



NVIDIA L40

Delivering Unprecedented Visual Computing Performance for the Data Center.

From virtual workstation application to large-scale modeling and simulation, modern visual computing and scientific workflows are growing in both complexity and quantity. Enterprises need data center technology that can deliver extreme performance and scale with versatile capabilities to conquer the diverse computing demands of these increasingly complex workloads.

The NVIDIA® L40 GPU delivers unprecedented visual computing performance for the data center, providing next-generation graphics, compute, and AI capabilities. Built on the revolutionary NVIDIA Ada Lovelace architecture, the NVIDIA L40 harnesses the power of the latest generation RT, Tensor, and CUDA cores to deliver groundbreaking visualization and compute performance for the most demanding data center workloads.

Accelerate next-generation workloads.

- > NVIDIA Omniverse™ Enterprise
- > Rendering and 3D Graphics
- > High-Performance Virtual Workstations with NVIDIA RTX™ Virtual Workstation (RTX vWS) Software
- > AI Training and Data Science
- > Streaming and Video Content

Powered by the NVIDIA Ada Lovelace architecture.

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 to deliver stunning real-time animations.



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, including **DLSS**, delivering upscaled resolution with better performance in select applications.

Large GPU memory

Tackle memory-intensive applications and workloads like data science, simulation, 3D modeling, and rendering with 48GB of ultra-fast GDDR6 memory. Allocate memory to multiple users with vGPU software to distribute large workloads among creative, data science, and design teams.

Data-center ready

Designed for 24x7 enterprise data center operations with power-efficient hardware and components, the NVIDIA L40 is optimized to deploy at scale and deliver maximum performance for a diverse range of data center workloads. The L40 includes secure boot with root of trust technology providing an additional layer of security, and is NEBS Level 3 compliant to meet the latest data center standards. Packaged in a dual-slot, passively cooled and power-efficient design, the L40 is available in a wide variety of NVIDIA-Certified Systems™ from leading OEM vendors.

Technical Specifications

	NVIDIA L40*
GPU Architecture	NVIDIA Ada Lovelace architecture
GPU Memory	48GB GDDR6 with ECC
Memory Bandwidth	864GB/s
Interconnect Interface	PCIe Gen4x16: 64GB/s bi-directional
NVIDIA Ada Lovelace architecture-based CUDA Cores	18,176
NVIDIA third-generation RT Cores	142
NVIDIA fourth-generation Tensor Cores	568
RT Core performance TFLOPS	209
FP32 TFLOPS	90.5
TF32 Tensor Core TFLOPS	90.5 181**
BFLOAT16 Tensor Core TFLOPS	181.05 362.1**
FP16 Tensor Core	181.05 362.1**
FP8 Tensor Core	362 724**
Peak INT8 Tensor TOPS	362 724**
Peak INT4 Tensor TOPS	724 1448**
Form Factor	4.4" (H) x 10.5" (L) - dual slot
Display Ports	4 x DisplayPort 1.4a
Max Power Consumption	300W
Power Connector	16-pin
Thermal	Passive
Virtual GPU (vGPU) software support	Yes
vGPU Profiles Supported	See Virtual GPU Licensing Guide ¹
NVENC NVDEC	3x 3x (Includes AV1 Encode & Decode)
Secure Boot with Root of Trust	Yes
NEBS Ready	Level 3
MIG Support	No
NVLink Support	No

* Preliminary specifications, subject to change

** With Sparsity.

¹ Coming in a future release of NVIDIA vGPU software.

Ready to get started?

To learn more about the NVIDIA L40 GPU, visit:

www.nvidia.com/L40

© 2022 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, NVIDIA-Certified Systems, NVIDIA Omniverse and NVIDIA RTX are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 2436245. NOV22

