NVIDIA Launches Inference Platforms for Large Language Models and Generative AI Workloads

Google Cloud, D-ID, Cohere Using New Platforms for Wide Range of Generative AI Services Including Chatbots, Text-to-Image Content, AI Video and More

GTC—NVIDIA today launched four inference platforms optimized for a diverse set of rapidly emerging generative AI applications — helping developers quickly build specialized, AI-powered applications that can deliver new services and insights.

The platforms combine NVIDIA's full stack of inference software with the latest NVIDIA Ada, NVIDIA Hopper™ and NVIDIA Grace Hopper™ processors — including the NVIDIA L4 Tensor Core GPU and the NVIDIA H100 NVL GPU, both launched today. Each platform is optimized for in-demand workloads, including AI video, image generation, large language model deployment and recommender inference.

"The rise of generative AI is requiring more powerful inference computing platforms," said Jensen Huang, founder and CEO of NVIDIA. "The number of applications for generative AI is infinite, limited only by human imagination. Arming developers with the most powerful and flexible inference computing platform will accelerate the creation of new services that will improve our lives in ways not yet imaginable."

Accelerating Generative AI's Diverse Set of Inference Workloads
Each of the platforms contains an NVIDIA GPU optimized for specific generative AI inference workloads as well as specialized software:

- **NVIDIA L4 for AI Video** can deliver 120x more AI-powered video performance than CPUs, combined with 99% better energy efficiency. Serving as a universal GPU for virtually any workload, it offers enhanced video decoding and transcoding capabilities, video streaming, augmented reality, generative AI video and more.
- **NVIDIA L40 for Image Generation** is optimized for graphics and AI-enabled 2D, video and 3D image generation. The L40 platform serves as the engine of NVIDIA Omniverse™, a platform for building and operating metaverse applications in the data center, delivering 7x the inference performance for Stable Diffusion and 12x Omniverse performance over the previous generation.
- **NVIDIA H100 NVL for Large Language Model Deployment** is ideal for deploying massive LLMs like ChatGPT at scale. The new H100 NVL with 94GB of memory with Transformer Engine acceleration delivers up to 12x faster inference performance at GPT-3 compared to the prior generation A100 at data center scale.
- **NVIDIA Grace Hopper for Recommendation Models** is ideal for graph recommendation models, vector databases and graph neural networks. With the 900 GB/s NVLink®-C2C connection between CPU and GPU, Grace Hopper can deliver 7x faster data transfers and queries compared to PCIe Gen 5.

The platforms' software layer features the NVIDIA AI Enterprise software suite, which includes NVIDIA TensorRT™, a software development kit for high-performance deep learning inference, and NVIDIA Triton Inference Server™, an open-source inference-serving software that helps standardize model deployment.

Early Adoption and Support
Google Cloud is a key cloud partner and an early customer of NVIDIA's inference platforms. It is integrating the L4 platform into its machine learning platform, Vertex AI, and is the first cloud service provider to offer L4 instances, with private preview of its G2 virtual machines launching today.

Two of the first organizations to have early access to L4 on Google Cloud include: Descript, which uses generative AI to help creators produce videos and podcasts, and WOMBO, which offers an AI-powered text to digital art app called Dream.

Another early adopter, Kuaishou provides a content community and social platform that leverages GPUs to decode incoming live streaming video, capture key frames, optimize audio and video. It then uses a transformer-based large-scale model to understand multimodal content and improve click-through rates for hundreds of millions of users globally.

"Kuaishou recommendation system serves a community having over 360 million daily users who contribute millions of UGC videos every day," said Yue Yu, senior vice president at Kuaishou. "Compared to CPUs under the same total cost of ownership, NVIDIA GPUs have been increasing the system end-to-end throughputs by 11x and reducing latency by 20%."

D-ID, a leading generative AI technology platform, elevates video content for professionals by using NVIDIA L40 GPUs to generate photorealistic digital humans from text — giving a face to any content while reducing the cost and hassle of video production at scale.

“L40 performance was simply amazing. With it, we were able to double our inference speed,” said Or Gorodissky, vice president of research and development at D-ID. “D-ID is excited to use this new hardware as part of our offering that enables real-time streaming of AI humans at unprecedented performance and resolution while simultaneously reducing our compute costs.”

Seyhan Lee, a leading AI production studio, uses generative AI to develop immersive experiences and captivating creative content for the film, broadcast and entertainment industries.

“The L40 GPU delivers an incredible boost in performance for our generative AI applications,” said Pinar Demirdag, co-founder of Seyhan Lee. “With the inferencing capability and memory size of the L40, we can deploy state-of-the-art models and deliver innovative services to our customers with incredible speed and accuracy.”

Cohere, a leading pioneer in language AI, runs a platform that empowers developers to build natural language models while keeping data private and secure.

“NVIDIA’s new high-performance H100 inference platform can enable us to provide better and more efficient services to our customers with our state-of-the-art generative models, powering a variety of NLP applications such as conversational AI, multilingual enterprise search and information extraction,” said Aidan Gomez, CEO at Cohere.

Availability
The NVIDIA L4 GPU is available in private preview on Google Cloud Platform and also available from a global network of more than 30 computer makers, including Advantech, ASUS, Atos, Cisco, Dell Technologies, Fujitsu, GIGABYTE, Hewlett Packard Enterprise, Lenovo, QCT and Supermicro.

The NVIDIA L40 GPU is currently available from leading system builders, including ASUS, Dell Technologies, GIGABYTE, Hewlett Packard Enterprise, Lenovo and Supermicro with the number of partner platforms set to expand throughout the year.

The Grace Hopper Superchip is sampling now, with full production expected in the second half of the year. The H100 NVL GPU also is expected in the second half of the year.

NVIDIA AI Enterprise is now available on major cloud marketplaces and from dozens of system providers and partners. With NVIDIA AI Enterprise, customers receive NVIDIA Enterprise Support, regular security reviews and API stability for NVIDIA Triton Inference Server, TensorRT and more than 50 pretrained models and frameworks.

Hands-on labs for trying the NVIDIA inference platform for generative AI are available immediately at no cost on NVIDIA LaunchPad. Sample labs include training and deploying a support chatbot, deploying an end-to-end AI workload, tuning and deploying a language model on H100 and deploying a fraud detection model with NVIDIA Triton™.

About NVIDIA
Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company’s invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling the creation of the metaverse. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, availability and performance of our products and technologies, including NVIDIA Ada, Hopper and Grace Hopper processors, NVIDIA L4 Tensor Core GPU, NVIDIA H100 NVL GPU, NVIDIA L4, NVIDIA L40, NVIDIA Omniverse, NVIDIA AI Enterprise, NVIDIA TensorRT, NVIDIA Triton Inference Server and NVIDIA LaunchPad; rise of generative AI requiring more powerful inference computing platforms; the benefits, impact, performance, availability and progress of collaboration with Google Cloud; the benefits, impact and performance of our products and technologies, including L4 and L40 GPU, as used by third parties, including Descript, WOMBO, Kuaishou, D-ID, Seyhan Lee and Cohere, are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners’ products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company’s website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

companies with which they are associated. Features, pricing, availability, and specifications are subject to change without notice.

Cliff Edwards
NVIDIA Corporation
+1-415-699-2755
cliffe@nvidia.com