MLPerf™ Inference v1.0 – CPU Based Benchmarks on Dell EMC PowerEdge R750 Server

Tech Note by

Vilmara Sanchez
Bhavesh Patel
Todd Mottershead

Summary

MLCommons™ Association has released the third round of results v1.0 for its machine learning inference performance benchmark suite MLPerf™. Dell EMC has participated in this effort by collaborating with several partners and using multiple configurations, spanning from Intel® CPU to accelerators such as GPU's and FPGA's. This blog is focused on the results for computer vision inference benchmarks (image classification and object detection), in the closed division / datacenter category, running on Dell EMC PowerEdge R750 in collaboration with Intel® and using its Optimized Inference System based on OpenVINO™ 2021.1.

- ✓ ***Servers**: PowerEdge R750*
- ✓ ***Processors**: 3rd Generation Intel® Xeon® Scalable Processors*
- ✓ ***Framework**: OpenVINO™ 2021.1*

Introduction

In this blog we present the MLPerf™ Inference v1.0 CPU based results submitted on PowerEdge R750 with Intel® processor using the Intel® optimized inference system based on OpenVINO™ 2021.1. Table 1 shows the technical specifications of this system.

Dell EMC PowerEdge R750 Server

System Name	PowerEdge R750
Status	Coming soon
System Type	Data Center
Number of Nodes	1
Host Processor Model Name	Intel(R) Xeon(R) Gold 6330 CPU @ 2.0GHz
Host Processors per Node	2
Host Processor Core Count	28
Host Processor Frequency	2.00 GHz
Host Memory Capacity	1TB 1 DPC 3200 MHz
Host Storage Capacity	1.5TB
Host Storage Type	NVMe

Table 1: Server Configuration Details

3rd Generation Intel® Xeon® Scalable Processor

The 3rd Generation Intel® Xeon® Scalable processor family is designed for data center modernization to drive operational efficiency and higher productivity, leveraged with built-in AI acceleration tools, to provide the seamless performance foundation for data center and edge systems. Table 2 shows the technical specifications for CPU's Intel® Xeon®.

Product Collection	3rd Generation Intel® Xeon® Scalable Processors
Code Name	Ice Lake
Processor Name	Gold 6330
Status	Launched
# of CPU Cores	28
# of Threads	56
Processor Base Frequency	2.0GHz
Max Turbo Speed	3.10GHz
Cache L3	42 MB
Memory Type	DDR4-2933
ECC Memory Supported	Yes

Table 2: Intel® Xeon® Processors technical specifications

MLPerf™ Inference v1.0 - Datacenter

The MLPerf™ inference benchmark measures how fast a system can perform ML inference using a trained model with new data in a variety of deployment scenarios. There are two benchmark suites, one for Datacenter systems and one for Edge. *Table 3* lists six mature models included in the official release v1.0 for Datacenter systems category and the vision models both image classification and object detection. The benchmark models highlighted below were run on PowerEdge R750.

Datacenter Benchmark Suite

Area	Task	Model	Dataset	QSL Size	Quality	Server latency constraint
Vision	Image classification	Resnet50-v1.5	ImageNet (224x224)	1024	99% of FP32 (76.46%)	15 ms
Vision	Object detection (large)	SSD-ResNet34	COCO (1200x1200)	64	99% of FP32 (0.20 mAP)	100 ms
Vision	Medical image segmentation	3D UNET	BraTS 2019 (224x224x160)	16	99% of FP32 and 99.9% of FP32 (0.85300 mean DICE score)	N/A
Speech	Speech-to-text	RNNT	Librispeech dev-clean (samples < 15 seconds)	2513	99% of FP32 (1 - WER, where WER=7.452253714852645%)	1000 ms
Language	Language processing	BERT	SQuAD v1.1 (max_seq_len=384)	10833	99% of FP32 and 99.9% of FP32 (f1_score=90.874%)	130 ms
Commerce	Recommendation	DLRM	1TB Click Logs	204800	99% of FP32 and 99.9% of FP32 (AUC=80.25%)	30 ms

Table 3: Datacenter Suite Benchmarks. Source: MLCommons™

Scenarios

The above models are deployed in a variety of critical inference applications or use cases known as “scenarios”, where each scenario requires different metrics, demonstrating production environment performance in the real practice. Below is the description of each scenario and the *Table 4* shows the scenarios required for each Datacenter benchmark included in this submission v1.0.

