VULTR

DATASHEET

Vultr Cloud GPU Accelerated by NVIDIA H100

NVIDIA H100 Tensor Core GPU: Unprecedented performance, scalability, and security for every workload

VULTR.COM

VULTR

Vultr Cloud GPU Accelerated by NVIDIA H100

Introduction

Introducing the unparalleled power and versatility of the NVIDIA HGX H100 on Vultr, a game-changing solution for businesses and developers seeking exceptional performance and flexibility. With the ability to deploy in over 30 cloud data center locations across the globe, users can capitalize on accelerated workloads, reduced latency, and optimized resource allocation. Harnessing the NVIDIA HGX H100's cutting-edge technology, customers can now access a truly borderless and agile environment for their most demanding projects, transcending the limits of traditional computing infrastructure and ushering in a new era of innovation and productivity.

Why it's important right now

The demand for AI, data science, and HPC workloads is growing exponentially, making it essential to harness the power of the NVIDIA HGX H100 on Vultr's infrastructure. As businesses and researchers increasingly rely on data-driven insights, the need for high-performance computing solutions is more critical than ever. By choosing Vultr's Cloud GPU accelerated by the NVIDIA HGX H100, you'll be equipped with the necessary tools to stay competitive and drive innovation in your industry.

Use Cases

Artificial Intelligence and Machine Learning

The NVIDIA HGX H100 is designed to accelerate AI training and inference, making it perfect for applications such as natural language processing, computer vision, and predictive analytics.

High-performance computing

The NVIDIA HGX H100 provides exceptional performance in scientific simulations, computational fluid dynamics, and other HPC workloads, enabling researchers and engineers to solve complex problems faster.

Data science and analytics

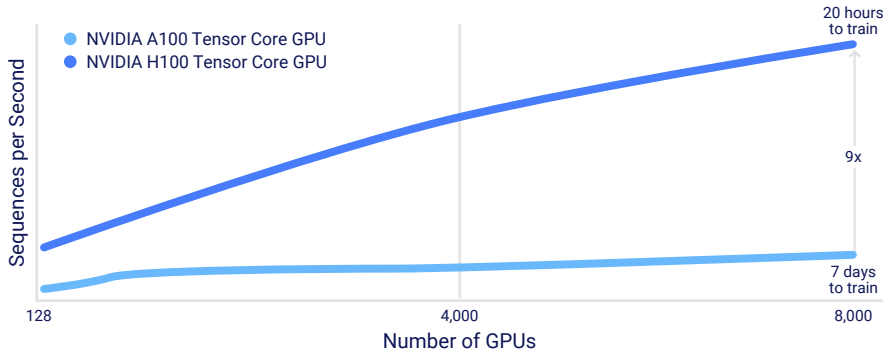
Utilize the NVIDIA HGX H100 for accelerated data processing, big data analysis, and real-time insights, enabling faster and more accurate decision-making.

Generative AI, large language model, and transformer model training

The NVIDIA HGX H100's superior computing capabilities enable faster training and iteration of complex models, significantly reducing the time required to develop and deploy cutting-edge AI applications. By leveraging the NVIDIA HGX H100, researchers and developers can effectively push the boundaries of natural language understanding and generation, ultimately enhancing AI-driven solutions across various industries.

Up to 9x higher AI training on largest models

Mixture of experts
395 billion parameters



Key benefits

Unprecedented performance: The NVIDIA HGX H100 is built on the cutting-edge Hopper™ architecture utilizing 8 NVIDIA H100 GPUs per system, delivering unprecedented performance in AI, data science, and HPC workloads. NVIDIA H100 GPUs feature a specialized Transformer Engine designed to handle language models with trillions of parameters. By integrating cutting-edge technological advancements, the NVIDIA HGX H100 is capable of accelerating large language models (LLMs) by an astounding 30 times compared to its predecessor, thereby providing unparalleled conversational AI performance.

Scalability: Vultr's infrastructure enables seamless scaling to support your growing needs. Whether you require a single GPU or 256 GPUs connected with NDR Infiniband we provide a flexible and scalable environment that can accommodate your requirements. This ensures consistent performance, even as your computational needs grow.

Broad application support: The NVIDIA HGX H100 offers extensive support for a wide range of applications, such as AI and machine learning, data analytics, scientific simulations, graphics rendering, video transcoding, and blockchain. Empower your business to drive innovation by leveraging cutting-edge technology designed to accelerate your digital transformation journey with optimal performance and adaptability.

Cost-effective: Leverage the power of the NVIDIA HGX H100 without incurring prohibitive hardware costs. Vultr Cloud GPU, accelerated by NVIDIA, offers a cost-effective solution that allows you to access the latest technology without upfront investment, making it an ideal choice for businesses of all sizes.

Easy deployment: Deploying and managing GPU resources on Vultr's platform is simple and straightforward, enabling you to focus on your core business activities. With our user-friendly management console and APIs, you can quickly provision and manage your NVIDIA H100 instances, allowing you to spend more time on your projects and less time managing your infrastructure.

Specifications

NVIDIA H100 SXM GPU	
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	80GB
GPU Memory Bandwidth	3.35TB/s
Decoders	7 NVDEC 7 JPEG
Interconnect	NVLink: 900GB/s PCIe Gen5: 128GB/s

** Shown with sparsity. Specifications ½ lower without sparsity.*

Learn more about Vultr Cloud GPU accelerated by NVIDIA H100

Contact us at vultr.com to get started. →