Written by Marco Attard
14 May 2020

A "kitchen keynote" by Nvidia CEO Jensen Huang reveals the first product based on next-generation Ampere architecture-- the A100, a datacentre GPU built for data analytics, scientific computing and cloud graphics.

The company claims the A100 is the largest 7nm-based processor, carrying over 54 billion transistors. It includes 3rd generation Tensor Cores with TF32, a math format to accelerate single-precision AI training, and uses structural sparsity acceleration for higher performance. A multi-instance GPU (MIG) allows customers to partition a single A100 into as many as seven independent GPUs, each with own resources, while 3rd generation NVLink technology doubles the high-speed connectivity between GPUs, turning A100 servers into one giant GPU.

In terms of raw numbers, the A100 consists of 3456 FP64 CUDA cores, 6912 FP32 CUDA cores, 432 Tensor cores, 108 streaming multiprocessors and 40GB of GPU memory, all within a 400W power envelope. The result, Nvidia claims, is a system able to push 6X the performance of previous generation Volta architecture for training and 7X the performance for interference.

Obviously Nvidia has a server based on the GPU-- namely the DGX A100, a machine the company claims is the first 5-petaflops server. Customers can divide a DGX A100 system into up to 56 applications, all running independently, allowing a single server to either scale up for computationally intensive tasks such as AI training or scale out for AI deployment or inference.

The company has also built the next-generation DGX SuperPOD, a combination of 140 DGX A100 systems and Mellanox networking bringing about 700 petaflops of AI performance (or the equivalent of one of the 20 fastest computers in the world). For the edge Nvidia offers the EGX A100, a system bringing Mellanx ConnectX-6 SmartNIC technology for secure networking.

A number of OEM partners will release A100-based servers in the near future, including Atos, Dell Technologies, Fujitsu, Gigabyte, H3C, Hewlett Packard Enterprise, Inspur, Lenovo, Quanta and Supermicro. In addition the technology will find use at a number of cloud providers, including Alibaba Cloud, Amazon Web Services, Baidu Cloud, Google Cloud and Tencent Cloud.

Go Nvidia CEO Introduces Ampere Architecture, A100 GPU in News-Packed "Kitchen Keynote"