



# NVIDIA L40S

Unparalleled AI and graphics performance for the data center.



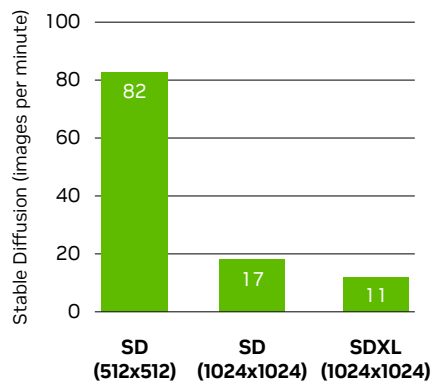
Generative AI is fueling transformative change, unlocking a new frontier of opportunities for enterprises across every industry. To transform with AI, enterprises need more compute resources, greater scale, and a broad set of capabilities to meet the demands of an ever-increasing set of diverse and complex workloads.

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 **gen AI**, LLM inference, small-model training and fine-tuning to 3D graphics, rendering, and video applications.

## Accelerate Next-Generation Workloads

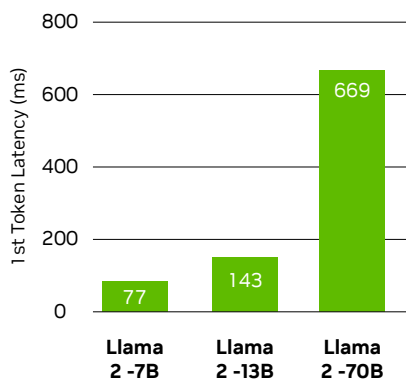
- > Generative AI
- > LLM inference
- > LLM fine-tuning and small-model training
- > NVIDIA Omniverse™ Enterprise
- > Rendering and 3D graphics
- > Streaming and video content

### Generative AI Image Generation



Measured performance; NVIDIA L40S  
 Stable Diffusion v2.1, TRT 8.6.1, BS:1, FP16 |  
 Stable Diffusion XL 1.0, TRT 8.6.1, BS:1, FP16

### Large Language Model (LLM) Inference



Measured performance; NVIDIA L40S  
 Llama 2-7B/13B/70B, ISL=2048, OSL=128,  
 BS=1: FP8.

## Powered by the NVIDIA Ada Lovelace Architecture

### 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.

## 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.

## Transformer Engine

Transformer Engine dramatically accelerates AI performance and improves memory utilization for both training and inference. Harnessing the power of the **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.

## Data Center Ready

The L40S GPU is optimized for 24/7 enterprise data center operations and designed, built, tested, and supported by NVIDIA to ensure maximum performance, durability, and uptime. The L40S GPU meets the latest data center standards, is Network Equipment-Building System (NEBS) Level 3 ready, and features secure boot with root of trust technology, providing an additional layer of security for data centers.

### Technical Specifications

<b>GPU Architecture</b>	NVIDIA Ada Lovelace Architecture
<b>GPU Memory</b>	48GB GDDR6 with ECC
<b>Memory Bandwidth</b>	864GB/s
<b>Interconnect Interface</b>	PCIe Gen4 x16: 64GB/s bidirectional
<b>NVIDIA Ada Lovelace Architecture-Based CUDA® Cores</b>	18,176
<b>NVIDIA Third-Generation RT Cores</b>	142
<b>NVIDIA Fourth-Generation Tensor Cores</b>	568
<b>RT Core Performance TFLOPS</b>	209
<b>FP32 TFLOPS</b>	91.6
<b>TF32 Tensor Core TFLOPS</b>	183   366*
<b>BFLOAT16 Tensor Core TFLOPS</b>	362.05   733*
<b>FP16 Tensor Core</b>	362.05   733*
<b>FP8 Tensor Core</b>	733   1,466*
<b>Peak INT8 Tensor TOPS</b>	733   1,466*
<b>Peak INT4 Tensor TOPS</b>	733   1,466*
<b>Form Factor</b>	4.4" (H) x 10.5" (L), dual slot
<b>Display Ports</b>	4x DisplayPort 1.4a
<b>Max Power Consumption</b>	350W
<b>Power Connector</b>	16-pin

<b>Thermal</b>	Passive
<b>Virtual GPU (vGPU) Software Support</b>	Yes
<b>vGPU Profiles Supported</b>	See the <a href="#">virtual GPU licensing guide</a>
<b>NVENC   NVDEC</b>	3x   3x (includes AV1 encode and decode)
<b>Secure Boot With Root of Trust</b>	Yes
<b>NEBS Ready</b>	Level 3
<b>MIG Support</b>	No
<b>NVIDIA® NVLink® Support</b>	No

\* With sparsity

## Ready to Get Started?

To learn more about the NVIDIA L40S, visit [www.nvidia.com/l40s](http://www.nvidia.com/l40s)

© 2024 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, CUDA, HGX, NVLink, and Omniverse 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. 3110647. FEB24

