Google Cloud and NVIDIA Expand Partnership to Advance AI Computing, Software and Services

NVIDIA Generative AI Technology Used by Google DeepMind and Google Research Teams Now Optimized and Available to Google Cloud Customers Worldwide

Google Cloud Next — Google Cloud and NVIDIA today announced new AI infrastructure and software for customers to build and deploy massive models for generative AI and speed data science workloads.

In a fireside chat at Google Cloud Next, Google Cloud CEO Thomas Kurian and NVIDIA founder and CEO Jensen Huang discussed how the partnership is bringing end-to-end machine learning services to some of the largest AI customers in the world — including by making it easy to run AI supercomputers with Google Cloud offerings built on NVIDIA technologies. The new hardware and software integrations utilize the same NVIDIA technologies employed over the past two years by Google DeepMind and Google research teams.

“We’re at an inflection point where accelerated computing and generative AI have come together to speed innovation at an unprecedented pace,” Huang said. “Our expanded collaboration with Google Cloud will help developers accelerate their work with infrastructure, software and services that supercharge energy efficiency and reduce costs.”

“Google Cloud has a long history of innovating in AI to foster and speed innovation for our customers,” Kurian said. “Many of Google’s products are built and served on NVIDIA GPUs, and many of our customers are seeking out NVIDIA accelerated computing to power efficient development of LLMs to advance generative AI.”

NVIDIA Integrations to Speed AI and Data Science Development
Google’s framework for building massive large language models (LLMs), PaxML, is now optimized for NVIDIA accelerated computing.

Originally built to span multiple Google TPU accelerator slices, PaxML now enables developers to use NVIDIA® H100 and A100 Tensor Core GPUs for advanced and fully configurable experimentation and scale. A GPU-optimized PaxML container is available immediately in the NVIDIA NGC™ software catalog. In addition, PaxML runs on JAX, which has been optimized for GPUs leveraging the OpenXLA compiler.

Google DeepMind and other Google researchers are among the first to use PaxML with NVIDIA GPUs for exploratory research.

The NVIDIA-optimized container for PaxML will be available immediately on the NVIDIA NGC container registry to researchers, startups and enterprises worldwide that are building the next generation of AI-powered applications.

Additionally, the companies announced Google’s integration of serverless Spark with NVIDIA GPUs through Google’s Dataproc service. This will help data scientists speed Apache Spark workloads to prepare data for AI development.

These new integrations are the latest in NVIDIA and Google’s extensive history of collaboration. They cross hardware and software announcements, including:

- **Google Cloud on A3 virtual machines powered by NVIDIA H100** — Google Cloud announced today its purpose-built Google Cloud A3 VMs powered by NVIDIA H100 GPUs will be generally available next month, making NVIDIA’s AI platform more accessible for a broad set of workloads. Compared to the previous generation, A3 VMs offer 3x faster training and significantly improved networking bandwidth.
- **NVIDIA H100 GPUs to power Google Cloud’s Vertex AI platform** — H100 GPUs are expected to be generally available on VertexAI in the coming weeks, enabling customers to quickly develop generative AI LLMs.
- **Google Cloud to gain access to NVIDIA DGX™ GH200** — Google Cloud will be one of the first companies in the world to have access to the NVIDIA DGX GH200 AI supercomputer — powered by the NVIDIA Grace Hopper™ Superchip — to explore its capabilities for generative AI workloads.
- **NVIDIA DGX Cloud Coming to Google Cloud** — NVIDIA DGX Cloud AI supercomputing and software will be available to customers directly from their web browser to provide speed and scale for advanced training workloads.
- **NVIDIA AI Enterprise on Google Cloud Marketplace** — Users can access NVIDIA AI Enterprise, a secure, cloud native software platform that simplifies developing and deploying enterprise-ready applications including generative AI, speech AI, computer vision, and more.
- **Google Cloud first to offer NVIDIA L4 GPUs** — Earlier this year, Google Cloud became the first cloud provider to offer NVIDIA L4 Tensor Core GPUs with the launch of the G2 VM. NVIDIA customers switching to L4 GPUs from CPUs for AI video workloads can realize up to 120x higher performance with 99% better efficiency. L4 GPUs are used widely for image and text generation, as well as VDI and AI-accelerated audio/video transcoding.

About Google Cloud
Google Cloud accelerates every organization's ability to digitally transform its business and industry. We deliver enterprise-grade solutions that leverage Google's cutting-edge technology, and tools that help developers build more sustainably. Customers in more than 200 countries and territories turn to Google Cloud as their trusted partner to enable growth and solve their most critical business problems.

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 industrial digitalization across markets. 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, performance, features and availability of our products and technologies, including NVIDIA GPUs, NVIDIA accelerated computing, NVIDIA H100 and A100 Tensor Core GPUs, the NVIDIA DGX GH200 AI supercomputer, NVIDIA DGX Cloud, NVIDIA AI Enterprise and NVIDIA L4 Tensor Core GPUs; NVIDIA's partnership with Google Cloud, including the benefits, impact, features and availability of Google Cloud offerings built on NVIDIA technologies; the inflection point where accelerated computing and generative AI have come together to speed innovation at an unprecedented pace; and NVIDIA's expanded collaboration with Google Cloud helping developers accelerate their work with infrastructure, software and services that supercharge energy efficiency and reduce costs 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.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements within are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

Cliff Edwards
NVIDIA Corporation
+1-415-699-2755
cliffe@nvidia.com
Google Cloud PR
press@google.com