Deep Learning Inference on P40 GPUs

Authors: Rengan Xu, Frank Han and Nishanth Dandapanthu. Dell EMC HPC Innovation Lab. Mar. 2017

Introduction to P40 GPU and TensorRT

Deep Learning (DL) has two major phases: training and inference/testing/scoring. The training phase builds a deep neural network (DNN) model with the existing large amount of data. And the inference phase uses the trained model to make prediction from new data. The inference can be done in the data center, embedded system, auto and mobile devices, etc. Usually inference must respond to user request as quickly as possible (often in real time). To meet the low-latency requirement of inference, NVIDIA® launched Tesla® P4 and P40 GPUs. Aside from high floating point throughput and efficiency, both GPUs introduce two new optimized instructions designed specifically for inference computations. The two new instructions are 8-bit integer (INT8) 4-element vector dot product (DP4A) and 16-bit 2-element vector dot product (DP2A) instructions. Deep learning researchers have found using FP16 is able to achieve the same inference accuracy as FP32 and many applications only require INT8 or lower precision to keep an acceptable inference accuracy. Tesla P4 delivers a peak of **21.8** INT8 TIOP/s (Tera Integer Operations per Second), while P40 delivers a peak of **47.0** INT8 TIOP/s. This blog only focuses on P40 GPU.

TensorRT™, previously called GIE (GPU Inference Engine), is a high performance deep learning inference engine for production deployment of deep learning applications that maximizes inference throughput and efficiency. TensorRT provides users the ability to take advantage of fast reduced precision instructions provided in the Pascal GPUs. TensorRT v2 supports the INT8 reduced precision operations that are available on the P40.

Testing Methodology

This blog quantifies the performance of deep learning inference using TensorRT on Dell's PowerEdge C4130 server which is equipped with 4 Tesla P40 GPUs. Since TensorRT is only available for Ubuntu OS, all the experiments were done on Ubuntu. Table 1 shows the hardware and software details. The inference benchmark we used was giexec in TensorRT sample codes. The synthetic images which were filled with random non-zero numbers to simulate real images were used in this sample code. Two classic neural networks were tested: AlexNet (2012 ImageNet winner) and GoogLeNet (2014 ImageNet winner) which is much deeper and complicated than AlexNet.

We measured the inference performance in images/sec which means the number of images that can be processed per second. To measure the performance improvement of the current generation GPU P40, we also compared its performance with the previous generation GPU M40. The most important goal of this testing is to measure the inference performance in INT8 mode, compared to FP32 mode. P40 uses the new Pascal architecture and supports the new INT8 instructions. The previous generation GPU M40 uses Maxwell architecture and does not support INT8 instructions. The theoretical performance of INT8, FP32 in both M40 and P40 is shown in Table 2. We measured the performance FP32 on both devices and both FP32 and INT8 on the P40.

Table 1: Hardware configuration and software details

Platform	PowerEdge C4130 (configuration G)
Processor	2 x Intel Xeon CPU E5-2690 v4 @2.6GHz (Broadwell)
Memory	256GB DDR4 @ 2400MHz
Disk	400GB SSD
GPU	4x Tesla P40 with 24GB GPU memory
Software and Firmware	
Operating System	Ubuntu 14.04
BIOS	2.3.3
CUDA and driver version	8.0.44 (375.20)
TensorRT Version	2.0 EA

Table 2: Comparison between Tesla M40 and P40

	Tesla M40	Tesla P40
INT8 (TIOP/s)	N/A	47.0
FP32 (TFLOP/s)	6.8	11.8

Performance Evaluation

In this section, we will present the inference performance with TensorRT on GoogLeNet and AlexNet. We also implemented the benchmark with MPI so that it can be run on multiple P40 GPUs within a node. We will also compare the performance of P40 with M40. Lastly we will show the performance impact when using different batch sizes.

Figure 1 shows the inference performance with TensorRT library for both GoogLeNet and AlexNet. We can see that INT8 mode is ~3x faster than FP32 in both neural networks. This is expected since the theoretical speedup of INT8 is 4x compared to FP32 if only multiplications are performed and no other overhead is incurred. However, there are kernel launches, occupancy limits, data movement and math other than multiplications, so the speedup is reduced to about 3x faster.