Offline scenario: represents applications that process the input in batches of data available immediately, and don’t have latency constraint for the metric performance measured as samples per second.

Server scenario: this scenario represents deployment of online applications with random input queries, the metric performance is queries per second (QPS) subject to latency bound. The server scenario is more complicated in terms of latency constraints and input queries generation, this complexity is reflected in the throughput-degradation results compared to offline scenario.

Scenario	Query Generation	Duration	Samples/query	Latency Constraint	Tail Latency	Performance Metric
Server	LoadGen sends new queries to the SUT according to a Poisson distribution	270,336 queries and 600 seconds	1	➢ 15ms image classification ➢ 100 ms object detection	99%	Maximum Poisson throughput parameter supported
Offline	LoadGen sends all samples to the SUT at start in a single query	1 query and 600 seconds	At least 24,576	None	N/A	Measured throughput

Table 4: MLPerf™ Inference Scenarios. Source: MLCommons™

Software Stack and System Configuration

The software stack and system configuration used for this submission is summarized in Table 5. Some of the settings that really mattered when looking at benchmark performance are captured in the table below.

OS	Ubuntu 20.10 (GNU/Linux 5.8.0-45-generic x86_64)
Intel® Optimized Inference SW for MLPerf™	MLPerf™ Intel OpenVino OMP CPP v1.0 Inference Build
ECC memory mode	ON
Host memory configuration	1TiB | 64G per memory channel (1DPC) with 2933mt/s
Turbo mode	ON
CPU frequency governor	Performance

Table 5: System Configuration

OpenVINO™ Toolkit

The OpenVINO™ 2021.1 toolkit is used to optimize and run Deep Learning Neural Network models on Intel® hardware. The toolkit consists of three primary components: inference engine, model optimizer, and intermediate representation. The Model Optimizer is used to convert the MLPerf™ reference implementation benchmarks from a framework into quantized INT8 models to run on Intel® architecture.

Benchmark Parameter Configurations

The benchmarks and scenarios submitted for this round are ResNet50-v1.5 and SSD-ResNet34 in offline and server scenarios. Both benchmarks required tunning certain parameters to achieve maximum performance. The parameter configurations and expected performance depend on the processor characteristics including number on CPUs used (number of sockets), number of cores, number of threads, batch size, number of requests, CPU frequency, memory configuration and the software accelerator. Table 6 shows the parameter setting used to run the benchmarks to obtain optimal performance and produce VALID results to pass Compliance tests.

Model	Scenario	OpenVINO params & batch size
ResNet50 INT8	Offline	nireq = 224, nstreams = 112, nthreads = 56, batch = 4
	Server	nireq = 28, nstreams = 14, nthreads = 56, batch = 1
SSD-ResNet34 INT8	Offline	nireq = 28, nstreams = 28, nthreads = 56, batch = 1
	Server	nireq = 4, nstreams = 2, nthreads = 56, batch = 1

Table 6: Benchmark parameter configuration

Results

From the scenario perspective, we benchmark the CPU performance by comparing server versus offline scenario and determine what is the delta. We also looked at results from our prior submission v0.7 to v1.0, so we can determine how the performance improved for Intel Xeon 3rd Generation compared to Intel Xeon 2nd .

