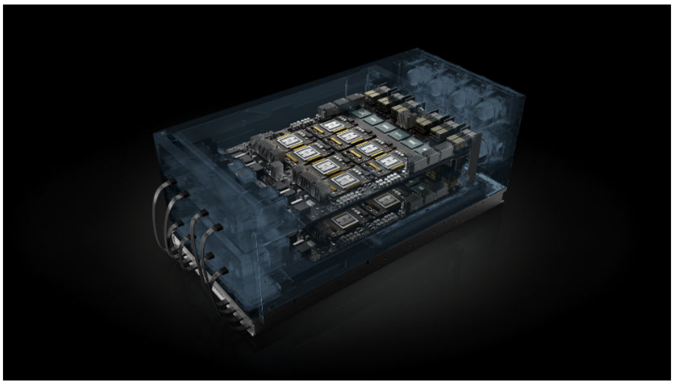
Nvidia Intros HGX-2 Server

Written by Marco Attard
01 June 2018

Nvidia presents an upgrade on the HGX-1 hyperscale GPU accelerator chassis-- the HGX-2, a server powered by no less than x16 Tesla V100 GPUs and x12 NVSwitches, making "the world's largest GPU."



According to the company, the HGX-2 is “the first unified computing platform for both artificial intelligence and high performance computing.” It offers up to 2 petaFLOPS performance for low-precision tensor operations, 250 teraFLOPS at single-precision and 125 teraFLOPS for double-precision.

The NVSwitches allow GPU-to-GPU communications reaching up to 300GB per second, and topping everything of is 0.5TB of GPU memory. The result is a unified system developers can consider as a single, huge GPU. It can handle a variety of tasks, including cloud, AI, deep learning or high-performance simulations, and Nvidia claims a single HGX-2 unit can replace 60 CPU-only servers.

“The world of computing has changed,” the company says. “CPU scaling has slowed at a time when computing demand is skyrocketing. NVIDIA’s HGX-2 with Tensor Core GPUs gives the industry a powerful, versatile computing platform that fuses HPC and AI to solve the world’s grand challenges.”

Nvidia will be pushing the HGX-2 through reseller partners-- specifically Lenovo, QCT, Supermicro, Wiwyn, Foxconn, Inventec, Quanta and Wistron, who will be shipping HGX-2-based systems later this year.

Go Nvidia HGX-2