NVIDIA Doubles Down: Announces A100 80GB GPU, Supercharging World's Most Powerful GPU for AI Supercomputing

Leading Systems Providers Atos, Dell Technologies, Fujitsu, GIGABYTE, Hewlett Packard Enterprise, Inspur, Lenovo, Quanta and Supermicro to Offer NVIDIA A100 Systems to World's Industries

SC20—NVIDIA today unveiled the NVIDIA® A100 80GB GPU — the latest innovation powering the NVIDIA HGX™ AI supercomputing platform — with twice the memory of its predecessor, providing researchers and engineers unprecedented speed and performance to unlock the next wave of AI and scientific breakthroughs.

The new A100 with HBM2e technology doubles the A100 40GB GPU's high-bandwidth memory to 80GB and delivers over 2 terabytes per second of memory bandwidth. This allows data to be fed quickly to A100, the world's fastest data center GPU, enabling researchers to accelerate their applications even faster and take on even larger models and datasets.

"Achieving state-of-the-art results in HPC and AI research requires building the biggest models, but these demand more memory capacity and bandwidth than ever before," said Bryan Catanzaro, vice president of applied deep learning research at NVIDIA. "The A100 80GB GPU provides double the memory of its predecessor, which was introduced just six months ago, and breaks the 2TB per second barrier, enabling researchers to tackle the world's most important scientific and big data challenges."

The NVIDIA A100 80GB GPU is available in NVIDIA DGX™ A100 and NVIDIA DGX Station™ A100 systems, also announced today and expected to ship this quarter.

Leading systems providers Atos, Dell Technologies, Fujitsu, GIGABYTE, Hewlett Packard Enterprise, Inspur, Lenovo, Quanta and Supermicro are expected to begin offering systems built using HGX A100 integrated baseboards in four- or eight-GPU configurations featuring A100 80GB in the first half of 2021.

Fueling Data-Hungry Workloads
Building on the diverse capabilities of the A100 40GB, the 80GB version is ideal for a wide range of applications with enormous data memory requirements.

For AI training, recommender system models like DLRM have massive tables representing billions of users and billions of products. A100 80GB delivers up to a 3x speedup, so businesses can quickly retrain these models to deliver highly accurate recommendations.

The A100 80GB also enables training of the largest models with more parameters fitting within a single HGX-powered server such as GPT-2, a natural language processing model with superhuman generative text capability. This eliminates the need for data or model parallel architectures that can be time consuming to implement and slow to run across multiple nodes.

With its multi-instance GPU (MIG) technology, A100 can be partitioned into up to seven GPU instances, each with 10GB of memory. This provides secure hardware isolation and maximizes GPU utilization for a variety of smaller workloads. For AI inferencing of automatic speech recognition models like RNN-T, a single A100 80GB MIG instance can service much larger batch sizes, delivering 1.25x higher inference throughput in production.

On a big data analytics benchmark for retail in the terabyte-size range, the A100 80GB boosts performance up to 2x, making it an ideal platform for delivering rapid insights on the largest of datasets. Businesses can make key decisions in real time as data is updated dynamically.

For scientific applications, such as weather forecasting and quantum chemistry, the A100 80GB can deliver massive acceleration. Quantum Espresso, a materials simulation, achieved throughput gains of nearly 2x with a single node of A100 80GB.

"Speedy and ample memory bandwidth and capacity are vital to realizing high performance in supercomputing applications," said Satoshi Matsuoka, director at RIKEN Center for Computational Science. "The NVIDIA A100 with 80GB of HBM2e GPU memory, providing the world's fastest 2TB per second of bandwidth, will help deliver a big boost in application performance."

Key Features of A100 80GB
The A100 80GB includes the many groundbreaking features of the NVIDIA Ampere architecture:

- **Third-Generation Tensor Cores:** Provide up to 20x AI throughput of the previous Volta generation with a new format TF32, as well as 2.5x FP64 for HPC, 20x INT8 for AI inference and support for the BF16 data format.
- **Larger, Faster HBM2e GPU Memory:** Doubles the memory capacity and is the first in the industry to offer more than 2TB per second of memory bandwidth.
- **MIG technology:** Doubles the memory per isolated instance, providing up to seven MIGs with 10GB each.
- **Structural Sparsity:** Delivers up to a 2x speedup inferencing sparse models.
- **Third-Generation NVLink® and NVSwitch™:** Provide twice the GPU-to-GPU bandwidth of the previous generation interconnect technology, accelerating data transfers to the GPU for data-intensive workloads to 600 gigabytes per second.

NVIDIA HGX AI Supercomputing Platform
The A100 80GB GPU is a key element in NVIDIA HGX AI supercomputing platform, which brings together the full power of NVIDIA GPUs, NVIDIA NVLink, NVIDIA InfiniBand networking and a fully optimized NVIDIA AI and HPC software stack to provide the highest application performance. It enables researchers and scientists to combine HPC, data analytics and deep learning computing methods to advance scientific progress.

Learn more about NVIDIA A100 80GB in the live NVIDIA SC20 Special Address at 3 p.m. PT today.

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, performance, features and abilities of the NVIDIA A100 80GB GPU and what it enables; the systems providers that will offer NVIDIA A100 systems and the timing for such availability; the A100 80GB GPU providing more memory and speed, and enabling researchers to tackle the world's challenges; the availability of the NVIDIA A100 80GB GPU; memory bandwidth and capacity being vital to realizing high performance in supercomputing applications; the NVIDIA A100 providing the fastest bandwidth and delivering a boost in application performance; and the NVIDIA HGX supercomputing platform providing the highest application performance and enabling advances in scientific progress 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.

Kristin Uchiyama
Enterprise and Edge Computing
+1-408-486-2248
kuchiyama@nvidia.com