Deep Learning at Scale on NVIDIA V100 Accelerators

HPC and AI Innovation Lab
Rengan Xu, Frank Han, Quy Ta

PMBS18 Workshop, Supercomputing 18
November 12, 2018

DELL EMC

Outline

- Background
- Deep Learning Training
  - Single GPU vs Multi-GPU
  - V100-SXM2 vs V100-PCIe
  - FP16 vs FP32
  - V100 vs P100
  - Storage Considerations
  - Node Interconnect Considerations
- Deep Learning Inference
- Conclusions and Future Work

DELL EMC

Background

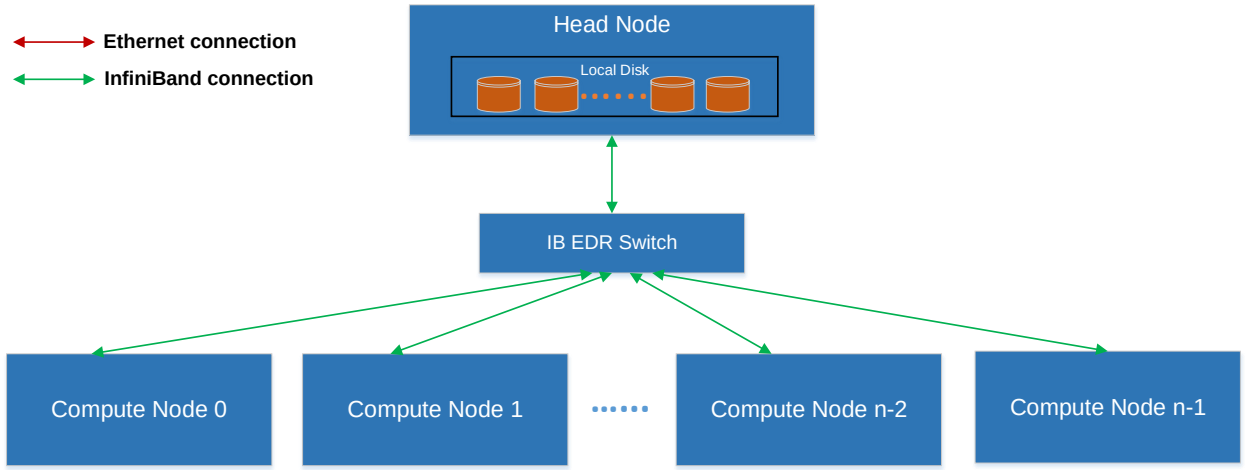
- V100 GPUs
  - V100-PCIe: GPUs are connected by PCIe buses
  - V100-SXM2: GPUs are connected by NVLink
- Deep Learning Frameworks
  - TensorFlow
  - Horovod: distributed framework for TensorFlow
  - MXNet
  - Caffe2
- Model and dataset
  - Model: resnet50
  - Dataset: ILSVRC2012

Description	V100-PCIe	V100-SXM2
CUDA Cores	5120	5120
GPU Max Clock Rate (HMz)	1380	1530
Tensor Cores	640	640
Memory Bandwidth (GB/s)	900	900
NVLink Bandwidth (GB/s)	N/A	300
Deep Learning (Tensor OPS)	112	120

DELL EMC

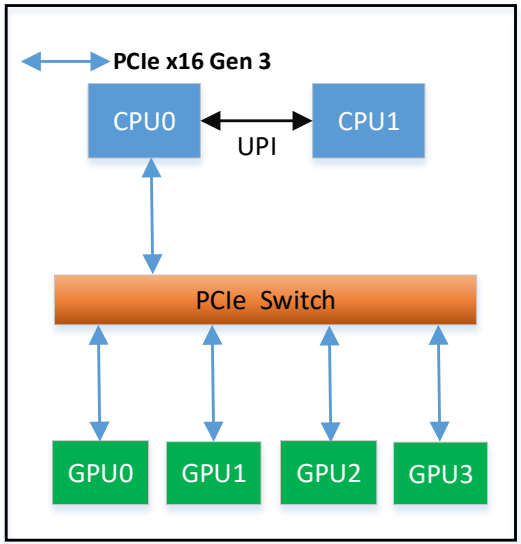
Architecture

- Head Node: Dell EMC PowerEdge R740xd
- Compute Nodes: Dell EMC PowerEdge C4140
- Storage: 9TB NFS through IPoIB on EDR InfiniBand
- Node Interconnect: EDR 100Gbps InfiniBand