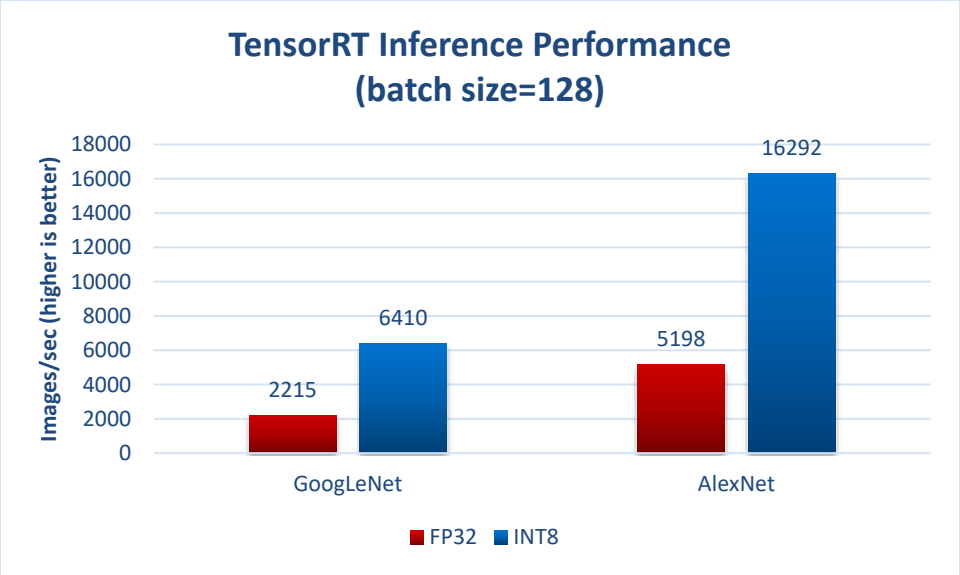

Figure 1: Inference performance with TensorRT library

Dell's PowerEdge C4130 supports up to 4 GPUs in a server. To make use of all GPUs, we implemented the inference benchmark using MPI so that each MPI process runs on each GPU. Figure 2 and Figure 3 show the multi-GPU inference performance on GoogLeNet and AlexNet, respectively. When using multiple GPUs, linear speedup were achieved for both neural networks. This is because each GPU processes its own images and there is no communications and synchronizations among used GPUs.

