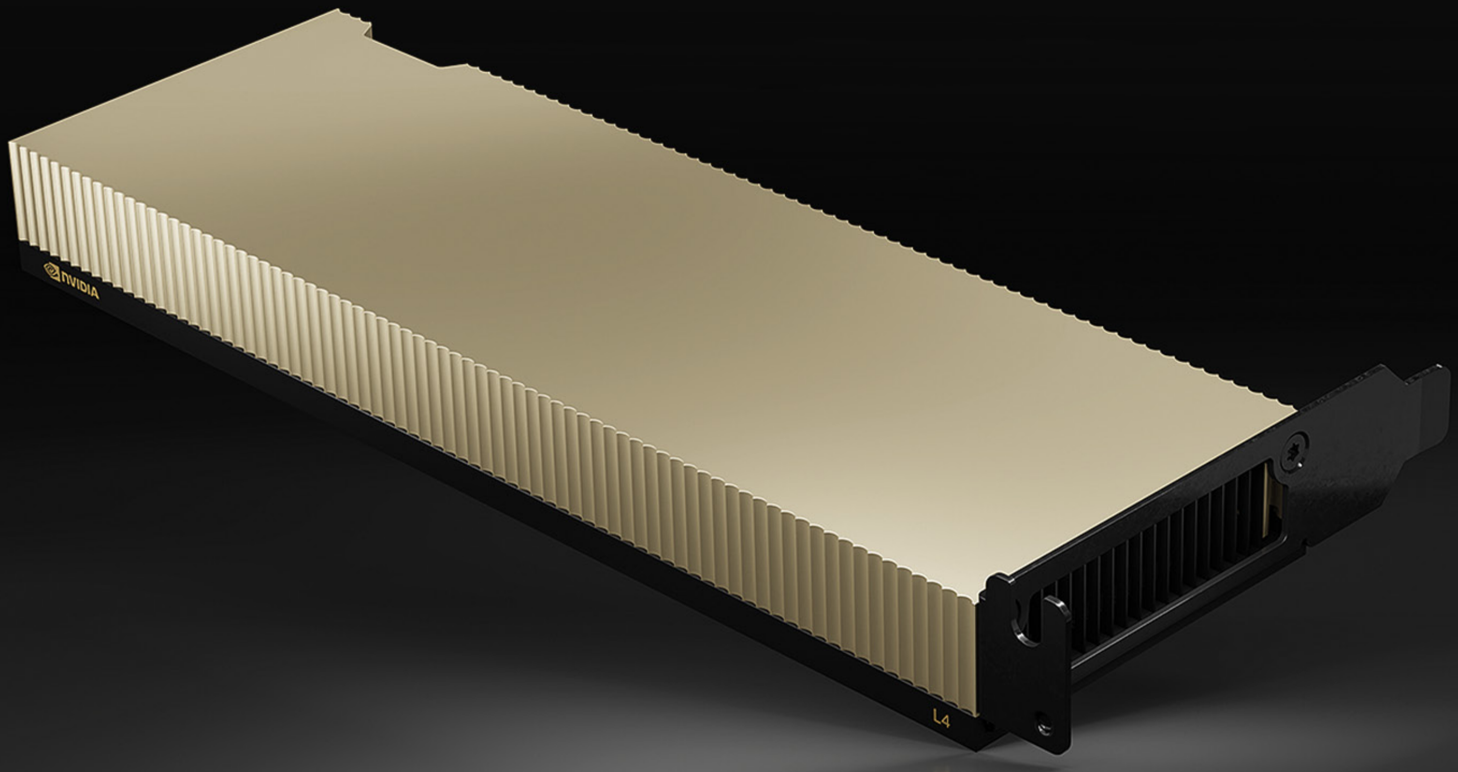