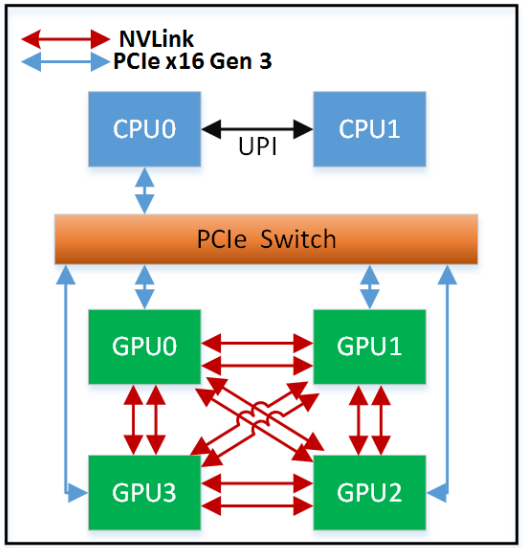


Dell EMC PowerEdge C4140 server

- Two types of compute nodes: nodes with V100-PCIe and V100-SXM2



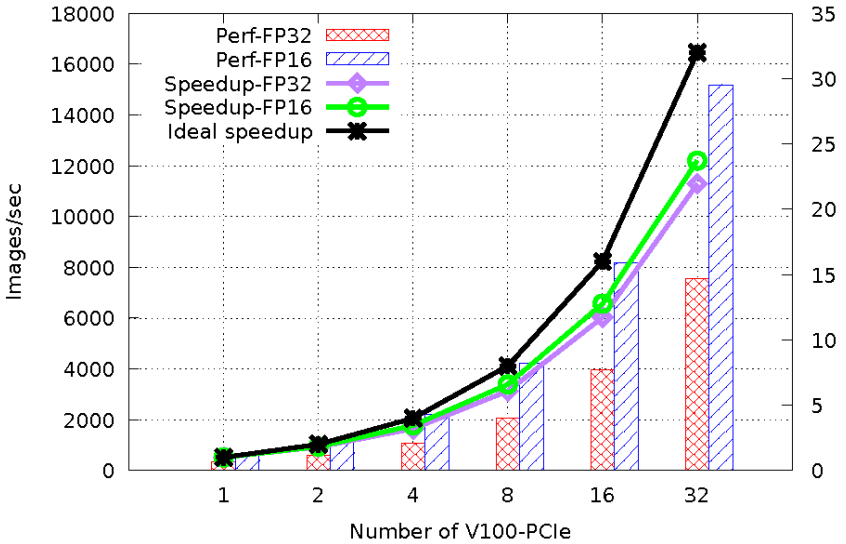
(a) V100-PCIe



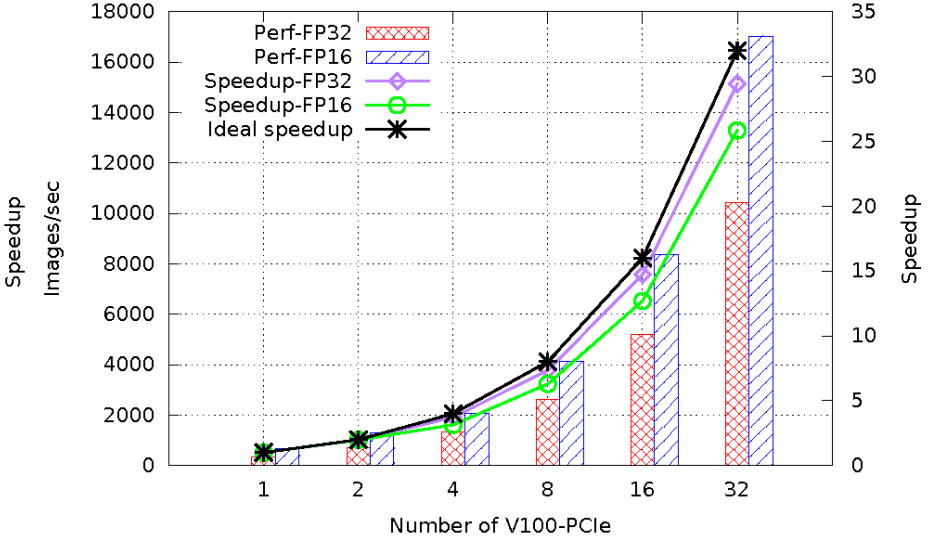
(b) V100-SXM2

Single GPU vs Multi-GPU

- TensorFlow scales well with 21.95x in FP32 and 23.72x in FP16
- MXNet scales the best with 29.43x in FP32 and 25.84 in FP16



