Google Cloud and NVIDIA Take Collaboration to the Next Level

AI leaders work to optimize Google Cloud to enable more generative AI startups to build next-generation applications.

Author: Dave Salvator

As generative AI and large language models (LLMs) continue to drive innovations, compute requirements for training and inference have grown at an astonishing pace.

To meet that need, Google Cloud today announced the general availability of its new A3 instances, powered by NVIDIA H100 Tensor Core GPUs. These GPUs bring unprecedented performance to all kinds of AI applications with their Transformer Engine — purpose-built to accelerate LLMs .

Availability of the A3 instances comes on the heels of NVIDIA being named Google Cloud's Generative AI Partner of the Year — an award that recognizes the companies' deep and ongoing collaboration to accelerate generative AI on Google Cloud.

The joint effort takes multiple forms, from infrastructure design to extensive software enablement, to make it easier to build and deploy AI applications on the Google Cloud platform.

At the Google Cloud Next conference, NVIDIA founder and CEO Jensen Huang joined Google Cloud CEO Thomas Kurian for the event keynote to celebrate the general availability of NVIDIA H100 GPU-powered A3 instances and speak about how Google is using NVIDIA H100 and A100 GPUs for internal research and inference in its DeepMind and other divisions.

During the discussion, Huang pointed to the deeper levels of collaboration that enabled NVIDIA GPU acceleration for the PaxML framework for creating massive LLMs. This Jax-based machine learning framework is purpose-built to train large-scale models, allowing advanced and fully configurable experimentation and parallelization.

PaxML has been used by Google to build internal models, including DeepMind as well as research projects, and will use NVIDIA GPUs. The companies also announced that PaxML is available immediately on the NVIDIA NGC container registry.

Today, there are over a thousand generative AI startups building next-generation applications, many using NVIDIA technology on Google Cloud. Some notable ones include Writer and Runway.

Writer uses transformer-based LLMs to enable marketing teams to quickly create copy for web pages, blogs, ads and more. To do this, the company harnesses NVIDIA NeMo , an application framework from NVIDIA AI Enterprise that helps companies curate their training datasets, build and customize LLMs, and run them in production at scale.

Using NeMo optimizations, Writer developers have gone from working with models with hundreds of millions of parameters to 40-billion parameter models. The startup's customer list includes household names like Deloitte, L'Oreal, Intuit, Uber and many other Fortune 500 companies.

Runway uses AI to generate videos in any style. The AI model imitates specific styles prompted by given images or through a text prompt. Users can also use the model to create new video content using existing footage. This flexibility enables filmmakers and content creators to explore and design videos in a whole new way.

Google Cloud was the first CSP to bring the NVIDIA L4 GPU to the cloud. In addition, the companies have collaborated to enable Google's Dataproc service to leverage the RAPIDS Accelerator for Apache

Spark to provide significant performance boosts for ETL, available today with Dataproc on the Google Compute Engine and soon for Serverless Dataproc.

The companies have also made NVIDIA AI Enterprise available on Google Cloud Marketplace and integrated NVIDIA acceleration software into the Vertex AI development environment.

Find more details about NVIDIA GPU instances on Google Cloud and how NVIDIA is powering generative AI, and see how organizations are running their mission-critical enterprise applications with NVIDIA NeMo on the GPU-accelerated Google Cloud .

Sign up for generative AI news to stay up to date on the latest breakthroughs, developments and technologies.

Original URL: https://blogs.nvidia.com/blog/2023/08/29/google-cloud-collaboration/