Datasheet

# NVIDIA L4 Tensor Core GPU

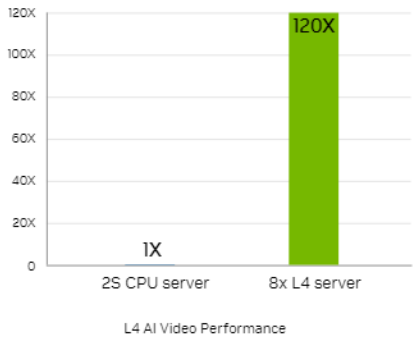
The breakthrough universal accelerator for efficient video, AI, and graphics.

## Accelerate Video, AI, and Graphics Workloads

The NVIDIA Ada Lovelace L4 Tensor Core GPU delivers universal acceleration and energy efficiency for video, AI, virtualized desktop, and graphics applications in the enterprise, in the cloud, and at the edge. With NVIDIA's AI platform and full-stack approach, L4 is optimized for inference at scale for a broad range of AI applications, including recommendations, voice-based AI avatar assistants, generative AI, visual search, and contact center automation to deliver the best personalized experiences.

As the most efficient NVIDIA accelerator for mainstream use, servers equipped with L4 power up to 120X higher AI video performance and 2.7X more generative AI performance over CPU solutions, as well as over 4X more graphics performance than the previous GPU generation. NVIDIA L4's versatility and energy-efficient, single-slot, low-profile form factor make it ideal for global deployments, including edge locations.

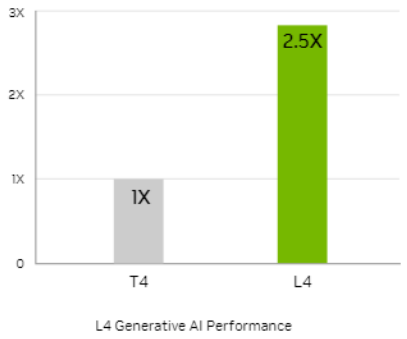
**Up to 120X Higher AI Video Performance**




L4 AI Video Performance

Measured performance: 8x L4 vs 2S Intel 8362 CPU server comparison, end-to-end video pipeline with CV-CUDA® decode, preprocessing, inference (SegFormer), postprocessing, encode, NVIDIA® TensorRT™ 8.6 vs CPU-only pipeline using OpenCV 4.7, PyTorch inference.

**2.5X More Generative AI Performance**

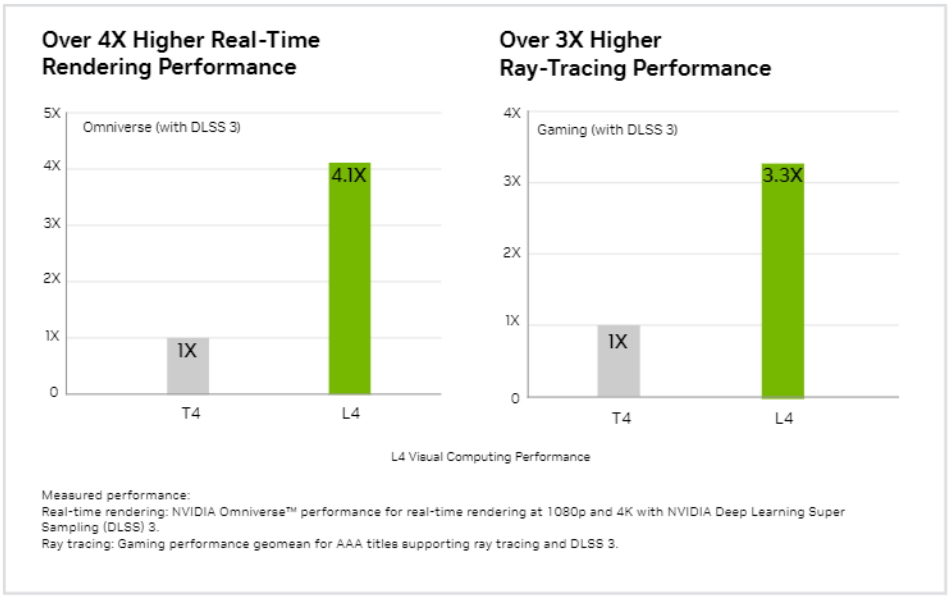



L4 Generative AI Performance

Measured performance: L4 vs T4 image generation, 512x512 stable diffusion v2.1, FP16, TensorRT 8.5.2.