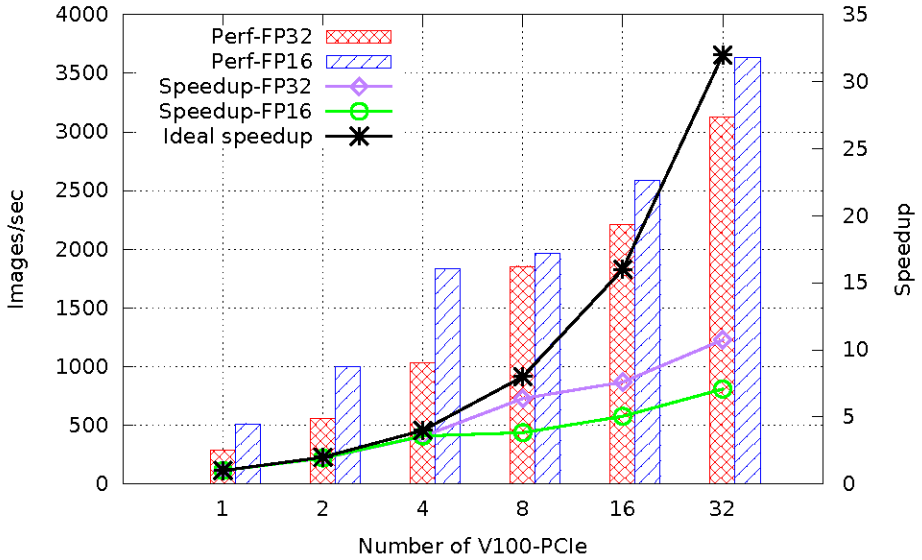
(a) TensorFlow + Horovod



(b) MXNet

Single GPU vs Multi-GPU

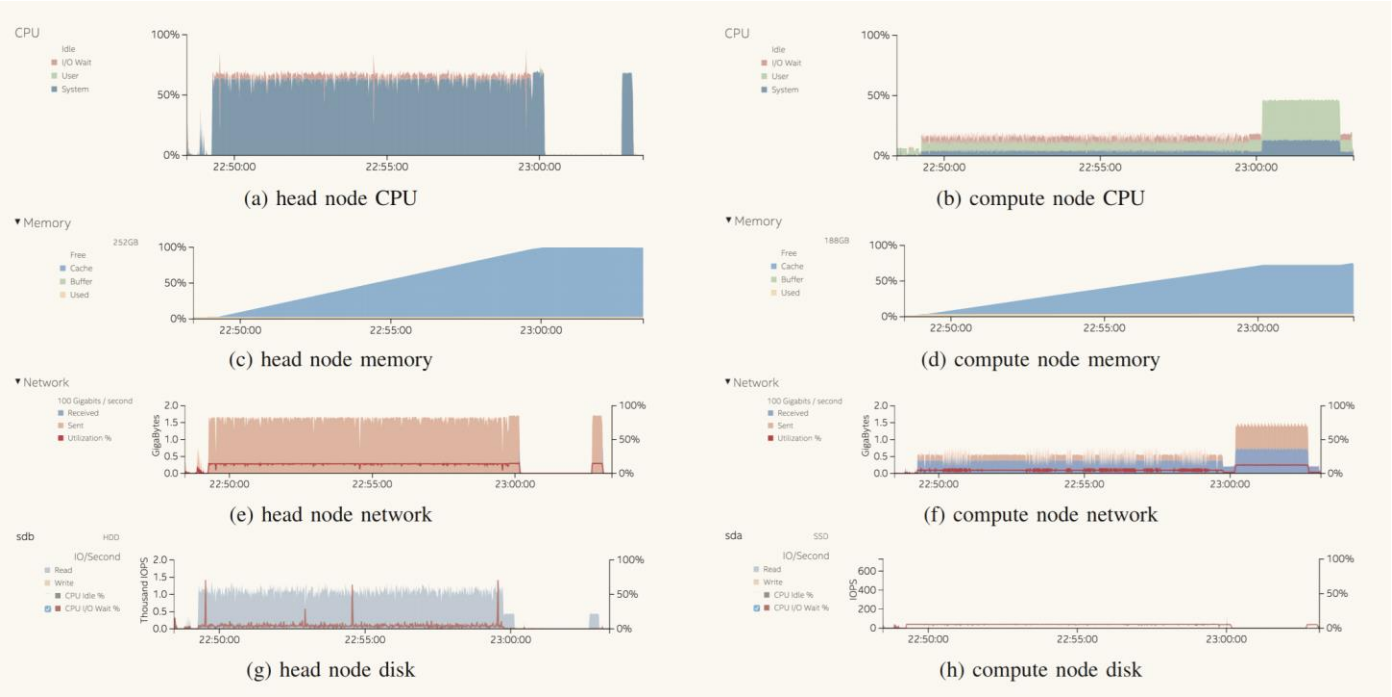
- Caffe2 scales well within a node: speedup of 3.55x in FP32 and 3.58 in FP16
- Caffe2 does not scale well with multiple nodes



