MLperf™ Performance Report

Silverton Consulting, Inc. StorInt™ Dispatch

This Storage Intelligence (StorInt™) dispatch covers the MLperf™ series of AI-ML-DL model training and inferencing benchmarks[1]. This report focuses on inferencing activity for data center environments.

MLperf v1.0 edge inferencing benchmark results

The MLperf v1.0 data center inferencing series of benchmarks takes standard AI Deep Learning models and runs inferencing activity against them measuring **server inference queries/sec and offline samples/sec.** The server queries/sec is generated by sending new queries to the server using a poisson distribution, while the offline samples/sec is generated by sending all the queries at once to a server.

There are 6 DNN models[2] represented in MLPerf 1.0 data center inferencing benchmark, and most of them have been pre-trained to 99% accuracy. These include, image classification (ImageNet-ResNet50), object detection large (COCO, SSD-ResNet34), medical image segmentation (3D-Unet, BraTS 2019), speech-to-text (LibriSpeech, dev-clean), (natural) language processing (SQuAD v1.1, BERT), and recommendation engine (DLRM, 1TB click logs).

One change from MLPerf v0.7 (discussed in prior reports) to new v1.0, is that they have added a **numerics type used** column to the benchmark results. We suppose that over time, different submissions could potentially use different numeric formats optimized for acceleration hardware in use. For this report, all models/all submissions used **INT8** as numeric type.

We start our discussion with the recommendation engine model (DLRM, 1TB click logs) with server queries/sec results, in Figure 1.

[1] All MLperf inferencing and training results are available at https://mlcommons.org/en/ as of 04/28/2021

[2] Reference versions of all models used in MLPerf benchmarks are available on GitHub (https://github.com/mlcommons)

