NVIDIA and Google Cloud Deliver Powerful New Generative AI Platform, Built on the New L4 GPU and Vertex AI

NVIDIA Inference Platform for Generative AI to Be Integrated Into Google Cloud Vertex AI; Google Cloud First CSP to Make NVIDIA L4 GPU Instances Available

GTC—NVIDIA today announced Google Cloud is integrating the newly launched L4 GPU and Vertex AI to accelerate the work of companies building a rapidly expanding number of generative AI applications.

Google Cloud, with its announcement of G2 virtual machines available in private preview today, is the first cloud services provider to offer NVIDIA's L4 Tensor Core GPU. Additionally, L4 GPUs will be available with optimized support on Vertex AI, which now supports building, tuning and deploying large generative AI models.

Developers can access the latest state-of-the-art technology available to help them get new applications up and running quickly and cost-efficiently. The NVIDIA L4 GPU is a universal GPU for every workload, with enhanced AI video capabilities that can deliver 120x more AI-powered video performance than CPUs, combined with 99% better energy efficiency.

"Surging interest in generative AI is inspiring a wave of companies to turn to cloud-based computing to support their business models," said Jensen Huang, founder and CEO of NVIDIA. "We are working with Google Cloud to help ensure that the capabilities they require are easily available and able to help fuel the incredible new tools and applications they will create."

"Generative AI represents a new era of computing — one that demands the speed, scalability and reliability we provide on Google Cloud," said Amin Vahdat, vice president of Systems & Services Infrastructure at Google Cloud. "As our customers begin to explore the possibilities of generative AI, we're proud to offer them NVIDIA's latest L4 GPU innovation as part of our workload-optimized Compute Engine portfolio."

Helping New Generative AI Applications Come to Life
Google Cloud provides the infrastructure for a wide variety of organizations offering generative AI applications, many of which are designed to help professionals do their work better and faster. Rapid inference is key to successfully running their applications.

Generative AI is also driving a number of new apps that help people connect and have fun. WOMBO, which offers an AI-powered text to digital art app called Dream, has had early access to NVIDIA's L4 inference platform on Google Cloud.

"WOMBO relies upon the latest AI technology for people to create immersive digital artwork from users' prompts, letting them create high-quality, realistic art in any style with just an idea," said Ben-Zion Benkhin, CEO at WOMBO. "NVIDIA's L4 inference platform will enable us to offer a better, more efficient image-generation experience for users seeking to create and share unique artwork."

Descript offers AI-powered editing features that let creators remove filler words, add captions and make social-media clips in a few clicks. Or they can use Descript's generative-AI voice cloning to fix audio mistakes — even create entire voiceover tracks — just by typing.

"Descript uses NVIDIA TensorRT to optimize models to accelerate AI inferencing," said Andrew Mason, CEO of Descript. "It allows users to replace their video backgrounds and enhance their speech to produce studio-quality content, without the studio."

Availability
NVIDIA L4 GPUs are available in private preview on Google Cloud. Apply for access.

Watch Huang discuss the integration of NVIDIA's inference platform for generative AI into Google Cloud in his GTC keynote.

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 L4 GPU, inference platform based on NVIDIA L4 GPU, and NVIDIA TensorRT; the benefits, impact, performance, availability and progress of integration of NVIDIA L4 GPU and inference platform based on NVIDIA L4 GPU by and collaboration with Google Cloud; surging interest in generative AI inspiring a wave of companies to turn to cloud-based computing to support their business model; and the benefits, impact and performance of our products and technologies, including NVIDIA Tensor RT and inference platform based on NVIDIA L4 GPU, as used by third parties, including WOMBO and Descript, 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.

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