DELL EMC

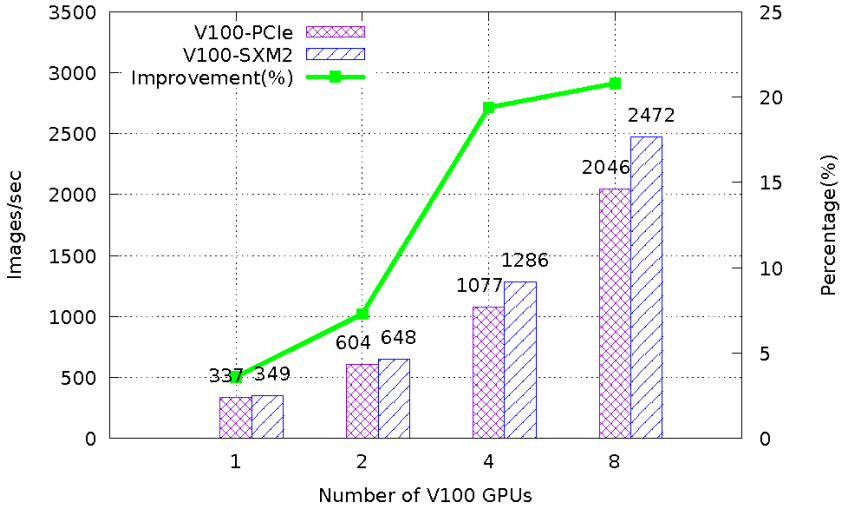
Caffe2 Performance Profiling

- ILSVRC2012 image database for Caffe2 is ~260GB
- This issue exists both in their Gloo and MPI implementations

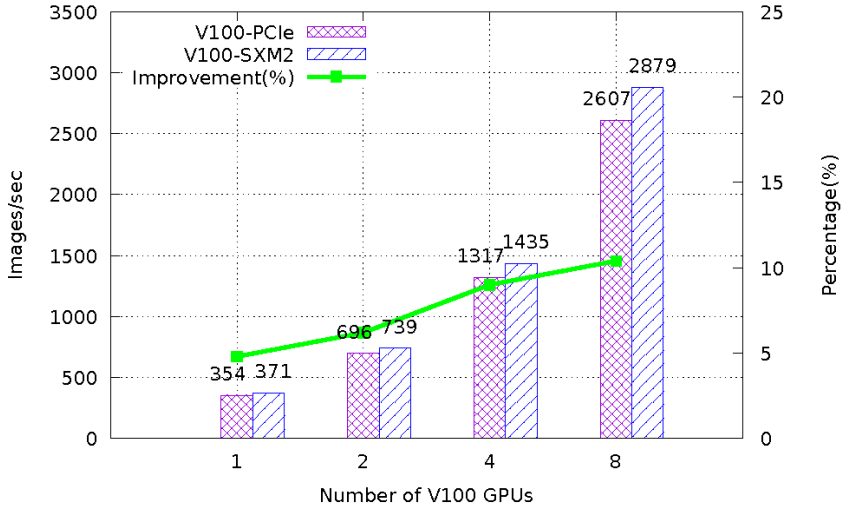


V100-SXM2 vs V100-PCIe

- The max clock rate for V100-SXM2 is ~10.9% higher than V100-PCIe
- For single GPU, the performance improvement of SXM2 is 3.6% - 6.9% better than PCIe
- With 2 nodes, the improvement is ~10% - 20%.



