VULTR

DATASHEET

Vultr Cloud GPU: Accelerated by NVIDIA GB300 NVL72

Harness unprecedented performance for the age of AI reasoning through a rack-scale solution combining NVIDIA Blackwell Ultra GPUs and NVIDIA Grace™ CPUs.

VULTR.COM

VULTR

Vultr Cloud GPU: Accelerated by NVIDIA GB300 NVL72

Harness unprecedented performance for the age of AI reasoning through a rack-scale solution combining NVIDIA Blackwell Ultra GPUs and NVIDIA Grace™ CPUs.

With 72 Blackwell Ultra GPUs and 36 NVIDIA Grace™ CPUs combined into a single rack-scale, liquid-cooled architecture, the NVIDIA GB300 NVL72 provides the compute capacity to accelerate the growing demands of fast-evolving AI and high-performance compute applications. Vultr's composable, global platform offers access to NVIDIA GB300 NVL72s through a simple-to-use platform with affordable and transparent pricing.

Why it's important right now

As AI is tasked with accomplishing ever-greater goals and workflows of increased complexity, multi-step reasoning has emerged as a key strategy for delivering the correct answer to a request. The large context windows involved in AI reasoning require compute capacity significantly greater than that for single-pass inference queries, demanding up to 100 times more compute. Scaling up infrastructure to meet these needs can be expensive and complicated.

Through Vultr Cloud GPU and Vultr Bare Metal, accelerated by NVIDIA GB300 NVL72, securing the compute capacity that these workloads need is a cost-effective and straightforward process. By leveraging the NVIDIA GB300 NVL72 on Vultr, customers can enjoy a new generation of acceleration while avoiding hardware management complexities and reducing the costs required to build a high-powered environment for the latest cutting-edge compute and AI applications.

Use cases

AI inference and training

NVIDIA GB300 NVL72s provide the compute scale and speed to support the growing needs of larger AI models. Each Blackwell Ultra GPU offers 279 GB of HBM3E memory, with 37 TB of fast memory per rack, delivering unprecedented performance – up to 50 times greater AI factory output compared to a NVIDIA Hopper-generation platform. Optimized for AI reasoning, the NVIDIA GB300 NVL72 also reduces time to answer, with a 30x overall increase in AI Factory output performance.

For diffusion-based video generation models, the NVIDIA GB300 NVL72 produces a 30x performance improvement vs Hopper GPUs, enabling real-time video generation.

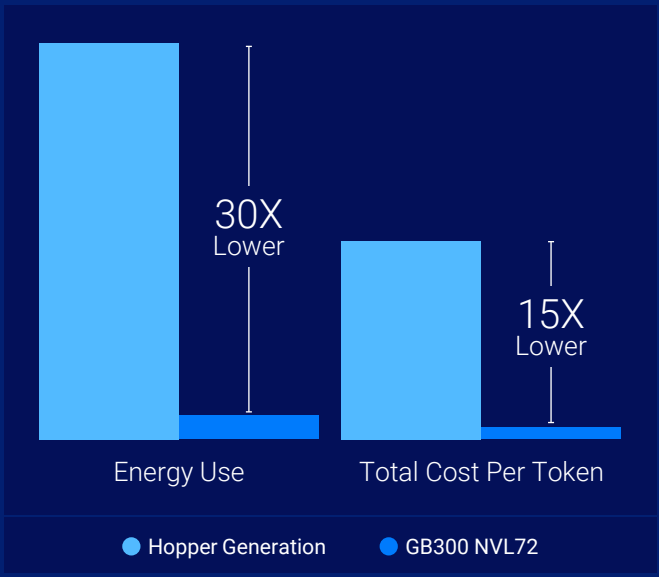
NVIDIA GB300 NVL72s at Vultr are also well-equipped for training large models.

Agentic AI

AI agents tasked with answering complex queries need a compute foundation that can support multi-step reasoning and large context windows efficiently and quickly. NVIDIA GB300 NVL72s are engineered to support these types of workloads with impressive results, and with 30x greater energy efficiency and 15x lower cost per token compared to NVIDIA Hopper-generation systems. With Vultr's global platform, research agents, copilots, autonomous systems, AI agents and more can be scaled worldwide.

High Performance Compute

The NVIDIA GB300 NVL72's Blackwell Tensor Core architecture produces high throughput and efficiency for HPC workloads. To accelerate data analytics and data science workloads, the Blackwell Decompression Engine can decompress data at rates of up to 800 GB/s, providing faster value generation and reduced time to insights when analyzing large quantities of data.



Specifications: NVIDIA GB300 NVL72	
Blackwell Ultra GPUs	72
NVIDIA Grace CPUs	36
Total Fast Memory	37 TB
Total Memory Bandwidth	576 TB/s
FP4 Tensor Core	1,440 petaFLOPS[1]
FP8/FP6 Tensor Core	720 petaFLOPS[1]
GPU Memory	279 GB HBM3E per GPU
Interconnect	5th Generation NVIDIA NVLink

[1] Specification in sparse

Key benefits

Exceptional performance

The NVIDIA GB300 NVL72 delivers 1.44 exaFLOPs sparse FP4 performance through a 72-GPU unified NVLink domain – 70x more FLOPs than available through an NVIDIA HGX™ H100. With 800 GB/s network connectivity per GPU, NVIDIA GB300 NVL72s are designed to scale beyond single compute units into AI factories for massive scale and efficiency.

Simple deployment and scaling

Vultr's console and API provide an easy-to-use, intuitive deployment platform for NVIDIA GB300 NVL72s.

Vultr GPU Enabled Images contain prepackaged NVIDIA software, including drivers and the NVIDIA CUDA toolkit, greatly accelerating the deployment and setup process. They can be deployed through a single click in the Vultr console.

Composable platform with full product suite

Vultr offers a composable platform designed to support a wide variety of applications without creating vendor lock-in. Customers can deploy a complete set of Vultr cloud solutions, including Vultr Cloud Compute, Vultr Cloud Storage, Vultr Cloud Networking, and a variety of partner solutions through the Vultr Cloud Alliance, or they can bring their own. Vultr Cloud GPU supports Vultr Kubernetes Engine for containerized workloads, compatible with Terraform and the Cluster API.

Get started with cost-effective, predictable pricing

When building for the future of AI, selecting the right cloud matters. Vultr offers unmatched price-to-performance, a global and composable platform, and more than a decade of experience providing compliant and secure enterprise-grade infrastructure. With the NVIDIA GB300 NVL72 on Vultr Cloud GPU, combine high-powered infrastructure with a standout cloud platform to achieve a performance edge for the future of AI.

Learn more about Vultr Cloud GPU

Contact us at vultr.com to get started. →