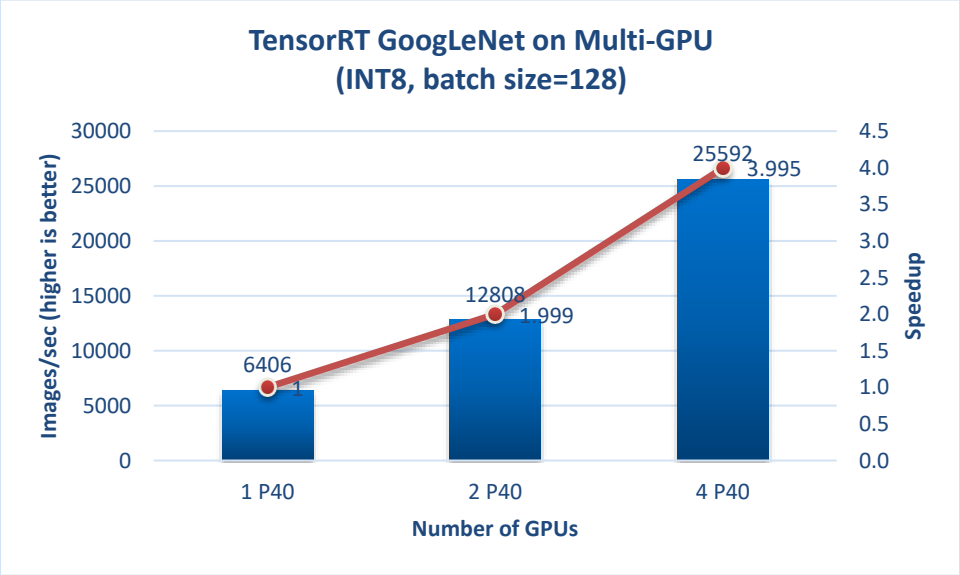

Figure 2: Multi-GPU inference performance with TensorRT GoogLeNet

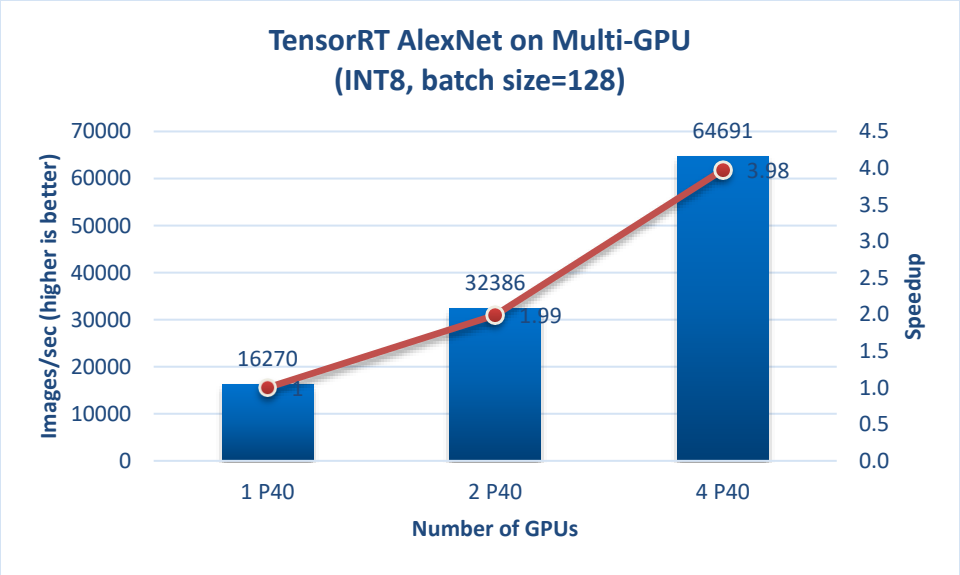

Figure 3: Multi-GPU inference performance with TensorRT AlexNet

To highlight the performance advantage of P40 GPU and its native support for INT8, we compared the inference performance between P40 with the previous generation GPU M40. The result is shown in Figure 5 and Figure 6 for GoogLeNet and AlexNet, respectively. In FP32 mode, P40 is 1.7x faster than M40. And the INT8 mode in P40 is 4.4x faster than FP32 mode in M40.

