VULTR

DATASHEET

Vultr Cloud GPU: Accelerated by NVIDIA HGX™ B300

Accelerate the era of AI reasoning with NVIDIA Blackwell Ultra GPUs' exceptional memory capacity and high-speed interconnect.

VULTR.COM

VULTR

Vultr Cloud GPU: Accelerated by NVIDIA HGX™ B300

Accelerate the era of AI reasoning with NVIDIA Blackwell Ultra GPUs' exceptional memory capacity and high-speed interconnect.

The NVIDIA HGX™ B300 provides the acceleration needed to power next-generation AI workloads and more. Containing eight NVIDIA Blackwell Ultra GPUs connected by high-speed interconnect, the NVIDIA HGX B300 delivers impressive performance for AI inference, training, and HPC applications. Through Vultr's composable, global platform, with 32 global cloud data center regions on six continents, access NVIDIA HGX B300s with predictable, transparent, and affordable pricing.

Why it's important right now

AI models have continued to scale up as they are applied to increasingly complex applications. Multi-step reasoning, required to address the requirements of complex workflows, necessitates a substantial upgrade in compute capacity to handle large context windows – calling for up to 100 times more compute than for single-pass inference queries. Obtaining the infrastructure required to support these workloads can be challenging and expensive.

Vultr Cloud GPU and Vultr Bare Metal, accelerated by NVIDIA HGX B300, provide a simple and cost-effective solution for meeting the needs of these evolving workloads without the complexities of hardware management. With the NVIDIA HGX B300 on Vultr, easily construct a high-powered infrastructure stack and empower cutting-edge AI and compute applications while reducing cost per token and decreasing energy usage compared to Hopper-generation GPUs.

Use cases

AI inference and training

NVIDIA HGX B300s are engineered to support the needs of growing AI and machine learning models. With 270 GB of HBM3E memory and 1.8 TB/s fifth-generation NVIDIA NVLink™ GPU-to-GPU bandwidth per GPU, and the second-generation Blackwell Transformer Engine, the NVIDIA HGX B300 delivers up to 1.5x more dense FP4 FLOPs than the NVIDIA HGX B200.

When optimized for AI factory output, AI inference workloads on NVIDIA HGX B300s can achieve up to 30 times greater AI factory productivity compared to Hopper-generation systems. NVIDIA HGX B300s are well-suited for other applications as well, such as real-time AI video generation.

The NVIDIA HGX B300 also excels in AI training, delivering 7 times more AI compute for training compared to the NVIDIA HGX Hopper platform. AI model training can leverage this additional computational power for up to 2.6x faster training for advanced large language models the size of DeepSeek R1.

Agentic AI

With its large-capacity GPU memory and high memory bandwidth, the NVIDIA HGX B300 excels at accelerating multi-step reasoning and agentic applications that require large context windows. AI agents require high-throughput performance to deliver the most appropriate responses to queries, and these advanced systems deliver with fast execution and efficiency. Vultr's global infrastructure provides options for worldwide scaling for autonomous systems, research agents, copilots, and more.

HPC and data analytics

With its high memory bandwidth and throughput, the NVIDIA HGX B300 delivers notable performance for HPC and data analytics workloads, reducing time to insights and accelerating value generation. The Blackwell Decompression Engine can decompress data at a rate up to 800 GB/s with support for the latest compression formats, greatly assisting in data processing for data analytics and data science.

Key benefits

High performance

NVIDIA HGX B300 includes eight NVIDIA Blackwell Ultra GPUs, 2.1 TB aggregate GPU memory, and 64 TB/s total memory bandwidth to deliver remarkable performance – up to 108 dense FP4 petaFLOPs.

Simple deployment and scalability

Vultr's simple and intuitive console, API, and deployment tools make deploying Vultr Cloud GPU or Vultr Bare Metal NVIDIA HGX B300 resources a quick and easy process.

Vultr GPU Enabled Images are available with prepackaged NVIDIA software, including the NVIDIA CUDA toolkit and drivers, through the Vultr console. These images can be deployed through a single click, accelerating time to value.

Composable full-service compute platform

The Vultr cloud platform is designed to be an open ecosystem, ensuring applications can be deployed without vendor lock-in. Vultr offers a comprehensive suite of complementary services, including Vultr Cloud Compute, Vultr Cloud Networking, and Vultr Cloud Storage. Or, deploy a broad set of partner solutions through the Vultr Cloud Alliance. For containerized workloads, Vultr Cloud GPU supports Vultr Kubernetes Engine, greatly assisting with application management.

Specifications: NVIDIA HGX B300

Blackwell Ultra GPUs	8
Total GPU Memory	2.1 TB
NVLink GPU-to-GPU Bandwidth	1.8 TB/s
Aggregate NVLink Bandwidth	14.4 TB/s
FP4 Tensor Core	144 petaFLOPS[1]
FP8/FP6 Tensor Core	72 petaFLOPS[1]
INT8 Tensor Core	307 TOPS[1]
GPU Memory	270 GB HBM3E per GPU
Decoders/GPU	7 NVDEC | 7 NVJPEG
Interconnect	5th Generation NVIDIA NVLink

[1] Specification in sparse

Get started with affordable, predictable pricing

When deploying GPUs, clouds are not all equivalent. Vultr's composable platform, global reach, transparent and cost-efficient pricing, and security and compliance posture stand apart. Leveraging NVIDIA HGX B300s through Vultr provides economic and performance advantages critical to conquering the next era of AI.

Learn more about Vultr Cloud GPU

Contact us at vultr.com to get started. →