# Comparing CPUs and GPUs for Deep Learning and Artificial Intelligence



## Read About

Intel® CPUs' and NVIDIA® GPUs' capabilities for deep learning

Inference performance comparisons using ResNet-50, BERT, and NCF benchmarks

Whether a CPU or GPU is better for aerospace and defense AI applications

## Introduction

At its Data-Centric Innovation Day in April 2019, Intel unveiled its new 2nd Generation Xeon® Scalable Processors (formerly known as Cascade Lake). The parts are divided across the Platinum, Gold, Silver, and Bronze lines. At the top of the line is the Platinum 9200, also known as Advanced Performance (AP). The 9282 has 56 cores per processor in a multi-chip module (two dies in one package, resulting in double the core count and double the memory). Measuring 76.0 x 72.5 mm, it's Intel's largest package to date. Focusing on density, high-performance computing, and advanced analytics, this packaged server can only be purchased from OEMS who buy from Intel and make modifications.

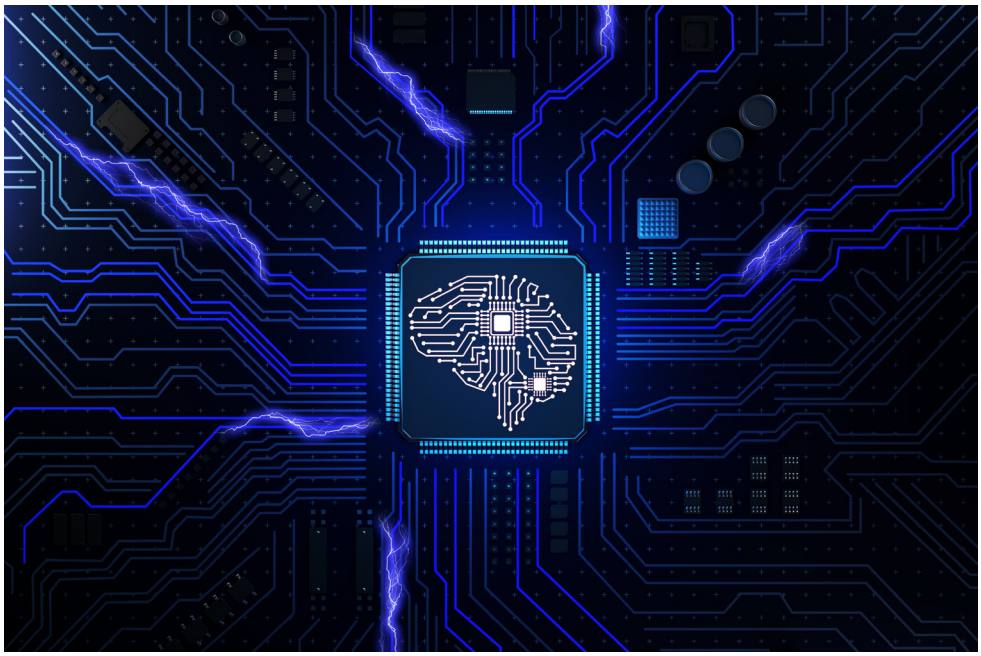

**Figure 1: Can a CPU do a GPU's job when it comes to deep learning and artificial intelligence?**

One of the new features of the 2nd Generation Xeon processors is Intel Deep Learning Boost (Intel DL Boost), also known as the Vector Neural Network Instruction (VNNI). VNNI combines three instructions into a single instruction, resulting in better use of computational resources and cache, while reducing the likelihood of bandwidth bottlenecks. Secondly, VNNI enables INT8 deep learning inference, which boosts performance with "little loss of accuracy." The 8-bit inference yields a theoretical peak compute gain of 4X over the 32-bit floating-point (FP32) operations.

Fast-forward to May 2019, when Intel announced that its new high-end CPU outperforms NVIDIA GPUs on ResNet-50, a popular convolutional neural network for computer vision. According to Intel, the company had "achieved leadership performance of 7878 images per second on ResNet-50 with our latest generation of Intel Xeon Scalable processors, outperforming 7844 images per second on NVIDIA Tesla V100, the best GPU performance as published by NVIDIA on its website including T4."

Employing the Xeon Platinum 9292, Intel achieved 7878 image/sec by creating 28 virtual instances of four CPU cores each (using a batch size of 11). An open-source deep learning framework, Intel Optimized Caffe, was used to optimize the ResNet-50 code. Intel recently added four general optimizations for the new INT8 inference:

1. Activation memory optimization
2. Weight sharing
3. Convolution algorithm tuning
4. First convolution transformation

NVIDIA wasted no time in replying to Intel's performance claims, releasing the statement, "It's not every day that one of the world's leading tech companies highlights the benefits of your products. Intel did just that last week, comparing the inference performance of two of their most expensive CPUs to NVIDIA GPUs." NVIDIA's detailed reply was a dual-prong response centering on power efficiency and performance per processor. NVIDIA's response can be summarized in Table 1.

