DELL EMC

PowerEdge Product Group

Direct from Development

Elastic AI Infrastructure using Dell EMC PowerEdge and Bitfusion FlexDirect

Tech Note by:

Ramesh Radhakrishnan, Dell EMC

Subbu Rama, Bitfusion

SUMMARY

Bitfusion FlexDirect disaggregates GPU accelerators and re-aggregates them in real-time over Ethernet, Infiniband RDMA or RoCE network, to create an elastic AI infrastructure.

Just like network attached storage, FlexDirect allows customers to do network attached GPUs.

FlexDirect on Dell EMC PowerEdge servers offers a seamless way for any machine in the network to access any arbitrary fraction of GPU or multiple GPUs, anytime.

bitfusion

Application performance demands have increasingly been outpacing Moore's Law in a variety of fields, particularly AI and deep learning. Co-processors like GPUs offer immense speedup to applications in fields like AI and deep learning, compared to CPUs. AI and deep learning applications requires truly elastic compute infrastructure from dev-test to model training and inference in order to achieve high utilization of infrastructure resources. In organizations, GPU accelerated servers are usually operated as silo-ed, stand-alone assets, causing increased CAPEX and OPEX as well as slow datacenter modernization.

The benefit of combining Dell EMC's powerful portfolio of compute, storage and networking with Bitfusion's FlexDirect software allows our customers to consolidate multiple silo-ed GPU clusters into a single shared resource pool, to decrease CAPEX and OPEX as well as increase productivity.

Composable Elastic AI Compute Platform

Bitfusion FlexDirect enables GPUs to be available as first-class resource on any machine in a PowerEdge Cluster that can be abstracted, partitioned, automated and shared much like traditional compute or storage resource. GPU accelerators can be partitioned into multiple virtual GPUs of any size and accessed remotely by any machine, over the network. With this, GPU accelerators are now part of a common infrastructure resource pool and available for use by anyone in the environment.

Organizations can scale the operations with policies and business logic (time of day policies, class of users, permission to access the top performance GPUs per user class, etc.) for AI development and production use cases. GPUs from different departments can be pooled to create bigger clusters to increase compute performance and infrastructure utilization.

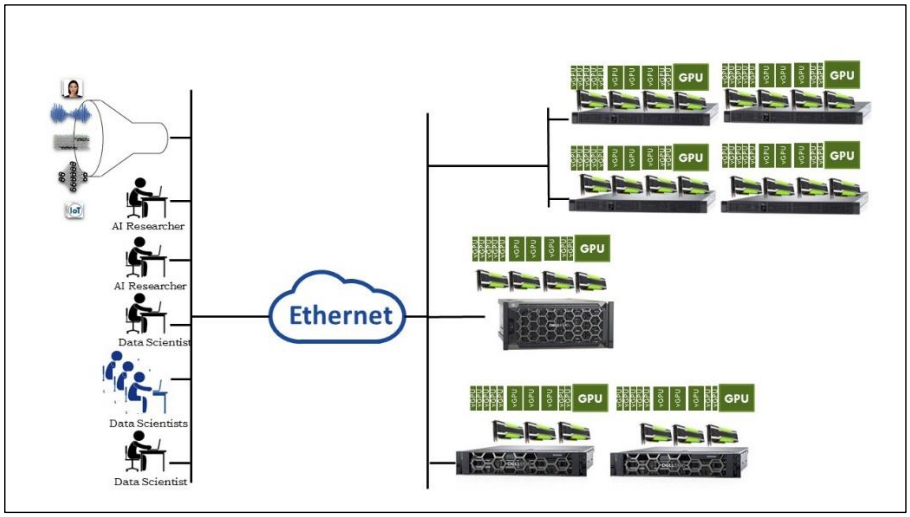

Figure 1: FlexDirect on Dell EMC PowerEdge Servers to create Elastic AI Infrastructure

DELL EMC

Increased Productivity across AI Development, Training and Deployment

IT regains the ability to assign GPU resources based on organization business priorities, and remotely pool together resources, while attaching them in real-time to the workloads, with known schedule and utilization plan. For example, GPU resources from Department A which completed intensive training and development schedule, can be reassigned to Department B which now experiences peak demand for GPUs for an urgent AI project.

