MLperf™ Performance Report

Silverton Consulting, Inc. StorInt™ Dispatch

This Storage Intelligence (StorInt™) dispatch covers the MLperf™ v1.1 series of AI-ML-DL model training and inferencing benchmarks[1]. This report focuses on inferencing activity for the Data Center computing environment and workloads, which consists of image classification, object detection-large, medical image segmentation, speech to text, natural language processing and recommendation inferencing. The main MLperf data center inferencing metric we use is number of server queries/second.

MLperf v1.1 Edge inferencing "closed-available" benchmark results

We start our discussion with data center medical image segmentation inferencing results in Figure 1.

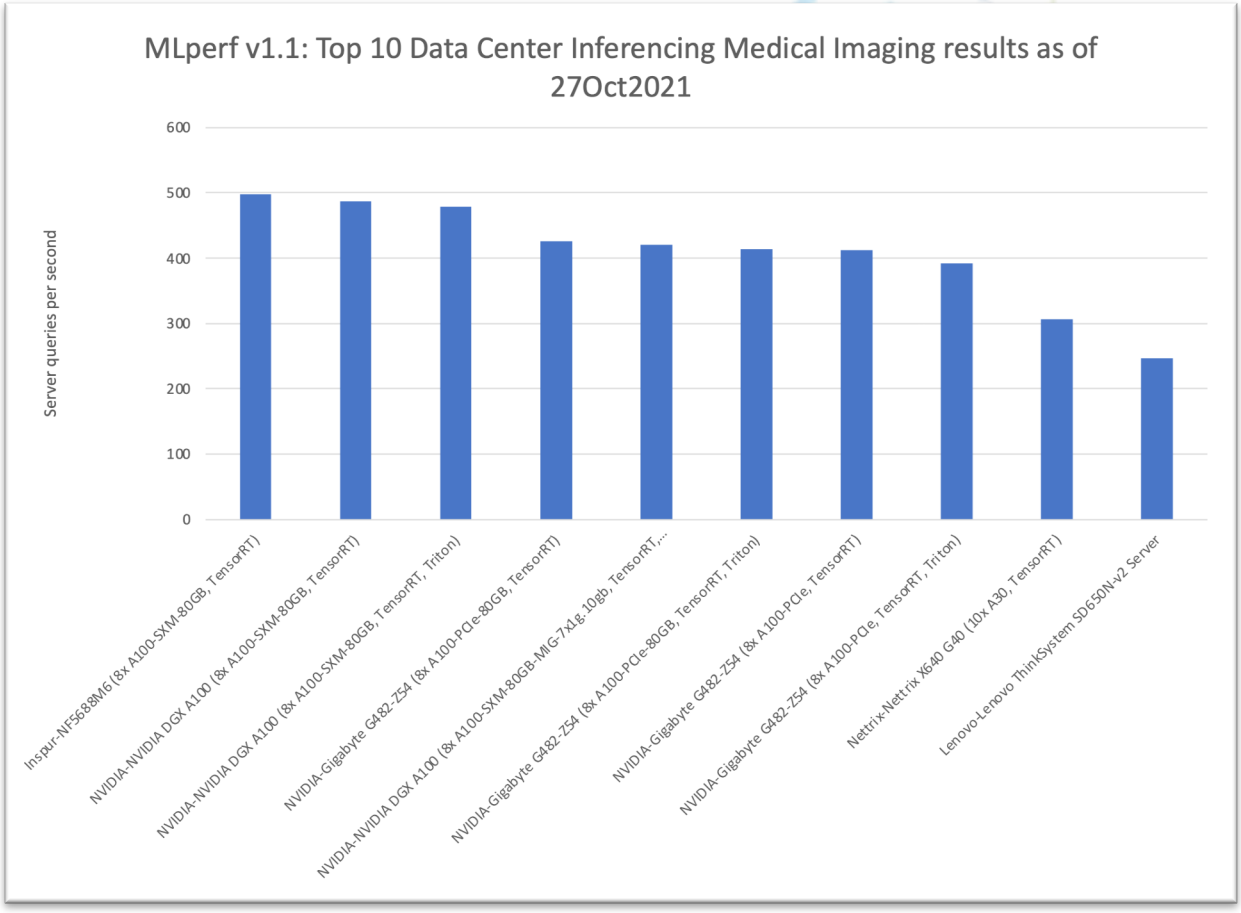

Figure 1 Top 10 MLperf v1.1 Data Center Medical Image Segmentation inferencing results