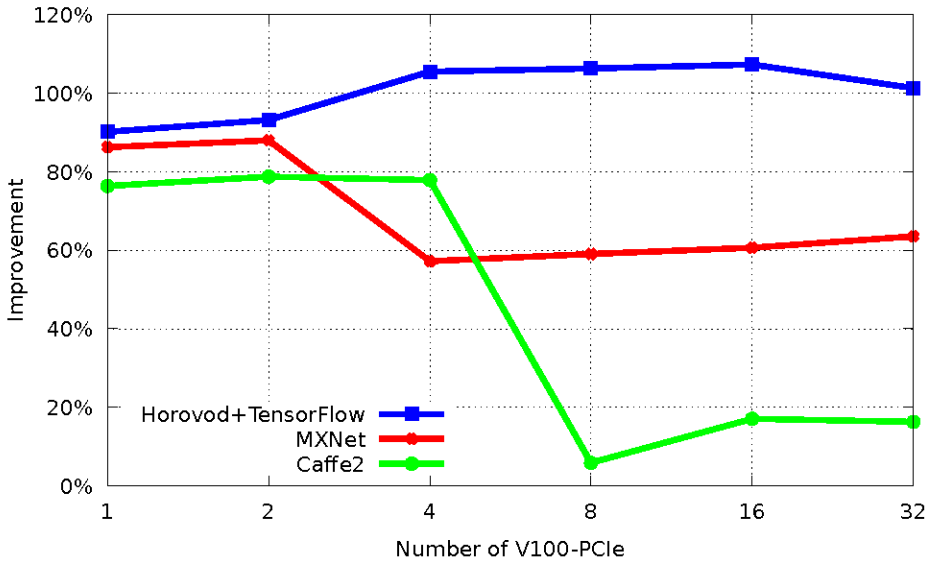
(a) TensorFlow + Horovod



(a) MXNet

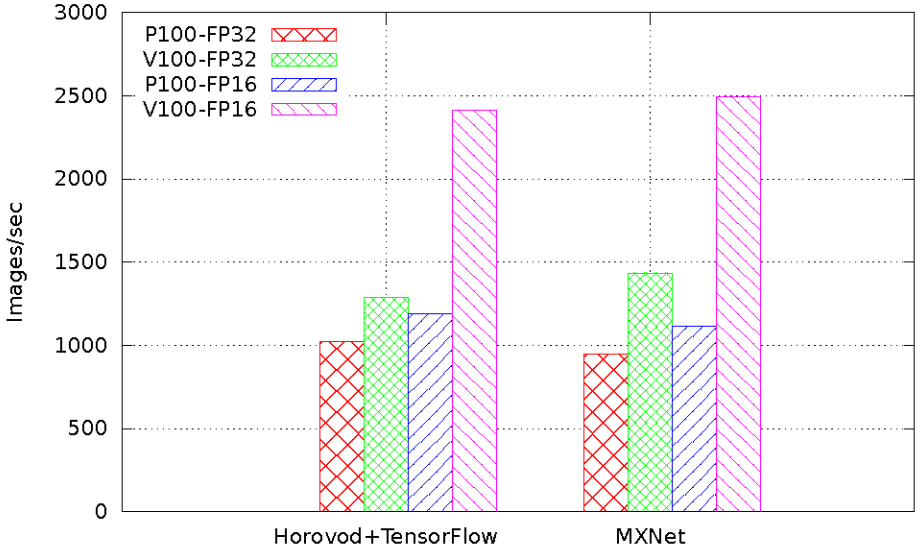
FP16 vs FP32

- Except Caffe2, FP16 is ~60% - 107% faster than FP32
- The performance gain comes from Tensor Cores



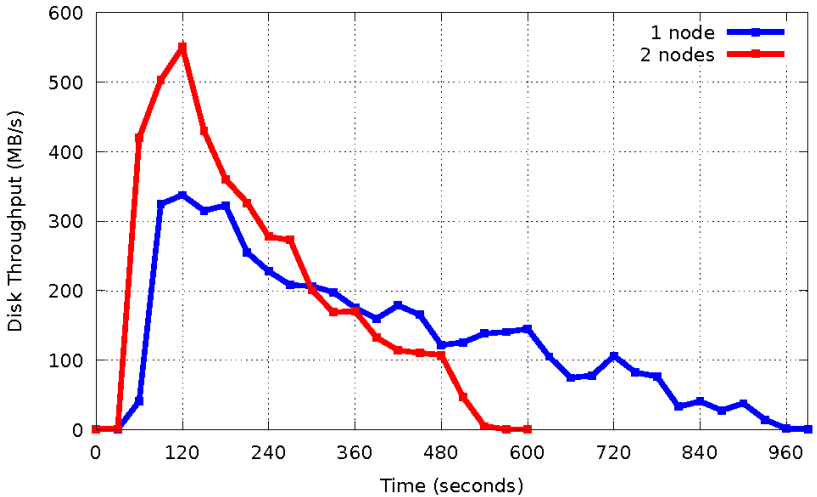
V100 vs P100

- The test was on 4 GPUs within one node. V100-SXM2 and P100-SXM2 were used.
- In FP32, V100 is 26% faster with TensorFlow, and 52% faster with MXNet
- In FP16, V100 is 103% faster with TensorFlow, and 123.8% faster with MXNet

