Google Cloud Press Corner

Google Cloud and NVIDIA Expand Partnership to Scale AI Development

New AI infrastructure offerings and integrations enable more open and accessible AI

SAN JOSE, Calif., March 18, 2024 /PRNewswire/ -- **GTC --** Google Cloud and NVIDIA today announced a deepened partnership to enable the machine learning (ML) community with technology that accelerates their efforts to easily build, scale and manage generative AI applications.

To continue bringing AI breakthroughs to its products and developers, Google announced its adoption of the new NVIDIA Grace Blackwell AI computing platform, as well as the NVIDIA DGX Cloud service on Google Cloud. Additionally, the NVIDIA H100-powered DGX Cloud platform is now generally available on Google Cloud.

Building on their recent collaboration to optimize theGemma family of open models, Google also will adoptNVIDIA NIM inference microservices to provide developers with an open, flexible platform to train and deploy using their preferred tools and frameworks. The companies also announced support for JAX on NVIDIA GPUs and Vertex AI instances powered by NVIDIA H100 and L4 Tensor Core GPUs.

"The strength of our long-lasting partnership with NVIDIA begins at the hardware level and extends across our portfolio - from state-of-the-art GPU accelerators, to the software ecosystem, to our managed Vertex AI platform," said Google Cloud CEO Thomas Kurian. "Together with NVIDIA, our team is committed to providing a highly accessible, open and comprehensive AI platform for ML developers."

"Enterprises are looking for solutions that empower them to take full advantage of generative AI in weeks and months instead of years," said Jensen Huang, founder and CEO of NVIDIA. "With expanded infrastructure offerings and new integrations with NVIDIA's full-stack AI, Google Cloud continues to provide customers with an open, flexible platform to easily scale generative AI applications."

The new integrations between NVIDIA and Google Cloud build on the companies' longstanding commitment to providing the AI community with leading capabilities at every layer of the AI stack. Key components of the partnership expansion include:

- **Adoption of NVIDIA Grace Blackwell:** The new Grace Blackwell platform enables organizations to build and run real-time inference on trillion-parameter large language models. Google is adopting the platform for various internal deployments and will be one of the first cloud providers to offer Blackwell-powered instances.
- **Grace Blackwell-powered DGX Cloud coming to Google Cloud:** Google will bring NVIDIA GB200 NVL72 systems, which combine 72 Blackwell GPUs and 36 Grace CPUs interconnected by fifth-generation NVLink, to its highly scalable and performant cloud infrastructure. Designed for energy-efficient training and inference in an era of trillion-parameter LLMs, it will also be available via DGX Cloud, an AI platform offering a serverless experience for enterprise developers building and serving LLMs. DGX Cloud is now generally available on Google Cloud A3 VM instances powered by NVIDIA H100 Tensor Core GPUs.
- **Support for JAX on GPUs:** Google Cloud and NVIDIA collaborated to bring the advantages of JAX to NVIDIA GPUs, widening access to large-scale LLM training among the broader ML community. JAX is a framework for high-performance machine learning that is compiler-oriented and Python-native, an easy to use and performant framework for LLM training. AI practitioners can now use JAX with NVIDIA H100 GPUs on Google Cloud through MaxText and Accelerated Processing Kit (XPK).
- **NVIDIA NIM on Google Kubernetes Engine (GKE):** NVIDIA NIM inference microservices, a part of the NVIDIA AI Enterprise software platform, will be integrated into GKE. Built on inference engines including TensorRT-LLM, NIM helps speed up generative AI deployment in enterprises, supports a wide range of leading AI models and ensures seamless, scalable AI inferencing.
- **Support for NVIDIA NeMo:** Google Cloud has made it easier to deploy the NVIDIA NeMo framework across its platform via Google Kubernetes Engine (GKE) and Google Cloud HPC Toolkit. This enables developers to automate and scale the training and serving of generative AI models, and it allows them to rapidly deploy turnkey environments through customizable blueprints that jump-start the development process. NVIDIA NeMo, part of NVIDIA AI Enterprise, is also available in the Google Marketplace, providing customers with another way to easily access NeMo and other frameworks to accelerate AI development.
- **Vertex AI and Dataflow expand support for NVIDIA GPUs:** To advance data science and analytics, Vertex AI now supports Google Cloud A3 VMs powered by NVIDIA H100 GPUs and G2 VMs powered by NVIDIA L4 Tensor Core GPUs. This provides MLOps teams with scalable infrastructure and tooling to confidently manage and deploy AI applications. Dataflow has also expanded support for accelerated data processing on NVIDIA GPUs.

