NVIDIA Teams With Microsoft to Build Massive Cloud AI Computer

Tens of Thousands of NVIDIA GPUs, NVIDIA Quantum-2 InfiniBand and Full Stack of NVIDIA AI Software Coming to Azure; NVIDIA, Microsoft and Global Enterprises to Use Platform for Rapid, Cost-Effective AI Development and Deployment

NVIDIA today announced a multi-year collaboration with Microsoft to build one of the most powerful AI supercomputers in the world, powered by Microsoft Azure's advanced supercomputing infrastructure combined with NVIDIA GPUs, networking and full stack of AI software to help enterprises train, deploy and scale AI, including large, state-of-the-art models.

Azure's cloud-based AI supercomputer includes powerful and scalable ND- and NC-series virtual machines optimized for AI distributed training and inference. It is the first public cloud to incorporate NVIDIA's advanced AI stack, adding tens of thousands of NVIDIA A100 and H100 GPUs, NVIDIA Quantum-2 400Gb/s InfiniBand networking and the NVIDIA AI Enterprise software suite to its platform.

As part of the collaboration, NVIDIA will utilize Azure's scalable virtual machine instances to research and further accelerate advances in generative AI, a rapidly emerging area of AI in which foundational models like Megatron Turing NLG 530B are the basis for unsupervised, self-learning algorithms to create new text, code, digital images, video or audio.

The companies will also collaborate to optimize Microsoft's DeepSpeed deep learning optimization software. NVIDIA's full stack of AI workflows and software development kits, optimized for Azure, will be made available to Azure enterprise customers.

"AI technology advances as well as industry adoption are accelerating. The breakthrough of foundation models has triggered a tidal wave of research, fostered new startups and enabled new enterprise applications," said Manuvir Das, vice president of enterprise computing at NVIDIA. "Our collaboration with Microsoft will provide researchers and companies with state-of-the-art AI infrastructure and software to capitalize on the transformative power of AI."

"AI is fueling the next wave of automation across enterprises and industrial computing, enabling organizations to do more with less as they navigate economic uncertainties," said Scott Guthrie, executive vice president of the Cloud + AI Group at Microsoft. "Our collaboration with NVIDIA unlocks the world's most scalable supercomputer platform, which delivers state-of-the-art AI capabilities for every enterprise on Microsoft Azure."

Scalable Peak Performance With NVIDIA Compute and Quantum-2 InfiniBand on Azure
Microsoft Azure's AI-optimized virtual machine instances are architected with NVIDIA's most advanced data center GPUs and are the first public cloud instances to incorporate NVIDIA Quantum-2 400Gb/s InfiniBand networking. Customers can deploy thousands of GPUs in a single cluster to train even the most massive large language models, build the most complex recommender systems at scale, and enable generative AI at scale.

The current Azure instances feature NVIDIA Quantum 200Gb/s InfiniBand networking with NVIDIA A100 GPUs. Future ones will be integrated with NVIDIA Quantum-2 400Gb/s InfiniBand networking and NVIDIA H100 GPUs. Combined with Azure's advanced compute cloud infrastructure, networking and storage, these AI-optimized offerings will provide scalable peak performance for AI training and deep learning inference workloads of any size.

Accelerating AI Development and Deployment
Additionally, the platform will support a broad range of AI applications and services, including Microsoft DeepSpeed and the NVIDIA AI Enterprise software suite.

Microsoft DeepSpeed will leverage the NVIDIA H100 Transformer Engine to accelerate transformer-based models used for large language models, generative AI and writing computer code, among other applications. This technology applies 8-bit floating point precision capabilities to DeepSpeed to dramatically accelerate AI calculations for transformers — at twice the throughput of 16-bit operations.

NVIDIA AI Enterprise — the globally adopted software of the NVIDIA AI platform — is certified and supported on Microsoft Azure instances with NVIDIA A100 GPUs. Support for Azure instances with NVIDIA H100 GPUs will be added in a future software release.

NVIDIA AI Enterprise, which includes the NVIDIA Riva for speech AI and NVIDIA Morpheus cybersecurity application frameworks, streamlines each step of the AI workflow, from data processing and AI model training to simulation and large-scale deployment.

Learn more about NVIDIA's AI technology on the Azure partner page.

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, performance and availability of our products and technologies, including NVIDIA A100 and H100 GPUs, NVIDIA Quantum-2 InfiniBand, NVIDIA AI Enterprise, and the NVIDIA H100 Transformer Engine; NVIDIA's collaboration with Microsoft, including the benefits and impact thereof; AI technology advances as well as industry adoption accelerating; the impacts of the breakthrough of foundation models; AI fueling the next wave of automation across enterprises and industrial computing and the impact thereof; future Azure instances being integrated with NVIDIA Quantum-2 400Gb/s InfiniBand networking and NVIDIA H100 GPUs; support for Azure instances with NVIDIA H100 GPUs being added in a future software release 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.

Cliff Edwards
NVIDIA Corporation
+1-415-699-2755
cliffe@nvidia.com