ResNet50-v1.5 in server and offline scenarios

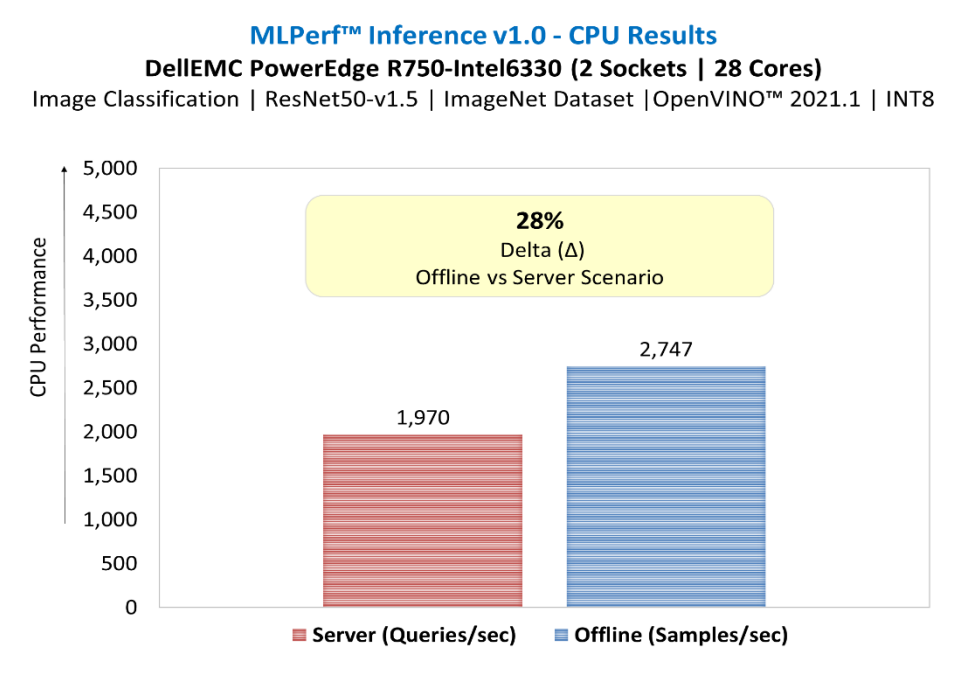

Figure 1: ResNet50-v1.5 in server and offline scenarios

SSD-ResNet34 in server and offline scenarios

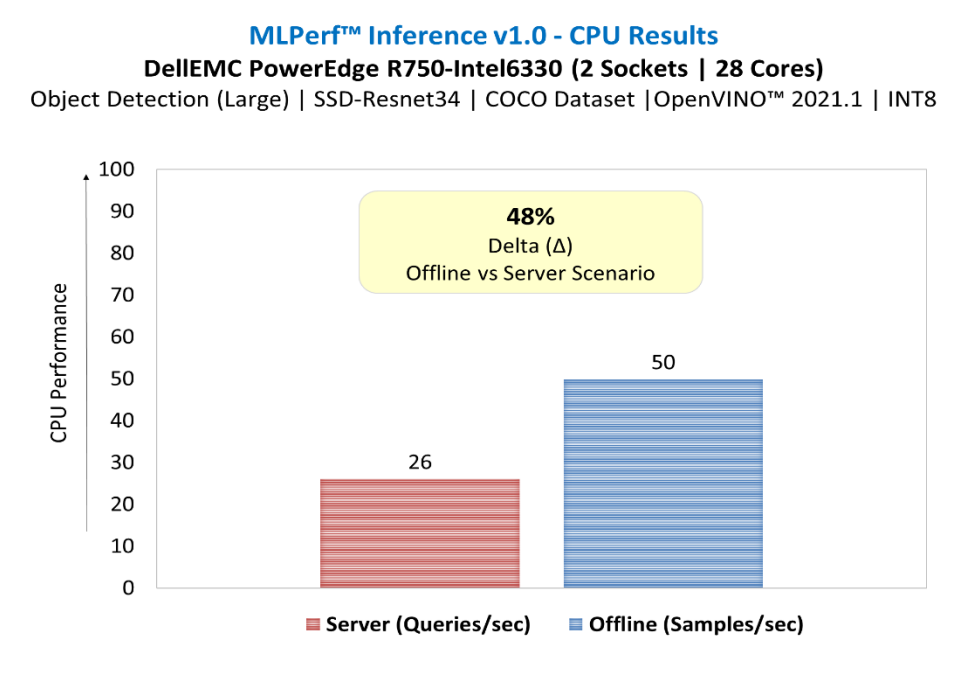

Figure 2: SSD-ResNet34 in server and offline scenario

Figure ***3*** illustrates the normalized server-to-offline performance for each model, scores close to 1 indicate that the model is delivering similar throughput in server scenario (constrained latency) as it is in offline scenario (unconstrained latency), scores close to zero indicate severe throughput degradation.