Google Cloud has long offered GPU VM instances powered by NVIDIA's cutting-edge hardware coupled with leading Google innovations. NVIDIA GPUs are a core component of the Google Cloud AI Hypercomputer - a supercomputing architecture that unifies performance-optimized hardware, open software, and flexible consumption models. The holistic partnership enables AI researchers, scientists, and developers to train, fine-tune, and serve the largest and most sophisticated AI models - now with even more of their favorite tools and frameworks jointly optimized and available on Google Cloud.

"Runway's text-to-video platform is powered by AI Hypercomputer. At the base, A3 VMs, powered by NVIDIA H100 GPUs gave our training a significant performance boost over A2 VMs, enabling large-scale training and inference for our Gen-2 model. Using GKE to orchestrate our training jobs enables us to scale to thousands of H100 GPUs in a single fabric to meet our customers' growing demand."

- *Anastasis Germanidis, CTO and Co-Founder of Runway*

"By moving to Google Cloud and leveraging AI Hypercomputer architecture with NVIDIA T4 GPUs, G2 VMs powered by NVIDIA L4 GPUs and Triton Inference Server, we saw a significant boost in our model inference performance while lowering our hosting costs 15% using novel techniques enabled by the flexibility that Google Cloud offers."

- *Ashwin Kannan, Sr Staff Machine Learning Engineer, Palo Alto Networks*

"Writer's platform all comes together through this extremely productive partnership with Google and NVIDIA. We're able to use NVIDIA GPUs optimally for training and inference. We leverage NVIDIA NeMo to build our industrial-strength models, which generate 990,000 words a second with over a trillion API calls per month. We're delivering the highest quality models that exceed those from companies with larger teams and bigger budgets – and all of that is possible with the Google and NVIDIA partnership. The benefits of their AI expertise are passed down to our enterprise customers, who can build meaningful AI workflows in days, not months or years."

- *Danny Leung, Director of Alliances, Writer*

Learn more about Google Cloud's collaboration with NVIDIA at GTC, the global AI conference,March 18-21 (booth #808).

Additional Resources

- Accelerate your generative AI journey using NVIDIA NeMo on Google Kubernetes Engine
- Announcing Cloud HPC Toolkit blueprint for AI/ML with NeMo on A3 VMs
- Keep up with the latest Google Cloud news on our newsroom and blog.
- Learn more about NVIDIA GPUs on Google Cloud here.
- Read about Google Cloud and NVIDIA's work together here.

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 infrastructure company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

About Google Cloud
Google Cloud is the new way to the cloud, providing AI, infrastructure, developer, data, security, and collaboration tools built for today and tomorrow. Google Cloud offers a powerful, fully integrated and optimized AI stack with its own planet-scale infrastructure, custom-built chips, generative AI models and development platform, as well as AI-powered applications, to help organizations transform. Customers in more than 200 countries and territories turn to Google Cloud as their trusted technology partner.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, features, and availability of NVIDIA's products and technologies, including NVIDIA Grace Blackwell platform, NVIDIA DGX Cloud, NVIDIA NIM, NVIDIA H100 GPUs, NVIDIA L4 Tensor Core GPUs, NVIDIA GB200 NVL72 systems, NVLink, NVIDIA AI Enterprise software platform, and NVIDIA NeMo; the benefits and impact of NVIDIA's partnership with Google Cloud, and the features and availability of its services and offerings; and enterprises looking for solutions that empower them to take full advantage of generative AI in weeks and months instead of years 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; NVIDIA's reliance on third parties to manufacture, assemble, package and test its products; the impact of technological development and competition; development of new products and technologies or enhancements to NVIDIA's existing product and technologies; market acceptance of NVIDIA's products or its partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of NVIDIA's 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.

SOURCE Google Cloud

For further information: press@google.com; Cliff Edwards, NVIDIA Enterprise PR, +1.415.699.2755
