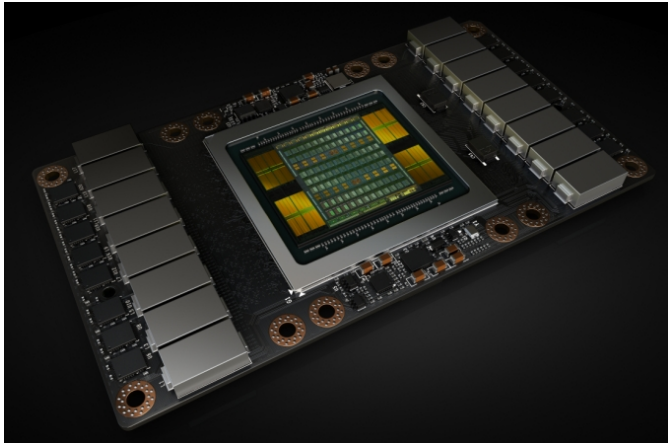


Nvidia Unveils First Volta-Based GPU

Written by Marco Attard
12 May 2017

Nvidia presents the Tesla V100-- the first GPU based on the company's Volta architecture, packing 3840 CUDA cores and 15 billion transistors on a 815mm slab of silicon.



Built using a 12-nanometer manufacturing process, the Tesla V100 is aimed at high performance computing applications. As the successor to the current Pascal GPU flagship, the Tesla P100, it features a redesigned streaming architecture promising a 50% increase in efficiency compared to Pascal, enabling "major boosts in FP32 and FP64 performance in the same power envelope."

In addition the GPU carries 672 tensor cores (TCs)-- a new kind of core designed for machine learning operations. The company claims these boost performance by 4x over Pascal, and make the V100 superior to the [Google dedicated tensor processing unit \(TPU\)](#).

In terms of raw numbers, the V100 features 7.5 TFLOP of double precision floating-point (FP64) performance, 15 TFLOPs of single precision (FP32) performance and 120 Tensor TFLOPs of mixed-precision matrix-multiply-and-accumulate. Such power is paired with 16GB of 4096-bit HBM2 memory and a 2nd generation version of NVLink technology allowing transfer speeds of up to 300GB/s. Clock speed reaches 1455MHz, while TDP is rated at 300W.

The Tesla V100 will first ship in an Nvidia compute server-- a DGX-1 rack-mount number packing eight cards. The server should ship on Q3 2017, followed by 250W PCIe slot and half-height 150W versions of the card.

Nvidia Unveils First Volta-Based GPU

Written by Marco Attard
12 May 2017

Go [Nvidia Tesla V100](#)