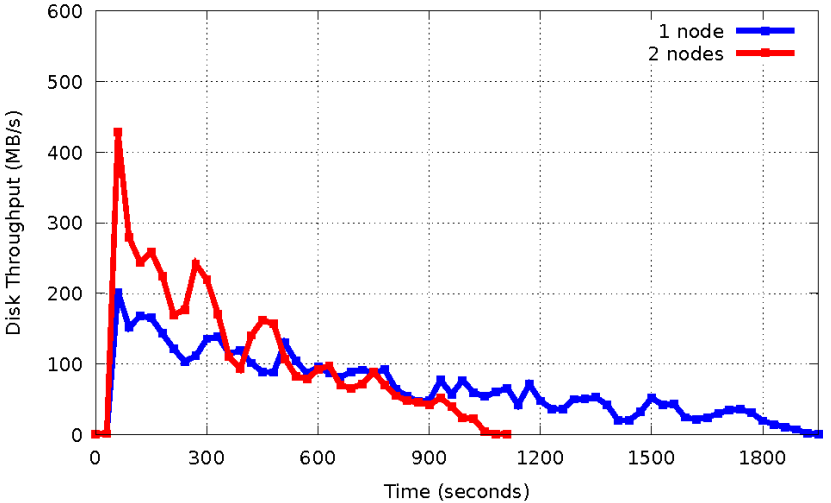


Storage Considerations - Horovd + TensorFlow

- The peak throughput almost doubles when doubling the number of nodes
- Two epochs. The read throughput is decreasing (1024 TFRecords files)
- Different MPI processes do not read distinct portions of the dataset.



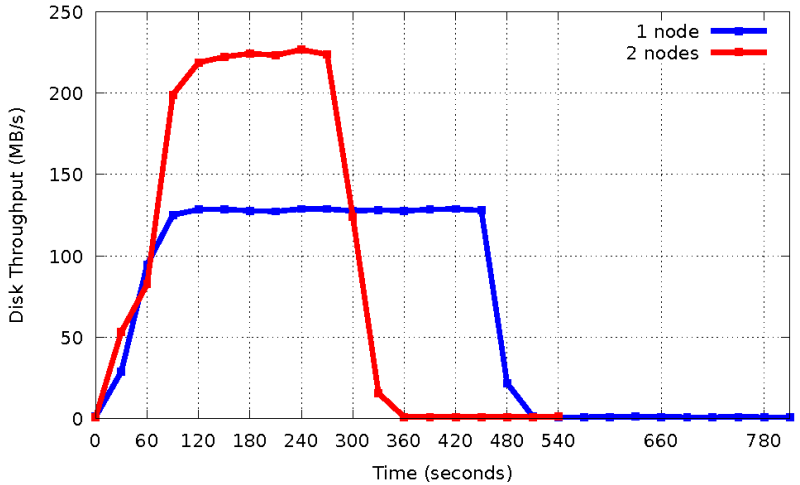
(a) Horovod+TensorFlow FP16



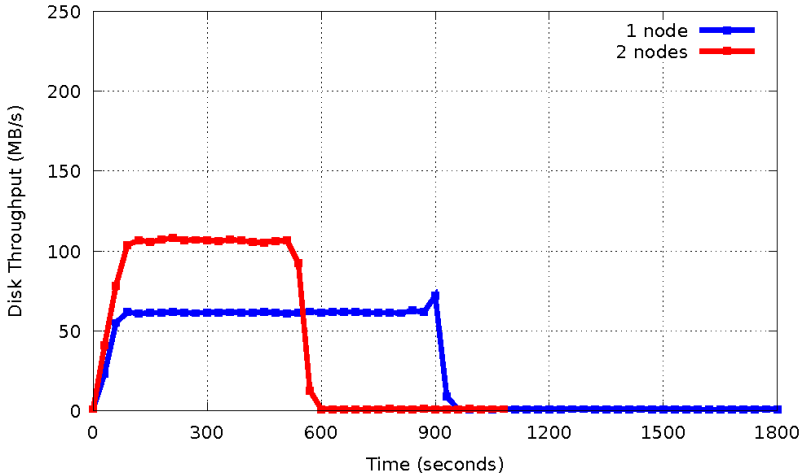
(b) Horovod+TensorFlow FP32

Storage Considerations - MXNet

- Two epochs were run
- The read throughput is consistent in the first epoch for MXNet, then the data is cached in memory
- One RecordIO database file



