NVIDIA Announces Scalable GPU-Accelerated Supercomputer in the Microsoft Azure Cloud

New Microsoft Azure NDv2 Supersized Instance Can Scale to Hundreds of Interconnected NVIDIA Tensor Core GPUs for Complex AI and High Performance Computing Applications

SC19 -- NVIDIA today announced the availability of a new kind of GPU-accelerated supercomputer in the cloud on Microsoft Azure.

Built to handle the most demanding AI and high performance computing applications, the largest deployments of Azure's new NDv2 instance rank among the world's fastest supercomputers, offering up to 800 NVIDIA V100 Tensor Core GPUs interconnected on a single Mellanox InfiniBand backend network. It enables customers for the first time to rent an entire AI supercomputer on demand from their desk, and match the capabilities of large-scale, on-premises supercomputers that can take months to deploy.

“Until now, access to supercomputers for AI and high performance computing has been reserved for the world's largest businesses and organizations," said Ian Buck, vice president and general manager of Accelerated Computing at NVIDIA. “Microsoft Azure's new offering democratizes AI, giving wide access to an essential tool needed to solve some of the world's biggest challenges."

Girish Bablani, corporate vice president of Azure Compute at Microsoft Corp., added, “As cloud computing gains momentum everywhere, customers are seeking more powerful services. Working with NVIDIA, Microsoft is giving customers instant access to a level of supercomputing power that was previously unimaginable, enabling a new era of innovation."

Dramatic Performance, Cost Benefits
The new offering -- which is ideal for complex AI, machine learning and HPC workloads -- can provide dramatic performance and cost advantages over traditional CPU-based computing. AI researchers needing fast solutions can quickly spin up multiple NDv2 instances and train complex conversational AI models in just hours.

Microsoft and NVIDIA engineers used 64 NDv2 instances on a pre-release version of the cluster to train BERT, a popular conversational AI model, in roughly three hours. This was achieved in part by taking advantage of multi-GPU optimizations provided by NCCL, an NVIDIA CUDA X™ library and high-speed Mellanox interconnects.

Customers can also see benefits from using multiple NDv2 instances to run complex HPC workloads, such as LAMMPS, a popular molecular dynamics application used to simulate materials down to the atomic scale in such areas as drug development and discovery. A single NDv2 instance can deliver an order of magnitude faster results than a traditional HPC node without GPU acceleration for specific types of applications, such as deep learning. This performance can scale linearly to a hundred instances for large-scale simulations.

All NDv2 instances benefit from the GPU-optimized HPC applications, machine learning software and deep learning frameworks like TensorFlow, PyTorch and MXNet from the NVIDIA NGC container registry and Azure Marketplace. The registry also offers Helm charts to easily deploy the AI software on Kubernetes clusters.

Availability and Pricing
NDv2 is available now in preview. One instance with eight NVIDIA V100 GPUs can be clustered to scale up to a variety of workload demands. See more details here.

About NVIDIA
NVIDIA’s (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at http://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance and availability of the Microsoft Azure NDv2 with NVIDIA V100 Tensor Core GPUs; and cloud computing gaining momentum everywhere 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.

Media Contacts

Kristin Uchiyama

+1-408-486-2248
kuchiyama@nvidia.com