The DGX-1 supercomputer uses Nvidia's Volta GPU to bring 400 servers into one box

There is no need to buy up to 400 servers if you already own a DGX-1 supercomputer using Nvidia's Volta GPU on your desk. This supercomputer will come to your house for $ 149,000 in the third quarter.

DGX-1 looks like a normal rack server and takes advantage of computing power from 8 Tesla V100 GPUs.

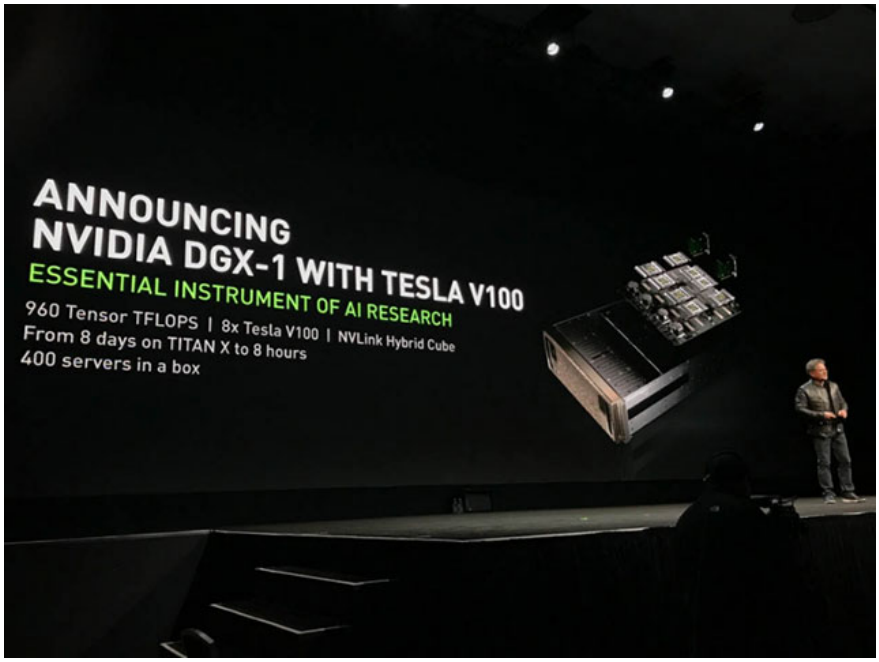
The GPU, based on the Volta architecture, was unveiled at the GPU technology conference in San Jose on Wednesday. 'You just need to plug in and work,' said Jen Hsun Huang, CEO of Nvidia. But DGX-1 with Tesla V100 is very expensive. At a price of $ 149,000, it can save many lives. However, Huang encouraged everyone to order and said that the goods would be delivered in the third quarter.

According to Nvidia, the new supercomputer uses a total of 40,960 CUDA cores, equivalent to the computing power of 800 CPUs. It replaces previous DGX-1, based on Pascal architecture with 250 2 socket servers. The system will have about 960 teraflops. Along with the GPU is a 20-core Intel Xeon E5-2698 v4S running at 2.2GHz. The system has four 1.92TB SSDs and runs on Ubuntu Linux. It will take 3,200 watts so users should not let it run all day.

Gamers don't need to be too excited by the DGX-1 Tesla V100 is too expensive for gamers, it is designed for more machine learning. GPUs can help with machine learning tasks in data centers and Nvidia's supercomputers are examples of how GPUs do applications such as image recognition or natural language processing. Huang thinks that the CPU is not enough to calculate, especially AI, but the GPU is very suitable.

Huang said the Tesla V100 in DGX-1 was five times faster than Pascal architecture. It will have new technology like NVLink 2.0, an internal connection with bandwidth up to 300Gbps. GPUs have more than 21 million transistors and 5120 cores. It also has 900Gbps bandwidth memory of HBM2.

Nvidia uses Tensor's cube-like core, working with conventional processing cores to improve deep learning. Nvidia focuses on the core structure to accelerate chain-matrix multiplication - the center of the efficient deep learning system.



Introduction about DGX-1 supercomputer

Huang is also very proud of the GPU providing 120 teraflops of deep learning, although it is difficult to verify this. Benchmarking tools for machine or deep learning applications are not yet available, although companies like Google are developing them.

Supercomputers work with many advanced computing and deep learning frameworks like CUDA, Tensor and Caffe2. The graphics were introduced as DGX Station, smaller than the new DGX-1 version, like a workstation with 4 GPU Tesla V100, half of DGX-1. It costs $ 69,000. Nvidia hasn't said anything about whether the product is sold globally.