New NVIDIA DGX System Software and Infrastructure Solutions Supercharge Enterprise AI

Adept AI, Hyundai Motor Group, KT and the University of Wisconsin-Madison are among the latest innovators to deliver AI built on DGX systems, Base Command and DGX SuperPOD.

Author: Tony Paikeday

At GTC today, NVIDIA unveiled a number of updates to its DGX portfolio to power new breakthroughs in enterprise AI development.

NVIDIA DGX H100 systems are now available for order. These infrastructure building blocks support NVIDIA's full-stack enterprise AI solutions.

With 32 petaflops of performance at FP8 precision, NVIDIA DGX H100 delivers a leap in efficiency for enterprise AI development. It offers 3x lower total cost of ownership and 3.5x more energy efficiency compared to the previous generation.

New NVIDIA Base Command software, which simplifies and speeds AI development, powers every DGX system — from single nodes to DGX SuperPODs.

Also unveiled was NVIDIA DGX BasePOD — the evolution of DGX POD — which makes enterprise data-center AI deployments simpler and faster for IT teams to acquire, deploy and manage.

Many of the world's AI leaders are building technological breakthroughs — from self-driving cars to voice assistants — using NVIDIA DGX systems and software, and the pace of innovation is not slowing down.

NVIDIA Base Command provides enterprise-grade orchestration and cluster management, and it now features a full software stack for maximizing AI developer productivity, IT manageability and workload performance.

The workflow management features of Base Command now include support for on-premises DGX SuperPOD environments, enabling businesses to gain centralized control of AI development projects with simplified collaboration for project teams, and integrated monitoring and reporting dashboards.

Base Command works with the NVIDIA AI Enterprise software suite, which is now included with every DGX system. The NVIDIA AI software enables end-to-end AI development and deployment with supported AI and data science tools, optimized frameworks and pretrained models .

Additionally, it offers enterprise-workflow management and MLOps integrations with DGX-Ready Software providers Domino Data Lab , Run.ai, Weights & Biases and NVIDIA Inception member Rescale. It also includes libraries that optimize and accelerate compute, storage and network infrastructure — while ensuring maximized system uptime, security and reliability.

DGX BasePOD provides a reference architecture for DGX systems that incorporates design best practices for integrating compute, networking, storage and software.

Customers are already using NVIDIA DGX POD to power the development of a broad range of enterprise applications. DGX BasePOD builds on the success of DGX POD with new industry solutions targeting the biggest AI opportunities, including natural language processing, healthcare and life sciences, and fraud detection.

Delivered as fully integrated, ready-to-deploy offerings through the NVIDIA Partner Network , DGX BasePOD solutions range in size, from two to hundreds of DGX systems, with certified high-performance storage from NVIDIA DGX storage technology partners including DDN , Dell, NetApp

, Pure Storage , VAST Data and WEKA .

Enterprises around the world choose NVIDIA DGX systems to power their most advanced AI workloads. Among the AI innovators developing mission-critical AI capabilities on DGX A100 systems:

ML research and product lab Adept is building an AI teammate powered by a large language model prototyped on NVIDIA DGX Foundry , and then scaled with NVIDIA A100 GPUs and NVIDIA Megatron on Oracle Cloud Infrastructure.

Hyundai Motor Group is using a 40-node DGX SuperPOD to explore hyperscale AI workloads.

Telecom company KT is developing a LLM with around 200 billion parameters for a variety of Korean-language applications, including the GiGA Genie smart speaker, using the NVIDIA NeMo Megatron framework, NVIDIA DGX SuperPOD and NVIDIA Base Command software.

The University of Wisconsin-Madison is quickly bringing AI to medical imaging devices using NVIDIA DGX systems with the Flywheel research platform and the NVIDIA Clara healthcare application framework. Using the NVIDIA Federated Learning Application Runtime Environment, or NVIDIA FLARE , in collaboration with other hospitals, the university is securely training AI models on DGX systems for medical imaging, annotation and classification.

Learn more about the AI breakthroughs powered by NVIDIA DGX systems by watching NVIDIA founder and CEO Jensen Huang’s GTC keynote in replay. And join the GTC session, “ Designing Your AI Center of Excellence ,” with Charlie Boyle, vice president of DGX systems at NVIDIA.

Original URL:
https://blogs.nvidia.com/blog/2022/09/20/dgx-system-software-and-infrastructure-updates/