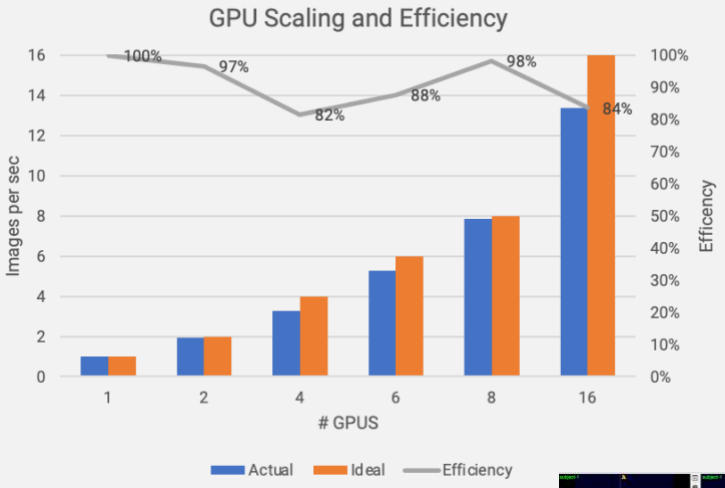
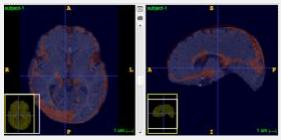
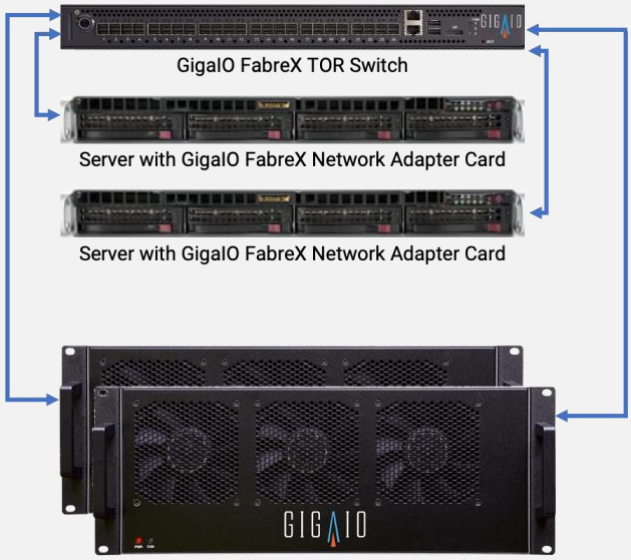
Composed Architecture Demonstrates Near-Linear Scaling from 1-16 GPUs Highly Efficient GPUDirect RDMA





Configuration



2 x GigaIO FabreX Accelerator Pooling Appliances with 16 NVIDIA GPUs

PCIe Fabric Unleashes GPU Performance and Scale for Medical AI Training and Inference Applications

Composed configuration -- The GigaIO configuration is a composed architecture with two 1U servers connected to the FabreX™ PCIe Network fabric. The configuration uses 16 NVIDIA GPUs distributed over two Accelerator Pooling Appliances. GPU to GPU communication uses GPUDirect RDMA, fully utilizing FabreX at PCIe latency and bandwidth to achieve scaling and efficiency.

AI Software Stack –This AI and Inferencing workload uses Distributed TensorFlow and NCCL Libraries. GPUs communicate with each other in the same PCIe domain using GPUDirect P2P DMA transfers and between PCIe domains using highly optimized GPUDirect RDMA, fully utilizing the inherent PCIe low latency and high bandwidth.

Results Summary – Test results demonstrate the dramatic impact of GPU scaling. The chart above shows the Ideal scaling and actual measured performance. With all 16 GPUs operating in parallel the FabreX efficiency is 84%, compared to the Ideal linear scaling at 100%.

Modern workloads perpetually grow and change, so data center architecture need to be flexible to support changing business needs. Deploying a FabreX composed architecture is flexible and easily reconfigured via software. FabreX delivers faster time-to-solution with exceptional scaling and efficiency together with increased utilization and sharing of resources, thus reducing CapEx or OpEx.

GigaIO FabreX is a Rack-Scale composable infrastructure solution that delivers the unlimited flexibility and agility of the cloud, at a fraction of the cost. Benefits include:

Improved system agility by disaggregating system resources on the fly and creating shared resource pools that can then be dynamically composed in real-time.

Slashed Total Cost of Ownership by enabling device sharing which increases resource utilization and eliminates over provisioning, resulting in reduced CapEx and OpEx.

Simplified and automated system set-up, administration and serviceability with freedom of choice for management tools from powerful CLI and Redfish APIs to ready-to-run, off-the-shelf enterprise-class orchestration software.

Seamless support for any PCIe-compliant device including servers, CPUs, memory, 3D-XPoint, storage, GPUs, FPGAs, specialty ASICS and NICs.

Blazing system performance with industry leading PCIe latency and bandwidth throughout the rack and beyond. As PCIe resources are added they immediately benefit from the native PCIe performance as all data transfers and buffers are completely eliminated.

Visit www.gigaio.com to discover more about GigaIO and FabreX, the industry's only pure PCIe Network Fabric.

In collaboration with University of Southern California