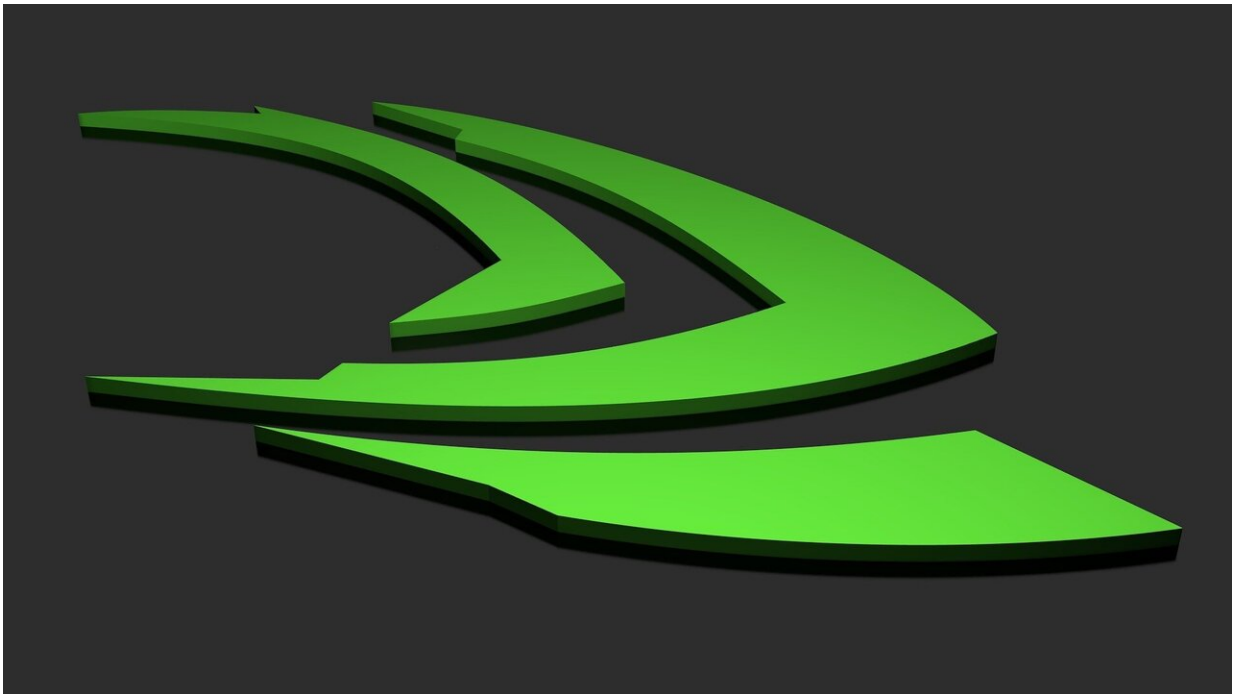# Nvidia unveils PCIe 4.0 version of Ampere A100 GPU

June 24 2020, by Peter Grad



Credit: Pixabay/CC0 Public Domain

Following on the heels of last month's display of its next generation, high end Ampere A100 GPU, Nvidia this week unveiled a PCIe 4.0 version of the graphics unit.

The card is designed for easier installation in personal computers, standardized servers and existing systems tackling artificial intelligence

and other supercomputing tasks. The A100 GPU is designed for custom-built systems constructed specifically for high-end tasks.

Otherwise, the latest Nvidia cards, built on the latest 7nm Ampere architecture, share nearly identical specs.

They each have 40 GB memory, 54 billion transistors, six HBM2 memory stacks, and 6,912 CUDA cores, which leverage the GPU to achieve greater efficiency in task execution.

They both also sport FP64 Tensor Cores capable of 19.5 teraFLOPS, the ability to process 19.5 trillion floating-point calculations per second.

One key difference is the thermal design point: the A100 is rated at 400W while the PCIe is 250W. The power reduction is attributed to the fact that the PCIe is designed specifically for lower TDP products, which generally have less room for cooling and lower power demands.

The difference, according to Nvidia, means a 10 percent reduction in performance in single-GPU workloads. In circumstances requiring continuous GPU capabilities or complex distributed workloads, the performance hit could reach 33 percent to 50 percent. Nvidia stressed, however, that the maximum power hit should be rare as only a limited number of tasks can tax the card to such an extent.

The A100 GPU offers impressive performance enhancement over previous models. It has the capacity to partition itself into seven virtual GPUs to tackle independent tasks and a linkage ability to combine several distinct GPUs into a single massive processing unit. Performance tests found the A100 achieving results 20 times better than its predecessor, which used the Volta microarchitecture.

Nvidia reported that the A100 recently broke a big data analytics

benchmark, completing a task that took the previous record-holder 4.7 hours in just 14.5 minutes.

Response to Nvidia's new GPUs has been swift.

"Adoption of Nvidia A100 GPUs into leading server manufacturers' offerings is outpacing anything we've previously seen," according to Ian Buck, [vice president](#) and general manager of accelerated computing at Nvidia.

Manufacturers that will incorporate the new units in their servers include Cisco, Dell Technologies, HPE, ASUS, Lenovo and Fujitsu. More than 30 systems are expected to be built this summer, with at least 20 more by the end of the year.

No pricing information has yet been released.

© 2020 Science X Network