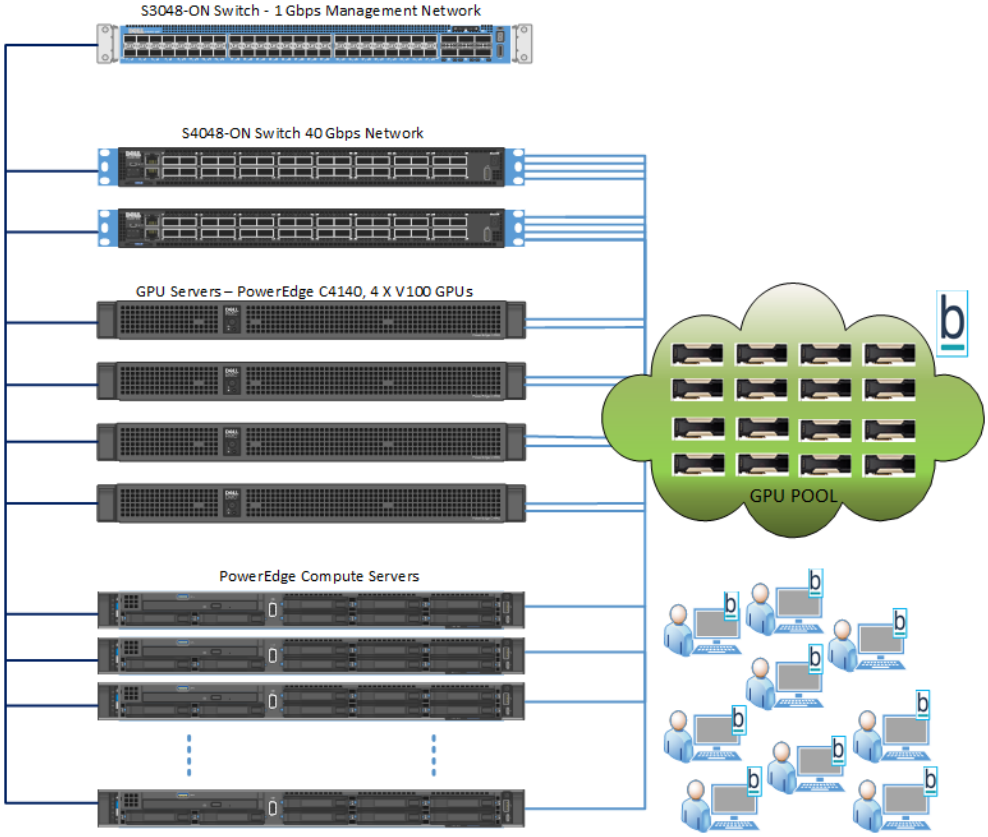

Figure 2: Dell EMC PowerEdge with FlexDirect Elastic AI Infrastructure Reference Architecture

Dell EMC and Bitfusion designed and validated the reference configuration shown in Figure 2 to help deployment of Elastic AI Infrastructure in customer datacenters. Integration of Bitfusion FlexDirect doesn't necessitate any changes to OS, drivers or AI frameworks. The intent of the tests were to prove the AI developer experience is the same, as if the GPUs are attached locally to the servers where the workloads are being executed, compared with executing the CUDA calls in a remote GPU (or GPUs). Standard AI benchmarks were used, with a variety of frameworks, models, batch sizes and network configurations namely 10G TCP, 10G RoCE, 40G RoCE and Infiniband EDR to simulate a range of customer environments. The results for AI model training using Tensorflow in a 40GbE RoCE environment are shown in Figure 3 and 4, respectively, on the next page.

Figure 3 shows the measurement of the performance for remote attach of GPUs (on PE-C4140) over the network compared against running the same workload locally on the GPU system. Figure 4 shows the performance of fractional GPUs (that can be shared) and shows how the aggregate performance is similar to using a full physical GPU. Across models, batch sizes and tests, Dell EMC PowerEdge with Bitfusion FlexDirect demonstrated that network attached full and fractional GPUs accomplish near native performance across the suite of benchmarks. Please contact Dell or Bitfusion to get the additional details regarding the performance benchmarking shared in this brief.