## Examining the NVIDIA GPUs

To better understand these claims and counterclaims, it's useful to first step back and take a quick review of NVIDIA's V100 and the T4. Supporting a memory bandwidth of 9000 GB/sec, the NVIDIA V100 has 640 Tensor cores plus 5120 CUDA cores with 16 GB HBM2. The NVIDIA V100's specified performance is 7.8 TFLOPs of double precision performance and 125 TFLOPS of Tensor performance. The next generation Tesla T4 boasts 320 Turing Tensor cores and 2560 CUDA cores with 16 GB GDDR6 memory for peak 8.1 TLFOPs of double precision. First introduced as part of the Volta architecture, Tensor cores are extremely efficient at mixed precision computation.

In its response, NVIDIA points out that, since there are actually two processors in a single Intel package, the comparison measurement should be on a per processor basis. This approach effectively halves Intel's inference score. Intel's own benchmark description – 28 virtual instances multiplied by four cores equals 112 cores – confirms that Intel was using both processors on the multi-chip die. Second, Intel's performance numbers are achievable only when ResNet-50 was written with Caffe, and then optimized by Intel's Optimization for Caffe. NVIDIA's TensorRT™ platform, which is able to take input from most all of the popular frameworks, including PyTorch, MxNet, TensorFlow, and Caffe, was used to create the deep learning inference applications used in its benchmarks. Third, Intel used 8-bit integer precision, whereas

**Table 1** ResNet-50 Inference Performance Comparison between Intel CPU and NVIDIA GPU

| | TWO-SOCKET INTEL XEON 9282 | NVIDIA V100 (VOLTA) | NVIDIA T4 (TURING™) |
|---|---|---|---|
| ResNet-50 Inference (images/sec) | 7,878 | 7,844 | 4,944 |
| # of Processors | 2 | 1 | 1 |
| Total Processor TDP* | 800 W | 350 W | 70 W |
| Energy Efficiency (using TDP) | 10 img/sec/W | 22 img/sec/W | 71 img/sec/W |
| Performance per Processor (images/sec) | 3,939 | 7,844 | 4,944 |
| GPU Performance Advantage | 1.0 (baseline) | 2.0x | 1.3x |
| GPU Energy-Efficiency Advantage | 1.0 (baseline) | 2.3x | 7.2x |

* Thermal Design Power

Source: Intel Xeon performance; NVIDIA GPU performance

NIVIDA used a mixture of 16-bit and 32-bit floating point. Next, while the Platinum 9282 has just been announced and is unavailable to most customers, the V100 was introduced back in 2017, and the T4 in 2018.

Finally, there is the issue of price. Current estimates for just one of the Xeon's 9282 processors is $30,000 to $50,000. While on Amazon, you can score a V100 for $6,479 or a T4 card for $2,799. Even if you still view Intel's dual socket as a single processing chip, the result of the comparison seems to be that Intel's CPU barely wins on performance, but it requires a lot more cooling, power and money.

## Evaluating the Benchmarks

In considering how to best compare these two platforms, another question comes to mind: did Intel pick the best benchmark for comparison? One measure of complexity for artificial intelligence (AI) models is the number of parameters. Released in 2015 by Microsoft Research Asia, ResNet-50 first successfully took the stage at the ImageNet competition that same year. Bidirectional Encode Representation from Transformers (BERT) is an AI model released by Google in 2018 for natural language processing tasks, such as question answering. Compared to ResNet-50, which has 25 million parameters, BERT uses 340 million parameters – an increase of 13x. Due to its additional complexity, NVIDIA claims that using the BERT benchmark would provide a better test.

**Table 2** BERT Inference Performance Comparison between Intel CPU and NVIDIA GPU

| | DUAL INTEL XEON GOLD 6240 | NVIDIA T4 (TURING) |
|---|---|---|
| BERT Inference* Question-Answering (sentences/sec) | 2 | 118 |
| Total Processor TDP | 300 W[150x2] | 70 W |
| Energy Efficiency (using TDP) | .007 sentences/ sec/W | 1.7 sentences/ sec/W |
| GPU Performance Advantage | 1.0 (baseline) | 59x |
| GPU Energy-Efficiency Advantage | 1.0 (baseline) | 240x |

*Sequence length of 128

The comparison of the BERT testing on a CPU to a GPU is summarized in Table 2. For this benchmark, notice that NVIDIA used a dual-socket Xeon Gold 6240 with 384 GB of RAM running at 2.6 GHZ. It was run with FP32 precision, inside an Intel TensorFlow (TF) Docker container. A batch size of four yielded the best CPU score, so the GPU used the same batch size. The GPU also used TensorFlow with automatic mixed precision enabled.

