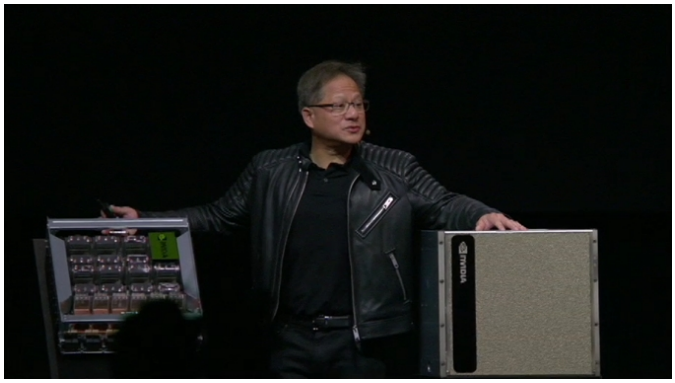
Written by Marco Attard
29 March 2018

Nvidia claims to offer "the world's most powerful AI system" in a more compact and energy efficient package with the DGX-2-- a turnkey HPC system promising to deliver up to 2 petaFLOPs of computing power.



As a followup to the DGX-1, the DGX-2 carries x16 fully interconnected Tesla V100 32GB GPUs (double the amount of the 2017 DGX-1). It also includes up to 1.5TB of system memory, 30/60TB of NVMe storage, InfiniBand or 100GBe networking and x2 Xeon Platinum processors. Tying the GPUs together is what Nvidia calls "NVSwitch," an interconnect fabric allowing the GPUs to communicate with each other in the system at 300GB/s, making 14TB/s of aggregate system bandwidth.

The company says developers can address the system and its 81920 CUDA cores like a "gigantic GPU," using the same software as one would with a single GPU system. In addition, the DGX-2 consumes up to 10kW of electricity, and weighs 159kg. As for further performance claims, Nvidia says the DGX-2 completes the training process for FAIRSEQ (a neural network model for language translation) in 2 days, down from the 15 days of the DGX-1.

Customers can already order the DGX-2.

Go Nvidia DGX-2