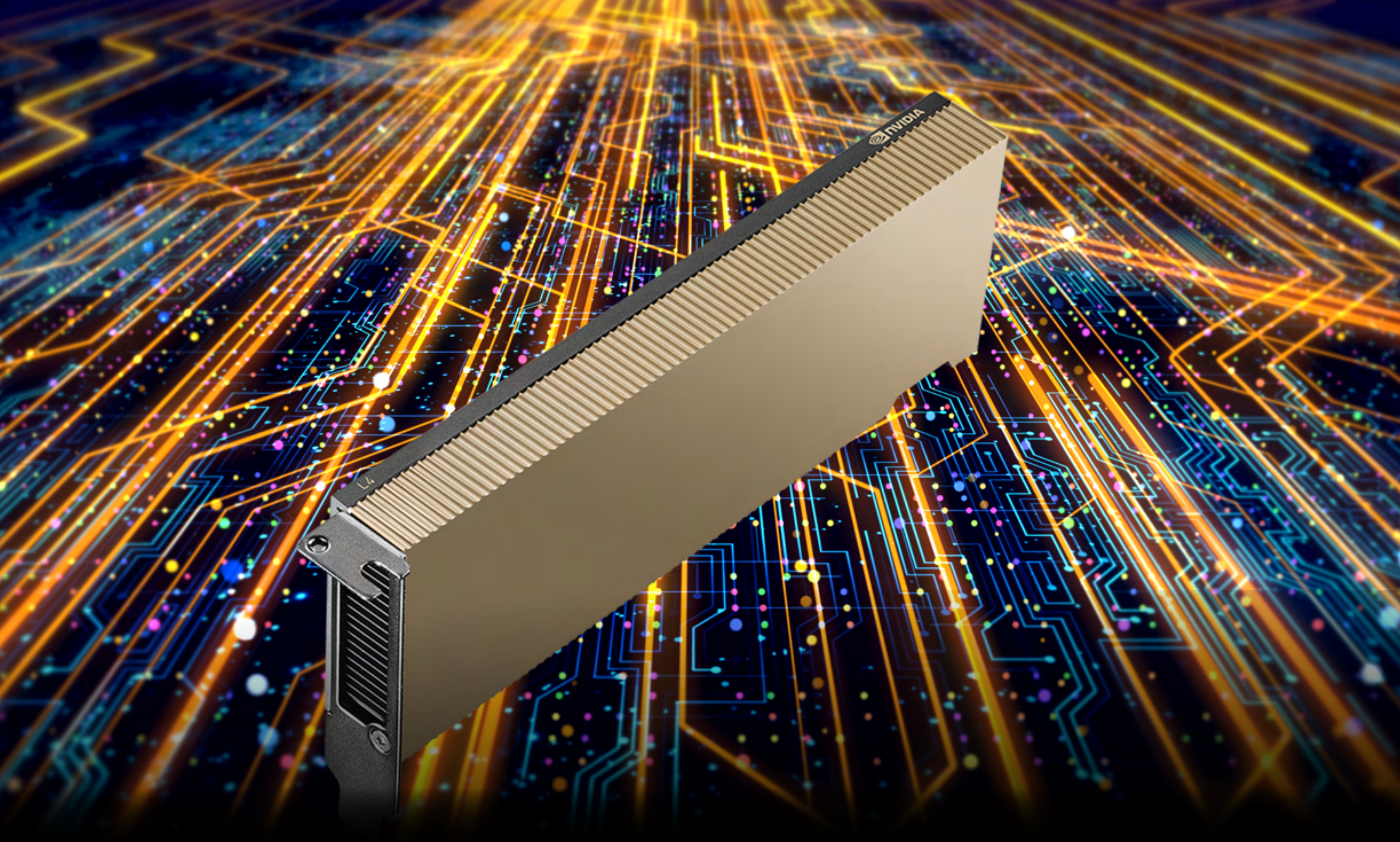


