Microsoft and NVIDIA Announce Major Integrations to Accelerate Generative AI for Enterprises Everywhere

March 18, 2024

- Microsoft Azure to Adopt NVIDIA Grace Blackwell Superchip to Accelerate Customer and First-Party AI Offerings
- NVIDIA DGX Cloud's Native Integration with Microsoft Fabric to Streamline Custom AI Model Development with Customer's Own Data
- NVIDIA Omniverse Cloud APIs First on Azure Power Ecosystem of Industrial Design and Simulation Tools
- Microsoft Copilot Enhanced with NVIDIA AI and Accelerated Computing Platforms
- New NVIDIA Generative AI Microservices for Enterprise, Developer and Healthcare Applications Coming to Microsoft Azure AI

GTC—At GTC on Monday, Microsoft Corp. and NVIDIA expanded their longstanding collaboration with powerful new integrations that leverage the latest NVIDIA generative AI and Omniverse™ technologies across Microsoft Azure, Azure AI services, Microsoft Fabric and Microsoft 365.

"Together with NVIDIA, we are making the promise of AI real, helping to drive new benefits and productivity gains for people and organizations everywhere," said Satya Nadella, Chairman and CEO, Microsoft. "From bringing the GB200 Grace Blackwell processor to Azure, to new integrations between DGX Cloud and Microsoft Fabric, the announcements we are making today will ensure customers have the most comprehensive platforms and tools across every layer of the Copilot stack, from silicon to software, to build their own breakthrough AI capability."

"AI is transforming our daily lives – opening up a world of new opportunities," said Jensen Huang, founder and CEO of NVIDIA. "Through our collaboration with Microsoft, we're building a future that unlocks the promise of AI for customers, helping them deliver innovative solutions to the world."

Advancing AI Infrastructure
Microsoft will be one of the first organizations to bring the power of NVIDIA Grace Blackwell GB200 and advanced NVIDIA Quantum-X800 InfiniBand networking to Azure, deliver cutting-edge trillion-parameter foundation models for natural language processing, computer vision, speech recognition and more.

Microsoft is also announcing the general availability of its Azure NC H100 v5 VM virtual machine (VM) based on the NVIDIA H100 NVL platform. Designed for mid-range training and inferencing, the NC series of virtual machines offers customers two classes of VMs from one to two NVIDIA H100 94GB PCIe Tensor Core GPUs and supports NVIDIA Multi-Instance GPU (MIG) technology, which allows customers to partition each GPU into up to seven instances, providing flexibility and scalability for diverse AI workloads.

Healthcare and Life Sciences Breakthroughs
Microsoft is expanding its collaboration with NVIDIA to transform healthcare and life sciences through the integration of cloud, AI and supercomputing technologies. By harnessing the power of Microsoft Azure alongside NVIDIA DGX™ Cloud and the NVIDIA Clara™ suite of microservices, healthcare providers, pharmaceutical and biotechnology companies, and medical device developers will soon be able to innovate rapidly across clinical research and care delivery with improved efficiency.

Industry leaders such as Sanofi and the Broad Institute of MIT and Harvard, industry ISVs such as Flywheel and SOPHiA GENETICS, academic medical centers like the University of Wisconsin School of Medicine and Public Health, and health systems like Mass General Brigham are already leveraging cloud computing and AI to drive transformative changes in healthcare and to enhance patient care.

Industrial Digitalization
NVIDIA Omniverse Cloud APIs will be available first on Microsoft Azure later this year, enabling developers to bring increased data interoperability, collaboration, and physics-based visualization to existing software applications. At NVIDIA GTC, Microsoft is demonstrating a preview of what is possible using Omniverse Cloud APIs on Microsoft Azure. Using an interactive 3D viewer in Microsoft Power BI, factory operators can see real-time factory data overlaid on a 3D digital twin of their facility to gain new insights that can speed up production.

NVIDIA Triton Inference Server and Microsoft Copilot
NVIDIA GPUs and NVIDIA Triton Inference Server™ help serve AI inference predictions in Microsoft Copilot for Microsoft 365. Copilot for Microsoft 365, soon available as a dedicated physical keyboard key on Windows 11 PCs, combines the power of large language models with proprietary enterprise data to deliver real-time contextualized intelligence, enabling users to enhance their creativity, productivity and skills.

From AI Training to AI Deployment
NVIDIA NIM ™ inference microservices are coming to Azure AI to turbocharge AI deployments. Part of the NVIDIA AI Enterprise software platform, also available on the Azure Marketplace, NIM provides cloud-native microservices for optimized inference on more than two dozen popular foundation models, including NVIDIA-built models that users can experience at ai.nvidia.com. For deployment, the microservices deliver pre-built, run-anywhere containers powered by NVIDIA AI Enterprise inference software — including Triton Inference Server, TensorRT™ and TensorRT-LLM — to hel developers speed time to market of performance-optimized production AI applications.