# Deep Learning Inference on P40 vs P4 with Skylake

**Authors: Rengan Xu, Frank Han and Nishanth Dandapanthula. Dell EMC HPC Innovation Lab. July. 2017**

This blog evaluates the performance, scalability and efficiency of deep learning inference on P40 and P4 GPUs on Dell EMC's PowerEdge R740 server. The purpose is to compare P40 versus P4 in terms of performance and efficiency. It also measures the accuracy differences between high precision and reduced precision floating point in deep learning inference.

## Introduction to R740 Server

The PowerEdge™ R740 is Dell EMC's latest generation 2-socket, 2U rack server designed to run complex workloads using highly scalable memory, I/O, and network options. The system features the Intel Xeon Processor Scalable Family (architecture codenamed Skylake-SP), up to 24 DIMMs, PCI Express (PCIe) 3.0 enabled expansion slots, and a choice of network interface technologies to cover NIC and rNDC. The PowerEdge R740 is a general-purpose platform capable of handling demanding workloads and applications, such as data warehouses, ecommerce, databases, and high performance computing (HPC). It supports up to 3 Tesla P40 GPUs or 4 Tesla P4 GPUs.

## Introduction to P40 and P4 GPUs

NVIDIA® launched Tesla® P40 and P4 GPUs for the inference phase of deep learning. Both GPU models are powered by NVIDIA Pascal™ architecture and designed for deep learning deployment, but they have different purposes. P40 is designed to deliver maximum throughput, while P4's is aimed to provide better energy efficiency. Aside from high floating point throughput and efficiency, both GPU models 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. Although many HPC applications require high precision computation with FP32 (32-bit floating point) or FP64 (64-bit floating point), deep learning researchers have found using FP16 (16-bit floating point) is able to achieve the same inference accuracy as FP32 and many applications only require INT8 (8-bit integer) 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. Other differences between these two GPU models are shown in Table 1. This blog uses both types of GPUs in the benchmarking.

Table 1: Comparison between Tesla P40 and P4

| | Tesla P40 | Tesla P4 |
|---|---|---|
| **CUDA Cores** | 3840 | 2560 |
| **Core Clock** | 1531 MHz | 1063 MHz |
| **Memory Bandwidth** | 346 GB/s | 192 GB/s |
| **Memory Size** | 24 GB GDDR5 | 8 GB GDDR5 |
| **FP32 Compute** | 12.0 TFLOPS | 5.5 TFLOPS |
| **INT8 Compute** | 47 TIOPS | 22 TIOPS |
| **TDP** | 250W | 75W |

## Introduction to NVIDIA TensorRT

NVIDIA 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 new INT8 operations that are available on both P40 and P4 GPUs, and to the best of our knowledge it is the only library that supports INT8 to date.

## Testing Methodology

This blog quantifies the performance of deep learning inference using NVIDIA TensorRT on one PowerEdge R740 server which supports up to 3 Tesla P40 GPUs or 4 Tesla P4 GPUs. Table 2 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 more complicated than AlexNet.

We measured the inference performance in images/sec which means the number of images that can be processed per second.

Table 2: Hardware configuration and software details

| Platform | PowerEdge R740 |
|---|---|
| **Processor** | 2 x Intel Xeon Gold 6150 |
| **Memory** | 192GB DDR4 @ 2667MHz |
| **Disk** | 400GB SSD |
| **Shared storage** | 9TB NFS through IPoIB on EDR Infiniband |
| **GPU** | 3x Tesla P40 with 24GB GPU memory, or<br><br>4x Tesla P4 with 8 GB GPU memory |
| **Software and Firmware** | |
| **Operating System** | RHEL 7.2 |
| **BIOS** | 0.58 (beta version) |
| **CUDA and driver version** | 8.0.44 (375.20) |
| **NVIDIA TensorRT Version** | 2.0 EA and 2.1 GA |

## Performance Evaluation

In this section, we will present the inference performance with NVIDIA TensorRT on GoogLeNet and AlexNet. We also implemented the benchmark with MPI so that it can be run on multiple GPUs within a server. Figure 1 and Figure 2 show the inference performance with AlexNet and GoogLeNet on up to three P40s and four P4s in one R740 server. In these two figures, batch size 128 was used. The power consumption of each configuration was also measured and the energy efficiency of the configurations is

plotted as a “performance per watt” metric. The power consumption was measured by subtracting the power when the system was idle from the power when running the inference. Both the images/sec and images/sec/watt metrics numbers are relative to the numbers on one P40. Figure 3 shows the performance with different batch sizes with 1 GPU, and both metrics numbers are relative to the numbers on P40 with batch size 1. In all figures, INT8 operations were used. The following conclusions can be observed:

- **Performance**: with the same number of GPUs, the inference performance on P4 is around half of that on P40. This is consistent with the theoretical INT8 performance on both types of GPUs: 22 TIOPS on P4 vs 47 TIOPS on P40 on single GPU. Also since inference with larger batch sizes gives higher overall throughput but consumes more memory, and P4 has only 8GB memory compared to P40 24GB memory, P4 could not complete the inference with batch size 2048 or larger.
- **Scalability**: the performance scales linearly on both P40s and P4s when multiple GPUs are used, because of no communication happens between the GPUs used in the test.
- **Efficiency** (performance/watt): the performance/watt on P4 is ~1.5x than that on P40. This is also consistent with the theoretical efficiency difference. Because the theoretical performance of P4 is 1/2 of P40 and its TDP is around 1/3 of P40 (75W vs 250W), therefore its performance/watt is ~1.5x than P40.

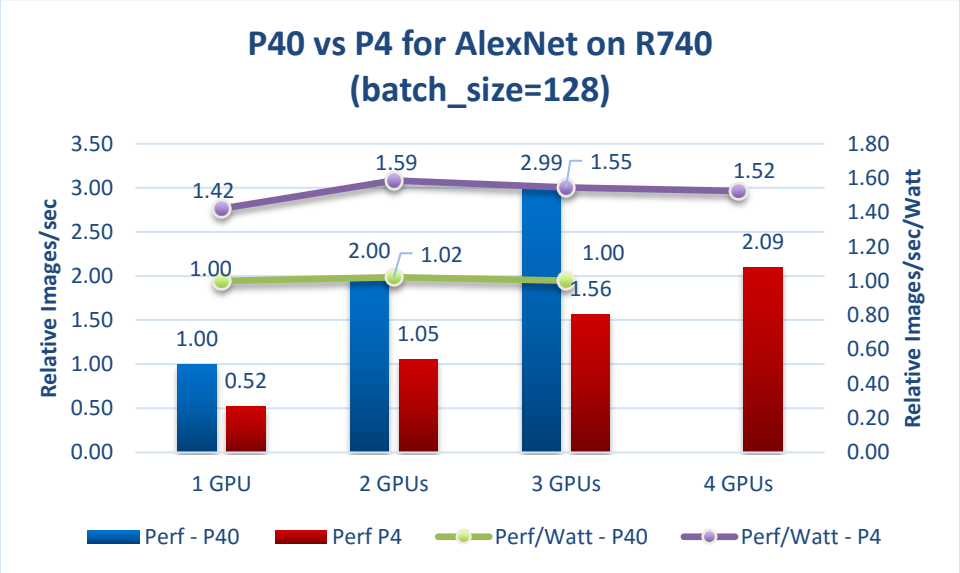


Figure 1: The inference performance with AlexNet on P40 and P4

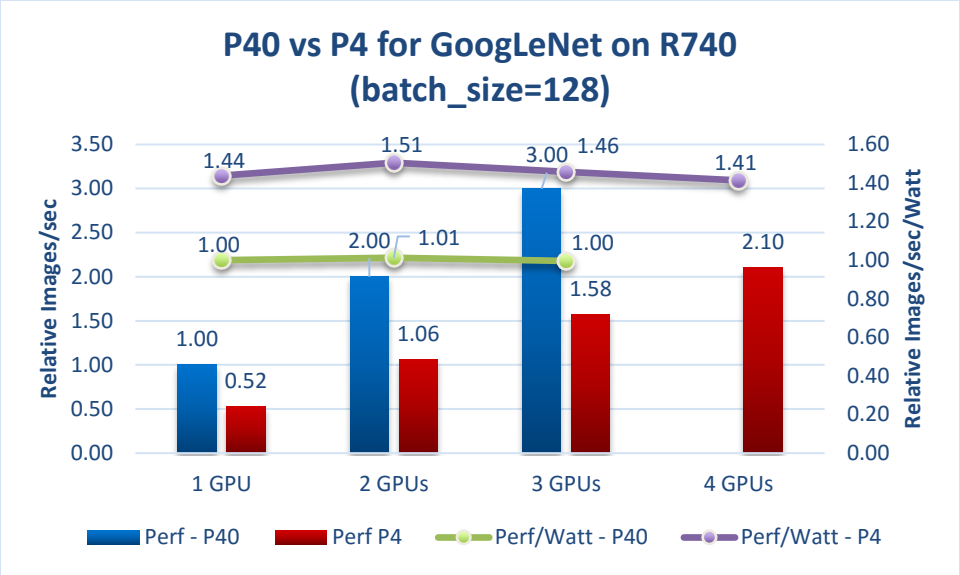


Figure 2: The performance of inference with GoogLeNet on P40 and P4

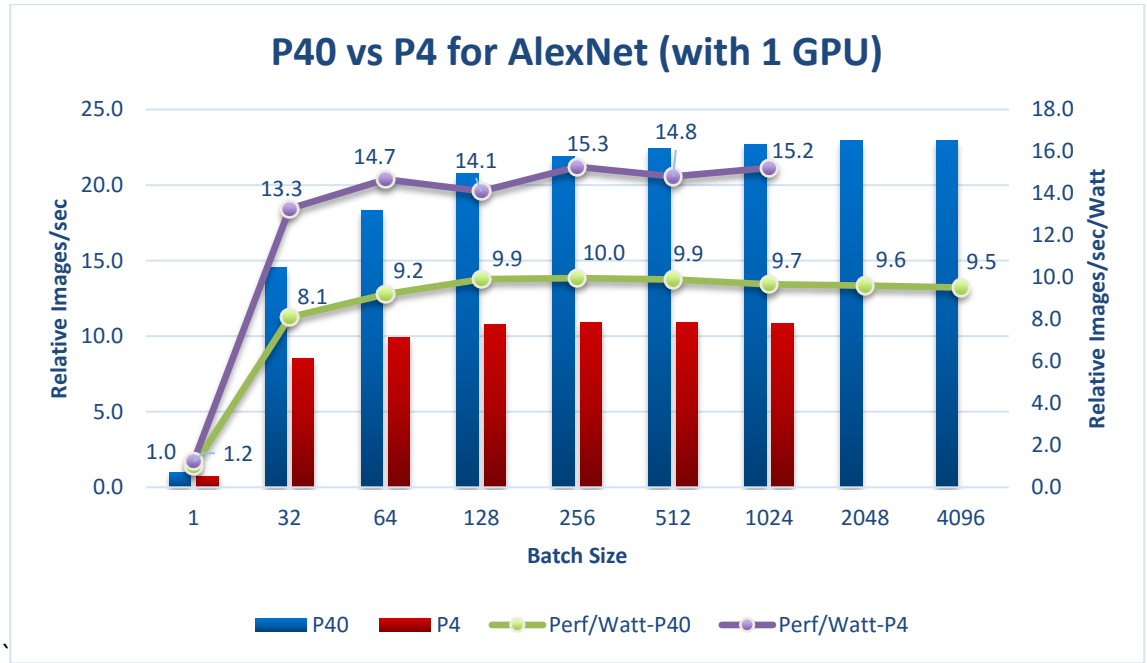


Figure 3: P40 vs P4 for AlexNet with different batch sizes

In our previous blog, we compared the inference performance using both FP32 and INT8 and the conclusion is that INT8 is ~3x faster than FP32. In this study, we also compare the accuracy when using both operations to verify that using INT8 can get comparable performance to FP32. We used the latest TensorRT 2.1 GA version to do this benchmarking. To make INT8 data encode the same information as FP32 data, a calibration method is applied in TensorRT to convert FP32 to INT8 in a way that minimizes the loss of information. More details of this calibration method can be found in the presentation "8-bit

Inference with TensorRT" from GTC 2017. We used ILSVRC2012 validation dataset for both calibration and benchmarking. The validation dataset has 50,000 images and was divided into batches where each batch has 25 images. The first 50 batches were used for calibration purpose and the rest of the images were used for accuracy measurement. Several pre-trained neural network models were used in our experiments, including ResNet-50, ResNet-101, ResNet-152, VGG-16, VGG-19, GoogLeNet and AlexNet. Both top-1 and top-5 accuracies were recorded using FP32 and INT8 and the accuracy difference between FP32 and INT8 was calculated. The result is shown in Table 3. From this table, we can see the accuracy difference between FP32 and INT8 is between 0.02% - 0.18% which means very minimum accuracy loss is achieved, while 3x speed up can be achieved.

Table 3: The accuracy comparison between FP32 and INT8

| Network | 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% |

## Conclusions

In this blog, we compared the inference performance on both P40 and P4 GPUs in the latest Dell EMC PowerEdge R740 server and concluded that P40 has ~2x higher inference performance compared to P4. But P4 is more power efficient and the performance/watt is ~1.5x than P40. Also with NVIDIA TensorRT library, INT8 can achieve comparable accuracy compared to FP32 while outperforming it with 3x in terms of performance.