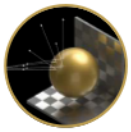
| Specifications | |
|---|---|
| **FP32** | 30.3 teraFLOPs |
| **TF32 Tensor Core** | 120 teraFLOPS* |
| **FP16 Tensor Core** | 242 teraFLOPS* |
| **BFLOAT16 Tensor Core** | 242 teraFLOPS* |
| **FP8 Tensor Core** | 485 teraFLOPS* |
| **INT8 Tensor Core** | 485 TOPs* |
| **GPU memory** | 24GB |
| **GPU memory bandwidth** | 300 GB/s |
| **NVENC \| NVDEC \| JPEG decoders** | 2 \| 4 \| 4 |
| **Max thermal design power (TDP)** | 72W |
| **Form factor** | 1-slot low-profile, PCIe |
| **Interconnect** | PCIe Gen4 x16 64GB/s |
| **Server options** | Partner and NVIDIA-Certified Systems with 1–8 GPUs |

* Shown with sparsity. Specifications 1/2 lower without sparsity.




Measured performance:
Real-time rendering: NVIDIA Omniverse™ performance for real-time rendering at 1080p and 4K with NVIDIA Deep Learning Super Sampling (DLSS) 3.
Ray tracing: Gaming performance geomean for AAA titles supporting ray tracing and DLSS 3.

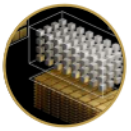
## Explore NVIDIA Ada Lovelace Architecture Breakthroughs



### Third-Generation RT Cores

NVIDIA made real-time ray tracing a reality with the invention of RT Cores, which are processing cores on the GPU specifically designed to tackle performance-intensive ray-tracing rendering.

Ada Lovelace's third-generation RT Cores have twice the ray-triangle intersection throughput, increasing RT-TFLOP performance by over 2X. NVIDIA Shader Execution Reordering (SER) improves performance over 3X, enabling deep immersive experiences for virtual worlds and unprecedented productivity for AI-based neural graphics and cloud gaming.



### Fourth-Generation Tensor Cores

The Ada Lovelace architecture Tensor Cores are designed to accelerate transformative AI technologies like intelligent chatbots, generative AI, natural language processing (NLP), computer vision, and NVIDIA DLSS 3. Ada Lovelace Tensor Cores unleash structured sparsity and 8-bit floating point (FP8) precision for up to 4X higher inference performance over the previous generation.[1] FP8 reduces memory pressure when compared to larger precisions and dramatically accelerates AI throughput.

### Advanced Video and Vision AI Acceleration

With an optimized AV1 stack, NVIDIA L4 takes video and vision AI acceleration to the next level, creating a broad array of new possibilities for use cases like real-time video transcoding, streaming, video conferencing, augmented reality (AR), virtual reality (VR), and vision AI. With four video decoders and two video encoders, combined with the AV1 video format, L4 servers can host over 1,000[2] concurrent video streams and over 120X more AI video end-to-end pipeline performance than CPU solutions.[3] On top of this, four JPEG decoders further speed up applications that need computer vision horsepower.

### Deep Learning Super Sampling (DLSS)

NVIDIA DLSS 3 is a revolutionary breakthrough in AI-powered graphics that massively boosts rendering performance. Powered by the new fourth-generation Tensor Cores and NVIDIA Optical Flow Accelerator (OFA) on L4, DLSS 3 uses AI to create additional high-quality frames for graphics-based workloads.

### Virtualization-Ready

With next-generation improvements in NVIDIA virtual GPU (vGPU) software and 1.5X more GPU memory than the previous generation, L4 increases workstation performance by 1.7X for mid- to high-end design workflows running on NVIDIA RTX™ Virtual Workstation (vWS) and accelerates productivity applications running on NVIDIA Virtual PC (vPC).

### Data Center Efficiency and Security

NVIDIA L4 is optimized for 24/7 enterprise data center operations and is designed, built, extensively tested, and supported by NVIDIA and partners for maximum performance, durability, and security. L4 features secure boot with root-of-trust technology, providing an additional layer of security for data centers.

