Case Study on Nvidia DGX Station

111603043 Chetan S. Paralikar

Artificial Intelligence(AI) and data scientists researchers require fast speed and accuracy from their machine learning and deep learning systems. Faster training and iteration means faster time to market and faster innovation. The NVIDIA DGX Station is a personal AI supercomputer for deep learning with fully integrated hardware and software that can be deployed quickly and easily.

The NVIDIA DGX Station is a groundbreaking deep learning and analytics supercomputer that delivers the immense computing capacity of 400 CPUs in a portable workstation that fits under an office desk. DGX Station is a whisper-quiet, water-cooled workstation that packs four NVIDIA Volta-powered Tesla V100 GPUs delivering up to 500 Tensor TFLOPS for deep learning applications.

DGX Station delivers almost 3x the performance for deep learning training and 3x the inferencing performance compared to today's fastest GPU workstations. The four Tesla V100 GPUs in the DGX Station are connected via NVIDIA's second-generation NVLink interconnect technology, delivering almost five times the IO bandwidth compared to PCIe based GPU workstations.

The DGX Station performance with a power of four-way Tesla V100. The Tesla V100 is 47x faster than the CPU based server. For performing the deep learning task the DGX station takes upto 15 hours to complete it, on the other hand the 4x GPU workstation takes 36 hours and 2x CPU server takes 711 hours to complete. NVIDIA DGX Station delivers 47x faster training.

There are number of DGX Station that can built the DGX server. In 2016, NVIDIA launched the first generation DGX-1 featuring eight NVIDIA Tesla P100 GPUs connected through NVIDIA's high-performance NVLink interconnect in a hybrid cube mesh network. Together with dual-socket Intel Xeon CPUs and four 100 Gb InfiniBand network interface cards, the P100-based DGX-1 provides exceptional performance for deep learning training. With up to 170 FP16 TFLOPS that significantly accelerate training time, the NVIDIA DGX-1 is the first AI supercomputer in a box.

DGX Station Specifications :

- GPUs - 4 NVIDIA Tesla V100 interconnected via NVLink
- TFLOPS - 500 Tensor TFLOPS, 15.7 FP32 TFLOPS
- Tensor Cores - 2560
- CPU - Intel Xeon E5-2698 v4 2.2 GHz (20-core)
- System Memory - 256 GB LRDIMM DDR4
- Storage - Data: 3 x 1.92 TB SSD RAID 0
  OS: 1 x 1.92 TB SSD
- Network - Dual 10 Gb LAN
- Display - 3x DisplayPort
- Acoustics - < 35 dB
- System Weight - 88 lbs / 40 kgs
- System Dimensions - 518 mm (D) x 256 mm (W) x 639 mm (H)
- Maximum Power - 1500 Watts
- Operation Temperature - 10^{o} C - 30^{o} C
- Operating System - Ubuntu Desktop Linux