NVIDIA H100 Tensor Core GPU Used on New Microsoft Azure Virtual Machine Series Now Generally Available

Microsoft Azure ND H100 v5 virtual machine series instance offers next-level performance at scale for LLMs, generative AI and other compute-intensive workloads.

Author: Dave Salvator

Microsoft Azure users can now turn to the latest NVIDIA accelerated computing technology to train and deploy their generative AI applications.

Available today, the Microsoft Azure ND H100 v5 VMs using NVIDIA H100 Tensor Core GPUs and NVIDIA Quantum-2 InfiniBand networking — enables scaling generative AI, high performance computing (HPC) and other applications with a click from a browser.

Available to customers across the U.S., the new instance arrives as developers and researchers are using large language models (LLMs) and accelerated computing to uncover new consumer and business use cases.

The NVIDIA H100 GPU delivers supercomputing-class performance through architectural innovations, including fourth-generation Tensor Cores , a new Transformer Engine for accelerating LLMs and the latest NVLink technology that lets GPUs talk to each other at 900GB/sec.

The inclusion of NVIDIA Quantum-2 CX7 InfiniBand with 3,200 Gbps cross-node bandwidth ensures seamless performance across the GPUs at massive scale, matching the capabilities of top-performing supercomputers globally.

ND H100 v5 VMs are ideal for training and running inference for increasingly complex LLMs and computer vision models. These neural networks drive the most demanding and compute-intensive generative AI applications, including question answering, code generation, audio, video and image generation, speech recognition and more.

The ND H100 v5 VMs achieve up to 2x speedup in LLMs like the BLOOM 175B model for inference versus previous generation instances, demonstrating their potential to further optimize AI applications.

NVIDIA H100 Tensor Core GPUs on Azure provide enterprises the performance, versatility and scale to supercharge their AI training and inference workloads. The combination streamlines the development and deployment of production AI with the NVIDIA AI Enterprise software suite integrated with Azure Machine Learning for MLOps, and delivers record-setting AI performance in industry-standard MLPerf benchmarks .

In addition, by connecting the NVIDIA Omniverse platform to Azure, NVIDIA and Microsoft are providing hundreds of millions of Microsoft enterprise users with access to powerful industrial digitalization and AI supercomputing resources.

Learn more about new Azure v5 instances powered by NVIDIA H100 GPUs .

Original URL: https://blogs.nvidia.com/blog/2023/08/07/microsoft-azure-nd-h100-v5-instance/