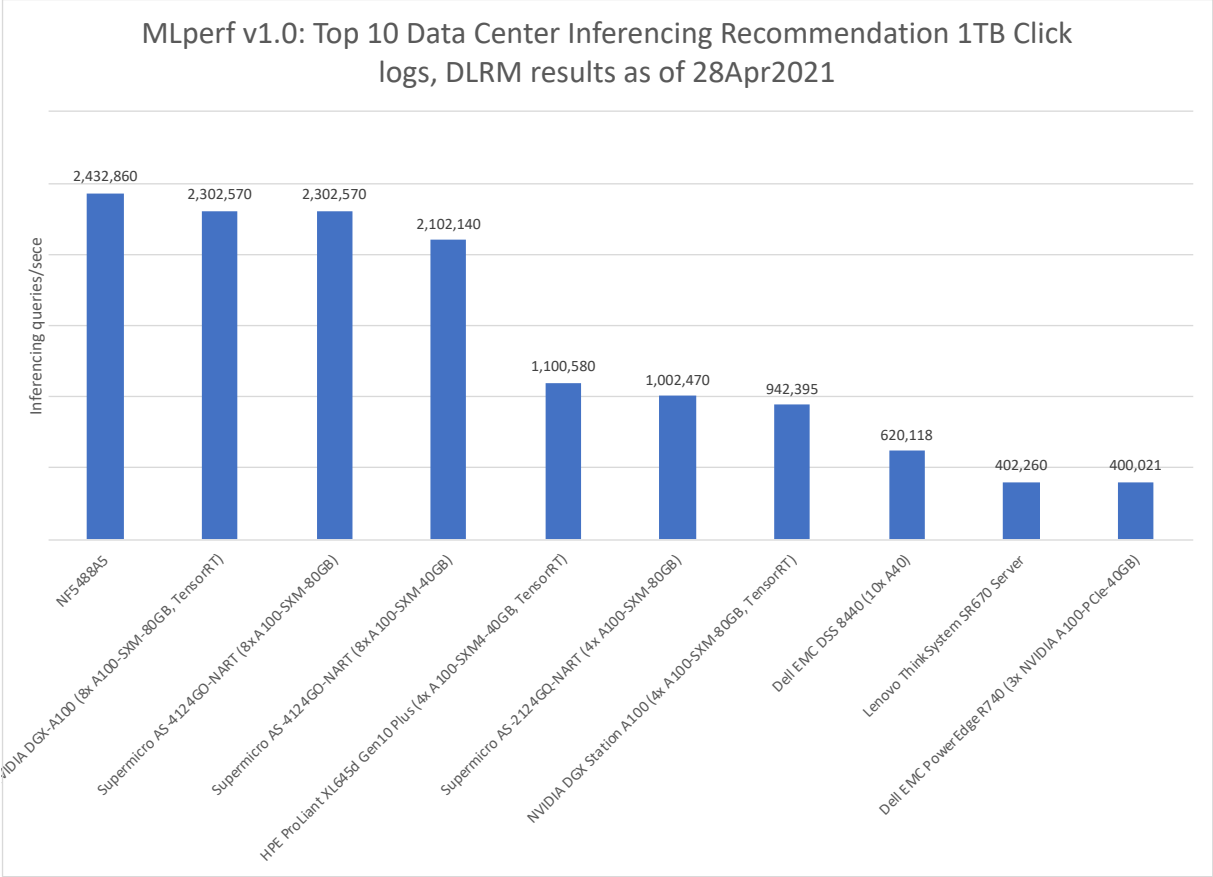

Figure 1 Top 10 MLperf v1.0 Data center inferencing recommendation engine results

In Figure 1, Inspur's NF5488A5 submission used 2 AMD EPYC 7742 (64-core) CPUs, 8-NVIDIA A100 SXM 80GB GPUs and TensorRT 7.2 with CUDA11.2. All of the others above used TensorRT 7.2.3 with CUDA 11.1 software.

The main differences between the first 4 (#1-4) submissions (>2M q/s) and the next performing tier (~1M q/s) three submissions (#5-7) were in the number of GPUs. The first 4 had 8 NVIDIA A100 SXM GPUs and the next three had 4 A100 SXM GPUs. The remaining 3 submissions (#8-10) used 10 NVIDIA A40 GPUs, 4-A100 PCIe GPUs, and 3 A100 PCIe GPUs, respectively.

The size of GPU memory seemed to result in some (10%) drop in performance. One can see this in the #3 and #4 SuperMicro submissions that used 80GB A100s and 40GB A100s respectively.

CPU counts also seemed to have some (10%) performance impact. This is evident between the #5 with 2 AMD EPYC 7742S CPUs and #7 with 1 AMD EPYC 7742 CPU.

Harder to untangle is core counts and processor type. The top 7 (#1-7) all used AMD EPYC and the bottom 3 (#8-10) used Intel Xeon CPUs. But the performance drop off could also be deo

ta core counts, as the AMD EPYC CPUs mostly had 64 cores while the Intel CPUs had 28 or fewer cores.

RANK	VENDOR	CPU TYPE	# CPUS	# CORES
1	Inspur	AMD EPYC 7742	2	64
2	NVIDIA	AMD EPYC 7742	2	64
3	Supermicro	AMD EPYC 7713	2	64
4	Supermicro	AMD EPYC 7742s	2	64
5	HPE	AMD EPYC 7742	1	64
6	Supermicro	AMD EPYC 7543	2	32
7	NVIDIA	AMD EPYC 7742	1	64
8	DellEMC	Intel Xeon Gold 6248 3GHz	2	20
9	Lenovo	Intel Xeon Plat. 8280 2.6GHz	2	28
10	DellEMC	Intel Xeon Gold 6248R 3GHz	2	24

In Figure 2, we present Top 10 MLPerf v1.0 data center inferencing results for image classification (ResNet50).