[1] All MLperf inferencing and training results are available at https://mlcommons.org/en/ as of 10/27/2021

In Figure 1, systems with eight (8) NVIDIA A100-SXM-80GB GPU chips took the top 3 spots, with Inspur NF5688M6 at #1 with 498 server queries/second (q/s), and NVIDIA DGX A100 systems at #2-3, with 487 and 479 q/s, respectively. The slower NVIDIA system used NVIDIA Triton to perform model inferencing, which for this benchmark was **2% slower** than running the model inferencing directly.

At #4, with 425 q/s, we see the use of NVIDIA A100-PCIe-80GB cards using eight GPUs, running on a Gigabyte G482-Z54 sever. The PCIe cards have always perform slower than the SXM chips, because they run at lower power. In this benchmark, the **PCIe cards ran ~15% slower than the SXM chips**. One also needs to consider that the SXM chips were running in a DGX A100 server while the cards were running in a Gigabyte G482-Z54 server.

But what's more interesting, is that at #5, with 421 q/s, we see NVIDIA MIG (multi-instance GPU) GPUs being used with Triton. This DGX A100 system had 8 A100-SXM-80GB GPU chips, each of which was configured as 7 GPUs with 10GB of memory each or 56 (virtual) A100-SXM-10GB GPUs. MIG necessarily introduces some overhead, and in this benchmark runs **~13% slower for the same number of GPU chips**. However, this is confounded by the fact that their MIG configured system could only make use of 70GB of the 80GB on each SXM chip.

In Figure 2, we show Natural Language Processing inferencing results.

