NVIDIA DGX Cloud Now Available to Supercharge Generative AI Training

Author: Tony Paikeday

NVIDIA DGX Cloud — which delivers tools that can turn nearly any company into an AI company — is now broadly available, with thousands of NVIDIA GPUs online on Oracle Cloud Infrastructure , as well as NVIDIA infrastructure located in the U.S. and U.K.

Unveiled at NVIDIA's GTC conference in March, DGX Cloud is 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.

"Generative AI has made the rapid adoption of AI a business imperative for leading companies in every industry, driving many enterprises to seek more accelerated computing infrastructure," said Pat Moorhead, chief analyst at Moor Insights & Strategy.

Generative AI could add more than $4 trillion to the economy annually, turning proprietary business knowledge across a vast swath of the world's industries into next-generation AI applications, according to recent estimates by global management consultancy McKinsey .

Nearly every industry can benefit from generative AI, with early pioneers already leading transformative change across their markets.

Healthcare companies use DGX Cloud to generate protein models to speed drug discovery and clinical reporting with natural language processing.

Financial service providers use DGX Cloud to forecast trends, optimize portfolios, build recommender systems and develop intelligent generative AI chatbots.

Insurance companies are building models to automate claims processing.

Software companies are using it to develop AI-powered features and applications.

And others are using DGX Cloud to build AI factories and digital twins of valuable assets.

DGX Cloud instances provide dedicated infrastructure enterprises rent on a monthly basis, ensuring customers can quickly and easily develop large, multi-node training workloads without having to wait for accelerated computing resources that are often in high demand.

"The availability of NVIDIA DGX Cloud provides a new pool of AI supercomputing resources, with nearly instantaneous access," Moorhead said.

This simple approach to AI supercomputing removes the complexity of acquiring, deploying and managing on-premises infrastructure. Providing NVIDIA DGX AI supercomputing paired with NVIDIA AI Enterprise software, DGX Cloud makes it possible for businesses everywhere to access their own AI supercomputer using a web browser.

Each instance of DGX Cloud features eight NVIDIA 80GB Tensor Core GPUs for 640GB of GPU memory per node. A high-performance, low-latency fabric ensures workloads can scale across clusters of interconnected systems, allowing multiple instances to act as one massive GPU. High-performance storage is integrated into DGX Cloud to provide a complete solution.

Enterprises manage and monitor DGX Cloud training workloads using NVIDIA Base Command Platform software. The platform provides a seamless user experience across DGX Cloud and on-premises NVIDIA DGX supercomputers, so enterprises can combine resources when needed.

And DGX Cloud includes NVIDIA AI Enterprise , the software layer of the NVIDIA AI platform, which provides over 100 end-to-end AI frameworks and pretrained models to accelerate data science pipelines and streamline the development and deployment of production AI.

Learn more about how to get started with DGX Cloud .

Original URL: https://blogs.nvidia.com/blog/2023/07/25/dgx-generative-ai/