VULTR

DATASHEET

Vultr Cloud GPU Accelerated by NVIDIA H200

NVIDIA H200 Tensor Core GPU: Extraordinary GPU power for supercharging AI and HPC workloads

VULTR.COM

VULTR

Vultr Cloud GPU Accelerated by NVIDIA H200

Introduction

The NVIDIA H200 Tensor Core GPU on Vultr offers remarkable memory capacity and bandwidth for AI and HPC workloads. With 141 GB of HBM3e memory and 4.8 TB/s memory bandwidth, the NVIDIA H200 provides significant memory capacity and bandwidth improvements over its predecessor and is ideal for the most demanding workloads. With 32 cloud data center locations spanning six continents, Vultr's infrastructure offers a scalable and composable environment, allowing users to seamlessly deploy anywhere and at any scale.

Why it's important right now

Rapid growth in AI, data science, and HPC has made high-performance GPUs essential. The NVIDIA H200 provides industry-leading memory and memory bandwidth as the first GPU with HBM3e enabling faster training and inference, which is crucial for modern AI and HPC workloads.

The updated memory and memory bandwidth nearly double the capacity and 1.4x the bandwidth of the NVIDIA H100 Tensor Core GPU. This upgraded memory accelerates GenAI and LLMs while advancing scientific computing for HPC. This reduces inference time and boosts throughput, leading to faster AI training and HPC applications. It improves energy efficiency and lowers the total cost of ownership.

Use Cases

Artificial Intelligence and Machine Learning: The NVIDIA H200 accelerates AI training and inference which is critical for large-scale AI projects. It offers the performance needed for deep learning models, including GenAI for large language models and computer vision for object detection, facial recognition, and autonomous driving. The NVIDIA H200's high memory bandwidth reduces training time and improves inference performance.

High-performance computing: For HPC, the NVIDIA H200 delivers exceptional performance in scientific simulations, enabling researchers to conduct complex analyses faster. It is ideal for weather forecasting, climate modeling, and computational fluid dynamics applications. The NVIDIA H200's increased memory bandwidth allows for efficient data transfer and manipulation, reducing bottlenecks in memory-intensive HPC workloads.

GenAI, LLMs, and transformer model training: The NVIDIA H200's high capacity and speed make it ideal for GenAI applications, specifically for training and deploying LLMs. Its memory capabilities support extensive model training and iteration, reducing deployment time for complex models. This makes it suitable for applications in natural language generation, deep learning art creation, and AI-driven content generation. The NVIDIA H200's enhanced performance facilitates rapid experimentation and fine-tuning, allowing developers to innovate quickly in AI and related fields.

Data science and analytics: In addition to AI, HPC, and GenAI, the NVIDIA H200 is a powerful choice for data science and analytics. Its high memory bandwidth allows for rapid data processing and analysis, enabling businesses to quickly extract insights from large datasets. This makes it suitable for real-time data analytics, business intelligence, and big data processing applications. The NVIDIA H200's capabilities enable faster analysis, facilitating improved decision-making and supporting advanced data-driven strategies in various industries.

Key benefits

High performance and memory capacity: The NVIDIA HGX H200 is designed for high-performance computing and AI, featuring 141 GB of HBM3e memory and a bandwidth of 4.8 TB/s. This performance level allows for rapid processing and can significantly reduce the time it takes to train AI models, improve AI inference performance or run complex scientific simulations.

Superior inference performance: The NVIDIA H200's performance is nearly double that of the NVIDIA H100 when deploying LLMs. This makes it ideal for applications that require fast inference, including chatbots, voice assistants, or other AI-driven solutions that rely on quick response times.

Advanced scalability: Vultr's global infrastructure allows businesses to scale anytime, anywhere seamlessly. Expand from a single GPU to multiple GPUs connected with NVIDIA Quantum-2 InfiniBand in a flexible, scalable environment that accommodates your requirements. Vultr's scalability ensures consistent performance as computational needs grow.

Reduced energy and TCO: The NVIDIA H200's energy efficiency reduces power consumption, leading to a lower total cost of ownership (TCO). This simultaneously minimizes operational costs while maintaining high performance and promotes sustainability.

Easy deployment: Vultr offers an intuitive, streamlined user interface with simple deployment tools and APIs, enabling quick provisioning and straightforward management of GPU resources. This ease of deployment results in less time spent on infrastructure setup and overhead and more resources reinvested into the business.

Broad application support: The NVIDIA HGX H200 offers extensive support for various applications, such as AI and machine learning, data analytics, and scientific simulations. Leverage cutting-edge technology designed to accelerate your digital transformation journey with optimal performance and adaptability and empower your business to drive innovation.

Specifications

NVIDIA H200 SXM GPU

**Shown with sparsity. Specifications 1/2 lower without sparsity.*

FP64: 34 TFLOPS

FP64 Tensor Core: 67 TFLOPS

FP32: 67 TFLOPS

TF32 Tensor Core: 989 TFLOPS*

BFLOAT16 Tensor Core: 1,979 TFLOPS*

FP16 Tensor Core: 1,979 TFLOPS*

FP8 Tensor Core: 3,958 TFLOPS*

INT8 Tensor Core: 3,958 TOPS

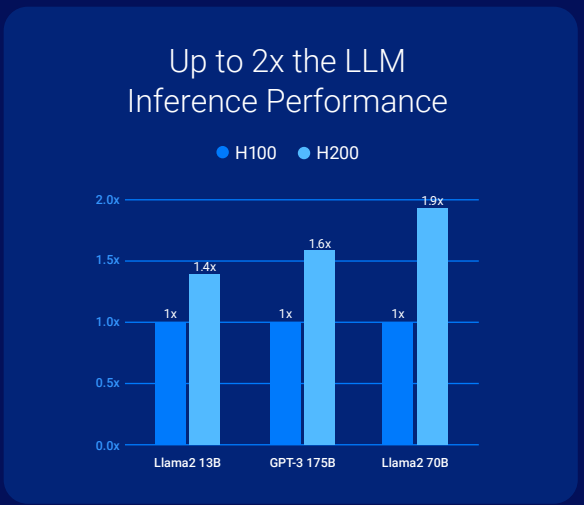
GPU Memory: 141GB

GPU Memory Bandwidth: 4.8TB/s

Decoders: 7 NVDEC 7 JPEG

Interconnect: NVLink: 900 GB/s PCIe Gen5: 128 GB/s

Projected performance, subject to change
LLM LoRA Fine-Tuning: 64 GPUs | H200 vs A100 | GPT3-175B
LLM Training: 1K GPUs | H200 vs A100 | GPT3-175B



Preliminary measured performance, subject to change

Llama2 13B: ISL 128, OSL 2K, | Throughput | H100 1x GPUs BS 64 | H200 1x GPU BS 128

GPT-3 175B: ISL 80, OSL 200 | x8 H100 GPUs BS 64 | x8 H200 GPUs BS 128

Llama2 70B: ISL 2K, OSL 128 | Throughput | H100 1x GPU BS 8 | H200 1x GPU BS 32

Learn more about
Vultr Cloud GPU accelerated by NVIDIA H200

Contact us at vultr.com to get started. →