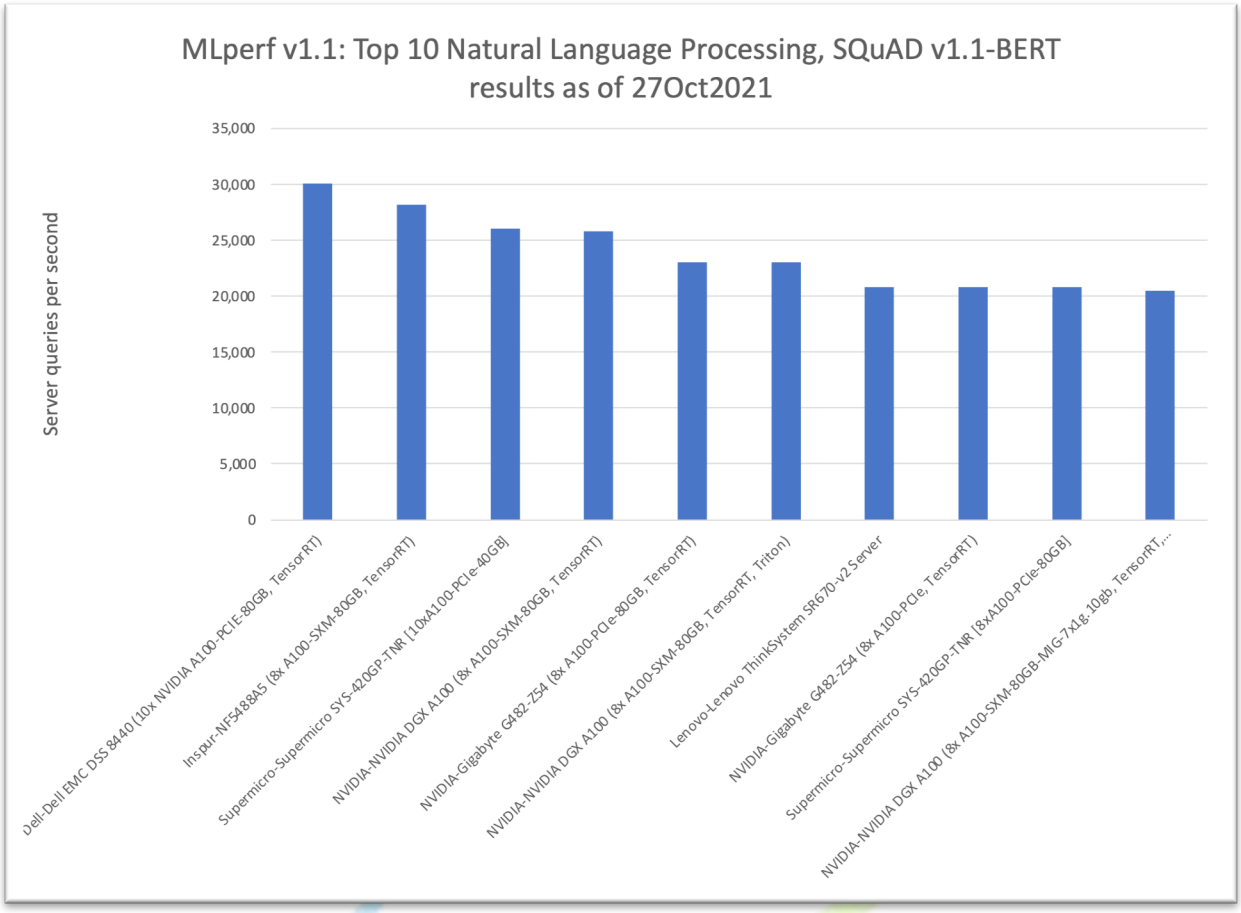

Figure 2 Top 10 MLperf v1.1 Natural Language Processing inferencing results

In Figure 2 we see more comparisons of A100 PCIe cards vs A100 SXM chips. For instance, the #1 Dell DSS 8440 used 10 NVIDIA A100-PCIe-80GB cards achieving ~30K server q/s, while the #2 Inspur NFS5488A5 system, with 8 NVIDIA A100-SXM-80GB chips, only reached ~28K q/s. Thus, for NLP inferencing, an A100 PCIe-80GB card can achieve 3.0K q/s about **17% slower** than a A100 SXM-80GB chip at 3.5K q/s. Again, there are different systems (Dell 8440 and Inspur NFS5488A) which confound this comparison of card vs. chip speed

Moreover, at #3, with ~26K q/s, the Supermicro SYS-4200GP system using 10 smaller memory, **A100-PCIe-40GB** cards. This shows that PCIe cards with ½ the memory only achieved 2.6K q/s, about **~15% less than the larger card.**

A MIG configured systems shows up at #10, with ~20K q/s, using 8 A100-SXM-80GB GPU chips, each configured as 7 (virtual) 10GB GPUs (total 56 SXM-10GB GPU [virtual]) chips running NVIDIA's Triton inferencing server. Like our previous analysis, for NLP inferencing, **the MIG configured chips run ~15% slower than the non-MIG** configurations with the proviso that the non-MIG system was able to use all the memory on the chips.

Significance

It's interesting to see the performance slowdown of MIG vs non-MIG configured systems seems similar (13% to 15% slower) for two entirely different workloads. And comparing A100 PCIe cards vs SXM chips also seems similar across the two workloads (15% to 17% slower). Yes, there are other factors to consider for both comparisons but seeing a similar slowdown across the two workloads indicates that SXM chips always faster than MIG configurations and PCIe cards.

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.

Silverton Consulting, Inc., is a U.S.-based Storage, Strategy & Systems consulting firm offering products and services to the data storage community

Newsletter signup QRcode