NVIDIA L4 GPU
Breakthrough Data Center Universal Accelerator for Efficient Video, AI, and Graphics



OVERVIEW

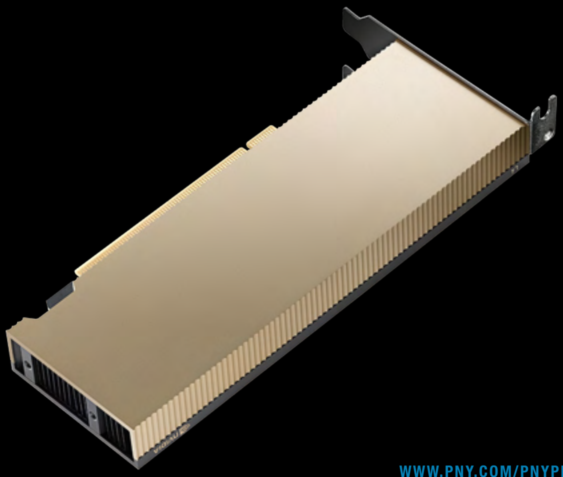
The NVIDIA Ada Lovelace L4 Tensor Core GPU delivers universal acceleration and energy efficiency for industry standard servers in the data center for video, artificial intelligence (AI), virtualized desktop, and graphics applications in the enterprise, the cloud, and at the edge. With NVIDIA's AI platform 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
- Contact center automation

The L4 GPU is NVIDIA's most efficient accelerator for mainstream data center use in PCIe-based servers with up to 120x higher AI video performance over central processing units (CPUs) and 2.7x more generative AI performance than CPU-based solutions. In addition, the L4 has over 4x the performance of the NVIDIA T4 GPU. This whitepaper spotlights how NVIDIA L4's versatility and energy-efficient, single-slot, low-profile design makes it ideal for global deployments, including edge locations.

NVIDIA L4 GPU FORM FACTOR AND FEATURES

The L4's efficient Small Form Factor (SFF) design fits into almost any server, making it an ideal upgrade for current installed base CPU-powered infrastructure. L4's passively cooled low-profile, single-slot design is optimized for mainstream deployment. Power consumption can be configured from a minimum of 40 W to a maximum of 72 W, with the latter as the default setting. The base GPU clock setting of 795 MHz can boost (temporarily) to 2,040 MHz depending on task and load requirements. L4 is compatible with PCIe Gen4 x16 or x8 slots, while PCIe Gen3 x16 is also implemented, along with support for lane and polarity reversal.



NVIDIA ADA LOVELACE ARCHITECTURE GPU ENHANCEMENTS

The L4 architecture is based on Ada's new generation SM (Streaming Multiprocessor) with 2x more FP32 operations per cycle per SM and up to 30.3 FP32 TFLOPS. The board is 2x more power efficient than prior generation Ampere architecture offerings and provides a larger L2 cache.

L4 GPU POWERED BY NVIDIA CORES

The L4 uses both 4th generation Tensor Cores and 3rd generation RT Cores which work in synergy to improve visual quality while reducing render times. For example, NVIDIA's AI denoising utilizes Tensor Cores to quickly 'fill in the blanks' after the RT Cores start rendering a scene.

4TH GENERATION TENSOR CORES

Ada Lovelace architecture 4th generation Tensor Cores are designed to accelerate transformative AI technologies. 4th generation Tensor Cores bring a new FP8 data format which can offer big boosts in performance. These enhanced Tensor Cores yield 120 TF32 TFLOPS, 242 FP16 TFLOPS, 485 FP8 TFLOPS, and 485 INT8 TOPS. These Tensor Core performance results are based on using fine-grained structural sparsity. These results make the L4 suitable for Generative AI and Natural Language Processing (NLP), while delivering up to 4x better performance than the NVIDIA T4's implementation. Tensor Cores work with 3rd generation RT Cores to deliver DLSS 3 (Deep Learning Super Sampling) to efficiently create high-resolution, photorealistic frames for imaging pipelines, and are also ideal for accelerating video analytics workflows.

3RD GENERATION RT CORES

NVIDIA RT Cores are specifically designed to tackle performance intensive real-time ray traced rendering. Ada Lovelace 3rd generation RT Cores have 2x the ray-triangle intersection throughout performance of the previous generation and use a displaced micro-mesh along with an opacity micro-map to deliver further performance and efficiency improvements. L4's RT cores use Shader Execution Reordering (SER), which is comparable in significance to out-of-order execution in central processing units (CPUs), and a first for GPUs. The L4 RT Cores deliver photorealistic real-time ray tracing in application viewports and works in conjunction with Tensor Cores to provide DLSS 3.

L4 ENERGY EFFICIENCY FEATURES

Enterprises' growing use of AI and video increases the demand for high performance, yet power efficient, cost-effective GPU accelerated computing. L4 GPUs provide high performance while lowering operating costs, including energy requirements and a reduced carbon footprint.