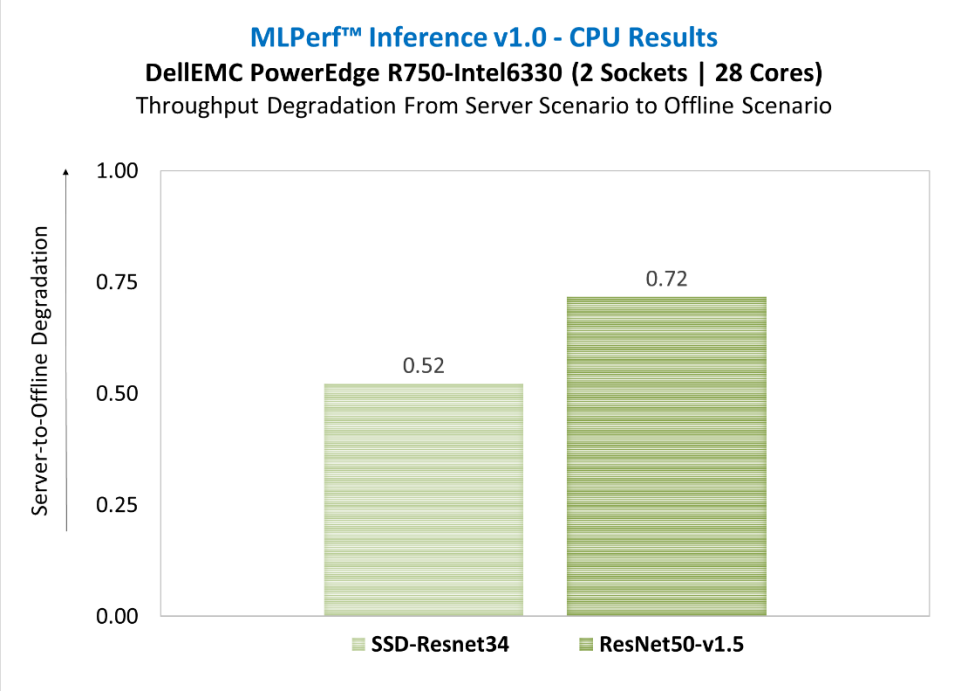

Figure 3: Throughput degradation from server scenario to offline scenario

Results submission v0.7 versus v1.0

In this section we compare the results from submission v0.7 versus this submission v1.0 to determine how the performance improved from servers with 2nd gen Xeon scalable processors vs. 3rd gen. The table below shows the server specifications used on each submission:

	Dell EMC Server for Submission v0.7	Dell EMC Server for Submission v1.0
System Name	PowerEdge R740xd	PowerEdge R750
Host Processor Model Name	Intel(R) Xeon(R) Platinum 8280M	Intel(R) Xeon(R) Gold 6330
Host Processor Generation	2nd	3rd
Host Processors per Node	2	2
Host Processor Core Count	28	28
Host Processor Frequency	2.70 GHz	2.00 GHz
Host Processor TDP	205W	205W
Host Memory Capacity	376GB - 2 DPC 3200 MHz	1TB - 1 DPC 3200 MHz
Host Storage Capacity	1.59TB	1.5TB
Host Storage Type	SATA	NVMe

