MLperf™ Performance Report

Silverton Consulting, Inc. StorInt™ Dispatch

This Storage Intelligence (StorInt™) dispatch covers the MLperf™ v2.0 series of AI-ML-DL model training and inferencing benchmarks[1]. This report focuses on **inferencing** activity for the **Data Center environment.** There are 6 AI models represented in the MLperf v2.0 Data Center inferencing benchmark ranging from Image Processing to Recommendation Engine. We focus below on just 2 of these, MLperf's Object Detection (large) and the Speech2Text Data Center inferencing benchmarks.

MLperf v2.0 Data Center inferencing benchmark results.

We start our discussion with DC Object Detection (large) inferencing server queries/second (q/s) performance results in Figure 1.

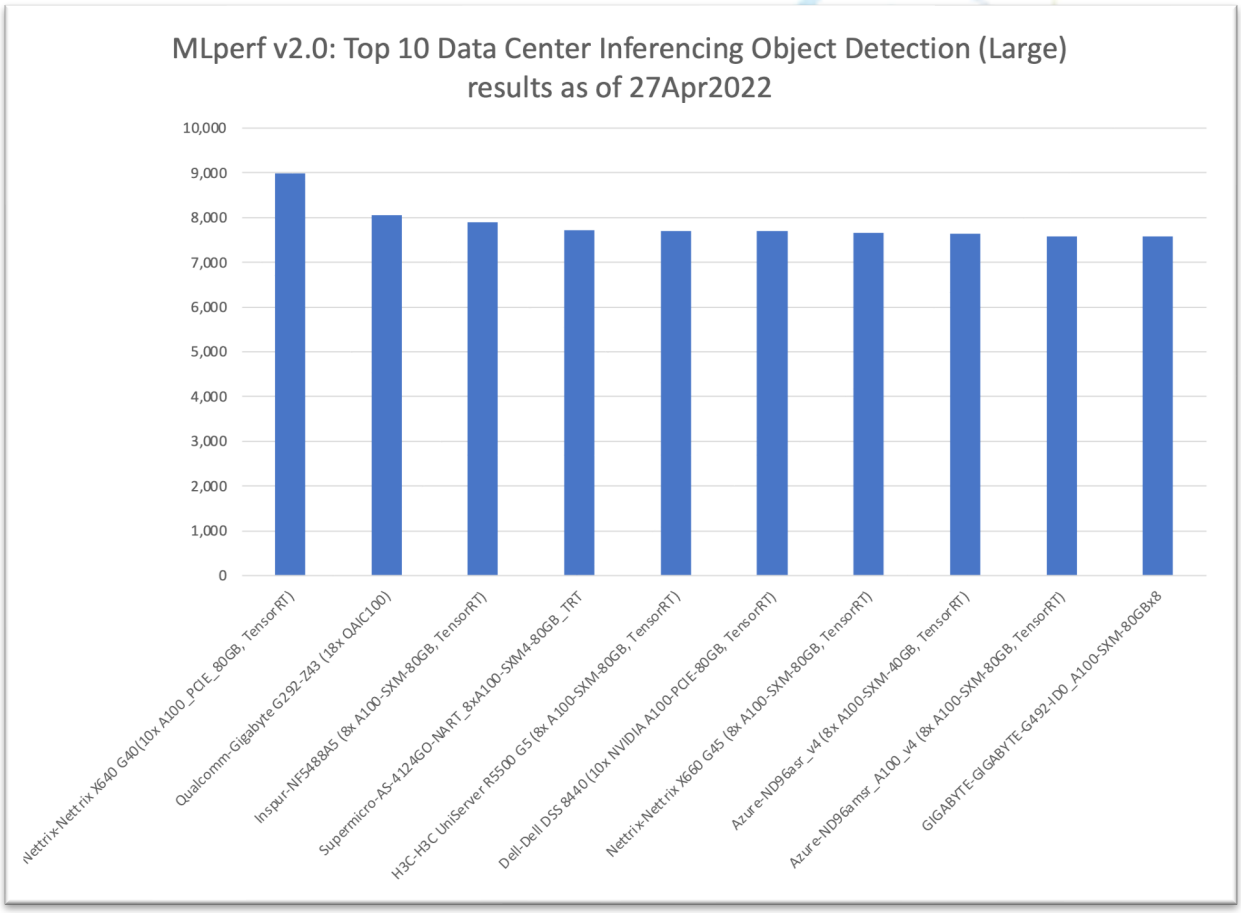

Figure 1 Top 10 MLperf v2.0 Data Center Object Detection (large) inferencing server q/s results

[1] All MLperf inferencing and training results are available at https://mlcommons.org/en/ as of 2022 Apr 27.