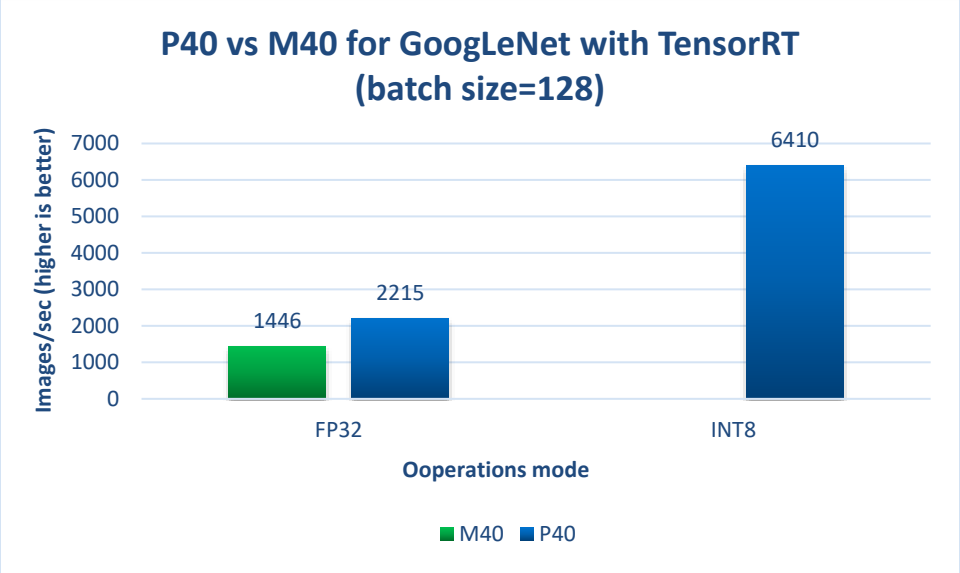

Figure 4: Inference performance comparison between P40 and M40

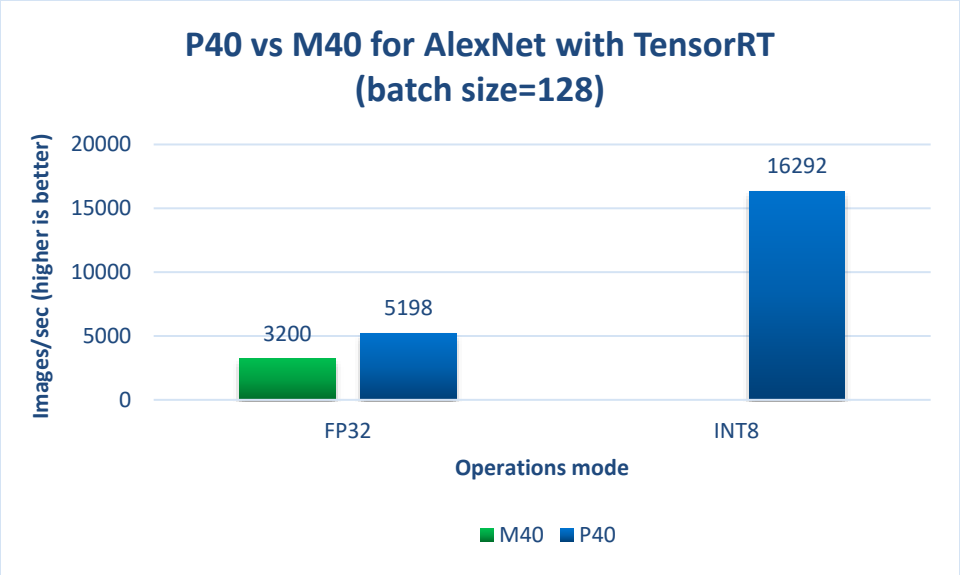

Figure 5: Inference performance comparison between P40 and M40

Deep learning inference can be applied in different scenarios. Some scenarios require large batch size and some scenarios even requires no batching at all (i.e. batch size is 1). Therefore we also measured the performance difference when using different batch sizes and the result is shown in Figure 6. Note that the purpose here is not comparing the performance of GoogLeNet and AlexNet, instead the purpose is to check how the performance changes with different batch sizes for each neural network. It can be seen that without batch processing the inference performance is very low. This is because the GPU is not assigned enough workloads to keep it busy. The larger the batch size is, the higher the inference performance is, although the rate of the speed increasing becomes slower. When batch size is 4096, GoogLeNet stopped running because the required GPU memory for this neural network exceeds the GPU memory limit. But AlexNet was able to run because it is a less complicated neural network than GoogLeNet and therefore it requires less GPU memory. So the largest batch size is only limited by GPU memory.

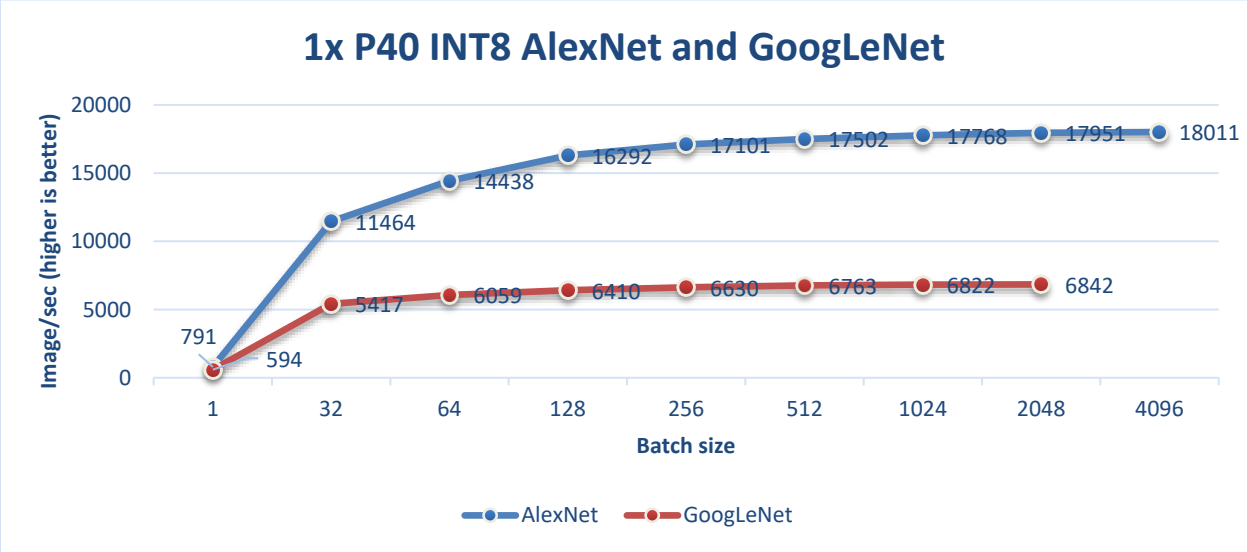

Figure 6: Inference performance with different batch sizes

Conclusions and Future Work

In this blog, we presented the inference performance in deep learning with NVIDIA® TensorRT library on P40 and M40 GPUs. As a result, the INT8 support in P40 is about **3x** faster than FP32 mode in P40 and **4.4x** faster than FP32 mode in the previous generation GPU M40. Multiple GPUs can increase the inferencing performance linearly because of no communications and synchronizations. We also noticed that higher batch size leads to higher inference performance and the largest batch size is only limited by GPU memory size. In the future work, we will evaluate the inference performance with real world deep learning applications.