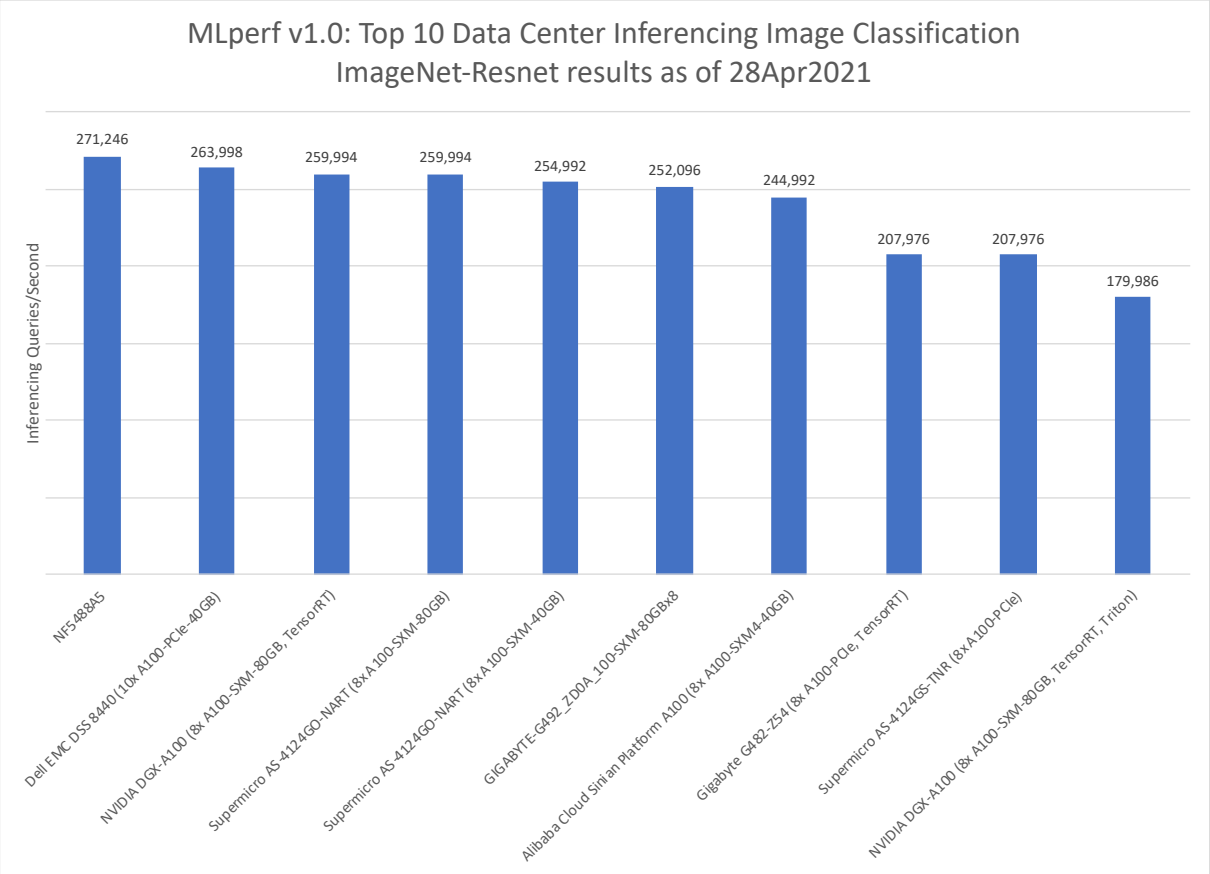

Figure 2 MLperf v1.0 data center ImageNet ResNet50 classification results

In Figure 2, once again the Inspur NFS5488A5 submission took top honors. Nine (#1, #3-10) of image classification top 10 submissions used 8 A100s except for the DellEMC DSS which used 10-A40s.

Core counts and processor type had no bearing on performance as the #2 had 20 cores, while eight of the rest had 64, with the Alibaba (#7) having 2 Intel Xeon Platinum 8269CY CPUs with 25 cores each.

What's perhaps most curious in this chart is that (seemingly) the same system performed so differently. The #3 NVIDIA DGX-A100 (8x A100-SXM-80GB, TensorRT) achieved 256.0K q/s while #10 NVIDIA DGX-A100 (8x A100-SXM-80GB, TensorRT) achieved only 180.0K q/s. As far can be told from the MLPerf website, there were no differences between the two submissions except for when they were submitted (Identification: 1.0-30 & 1.0-31). It's possible this is just due to query randomization but more likely it's a typo somewhere in the MLPerf information.

Next, we turn to MLperf 1.0 data center inferencing results for medical imaging segmentation in Figure 3.