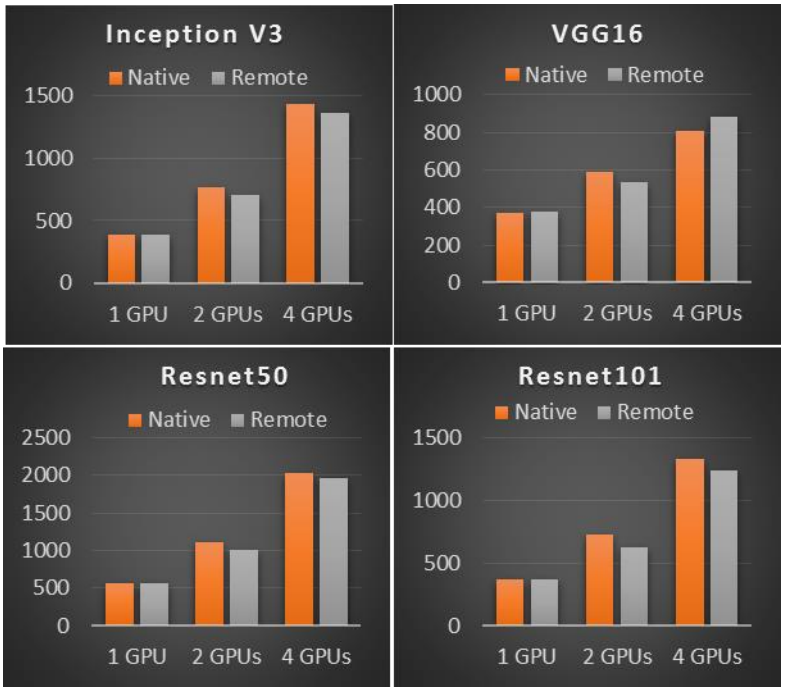

Figure 3: Performance comparison of Network Attached Remote GPUs with FlexDirect on Dell EMC PowerEdge Servers vs. Native GPU execution Performance is represented on y-axis (images/sec, higher is better), for training different CNN models using Tensorflow.

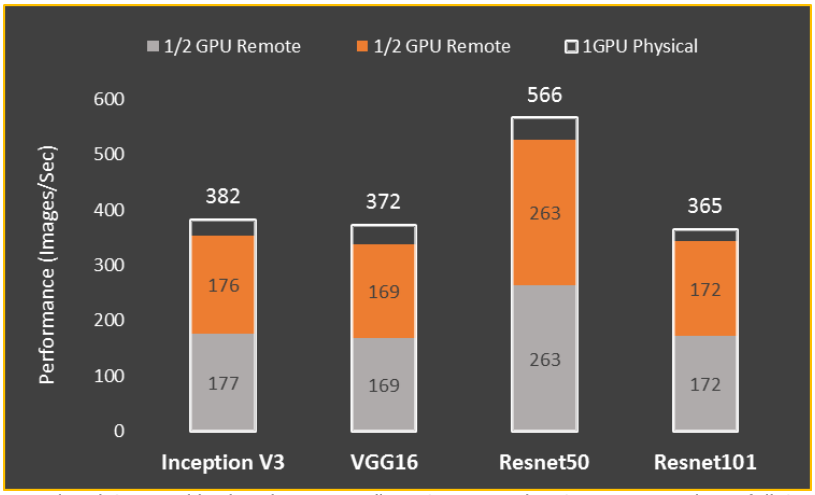

Figure 4: Fractional GPUs with FlexDirect on Dell EMC PowerEdge Servers vs. using a full GPU. The aggregate performance of fractional GPUs (which can be assigned to different users) is compared against a full GPU when training different model types using TensorFlow.

Conclusions

Bitfusion FlexDirect with Dell EMC PowerEdge Servers can bring flexibility and composability to the IT infrastructure along with reducing overall Capex and Opex with shared GPU pools across multiple organizations, business needs and use-cases. In addition, Bitfusion FlexDirect on Dell EMC PowerEdge Servers can increase user and organization productivity by meeting GPUs consumption from a consolidated resource pool based on real-time workload needs, to any machine in the environment.

Learn More

Visit https://www.bitfusion.io/product/flexdirect to get more information on Elastic Network Attached GPUs.
Visit http://www.dellemc.com/AI to get more information on Dell EMC PowerEdge Solutions for AI.