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 edge environments.

MLperf v0.7 edge inferencing benchmark results

The MLperf v0.7 edge inferencing series of benchmarks takes standard AI Deep Learning models and runs inferencing activity against them measuring **single stream average latency**, **multiple streams inferences**/sec and **offline activity** (batched inferencing, inferences/sec). Most of the edge models are trained to 99% accuracy.

There are 6 separate inferencing models in use by MLperf v0.7 for edge submissions. These include, image classification (ImageNet-ResNet), object detection small (COCO, SSD-small), object detection large (COCO, SSD-large), medical imaging (BraTS 2019, 3D-Unet trained to 99% and 99.9% accuracy), speech-to-text (LibriSpeech, RNN-T) and natural language processing (SQuAD v1.1, BERT). Reference versions of all models used in MLperf benchmarks are available on GitHub (https://github.com/mlcommons).

For example, the natural language processing BERT model has a number of different inferences it can perform. The MLperf v0.7 edge inferencing benchmark is run against the Stanford Question Answering Dataset (SQuAD v1.1) that determines whether two sentences contain a question and its answer.

We start our discussion with the medical imaging BraTS 2019, 3D-Unet, single stream latency (in msec), in Figure 1.

[1] All MLperf inferencing and training results are available at https://mlcommons.org/en/ as of 03/30/2021

© 2014-2020 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

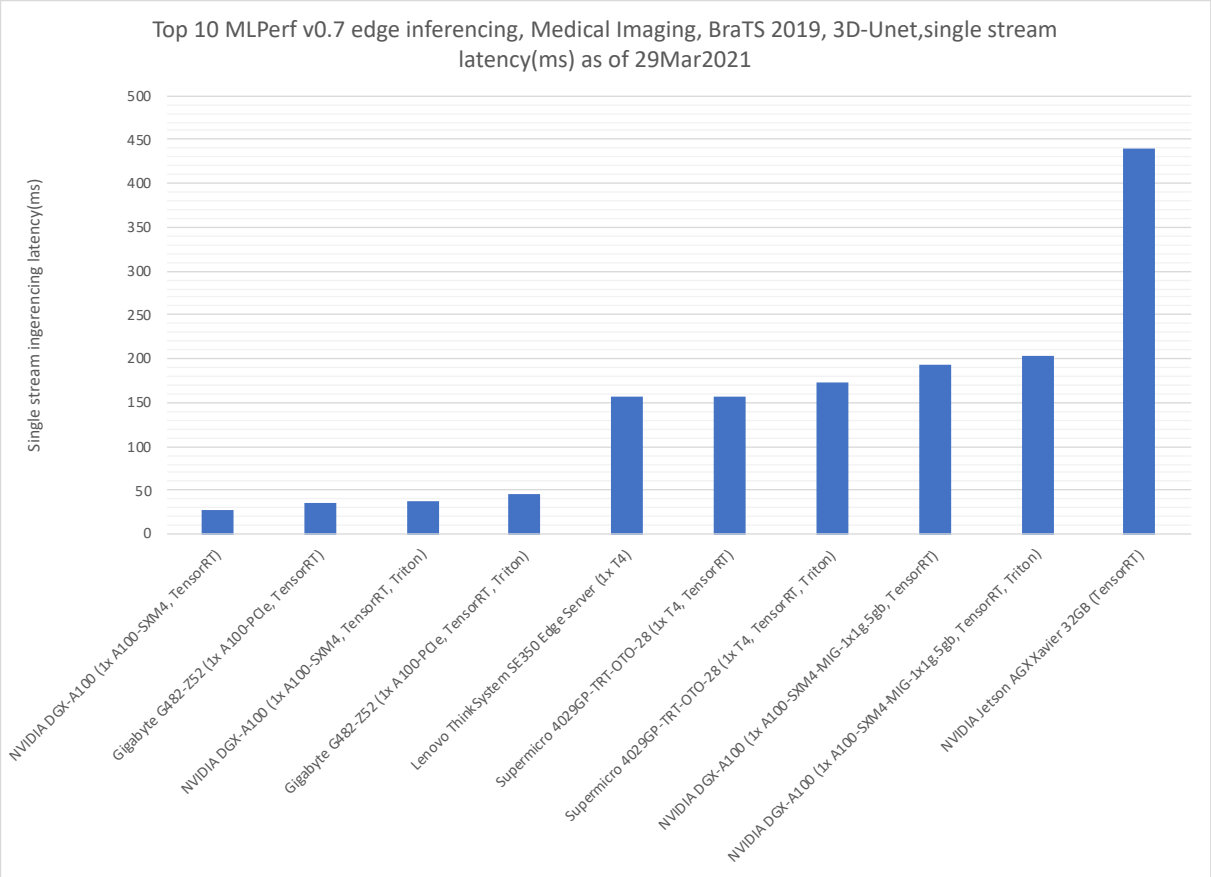

Figure 1 Top 10 MLperf v0.7 Edge Inferencing Medical Imaging results

In Figure 1 (lower is better), NVIDIA A100 GPUs took the top 4 slots, with NVIDA T4 GPUs taking the next 3, followed by 2 NVIDA A100 in Multi-instance GPU (MIG) mode systems followed at #10 by a NVIDIA Jetson AGX (embedded GPU). Latencies spanned 27.42msec (#1) to 439.27msec for #10.

The top ten systems above all used a single GPU accelerator. However, 8 out of the 10 systems had dual CPU chips (Lenovo and NVIDIA Jetson being the exceptions). And all systems used X86 processors except the NVIDIA Jetson which used ARMv8.2. The three NVIDIA T4s submission used Intel CPUs while the other X86 processor submissions used AMD EPYC CPUs. And all the above systems used TensorRT for their inferencing engine.