Adding L4 to existing data center infrastructure dramatically scales the number of users that can be supported. With the L4, up to 99% better energy efficiency and significantly lower total cost of ownership (TCO) can be realized over traditional CPU-based infrastructure.

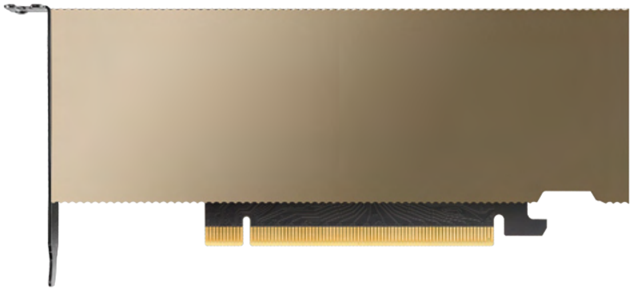
Using the L4 GPU reduces square footage required by racks, lowers networking hardware requirements, provides dramatically reduced power consumption, slashes HVAC costs, and delivers a significantly reduced carbon footprint. Energy saved by transitioning from CPUs to L4 GPUs in a 2MW data center can power over 2,000 homes for one year or the carbon offset from 172,000 trees grown over 10 years.

L4 UTILIZES NVIDIA AI ENTERPRISE SOFTWARE

The L4 GPU is an integral part of the NVIDIA data center platform. NVIDIA's data center platform accelerates over 3,000 applications and is available everywhere at scale, from data center to edge to cloud.

NVIDIA AI Enterprise software is a license addition for the L4 Tensor Core GPU. NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software optimized to help organizations excel at AI. Using NVIDIA AI Enterprise and L4 GPUs helps organizations simplify the building of an AI-ready platform, accelerates AI development and deployment, and delivers performance, security, and scalability.

The L4 GPU is optimized for 24/7 enterprise data center operations and is designed, built, and extensively tested by NVIDIA for maximum performance, durability, stability, and security. L4 implements secure boot with root-of trust technology, providing an additional layer of security for data centers – an essential requirement for mission-critical use cases, workflows, and IT infrastructure.



L4 HAS INCREASED MEMORY CAPACITY

L4 is enhanced with a larger GPU memory capacity of 24GB GDDR6 with optional ECC (Error Correction Code) which is 1.5x more memory than the NVIDIA T4. The default ECC setting is set to off but is user configurable. Other memory features include 251 MHz memory clock, 192-bit memory bus width, and 300 GB/sec peak memory bandwidth.

L4 VIDEO AND AVI FEATURES

L4 is enhanced with impressive video analytics, encoding, decoding, and transcoding features. It provides full AV1 support using 2x NVENC encode engines and 4x NVDEC decode engines along with fully symmetrical AV1 encode and decode support. AV1 is 40% more efficient than H.264 at full high definition (FHD) resolution. 4x JPEG decoders are also provided for static image processing tasks.

These capabilities enable the L4 to accelerate the latest transformative AI, graphics, and video applications including real-time video transcoding. Its data center ready video streaming engine is ideal for video conferencing stream encoding and implements hardware for full AI-enhanced video pipelines.

L4 can combine graphics, video, and be used for NLP inference. L4 can utilize generative AI and transformers for image and video generation, as well as chatbots and other NLP (Natural Language Processing) tasks. L4 is equally adept using JPEG decoders for computer vision use cases. In addition, L4 can support augmented reality (AR), enhanced reality (ER), virtual reality (VR), and extended reality (XR) applications.

An L4-based server can host over 1,000 concurrent video streams and provide over 120x more AI video end-to-end pipeline performance than CPU-based solutions. L4 is an ideal foundation for multi-platform streaming, or simultaneous broadcasting on more channels or across an ever-expanding array of social media apps.

L4 SUPPORTS GPU VIRTUALIZATION

Additional GPU memory (24GB) in the L4 GPU makes it ideal for virtual GPU (vGPU) use case scenarios. 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). Using the L4 GPU allows IT to quickly provision or deprovision users and shift data center assets to where they're needed most.