In Figure 1, we see the top solution, the Netrix X640 G40, achieved almost 9K object detection server q/s and used 10 A100 PCIe-80GB GPUs. The #2 solution above, the Qualcomm Gigabyte G292-Z43 with 18 (A100 PCIe-HHHL) QAIC100s, achieved a little over 8K server q/s and the #3 ranked system, the Inspur NFS488A5 with 8 A100 SXM-80GB GPUs achieved just a little under 8K server q/s.

In prior MLperf reports we discussed how the A100 PCIe GPUs are slower than the A100 SXM GPUs. But the Netrix (#1) had 10 A100 PCIe GPUs while the Inspur (#3) only had only 8 A100 SXM GPUs with the same GPU memory. That alone could easily explain the reduction in q/s performance.

But more interesting, is the #6 Dell DSS8440 System with 10 A100 PCIe-80GB GPUs vs. #1 Netrix. The Dell system was only able to achieve 7.7K, almost 17% less server q/s with the same GPU hardware.

The Dell system had 2 Intel 6248R CPUs while the Netrix X640 system had 2 Intel 8380 CPUs which could explain the performance reduction. No information is supplied as to system DRAM, bus speeds, or storage, so we can't tell whether any of those were factors. Both systems used PCIe-80GB A100s and both had 10 GPUs, so we must conclude that object detection server q/s performance has some dependence on CPU performance.

The other interesting comparison is the #2 Qualcomm Gigabyte systems which had 18 A100 PCIe HHHL (no memory size info) GPUs but still came in 2nd place with ~13% slower server q/s performance than the #1 system. There were two confounding factors with the Qualcomm system. One is that they used 2 AMD EPYC 7713 CPUs and the other is they used Qualcomm Cloud AI SDK v1.6.80 software, different than #1 (TensorRT & CUDA 11.6 software) and most other submissions. The only other top 10 system, not using TensorRT, was #4 SuperMicro solution which used CUDA 11.6 (alone without TensorRT).

We have already concluded that CPU power is one factor on object detection server q/s performance. But with 18 A100 PCIe GPUs, we would think it should have been able to do (taking just the Dell level of A100 GPU performance) almost 14K server q/s. The fact that it only attained ~8K server q/s doesn't speak well to their software. Note, there was no information on the memory configuration of the Qualcomm A100 GPUs, so that could also be a factor.

In Figure 2, we show Data Center Speech2Text inferencing server q/s performance results.

