Deep Learning Performance on R740 with V100-PCIe GPUs

Authors: Rengan Xu, Frank Han, Nishanth Dandapanthula

Dell EMC HPC Innovation Lab. February 2018

Overview

The Dell EMC PowerEdge R740 is a 2-socket, 2U rack server. The system features the Intel Skylake processors, up to 24 DIMMs, and up to 3 double width or 6 single width GPUs. In our previous blog Deep Learning Inference on P40 vs P4 with SkyLake, we presented the deep learning inference performance on Dell EMC's PowerEdge R740 server with P40 and P4 GPUs. This blog will present the performance of the deep learning training performance on single R740 with multiple V100-PCIe GPUs. The deep learning frameworks we benchmarked include Caffe2, MXNet and Horovod+TensorFlow. Horovod is a distributed framework for TensorFlow. We used Horovod because it has better scalability implementation (using MPI model) than TensorFlow, which has been explained in the article "Meet Horovod: Uber's Open Source Distributed Deep Learning Framework for TensorFlow". Table 1 shows the hardware configuration and software details we tested. To test the deep learning performance and scalability on R740 server, we used the same neural network, the same dataset and the same measurement as in our other deep learning blog series such as Scaling Deep Learning on Multiple V100 Nodes and Deep Learning on V100.

Table 1: The hardware configuration and software details

Platform	**PowerEdge R740**
CPU	2 x Intel Xeon 6150 @2.7GHz (Skylake)
Memory	192GB DDR4 @ 2667MHz
Shared Storage	9TB NFS through IPoIB on EDR Infiniband
GPU	V100-PCIe
Software and Firmware	
Operating System	RHEL 7.3 x86_64
Linux Kernel	3.10.0-514.26.2.el7.x86_64
BIOS	2.4.2
CUDA compiler and GPU driver	CUDA 9.0 (387.26)
NCCL	2.0
Python	2.7.5
Deep Learning Libraries and Frameworks	
CUDNN	7.0
Caffe2	0.8.1
MXNet	0.11.1
TensorFlow	1.4.0
Horovod	1.11.1

Performance Evaluation

The Figure 1, Figure 2 and Figure 3 show the Resnet50 performance and speedup of multiple V100 GPUs with Caffe2, MXNet and TensorFlow, respectively. We can obtain the following conclusions based on these results:

- Overall the performance of Resnet50 scales well on multiple V100 GPUs within one node. With 3 V100:
  - Caffe2 achieved the speedup of 2.61x and 2.65x in FP32 and FP16 mode, respectively.
  - MXNet achieved the speedup of 2.87x and 2.82x in FP32 and FP16 mode, respectively.
  - Horovod+TensorFlow achieved the speedup of 2.12x in FP32 mode. (FP16 still under development)
- The performance in FP16 mode is around 80%-90% faster than FP32 for both Caffe2 and MXNet. TensorFlow still has not supported FP16 yet, so we will test its FP16 performance once this feature is supported.

