

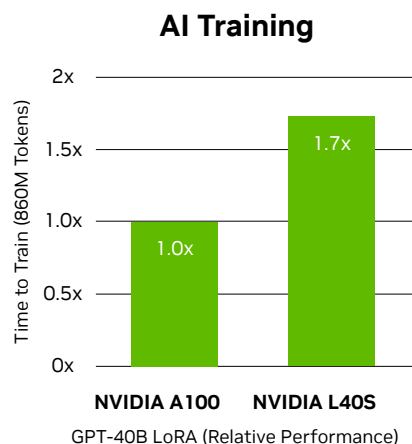


# NVIDIA L40S

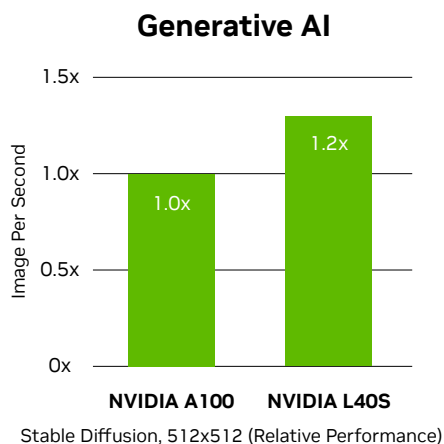
Unparalleled AI and graphics performance for the data center.

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 generative AI and model training and inference to 3D graphics, rendering, and video applications.



Fine-tuning LoRA (GPT-40B): global train batch size: 128 (sequences), seq-length: 256 (tokens). NVIDIA HGX™ A100 8 GPU vs. Two systems with 4x L40S GPUs. Performance based on prerelease build, subject to change.



Stable Diffusion v2.1. Relative speedup for 512 x 512 resolution image generation. NVIDIA HGX A100 8 GPU vs. Two systems with 4x L40S GPUs. Performance based on prerelease build, subject to change.

## Accelerate Next-Generation Workloads

- > Generative AI
- > Large language model (LLM) training and inference
- > NVIDIA Omniverse™ Enterprise
- > Rendering and 3D graphics
- > Streaming and video content

## 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	
PNY Part Number	NVL40STCGPU-KIT
GPU Architecture	NVIDIA Ada Lovelace Architecture
GPU Memory	48GB GDDR6 with ECC
Memory Bandwidth	864GB/s
Interconnect Interface	PCIe Gen4 x16: 64GB/s bidirectional
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	91.6
TF32 Tensor Core TFLOPS	183   366*
BFLOAT16 Tensor Core TFLOPS	362.05   733*
FP16 Tensor Core	362.05   733*
FP8 Tensor Core	733   1,466*
Peak INT8 Tensor TOPS	733   1,466*
Peak INT4 Tensor TOPS	733   1,466*
Form Factor	4.4" (H) x 10.5" (L), dual slot
Display Ports	4x DisplayPort 1.4a
Max Power Consumption	350W
Power Connector	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.pny.com/L40S](http://www.pny.com/L40S)

© 2023 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. 2841316. AUG23

