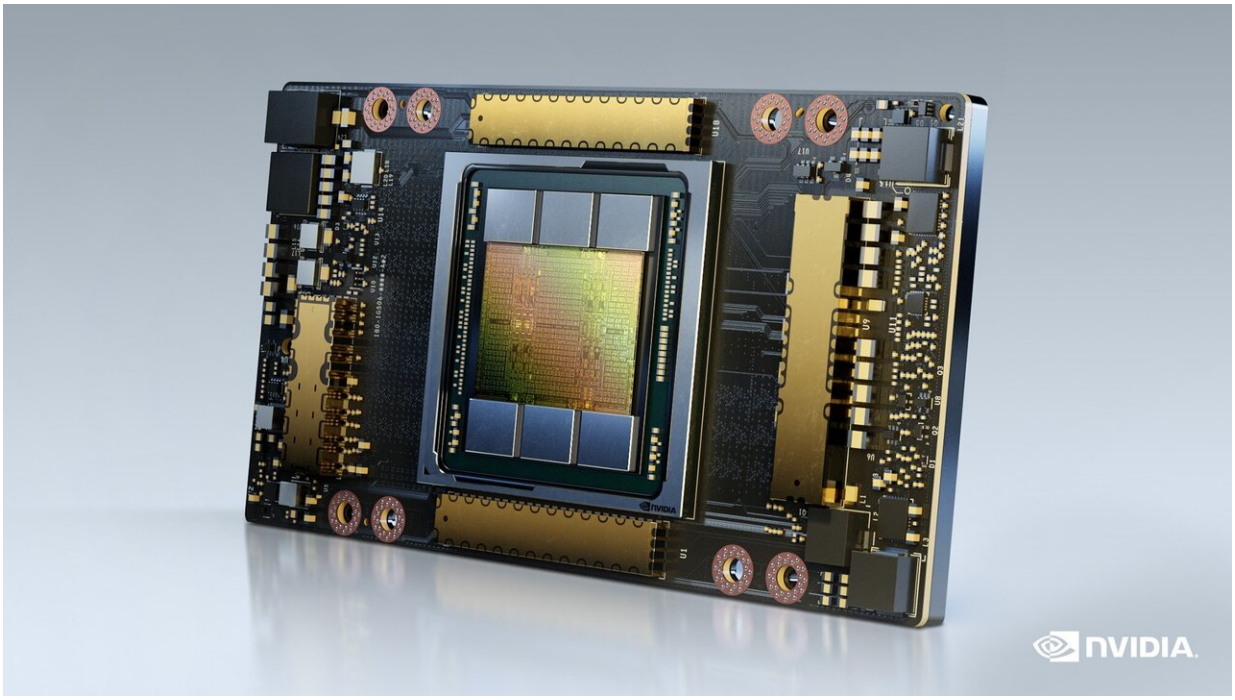
NVIDIA's latest Ampere 80GB graphics processing unit boasts 2TB memory bandwidth

November 17 2020, by Peter Grad



Credit: NVIDA

NVIDA has surpassed the 2 terabyte-per-second memory bandwidth mark with its new GPU, the Santa Clara graphics giant announced Monday.

The top-of-the-line A100 80GB GPU is expected to be integrated in multiple GPU configurations in systems during the first half of 2021.

Earlier this year, NVIDIA unveiled the A100 featuring Ampere architecture, asserting that the GPU provided "the largest leap in performance" ever in its lineup of graphics hardware. It said AI training on the GPU could see performance boosts of 20 times the speed of its earlier generation units.

The new A100 80GB model now doubles the high-bandwidth memory from 40GB to 80GB, and boosts the bandwidth speed of the overall array by .4 TB per second, from 1.6TB/s to 2TB/s.

It features a 1.41GHz boost clock, 5120-bit memory bus, 19.5 TFLOPS of single-precision performance and 9.7 TFLOPS of double-precision performance.

"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."

According to 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."

The GPU will be sought by companies engaged in data-intensive analysis, cloud-based computer rendering and scientific research such as,

for instance, weather forecasting, quantum chemistry and protein modeling.

BMW, Lockheed Martin, NTT Docomo and the Pacific Northwest Laboratory are currently utilizing NVIDIA's DGX Stations for AI projects.

NVIDIA provided performance results on a number of testing benchmarks. The GPU achieved a threefold performance improvement in AI deep learning, and a doubling of speed in big data analytics.

Atos, Dell Technologies, Fujitsu, Hewlett-Packard Enterprise, Lenovo, Quanta and Supermicro are expected to offer systems using HGX A100 baseboards with four- or eight-GPU configurations.

]NVIDIA said the new GPU provides offers "data center performance without a data center."

Customers seeking individual A1000 GPUs on a PCIe card are limited to the 40GB VRAM version only, at least for the time being.

NVIIDA has not announced pricing yet. The original DGX A100 unveiled last spring had a price tag of $199,000.

More information: nvidianews.nvidia.com/news/nvi … or-ai-supercomputing