Table 7: Server Configurations used for submission v0.7 and v1.0

ResNet50-v1.5 in Offline Scenario | Submission v0.7 vs. v1.0

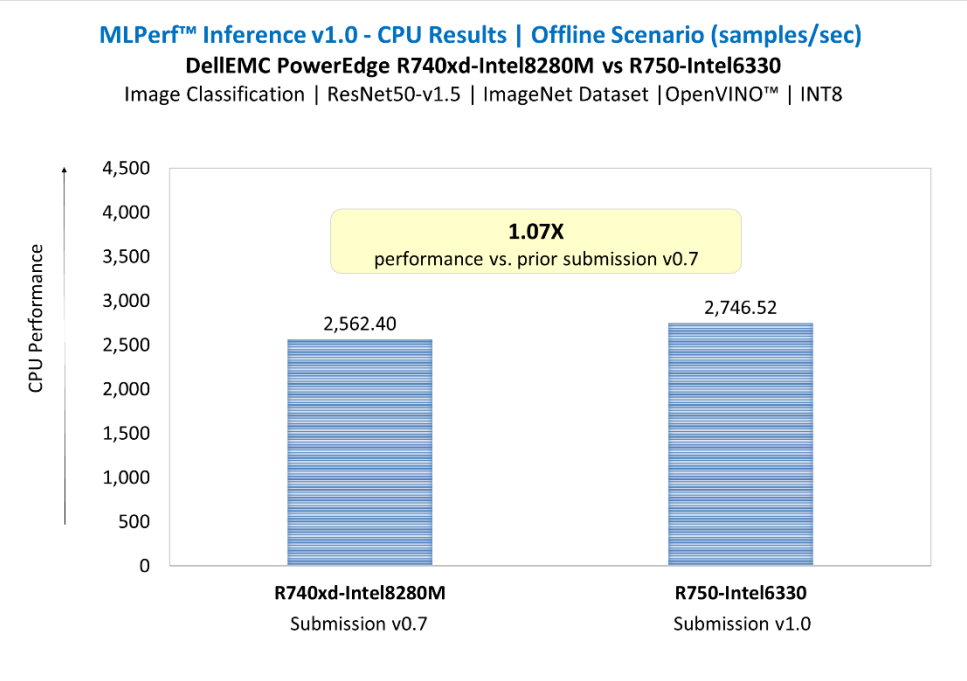

Figure 4: ResNet50-v1.5 in Offline Scenario | Submission v0.7 vs. v1.0

ResNet50-v1.5 in Server Scenario | Submission v0.7 vs. v1.0

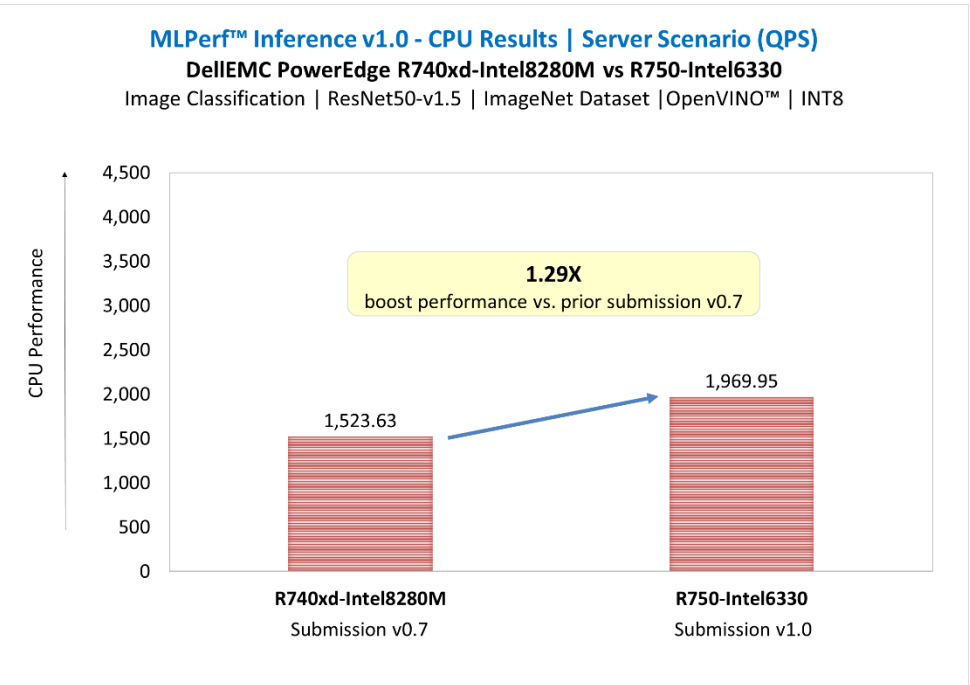

Figure 5: ResNet50-v1.5 in Server Scenario | Submission v0.7 vs. v1.0

SSD-ResNet34 in Offline Scenario | Submission v0.7 vs. v1.0

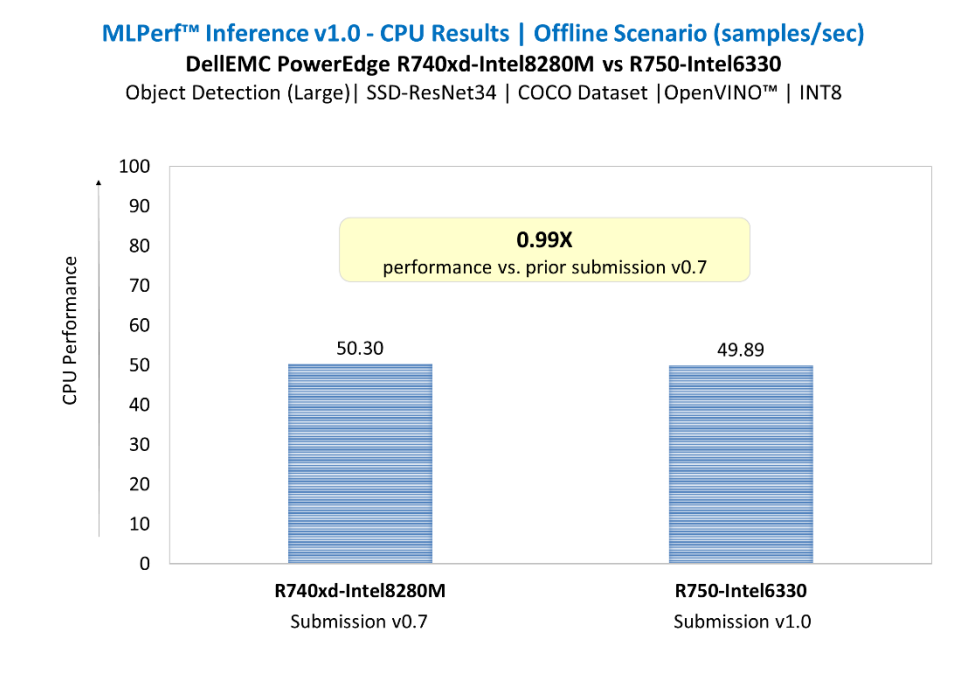

Figure 6: SSD-ResNet34 in Offline Scenario | Submission v0.7 vs. v1.0

SSD-ResNet34 in Server Scenario | Submission v0.7 vs. v1.0

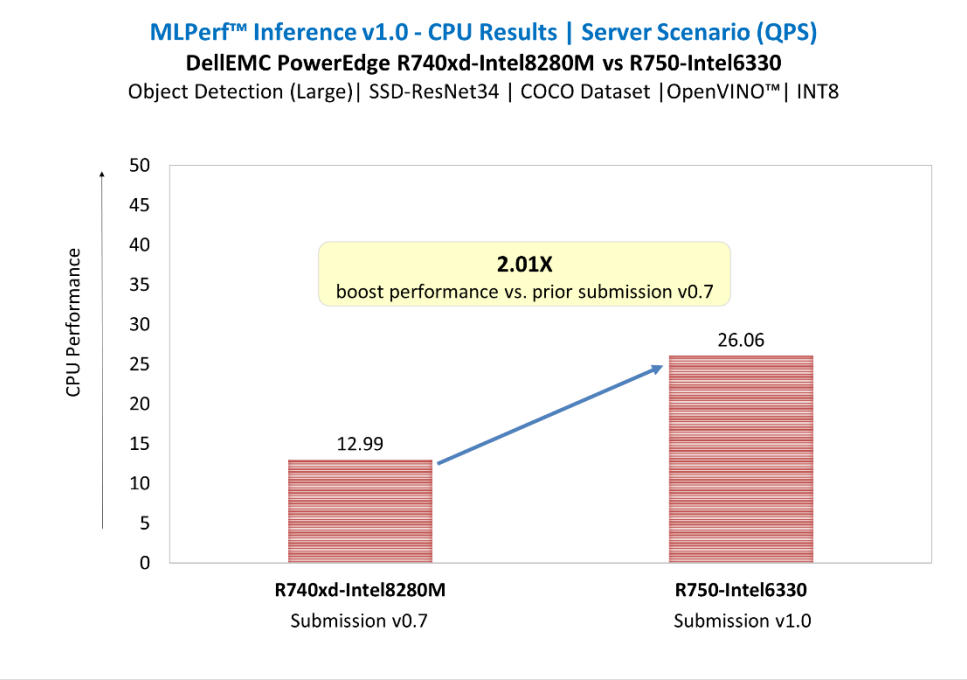

Figure 7: SSD-ResNet34 in Server Scenario | Submission v0.7 vs. v1.0

Conclusion

Both the Gold 6330 and the previous generation Platinum 8280 were chosen for this test because they have 28 cores and a memory interface that operates at 2933Mt/s. Customers with more demanding requirements could also consider higher performing variants of the 3rd Gen Intel® Xeon® scalable processor family up to the 40 core Platinum 8380 which uses a memory interface capable of 3200MT/s.

- The two-socket (dual CPU) server Dell EMC PowerEdge R750 equipped with 3rd Gen Intel® Xeon® scalable processors delivered:
    - Up to **1.29X** boost performance for image classification and up to **2.01X** boost performance for object detection large in server scenario, compared to prior submission of PowerEdge R740xd equipped with 2nd Gen Intel® Xeon® processors.
    - For ResNet50-v1.5 benchmark, there was a loss degradation around 28% from server scenario (constrained latency) to offline scenario (unconstrained latency). For SSD-ResNet34 benchmark, the loss was around 48%. These results demonstrate the complexity of server scenario in terms of latency constraints and input queries generation. The throughput degradation from server scenario is an indication of how well the system handles the latency constraint requirements, and it could be related to several factors such as the hardware architecture, the batching management, the inference software stack used to run the benchmarks. It is recommended to conduct performance analysis of the system including both scenarios.

- PowerEdge R750 server drives enhanced performance to suite computer vision inferencing tasks, as well as other complex workloads such as database and advanced analytics, VDI, AI, DL, and ML in datacenters deployments; it is an ideal solution for data center modernization to drive operational efficiency, lead higher productivity, and maximize total cost of ownership (TCO).

Citation

```
@misc{reddi2019mlperf,
        title={MLPerf™ Inference Benchmark},
        author={Vijay Janapa Reddi and Christine Cheng and David Kanter and Peter Mattson and Guenther Schmuelling
        and Carole-Jean Wu and Brian Anderson and Maximilien Breughe and Mark Charlebois and William Chou and Ramesh
        Chukka and Cody Coleman and Sam Davis and Pan Deng and Greg Diamos and Jared Duke and Dave Fick and J. Scott
        Gardner and Itay Hubara and Sachin Idgunji and Thomas B. Jablin and Jeff Jiao and Tom St. John and Pankaj
        Kanwar and David Lee and Jeffery Liao and Anton Lokhmotov and Francisco Massa and Peng Meng and Paulius
        Micikevicius and Colin Osborne and Gennady Pekhimenko and Arun Tejusve Raghunath Rajan and Dilip Sequeira
        and Ashish Sirasao and Fei Sun and Hanlin Tang and Michael Thomson and Frank Wei and Ephrem Wu and Lingjie
        Xu and Koichi Yamada and Bing Yu and George Yuan and Aaron Zhong and Peizhao Zhang and Yuchen Zhou},
        year={2019},
        eprint={1911.02549},
        archivePrefix={arXiv},
        primaryClass={cs.LG}
}
```