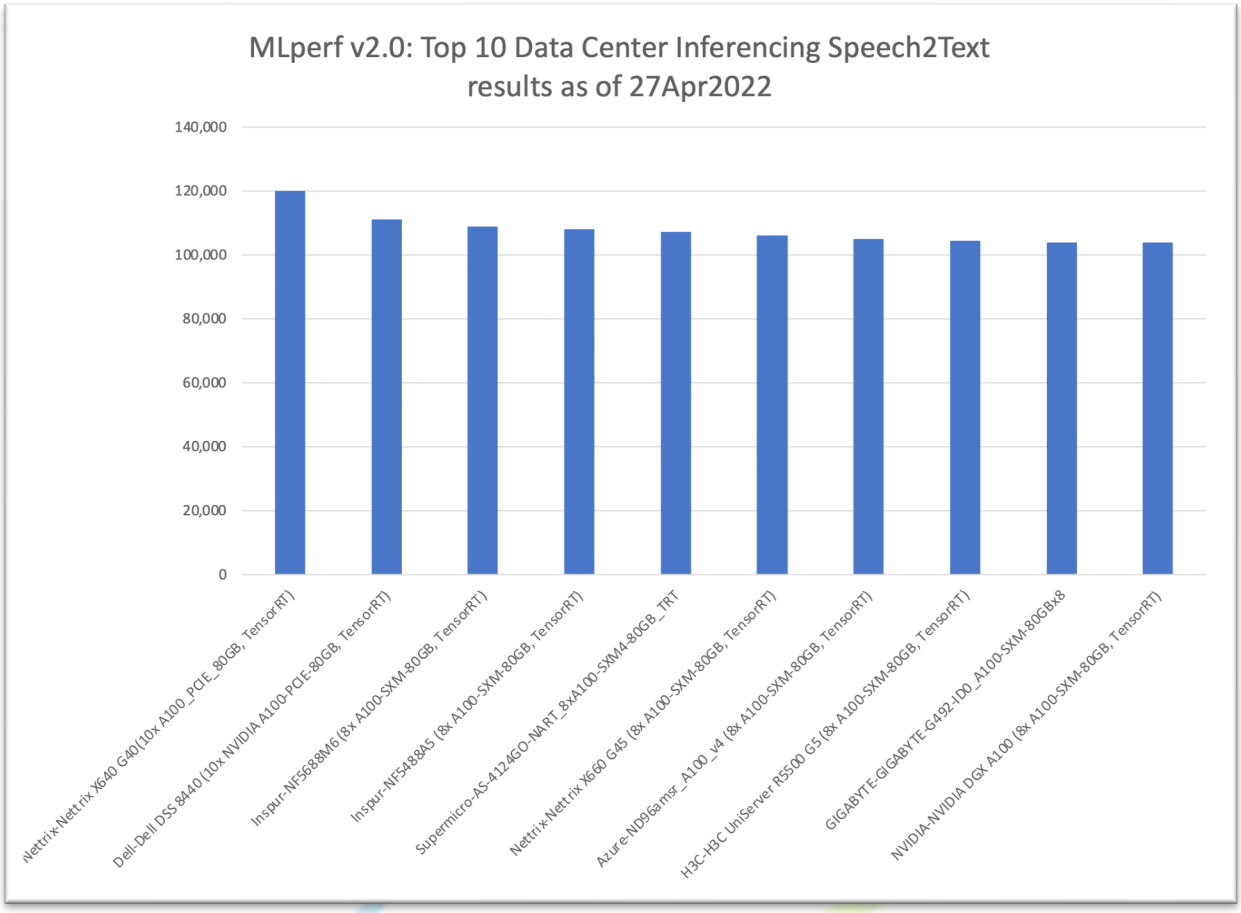

Figure 2 Top 10 MLperf v2.0 Data Center Speech2Text inferencing server q/s performance results

In Figure 2, once again the Netrix X640 G40 with 10-A100 PCIe-80GB GPUs came in #1 with just over 120K server q/s. At #2, we see the Dell DSS 8440 with 10-A100 PCIe-80GB GPUs, that achieved just over 111K server q/s and at #3, the Inspur NFS688M6 with 8-A100 SXM-80GB GPUs that achieved just over 109K server q/s on Speech2Text inferencing performance.

We've discussed the Netrix X640 vs Dell DSS 8440 systems and their respective CPU differences above. But the more interesting comparison here is between the #3 Inspur NFS688M6 and the #4 Inspur NF5488A5 which achieved a little over 108K server q/s performance.

The only difference between the two Inspur systems besides their names, seems to be their CPUs, in this case, the Intel 8358 CPUs for the #3 system and AMD EPYC 7742 for the #4 ranked system.

So, the AMD vs Intel CPU decreases Speech2Text inferencing server q/s performance by less than 1%. We don't compare pricing here, but it might be interesting to know the difference in

prices between these two, to compute their respective CPU price/performance on Speech2Text inferencing server q/s.

Significance

MLperf keeps releasing new versions of their benchmarks and vendors seem more than willing to measure their system performance using them. With v2.0 of data center inferencing, we see CPU choice can have a 1-10% impact on server query per second performance. But with the one data point, for Object Detection (large) inferencing, software choice can have a more significant (75% hit) impact on server performance.

If MLperf were listening, we would ask that they add more information in their reports on the system, storage and interface configurations used for each submission.

As always, suggestions on how to improve any of our performance analyses are welcomed.

This report was sent out to subscribers as part of 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.