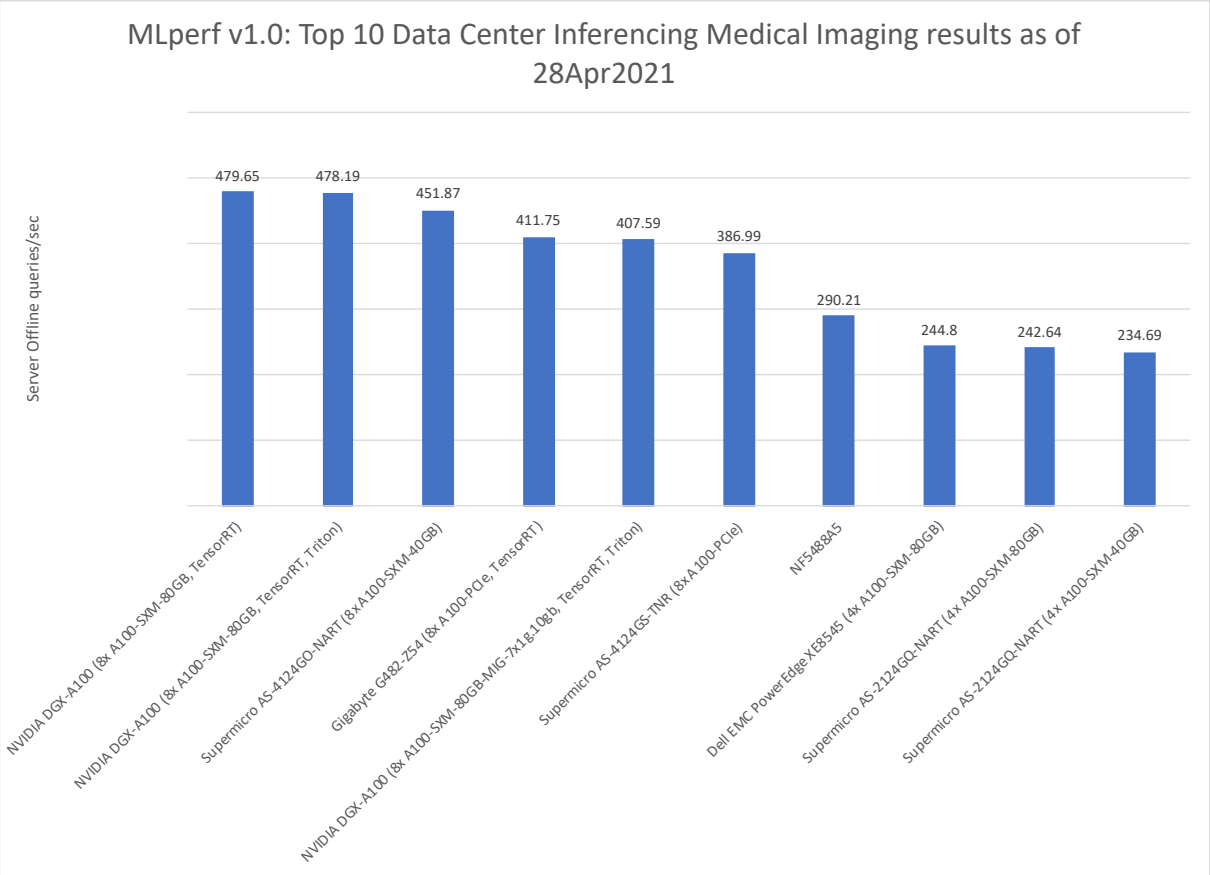

Figure 3 Top 10 MLperf v1.0 data center inferencing medical image segmentation results

In Figure 3, the two NVIDIA DGX systems (Identification: 1.0-30 & 1.0-31) came in at #1 and 2 which one would expect with only a slight difference in performance (479.7 vs 478.2 q/s). The bottom 3 (#8-10) submissions only used 4-A100s which might explain their lack of performance (all under 245.0 q/s). The medical imaging segmentation only reports on offline queries/sec.

It's interesting to note the Inspur NFS5488A5 submission only ranked #7 for medical imaging. This is likely due to some software difference (fix?) in TensorRT 7.2/ CUDA 11.2 vs TensorRT 7.2.3/CUDA 11.1 used by all other submissions.

The other substantial difference here is between the top 2 (#1-2) NVIDIA DGX-A100 Systems and the #5 ranked NVIDIA DGX-A100 system which was the result of the latter's use of MIG (multi-instance GPU support). Recall that MIG virtualizes a GPU so to appear as multiple (virtual) GPUs rather than one. In this case, we believe the #5 submission carved up each 80GB A100 into 7-10GB virtual A100s which cost it about 15% in offline q/s performance.

Significance

AI-ML-DL work is being widely adopted in enterprise data centers. Witness the latest VMware software now supports NVIDIA MIG.

We realize v1.0 reported numeric types may have more bearing on future submissions, but as all submissions reported here used GPUs and INT8 numerics, it didn't matter this time.

This is only our third performance report analyzing MLperf. We plan to discuss different training and inferencing workloads in future reports. If there is something we missed or have an error in any of our analysis, please let us know and we would gladly fix it.

This report was sent out to subscribers as part our **free, monthly Storage Intelligence e-newsletter**. If you are interested in receiving future storage performance analyses along with recent product announcement summaries, please use the QR code (below right) to sign up for your own copy.