstanding performance and incredible energy efficiency, cutting through burdensome constraints while adding more memory and accelerators as compared to NVIDIA Hopper-generation GPUs.

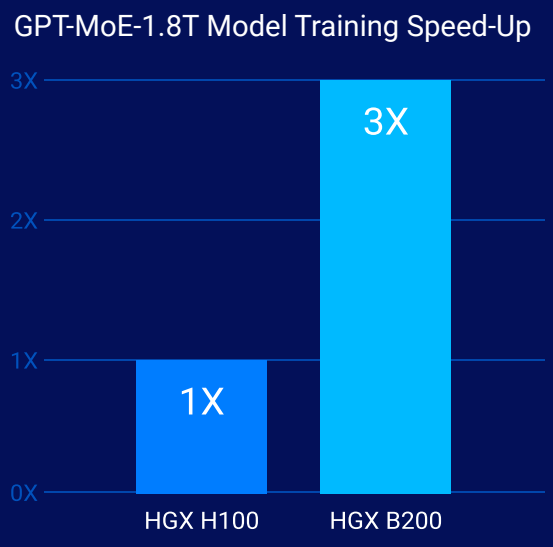
Use cases

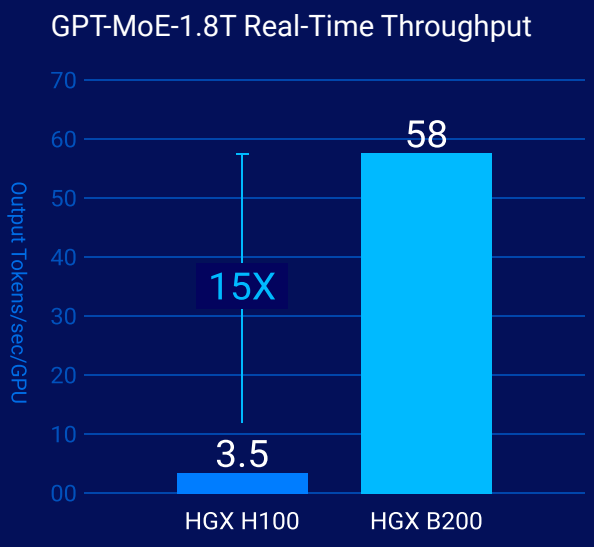
Artificial intelligence training and inference

The NVIDIA Blackwell architecture in the NVIDIA HGX™ B200 is designed to handle the quest for ever-larger and more accurate AI and machine learning models. The second generation NVIDIA Transformer Engine leverages Blackwell Tensor Core technology combined with up to 192 GB of HBM3E memory per GPU to enable FP4 AI, doubling the performance and size of models that the NVIDIA HGX™ B200 can support while maintaining high accuracy. For training, this means that trillion-parameter models can be trained 3x faster than NVIDIA Hopper-generation GPUs. For inference, this means 15x faster real-time inference performance while consuming 12x less energy, which is critical for building sustainable AI solutions. Integrated with Vultr Kubernetes Engine, these models can be deployed and scaled seamlessly across global infrastructure, ensuring efficient orchestration and meeting demand with minimal complexity.

High-performance computing and advanced data analytics

With high memory capacity and memory bandwidth, the NVIDIA HGX™ B200 is built to speed up HPC and analytics workloads while reducing costs. The new NVIDIA Blackwell dedicated Decompression Engine accelerates data queries for exceptional performance in data analytics and data science, supporting the latest compression formats. The NVIDIA HGX™ B200 performs 6x faster than CPUs and 2x faster than NVIDIA H100s for query benchmarks. Combined with Vultr Cloud GPU, HPC workloads leverage low-latency connections across Vultr's global infrastructure, driving faster insights and innovation across industries.





Key benefits

High performance

With 8 NVIDIA Blackwell GPUs, 1.8 TB/s 5th Generation NVIDIA NVLink™ Interconnect, and a remarkable combined 1.4 TB of GPU memory and 60 TB/s memory bandwidth, the NVIDIA HGX™ B200 has the power to deliver exceptional performance. At FP4 precision, the NVIDIA HGX™ B200 delivers 144 petaFLOPs.

Reduced energy usage

For inference workloads, the NVIDIA HGX™ B200 requires up to 12x less energy than Hopper-generation GPUs for similar performance. As sustainability concerns around AI energy usage mount, the NVIDIA HGX™ B200 plays an essential role in building AI not just for today, but also for tomorrow.

Simple deployment

Vultr's streamlined, intuitive user interface and deployment tools make it easy and quick to provision GPU resources, accelerating time to value for your workloads. Vultr GPU Enabled Images offer core NVIDIA software, such as the NVIDIA CUDA toolkit and NVIDIA drivers, packaged and ready to deploy with a single click from the Vultr control panel.

Scalability

Vultr's cloud infrastructure enables seamless scaling for your GPU requirements, ensuring consistent performance as computational demands evolve.

Broad application support

The NVIDIA HGX™ B200 offers support for an extensive range of applications, from AI and machine learning to data analytics, scientific simulations, high-performance compute workloads, and other demanding use cases.

Specifications

NVIDIA HGX B200	
Blackwell GPUs	8
Fast Memory	Up to 1.5 TB
Aggregate Memory Bandwidth	Up to 64 TB/s
Aggregate NVLink Bandwidth	14.4 TB/s
FP4 Tensor Core	144 petaFLOPs*
FP8 Tensor Core	72 petaFLOPs*
INT8 Core	72 petaLOPs*
GPU Memory	Up to 192 GB HBM3e per GPU
Decoders/GPU	2x7 NVDEC | 2x7 NVJPEG
All-to-All Bandwidth	5th Generation NVIDIA NVLink™: 14.4 TB/s

** With sparsity*

Learn more about
Vultr Cloud GPU

Contact us at vultr.com to get started. →