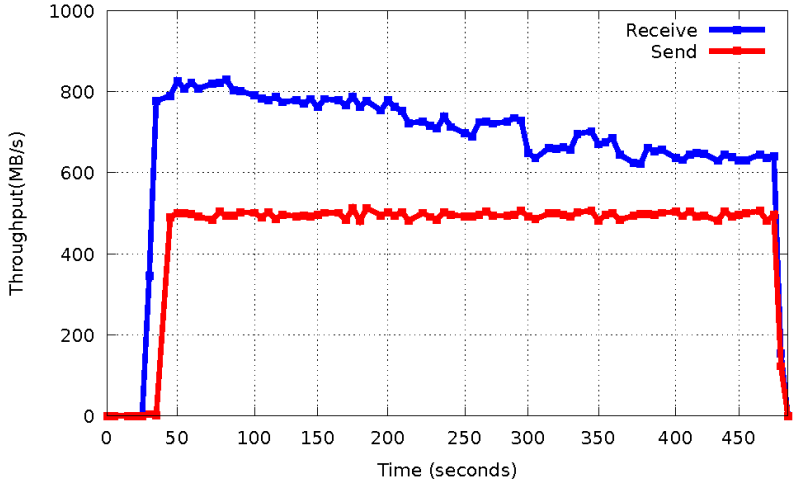
(a) MXNet FP16



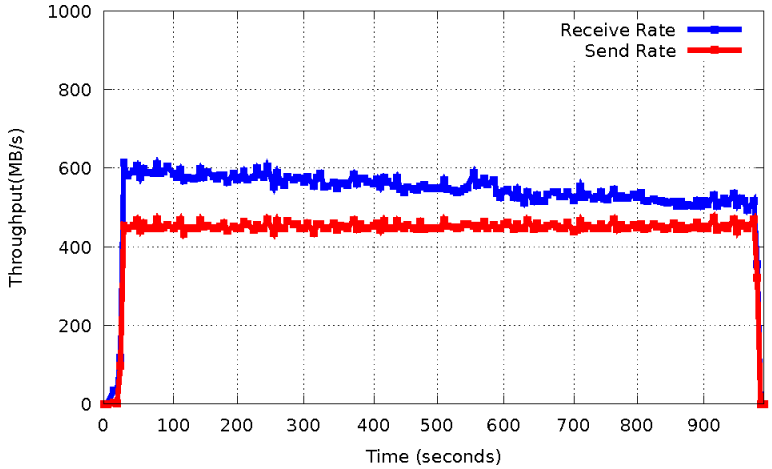
(b) MXNet FP32

Node Interconnect Considerations - Horovod+TensorFlow

- Testing on 2 V100-SXM2 compute nodes
- The sent data includes the gradients exchange. Slightly higher with FP16: ~450 MB/s in FP32 and 480 MB/s in FP16
- The received data includes both the gradients exchange and input data from storage
- The data shuffle uses a buffer to sample from the whole dataset. Buffer size=10,000, training samples=1,281,167



