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

- 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 help developers speed time to market of performance-optimized production AI applications.

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 Microsoft
Microsoft (Nasdaq "MSFT" @microsoft) enables digital transformation for the era of an intelligent cloud and an intelligent edge. Its mission is to empower every person and every organization on the planet to achieve more.

Note to editors: For more information, news and perspectives from Microsoft, please visit the Microsoft News Center at http://news.microsoft.com. Web links, telephone numbers and titles were correct at time of publication but may have changed. For additional assistance, journalists and analysts may contact Microsoft's Rapid Response Team or other appropriate contacts listed at https://news.microsoft.com/microsoft-public-relations-contacts.

NVIDIA Forward-Looking Statements
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 Superchip, NVIDIA DGX Cloud, NVIDIA Omniverse Cloud APIs, NVIDIA AI and Accelerated Computing Platforms, and NVIDIA Generative AI Microservices; the benefits and impact of NVIDIA's collaboration with Microsoft, and the features and availability of its services and offerings; AI transforming our daily lives, the way we work and opening up a world of new opportunities; and building a future that unlocks the promise of AI for customers and brings transformative solutions to the world through NVIDIA's continued collaboration with Microsoft 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 NVIDIA's 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 NVIDIA 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.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements above are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

Microsoft Media Relations
rapidresponse@we-worldwide.com
Natalie Hereth
NVIDIA Corporation
nhereth@nvidia.com