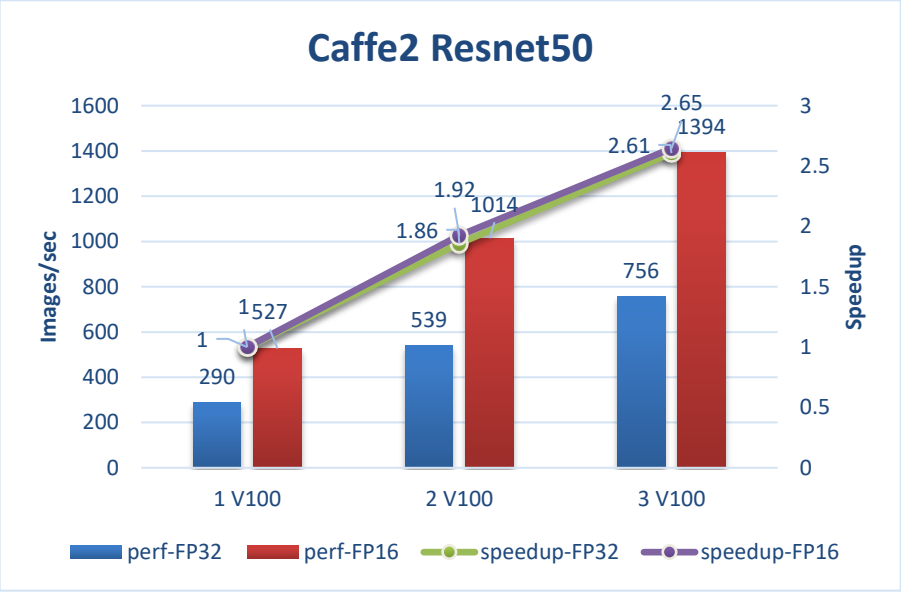

Figure 1: Caffe2: Performance and speedup of V100

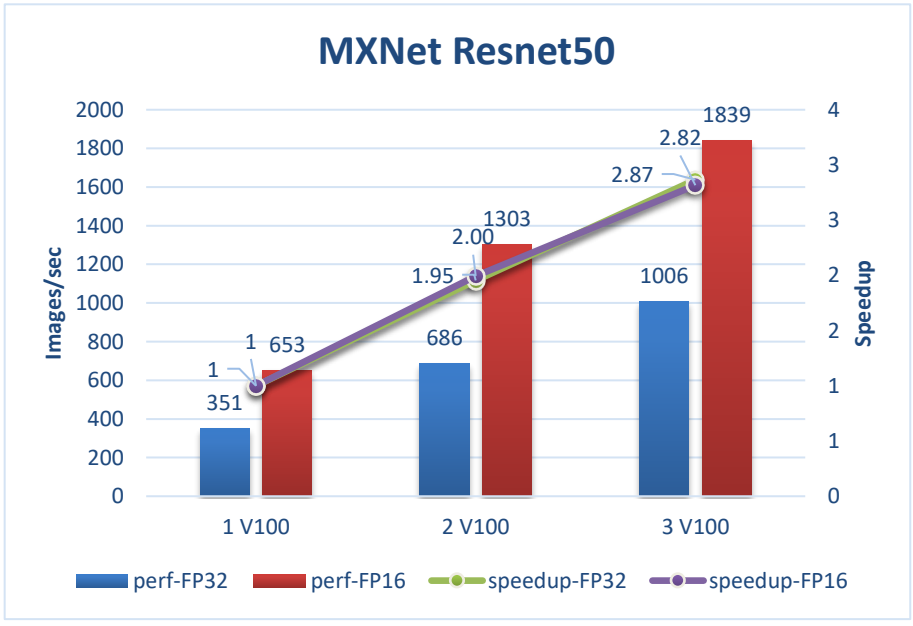

Figure 2: MXNet: Performance and speedup of V100

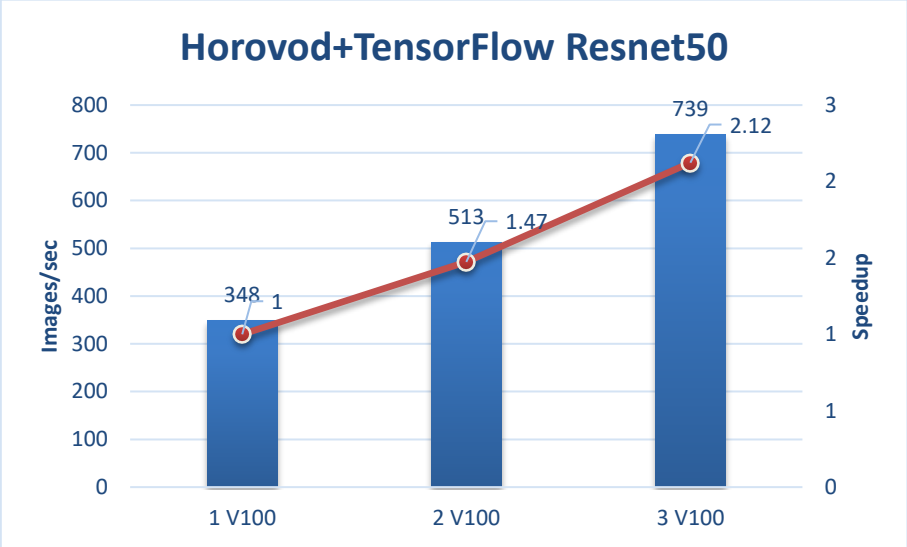

Figure 3: TensorFlow: Performance and speedup of V100

Conclusions

In this blog, we presented the deep learning performance and scalability of popular deep learning frameworks like Caffe2, MXNet and Horovod+TensorFlow. Overall the three frameworks scale as expected on all GPUs within single R740 server.