VULTR

DATASHEET

Vultr Cloud GPU Accelerated by NVIDIA L40S

Unparalleled AI and graphics performance for the data center

VULTR.COM

VULTR

Vultr Cloud GPU, Accelerated by NVIDIA L40S

Built to power the most demanding AI and graphics-intensive workloads for the data center, the NVIDIA L40S GPU is the most-powerful universal GPU.

Introduction

The NVIDIA L40S GPU is the most powerful universal GPU for the data center, delivering end-to-end acceleration for the next generation of AI-enabled applications — from generative AI and model training and inference to 3D graphics, rendering, and video applications.

With Vultr Cloud GPU, accelerated by NVIDIA's computing platform, the NVIDIA L40S GPU can be harnessed through Vultr Cloud GPU or as an 8-GPU bare-metal server. Get up to speed quickly powered by Vultr GPU Enabled Images, or enjoy the flexibility of direct access to NVIDIA L40S GPUs through Vultr Bare Metal. Experience greater control and the ability to supply your own drivers for maximum software compatibility.

Why it's important right now

Rapid developments and continuous breakthroughs in AI are fueling transformative change, spanning all industries and revolutionizing the workflows of scientists, engineers, creators, and more. On top of the demand for accelerated computing to power traditional AI applications, such as machine learning, deep learning, natural language processing, and computer vision, a new model emerged, unlocking a frontier of opportunities — GenAI.

To transform with AI, enterprises must deploy more compute resources at a larger scale. With existing pressures to boost performance, efficiency, and ROI, modern data centers need universal computing solutions that provide accelerated compute, graphics, and video processing capabilities for an ever-increasing set of complex and diverse workloads.

Use cases

Generative AI

With next-generation AI, graphics, and media acceleration capabilities, the NVIDIA L40S GPU delivers up to 1.7x training and 1.5x inference performance versus the previous generation NVIDIA A100 Tensor Core GPU. With breakthrough performance and 48 gigabytes (GB) of memory capacity, the NVIDIA L40S GPU is the ideal platform for accelerating multimodal GenAI workloads.

LLM training and interference

NVIDIA fourth-generation Tensor Cores with support for FP8 deliver exceptional AI computing performance to accelerate training and inference of state-of-the-art LLM and GenAI models.

Rendering and 3D graphics

With third-generation RT Cores that deliver up to 2x the real-time ray-tracing performance of the previous generation to power the creation of stunning visual content and high-fidelity creative workflows, from interactive rendering to real-time virtual production.

The highest performance universal GPU for AI, graphics, and video

Fine Tuning LLM	AI Training	AI Inference
4 hrs	1.7x	1.5x
GPT-175B 860M Tokens[1]	Performance vs. HGX A100[2]	Performance vs. A100 80 GB SXM[3]

GPT3 Training	Image GenAI	Full Video Pipeline
<4 days	>82	184
GPT-175 300B Tokens[4]	Images per minute[5]	AV1 Encode Streams[6]

Features

NVIDIA fourth-generation Tensor Cores

Hardware support for structural sparsity and optimized TF32 format provides out-of-the-box performance gains for faster AI and data science model training. Accelerate AI-enhanced graphics capabilities with DLSS to upscale resolution with better performance in select applications.

NVIDIA third-generation RT Cores

Enhanced throughput and concurrent ray-tracing and shading capabilities improve ray-tracing performance, accelerating renders for product design and architecture, engineering, and construction workflows. See lifelike designs in action with hardware-accelerated motion blur and stunning real-time animations.

NVIDIA Transformer Engine

NVIDIA Transformer Engine dramatically accelerates AI performance and improves memory utilization for both training and inference. Harnessing the power of the NVIDIA Ada Lovelace fourth-generation Tensor Cores, Transformer Engine intelligently scans the layers of transformer architecture neural networks and automatically recasts between FP8 and FP16 precisions to deliver faster AI performance and accelerate training and inference.

Specifications

NVIDIA L40S GPU	
GPU Architecture	NVIDIA Ada Lovelace Architecture
GPU Memory	48GB GDDR6 with ECC
Memory Bandwidth	864GB/s
NVIDIA Ada Lovelace Architecture-based CUDA® Cores	18,176
NVIDIA Third-generation RT Cores	142
NVIDIA Fourth-generation RT Cores	568
NVENC | NVDEC	3x | 3x (includes AV1 encode and decode)

Preliminary performance projections, subject to change

1. Fine-Tuning LoRA (GPT-175B), bs: 128, sl: 256; 64 GPUs: 16 systems with 4xL40S
2. Fine-Tuning LoRA (GPT-40B), bs: 128, sl: 256; Two systems with 4x L40S, vs HGX A100 8 GPU
3. Hugging Face SWIN Base Inference (BS=1,Seq 224); L40S vs. A100 80GB SXM
4. GPT 175B, 300B tokens, Foundational Training; 4K GPUs; 1000 systems with 4xL40S
5. Image Generation, Stable Diffusion v2.1, 512 x 512 resolution; 1xL40S
6. Concurrent Encoding Streams; 720p30; 1xL40S

Learn more about Vultr Cloud GPU accelerated by NVIDIA L40S

Contact us at vultr.com to get started. →