

Explore GPU Variants

A comprehensive guide to Vultr's diverse GPU options designed for high-performance computing tasks like AI, machine learning, and 3D rendering.

Contents

01	Introduction	3
02	GPU Specifications and Performance Overview	3

Explore GPU Variants for Vultr Cloud GPU

Introduction

Vultr offers a wide range of Vultr Cloud GPU options to meet different user needs. These powerful GPUs help users handle demanding tasks such as machine learning, artificial intelligence, 3D rendering, and complex simulations. With Vultr's resources, developers can create and launch new applications more effectively.

This article explains the four different Vultr Cloud GPU offerings which are A16, A40, A100 Tensor Core and L40S, you will get an understanding of the core differences between the GPU variants to help users make the best choice for a specific workload.

GPU Specifications and Performance Overview

This section provides the necessary technical and performance data to compare the 4 different GPUs based on their CUDA Cores, Tflops performance, parallel processing.

Below are the stats of the different Cloud GPUs Vultr offers to compare performances:

Variant	Implementation	CUDA cores	Tensor cores	TF32 tflops with sparsity	GPU memory	Architecture type
	vGPU	5120	160	72	64 GB	

Variant	Implementation	CUDA cores	Tensor cores	TF32 tflops with sparsity	GPU memory	Architecture type
NVIDIA A16 GPU						NVIDIA Ampere
NVIDIA A40 GPU	vGPU	10752	336	149.6	48 GB	NVIDIA Ampere
NVIDIA A100 Tensor Core GPU	vGPU	6912	432	312	80 GB	NVIDIA Ampere
NVIDIA L40S GPU	Passthrough	18716	568	366	48 GB	NVIDIA ADA Lovelace

Key Terms

1. **vGPU:** Allows a physical GPU to be shared among multiple virtual machines (VMs), enabling high-performance graphics and compute workloads in a virtualized environment. This is particularly useful for tasks like virtual desktop infrastructure (VDI), graphics-intensive applications, and AI workloads in shared setups.
2. **Passthrough:** Virtualization technique that allows a physical GPU to be directly assigned to a single virtual machine (VM), providing the VM with exclusive access to the GPU's resources. Unlike vGPU, where a GPU is shared among multiple VMs, passthrough dedicates the entire GPU to one VM for maximum performance and compatibility.
3. **CUDA Cores:** These are specific type of processing unit designed to work with NVIDIA's CUDA programming model, they play a fundamental role in parallel processing and accelerating various computing tasks focused on graphics rendering. They often use a Single Instruction, Multiple Data (SIMD) architecture so that a single instruction is executed simultaneously on multiple data elements, resulting a high throughput in parallel computing.

4. **Tensor Cores:** These are specialized hardware component in the NVIDIA GPUs made for accelerating matrix-based computations that are commonly used in deep learning and many artificial intelligence workloads. They are optimized for mathematical operations involved in neural network training and inference by taking advantage of their mixed precision computing, where certain part of the calculations with higher precision and the rest with half precision while maintaining the accuracy in results by using error correction and accumulation.

GPU Briefs

- **NVIDIA A16:** Vultr Cloud GPU, powered by the NVIDIA A16, is a flexible and reliable solution for low-latency VDI, efficient transcoding, and AI inference. This technology helps businesses provide high-performance, secure, and scalable computing solutions for different applications. With improved computational power, organizations can increase productivity, protect data more effectively, and use cost-efficient infrastructure. [Learn more about Vultr Cloud GPU A16.](#)
- **NVIDIA A40:** Vultr Cloud GPU, powered by the NVIDIA A40, is designed for workloads that require advanced visualization, such as 3D rendering, virtual production, and complex simulations. This GPU offers high memory capacity and processing power, making it ideal for data-intensive tasks. Businesses can leverage the A40 to manage large datasets, perform real-time rendering, and support demanding creative and scientific workflows. It also enables scalable, secure infrastructure that helps improve efficiency and reduce operational costs. [Learn more about Vultr Cloud GPU A40.](#)
- **NVIDIA A100 Tensor Core:** Vultr Cloud GPU, powered by the NVIDIA A100, is designed for tasks such as AI training, deep learning, and large-scale data processing. With its tensor cores and large memory, the A100 helps businesses handle complex computations, work with large datasets, and improve AI model efficiency. It is suitable for research, scientific work, and other tasks that need strong parallel processing. This GPU also allows organizations to scale their infrastructure while managing costs. [Learn more about Vultr Cloud GPU A100.](#)

- **NVIDIA L40S:** Vultr Cloud GPU, powered by the NVIDIA L40S, is optimized for advanced graphics, AI inference, and real-time processing. The L40S excels in handling tasks like 3D visualization, virtual environments, and video processing, making it suitable for creative industries and high-end computing projects. With its efficient performance and flexibility, the L40S can help businesses deliver smoother graphics and faster data processing while keeping infrastructure scalable and cost-effective. [Learn more about Vultr Cloud GPU L40S.](#)



VULTR

