NVIDIA Launches DGX Cloud, Giving Every Enterprise Instant Access to AI Supercomputer From a Browser

Oracle Cloud Infrastructure First to Run NVIDIA AI Supercomputing Instances; Microsoft Azure, Google Cloud and Others to Host DGX Cloud Soon

GTC—NVIDIA today announced NVIDIA DGX™ Cloud, an AI supercomputing service that gives enterprises immediate access to the infrastructure and software needed to train advanced models for generative AI and other groundbreaking applications.

DGX Cloud provides dedicated clusters of NVIDIA DGX AI supercomputing, paired with NVIDIA AI software. The service makes it possible for every enterprise to access its own AI supercomputer using a simple web browser, removing the complexity of acquiring, deploying and managing on-premises infrastructure.

Enterprises rent DGX Cloud clusters on a monthly basis, which ensures they can quickly and easily scale the development of large, multi-node training workloads without having to wait for accelerated computing resources that are often in high demand.

“We are at the iPhone moment of AI. Startups are racing to build disruptive products and business models, and incumbents are looking to respond,” said Jensen Huang, founder and CEO of NVIDIA. “DGX Cloud gives customers instant access to NVIDIA AI supercomputing in global-scale clouds.”

NVIDIA is partnering with leading cloud service providers to host DGX Cloud infrastructure, starting with Oracle Cloud Infrastructure (OCI). Its OCI Supercluster provides a purpose-built RDMA network, bare-metal compute and high-performance local and block storage that can scale to superclusters of over 32,000 GPUs.

Microsoft Azure is expected to begin hosting DGX Cloud next quarter, and the service will soon expand to Google Cloud and more.

Industry Titans Adopt NVIDIA DGX Cloud to Speed Success
Amgen, one of the world’s leading biotechnology companies, insurance technology leader CCC Intelligent Solutions (CCC), and digital-business-platform provider ServiceNow are among the first AI pioneers using DGX Cloud.

Amgen is using DGX Cloud with NVIDIA BioNeMo™ large language model software to accelerate drug discovery, including NVIDIA AI Enterprise software, which includes NVIDIA RAPIDS™ data science acceleration libraries.

“With NVIDIA DGX Cloud and NVIDIA BioNeMo, our researchers are able to focus on deeper biology instead of having to deal with AI infrastructure and set up ML engineering,” said Peter Grandsard, executive director of Research, Biologics Therapeutic Discovery, Center for Research Acceleration by Digital Innovation at Amgen. “The powerful computing and multi-node capabilities of DGX Cloud have enabled us to achieve 3x faster training of protein LLMs with BioNeMo and up to 100x faster post-training analysis with NVIDIA RAPIDS relative to alternative platforms.”

CCC, a leading cloud platform for the property and casualty insurance economy, is using DGX Cloud to speed and scale the development and training of its AI models. These models power the company’s innovative auto claims resolution solutions, helping to accelerate the intelligent automation of the industry and improve the claims experience for millions of business users and their consumers every day.

ServiceNow is using DGX Cloud with on-premises NVIDIA DGX supercomputers for flexible, scalable hybrid-cloud AI supercomputing that helps power its AI research on large language models, code generation, and causal analysis. ServiceNow also co-stewards the BigCode project, a responsible open-science generative AI initiative, which is trained on the Megatron-LM framework from NVIDIA.

Open a Browser to NVIDIA AI Supercomputing and Software
Enterprises manage and monitor DGX Cloud training workloads using NVIDIA Base Command™ Platform software, which provides a seamless user experience across DGX Cloud, as well as on-premises NVIDIA DGX supercomputers. Using Base Command Platform, customers can match their workloads to the right amount and type of DGX infrastructure needed for each job.

DGX Cloud includes NVIDIA AI Enterprise, the software layer of the NVIDIA AI platform, which provides end-to-end AI frameworks and pretrained models to accelerate data science pipelines and streamline the development and deployment of production AI. New pretrained models, optimized frameworks and accelerated data science software libraries, available in NVIDIA AI Enterprise 3.1 released today, give developers an additional jump-start to their AI projects.

Each instance of DGX Cloud features eight NVIDIA H100 or A100 80GB Tensor Core GPUs for a total of 640GB of GPU memory per node. A high-performance, low-latency fabric built with NVIDIA Networking ensures workloads can scale across clusters of interconnected systems, allowing multiple instances to act as one massive GPU to meet the performance requirements of advanced AI training. High-performance storage is integrated into DGX Cloud to provide a complete solution for AI supercomputing.

DGX Cloud features support from NVIDIA experts throughout the AI development pipeline. Customers can work directly with NVIDIA engineers to optimize their models and quickly resolve development challenges across a broad range of industry use cases.

Availability
DGX Cloud instances start at $36,999 per instance per month. Organizations can contact their NVIDIA Partner Network representative for additional details.

Watch Huang discuss NVIDIA DGX Cloud in his GTC keynote on demand, and tune in to the GTC panel with NVIDIA DGX Cloud pioneers.

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, specifications, and availability of our products, technologies, and services, including DGX Cloud, NVIDIA DGX, NVIDIA AI software, NVIDIA BioNeMo, NVIDIA AI Enterprise, NVIDIA RAPIDS, NVIDIA Base Command, and NVIDIA GPUs; the iPhone moment of AI; startups racing to build disruptive products and business models, and incumbents looking to respond; leading cloud service providers hosting DGX Cloud infrastructure; and third parties using DGX Cloud, including the benefits and impact thereof, 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.

Shannon McPhee
+1-310-920-9642
smcphee@nvidia.com