PowerEdge Servers and Deep Learning Domains: The Impact of Scaling Accelerators

Tech Note by

Matt Ogle
Ramesh Radhakrishnan

Summary

With deep learning principles becoming a widely accepted practice, customers are keen to understand how to select the most optimized server, based on GPU count, to accommodate varying machine and deep learning workloads.

This tech note delivers test results that portray how scaling NVIDIA GPU's on PowerEdge server configurations will impact performance for various deep learning domains, and how these results outline general guidelines for constructing an optimized deep learning platform.

Evaluating Performance Using MLPerf Benchmark

To accurately harvest Artificial Intelligence (AI) performance data it is critical to select a benchmark that is qualified to accurately test multiple domain types. MLPerf is a new and broad Machine Learning (ML) and Deep Learning (DL) benchmark suite that is gaining popularity and adoption for its multi-domain capabilities and representative models. The current version (v0.5) covers five domains associated with AI subsets, as seen in *Figure 1*: image classification, object detection, language translation, reinforcement learning and recommendation.

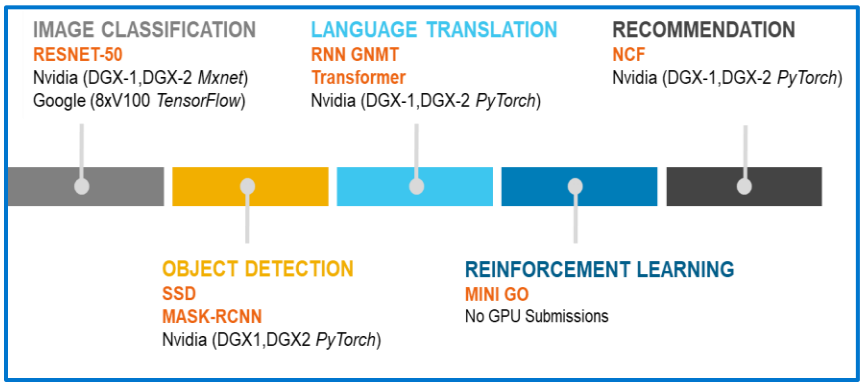

Figure 1: Domains covered within the MLPerf v0.5 benchmark

For each given domain, MLPerf will measure performance by assessing and comparing total training times; the amount of time that it takes to train a neural net model for a given domain to reach target accuracy. Dell EMC team benchmarked various PowerEdge servers that have GPU compatibility to help customers pick the appropriate GPU infrastructure that will achieve their requirements. We used multi-GPU training to highlight the shortest amount of training time needed to reach target accuracy the fastest for the MLPerf.

Server	# of CPU's	# of GPU's	GPU Type	GPU Interconnect
DSS 8440	2	8	V100 (16GB)	PCIe
PE T640	2	4	V100 (32GB)	PCIe
PE R740	2	3	V100 (16GB)	PCIe
Precision 5820	1	2	GV100 (32GB)	PCIe

Figure 2: PowerEdge CPU & GPU details for each tested configuration

Every benchmark ran on single node PowerEdge servers, as seen in Figure 2. Each server was loaded with either 2, 3, 4 or 8 Tesla V100 PCIe GPU's, and these configurations ran until the unique domain being tested reached the target accuracy. By comparing these configurations, we can deduce the performance increase per domain when additional GPU's are included.

MLPerf scores were calculated by exhibiting the total training times of each configuration relative to the reference accelerator, one NVIDIA Pascal P100. Each score indicates that the Tesla GV/V100 server is that many times faster than the Pascal P100. This methodology ensure consistency amongst each platform so that each scaled score remains accurate.

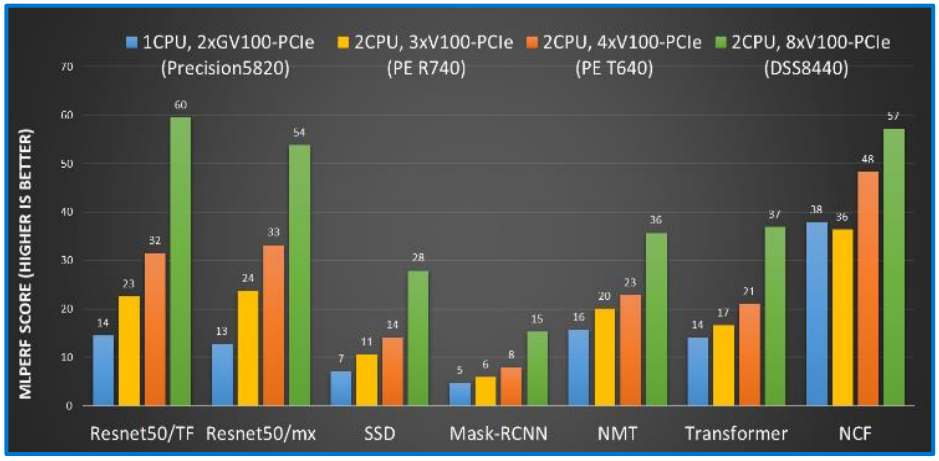

Figure 3: MLPerf benchmark scores calculated against the reference accelerator (one NVIDIA Pascal P100)

The first notable observation is the variance in training times for each domain. Recommendation, Reinforcement Learning and Language Translation DL consistently require the most training time for completion, while Object Detection and Image Classification appear to take half as long. This illustrates the varying learning difficulties associated with each DL domain. Furthermore, we learn from observing Figure 3 that Image Recognition (Resnet50) and Object Detection (Mask-RCNN) domains scale linearly; we can assume that when the GPU count increases than the speedup times decrease at a linear rate. Translation (NMT) and Recommendation (NCF) domains, on the other hand, were not as predictable. The bar graphs for Translation scores almost seems to scale quadratically and the Recommendation scores appear to not scale beyond 2 GPU's (it is an artifact of the dataset being too small which is being fixed in a later version of MLPerf).

Recommendations

1. **The training times and scaling behavior vary between different domains and models**
   - Using superior accelerators would be advantageous for the domains that require the most time
   - In order to pick the appropriate server and number of GPU's, it is useful to understand the models and domains being used.

2. **Increasing GPU's scales performance at a near linear rate for Image Recognition and Object Detection domains**
   - Servers with higher GPU counts will linearly reduce training time for these domains. Scaling to 4 GPUs using NVLink appears to be the sweet spot from an efficiency stand point.

3. **Increasing GPU's does not scale performance at a linear rate for Translation and Recommendation domains**
   - Servers with higher GPU counts will *not* linearly reduce training times for these domains due to data set or computation/communication ratios. However, using larger GPU counts is still useful to meet time to solution as the training time is reduced across these models.

Conclusion

Optimizing a platform for ML/DL workloads goes far beyond scaling the accelerators; every variable must be considered and there are a plethora of them. Fortunately, Dell EMC is committed to designing PowerEdge servers with GPU counts that cater to specific ML/DL domains, thereby reducing these variables for a smooth and simple customer experience. This tech note provided insight on how the accelerator model, accelerator count, and domain type are influenced by unique PowerEdge server models, and more importantly how customers can make the best decisions to perform their required ML/DL workloads at full throttle.

PowerEdge DfD Repository
For more technical learning

Contact Us
For feedback and requests

Follow Us
For PowerEdge news