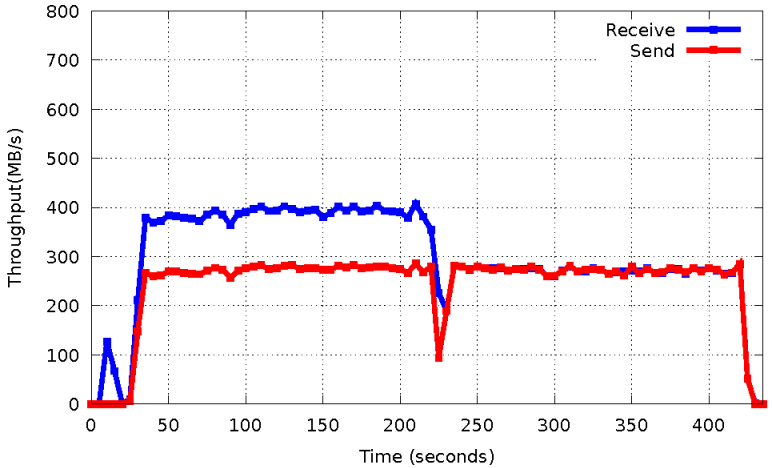
(a) FP16



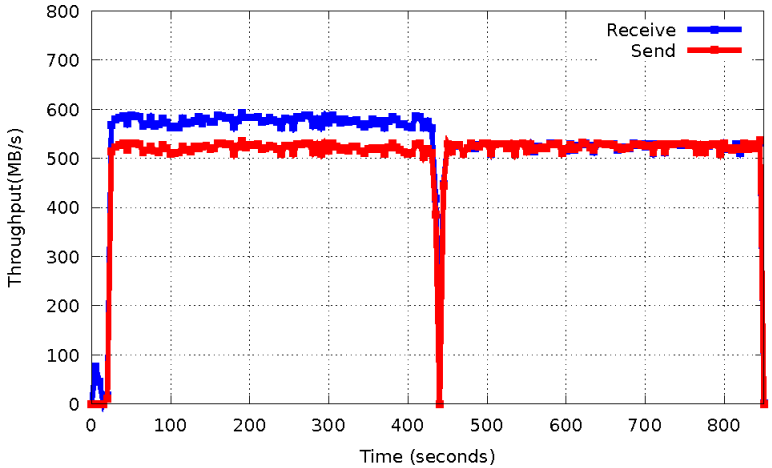
(b) FP32

Node Interconnect Considerations - MXNet

- Testing on 2 V100-SXM2 compute nodes
- The sent data includes the gradients exchange. Much lower with FP16: ~280 MB/s in FP16 and 520MB/s in FP32
- The received data includes both the gradients exchange and input data from storage
- The sudden drop is data shuffle after one epoch



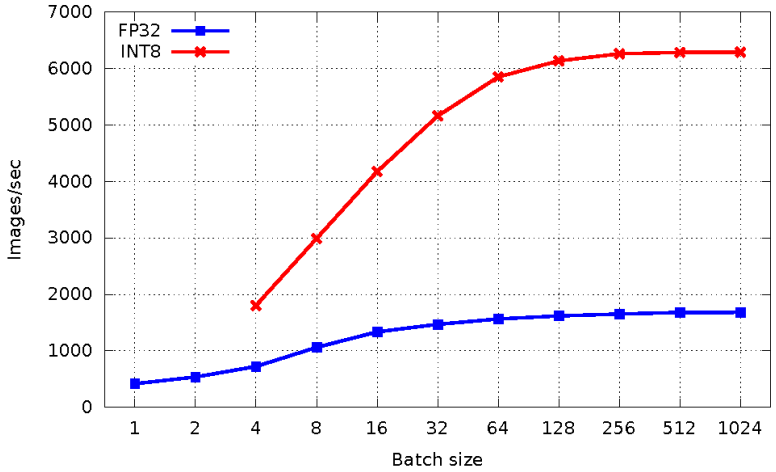
(a) FP16



(b) FP32

Deep Learning Inference - Performance

- Inference with TensorRT 4.0.0 and on one V100-PCIe
- INT8 works only if the batch size is evenly divisible by 4
- INT8 is 2.5x – 3.5x faster than FP32 for batch size < 64
- INT8 is ~3.7x faster than FP32 for batch size >= 64



Inference performance with INT8 vs FP32 for Resnet50 model

Deep Learning Inference - Accuracy

- A calibration is needed to make INT8 to encode the same information as FP32.
- The accuracy differences between INT8 and FP32 is only 0.02% - 0.18%
- NO accuracy lost for INT8, while achieving 3x speedup over FP32

Neural Network Model	FP32		INT8		Difference	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
ResNet-50	72.90%	91.14%	72.84%	91.08%	0.07%	0.06%
ResNet-101	74.33%	91.95%	74.31%	91.88%	0.02%	0.07%
ResNet-152	74.90%	92.21%	74.84%	92.16%	0.06%	0.05%
VGG-16	68.35%	88.45%	68.30%	88.42%	0.05%	0.03%
VGG-19	68.47%	88.46%	68.38%	88.42%	0.09%	0.03%
GoogLeNet	68.95%	89.12%	68.77%	89.00%	0.18%	0.12%
AlexNet	56.82%	79.99%	56.79%	79.94%	0.03%	0.06%

DELL EMC

Conclusions and Future Work

- Conclusions
  - The Resnet50 training with Caffe2 is not stable in multi-node when the data is not cached into memory
  - The Resnet50 training with TensorFlow and MXNet scales well with more GPUs and nodes
  - The disk throughput almost doubles when doubling the number of nodes, but the throughput of the gradient exchange does not
  - Network throughput reveals different implementations in TensorFlow and MXNet
  - Inference with INT8 can achieve the same accuracy as FP32, but 3.7x faster
- Future Work
  - Recurrent Neural Networks
    - Neural Machine Translation (text2text)
    - Automatic Speech Recognition (speech2text)
    - Speech Synthesis (text2speech)
  - Accuracy study for large-scale runs (currently the accuracy drops with more GPUs)
  - Model parallel neural network implementations

DELL EMC


DELL EMC