Some would argue that using the Xeon Gold 6420 with 18 cores instead of the Xeon Platinum 9282 is not exactly comparing apples to apples when considering the previous ResNet-50 test. The price and limited availability of Xeon 9282 may have helped determine why NVIDIA did not use a Xeon Platinum 9282 for the test. Although this benchmark uses dual Xeon processors, NVIDIA did not apply the "performance per processor" metric and did not compare the Xeon Gold to the powerhouse Volta 100. Perhaps these choices were also made to compensate for lack of access to a Xeon 9282.

Continuing to press its point that GPUs are better, NVIDIA added a comparison of a recommender system known as Neural Collaborative Filtering (NCF) from the MLPerf training benchmark. Used in applications such as suggesting posts you might like on Facebook or Twitter and recommending content for you on YouTube or Netflix, NCF leverages users' prior interactions to the offered items to formulate its recommendations.

**Table 3** NCF Inference Performance Comparison between Intel CPU and NVIDIA GPU

| | SINGLE INTEL XEON GOLD 6240 | NVIDIA T4 (TURING) |
|---|---|---|
| Recommender Inference Throughput* [MovieLens] (thousands of samples/ sec) | 2,860 | 27,800 |
| Total Processor TDP | 150 W | 70 W |
| Energy Efficiency (using TDP) | 19 samples/sec/W | 397 samples/ sec/W |
| GPU Performance Advantage | 1.0 (baseline) | 10x |
| GPU Energy-Efficiency Advantage | 1.0 (baseline) | 20x |

*Batch size: 2048 for CPU and 1,048,576 for the GPU

The results captured in Table 3 were yielded from the systems used for the BERT benchmark. On the CPU, the Intel Benchmark for NCF on TensorFlow provided the benchmark results. The cynical might notice that this test only used one Xeon processor; NVIDIA notes that the tests used a single-socket CPU configuration because it yields a better score than using dual processor.

## Applying These Findings to Aerospace & Defense

While all of this competitive posturing is interesting, the real question is how the results apply to the embedded military and aerospace industries. Included among the 50 new Intel SKUs are three "Gold" chips that support long life cycles and better thermal performance:

+ 6238T (22 cores, 1.9 GHZ, 125 TDP)
+ 6230T (20 cores, 2.1 GHz, 125 TDP)
+ 5220T (18 cores, 2.2 GHZ, and 105 TDP)

Fortunately, the Gold 6238T is very similar to the Gold 6240 that NVIDIA used in its benchmarking.

**Figure 2: The VPX3-4935 3U OpenVPX module features NVIDIA Turing GPU technology and is designed for intense processing and AI in 3U HPEC systems.**

Curtiss-Wright, in partnership with Wolf Advanced Technology, offers OpenVPX™ Turing products based on the TU104 GPU. The TU104 GPU is the foundation of NVIDIA's Quadro® and Tesla products, including the Tesla T4 compared in the NVIDIA benchmarks and the Quadro RTX5000E featured in the VPX3-4935 3U and VPX6-4955 6U OpenVPX boards. The Tesla T4 has 2560 CUDA cores and 320 Tensor cores, while the RTX5000E boasts 3072 CUDA cores plus 384 Tensor cores. At 11.2 peak TFLOPs FP32, the RTX5000E improves on the performance of the T4 (8.1 TFLOPs).

**Figure 3: For 6U OpenVPX systems, the VPX6-4955 delivers processing power from not one, but two NVIDIA Turing GPUs to support deep learning and AI applications.**

## Conclusion

Yes, with the introduction of the VNNI in its new processors, Intel has achieved up to a 14x improvement in inference performance over its 1st Generation Xeon processors. Even NVIDIA conceded that "Intel's latest Cascade Lake CPUs include new instructions that improve inference, making them the best CPUs for inference." However, NVIDIA next added that CPUs are not in the same league with NVIDIA GPUs, which feature dedicated deep learning optimized Tensor cores. The benchmarks in this white paper support this observation. Since GPUs were designed to perform parallel processing, it's no surprise that they are more efficient than CPUs in AI training and inference applications. In the end, the Xeon CPUs are general-purpose processors that can do a decent job at inference if needed. NVIDIA GPUs, though, which are designed to be the ideal solution for inference workloads, are faster and more energy efficient for parallel workloads and AI applications.

Therefore, we can only conclude that the answer is no, you should not send a CPU to do a GPU's job.

## Author

**Tammy Carter, MSCS**
Senior Product Manager
Curtiss-Wright Defense Solutions

# Learn More

### Curtiss-Wright Products

- VPX3-4935 GPGPU Processor with NVIDIA Quadro Turing TU104/ RTX5000E
- VPX6-4955 GPGPU Processor with Dual NVIDIA Quadro Turing TU104/ RTX5000E GPUs

### Curtiss-Wright White Papers

- Machine Learning and Artificial Intelligence in Defense and Aerospace Applications - What You Need to Know
- Enabling AI at the Network Edge of the Battlefield