OSS DEEP LEARNING APPLIANCES

Ultimate Performance for Deep Learning Training

The OSS-VOLTA4 and OSS-VOLTA8 are purpose-built for deep learning applications with fully integrated hardware and software. The OSS-VOLTA8 is a 896 TeraFLOP engine with 80GB/s NVLink for the largest deep learning models. The OSS-PASCAL4 provides 21.2 TeraFLOPS of double precision performance with an 80GB/s GPU peer-to-peer NVLink. These systems are tuned for out-of-the-box operation and quick and easy deployment.

SOFTWARE PLATFORM: DEEP LEARNING ENVIRONMENT

The OSS Deep Learning Appliances come with a choice of machine learning frameworks such as Caffe2, Pytorch, Mxnet, Microsoft Cognitive Toolkit, Tensorflow and Theano. They also come with a choice of machine learning libraries such as MLPython, NVIDIA cuDNN, DIGITS and CaffeOnSpark. GPU drivers, CUDA drivers, CUB and NCCL are supporting elements for the OSS-VOLTA4 and OSS-VOLTA8.

HARDWARE: ULTIMATE DEEP LEARNING TURBO ENGINES

The OSS-VOLTA4 and OSS-VOLTA8 have the latest Tesla V100 SXM GPUs for up to 125 TeraFLOPS of single precision performance. They utilize NVLink with speeds up to 80GB/s peer-to-peer between GPUs. These GPU accelerated servers have dual v4 Broadwell CPUs and up to 2TB DDR4 memory. The OSS-VOLTA4 and OSS-VOLTA8 can integrate into the GPUltima rack-level solution using 100Gb EDR Infiniband interfaces to large-scale multi-root peer-to-peer RDMA networks.

NVLink ADVANTAGES

All GPUs are capable of Peer-to-Peer direct access to all other GPUs' memory as well as direct transfer operations via NVLink at high Bandwidth. These GPU Accelerated servers provide high performance for collective communications. The PCIe bandwidth is fully available for host and/or NIC communication during inter-GPU communication.

GPU MANAGEMENT AND MONITORING FROM Bright Computing

The GPU management and monitoring is pre-installed and provides both health and workload management. It samples all of the metrics provided by all of the NVIDIA GPUs and automatically performs health checks on every GPU. It is integrated with all of the popular HPC workload managers and automatically configures GPUs within the workload manager. The user jobs are automatically directed to available GPUs. Health checks can be designated as pre-job health checks and provides job-level metrics.

NVIDIA PREFERRED SOLUTION PROVIDER

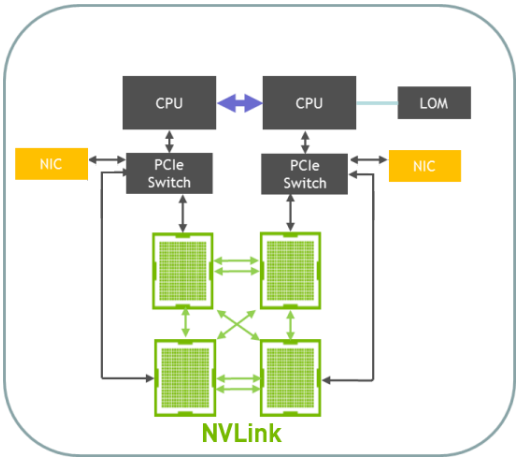
sales@onestopsystems.com | 877.438.2724 | www.onestopsystems.com

GPU ACCELERATED SERVER WITH 4 NVIDIA V100 SXM2 AND NVLink

OSS-VOLTA4

FEATURES

- 1U Chassis
- Dual Intel Xeon 3.2GHz CPUs
- Up to 1TB DDR4 LRDIMM System Memory
- Two 2.5" 1.9TB SATA SSDs
- Four Volta GPU SXM2 with 80GB/s NVLink
- Three x16 PCIe 3.0 slots
- One x8 PCIe 3.0 slot
- Two 2000W Titanium Power Supplies

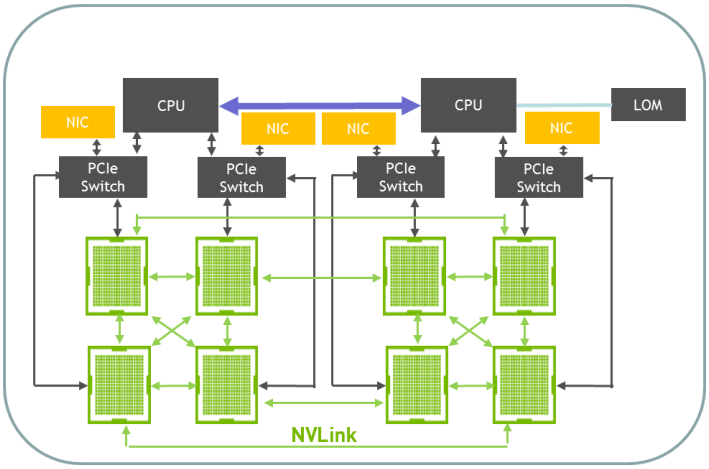


GPU ACCELERATED SERVER WITH 8 NVIDIA V100 SXM2 AND NVLink

OSS-VOLTA8

FEATURES

- 4U Chassis
- Dual Intel Xeon 3.2GHz CPUs
- Up to 2TB DDR4 LRDIMM System Memory
- Four 2.5" 1.6TB NVMe SSDs
- Eight Volta GPU SXM2 with 80GB/s NVLink
- Four x16 PCIe 3.0 slots
- Two x8 PCIe 3.0 slot
- Four 2000W Titanium Power Supplies



Software Bundle

- Choice of Operating System
  - CentOS7
  - RHEL7
  - SL7
- Choice of Machine Learning Framework
  - Caffe2
  - Pytorch
  - Mxnet
  - Microsoft Cognitive Toolkit
  - Tensorflow
  - Theano
- MLPython
- ML Dependencies (400MB Python)
- cuDNN (5.0 & 5.1)
- DIGITS
- Caffe on Spark
- CUDA & NVIDIA driver
- CUB (CUDA building blocks)
- NCCL
- GPU Management
  - Health Management
  - Workload Integration

sales@onestopsystems.com | 877.438.2724 | www.onestopsystems.com