1. L4's FP8 compared to T4's FP16.

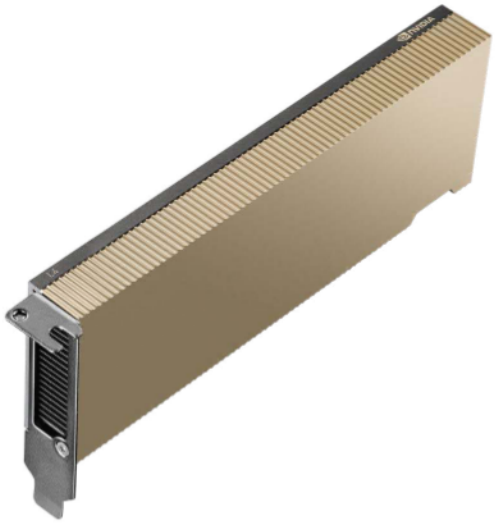
2. 8x L4 AV1 low-latency P1 preset encode at 720p30.

3. 8x L4 vs 2S Intel 8362 CPU server performance comparison: end-to-end video pipeline with CV-CUDA pre- and postprocessing, decode, inference (SegFormer), encode, TRT 8.6 vs CPU only pipeline using OpenCV.

## Accelerate Workloads Efficiently and Sustainably

The NVIDIA L4 is an integral part of the NVIDIA data center platform. Built for AI, video, virtual workstations, graphics, simulation, data science, and data analytics, the platform accelerates over 3,000 applications and is available everywhere at scale, from data center to edge to cloud, delivering both dramatic performance gains and energy-efficiency opportunities.

As AI and video become more pervasive, the demand for efficient, cost-effective computing is increasing more than ever. NVIDIA L4 Tensor Core GPUs deliver up to 120X better AI video performance, resulting in up to 99 percent better energy efficiency and lower total cost of ownership compared to traditional CPU-based infrastructure. This lets enterprises reduce rack space and significantly lower their carbon footprint, while being able to scale their data centers to many more users. The energy saved by switching from CPUs to NVIDIA L4s in a 2 megawatt (MW) data center can power over 2,000 homes for one year or match the carbon offset of 172,000 trees grown over 10 years.[4,5]



## Enterprise Ready: AI Software Streamlines Development and Deployment

Enterprise adoption of AI is now mainstream, and organizations require end-to-end, AI-ready infrastructure that will future-proof them for this new era. NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software optimized to help every organization excel at AI and certified to deploy anywhere, from the enterprise data center to the cloud. It comes with included global enterprise support to ensure AI projects stay on track.

4. 8x L4 vs 2S Intel 8362 CPU server comparison: end-to-end video pipeline with CV-CUDA pre- and postprocessing, decode, inference (SegFormer), encode, TRT 8.6 vs. CPU-only pipeline using OpenCV 4.7, PyT inference.

5. Results from EPA calculator using 1.677MW savings. www.epa.gov/energy/greenhouse-gas-equivalencies-calculator

Optimized to streamline AI development and deployment, NVIDIA AI Enterprise includes proven, open-source containers and frameworks that are certified to run on common data center platforms and mainstream NVIDIA-Certified Systems™ with NVIDIA L4 Tensor Core GPUs. Since support is included, organizations get the transparency of open source and the assurance of global NVIDIA Enterprise Support with AI expertise for both their AI practitioners and IT administrators.

NVIDIA AI Enterprise software is a license addition for NVIDIA L4 Tensor Core GPUs, making AI accessible to nearly every organization with the highest performance in training, inference, and data science. NVIDIA AI Enterprise together with NVIDIA L4 simplifies the building of an AI-ready platform, accelerates AI development and deployment, and delivers performance, security, and scalability to gather insights faster and achieve business value sooner.

Learn about all the AI workloads you can run on L4 with free, hands-on **NVIDIA AI Enterprise labs** through NVIDIA LaunchPad.

## Ready to Get Started?

To learn more about the NVIDIA L4 Tensor Core GPU, visit:
**www.nvidia.com/l4**



NVIDIA.