

Applied Computing GPU cluster with the most advanced Nvidia Tesla GPUs are built to accelerate research in multiple disciplines, such as artificial intelligence, deep learning, high performance computing, data science, visualization, and life science. Currently Applied Computing GPU cluster consists of five GPU servers, each with 2 Intel Xeon Silver 4212 Processors, 32 GB DDR4 memory, 2TB SSD storage, and 4 Nvidia Tesla V100 32GB GPUs. Each V100 GPU, equipped with 640 Tensor Cores, offers the equivalent performance of up to 32 CPUs. The servers in the GPU cluster are connected by InfiniBand switch, which enables the high interconnect performance with a non-blocking switching capacity of 7Tb/s to achieve high throughput and low latency.

The latest machine learning/deep learning frameworks (e.g., Tensorflow, Pytorch, MXNet, Keras, Caffe), libraries (e.g., CUDA toolkit, OpenCV, Anaconda, NVIDIA Digits), and Nvidia drivers are installed in the cluster, through a developer-ready container environment. The flexibility of containerization makes the developers easy to deploy custom deep learning environments on the GPU cluster. The GPU cluster is managed by a Linux server administration tool, Cockpit, which provides a user-friendly interface to manage and administer a variety of server resources, such as containers, virtual machines, user accounts, network interfaces, and storage devices.