L4 HAS ACCESS TO NVIDIA'S AI-ACCELERATED SOFTWARE STACK

Enterprise adoption of AI is now mainstream, and organizations require AI-ready infrastructure. The L4 GPU is supported in NVIDIA's AI-accelerated software stack and tools including:

- NVIDIA CUDA-X AI is a collection of libraries and tools which allow dramatically higher performance across multiple application domains, from AI to data science
- NVIDIA OptiX™ Ray Tracing Engine is an application framework for achieving optimal ray tracing performance on the GPU
- NVIDIA Omniverse™ Enterprise with OptiX is used to build and operate metaverse applications, virtual worlds, and digital twins
- CloudXR™ is NVIDIA's solution for streaming virtual reality (VR), augmented reality (AR), and mixed reality (MR) content from OpenVR XR applications
- NVIDIA RTX vWS virtual workstation software lets users access a high-performance, GPU-accelerated virtual workstation from anywhere to aid in remote collaboration

The L4 GPU is also supported in a wide array of data science applications, software development kits (SDKs), and frameworks, including the assets available on NVIDIA's own NGC (containers). All major third-party virtualization hypervisors also support the L4 GPU for virtualized deployment of these solutions and tools.

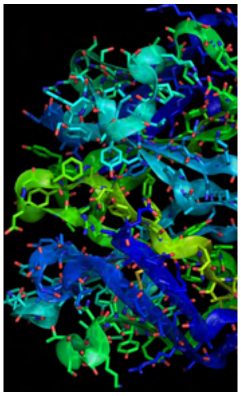
USE CASE: SPEAKING OF SPEECH

Conversational AI applications are now mainstream. Speech generates hundreds of billions of minutes of data every day. Online meetings generate 200 million minutes daily. Contact, Call, and Customer Service Centers generate 500 million minutes daily. Consumer applications generate 1.8 billion minutes daily. The NVIDIA L4 GPU provides up to 28x faster Natural Language Processing (NLP) than CPUs can achieve.

USE CASE: GPU ACCELERATED SCIENTIFIC MODELING AND SIMULATION

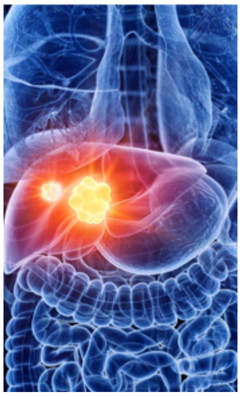
AI simulations are used in areas such as life sciences, radiology, genomics, weather and climate modeling, and particle physics.

Life Sciences



Protein Similarity Search

Radiology



Laser Electron Accelerators

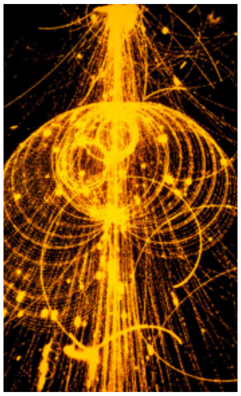
Genomics

Rapid DNA or RNA Sequencing

Weather and Climate

FourCastNet

Particle Physics



Beyond the Standard Model

L4 simulation performance is faster than CPU performance.

- Molecular Dynamics Simulations – AMBER software to simulate and analyze biomolecular interactions. One of the features of AMBER is the ability to use GPUs to massively accelerate these simulations, where the L4 is up to 46x faster vs CPU nodes
- Molecular Dynamics – NAMD (NAnoscale Molecular Dynamics) for high-performance simulation of large biomolecular systems : L4 is up to 13x faster vs CPU
- Fusion Physics GTC (Gyrokinetic Toroidal Code): L4 is up to 14x faster vs CPU

Summary

The NVIDIA Ada Lovelace architecture L4 Tensor Core GPU delivers universal acceleration and energy efficiency for video, AI, virtual workstations, and graphics in the enterprise, in the cloud, and at the edge. With NVIDIA's AI and graphics platforms and comprehensive full-stack software support, the NVIDIA L4 GPU is optimized for next-generation video and inference at scale for a broad range of AI applications to deliver the best in personalized experiences.

L4's high server density positions it to deliver massive performance improvements across all major use cases and deployment scenarios, to tackle ever more complex problems with greater speed and efficiency.

For additional information on the barrier-breaking NVIDIA L4 Tensor Core GPU, contact PNY at GOPNY@PNY.COM or visit our NVIDIA L4 product landing page.