The differences between A100 PCIe and A100 SXM4 seems to be that the PCIe has graphics port support. Gigabyte servers come from a Taiwanese company that provides AI workstations using NVIDIA GPUs.

One item of interest is that the use of MIG degraded performance by ~6.2X in latency (avg 32 msec for the #1&3 systems vs. avg. 198 msec for #8&9 systems) even though all GPUs were used in single stream mode. This could be due to the memory allocated to the MIG GPU

© 2014-2021 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

instance. The name of the submissions implies the A100 was partitioned into 1 GPU (1g) with 5GB of memory (5gb). A normal A100 has 40GB (or 80GB) of GPU memory. With only 5GB of memory allocated to the GPU instance, the single A100 could have supported 6 more (or 7 total) GPU instances running the same AI inferencing workloads.

In Figure 2 we present Top 10 MLPerf v0.7 edge inferencing results for object detection large, COCO, SSD-large results.

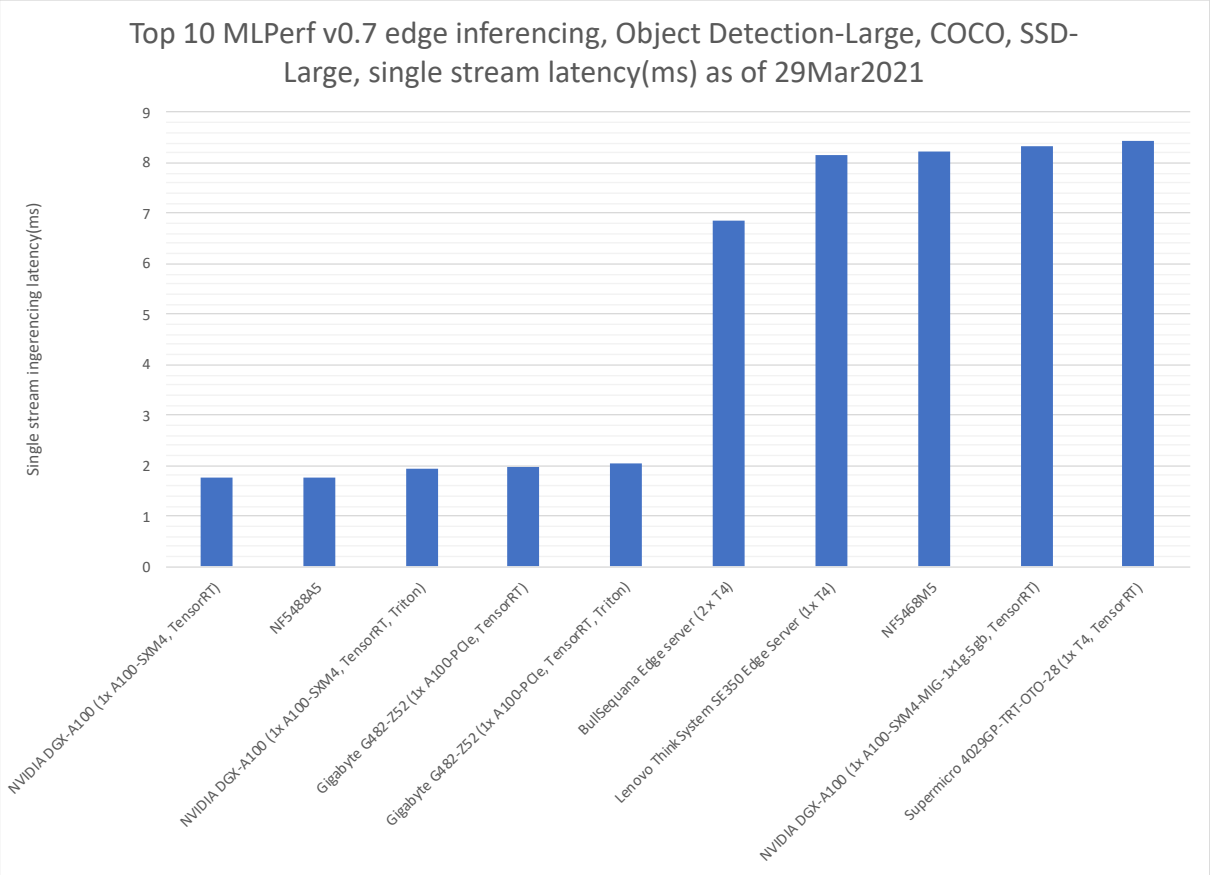

Figure 2 MLperf v0.7 edge inferencing object detection large COCO SSD-large performance results

In Figure 2, NVIDIA A100s took the top 5 slots (the NF5488A5 is an Inspur system using an NVIDIA A100). Latencies for object detection large, single stream were much closer than medical imaging above. For large object detection, latencies ran from 1.76 to 8.42 msec.

The surprising thing is that the top 3 submissions are all SXM4 A100s while the next 2 are PCIe versions. And 4 out of the bottom 5 (in top 10) used NVIDIA T4 GPUs while there was one (#9) using A100 in MIG mode (#9). All the T4 systems in this top 10 used Intel CPUs and all the others used AMD EPYC CPUs.

On this workload the MIG result (#9, 8.34 msec) single stream latency was only ~4.5X slower than the non-MIG A100 SXM4 systems (avg 1.8msec).

© 2014-2021 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

Next, we turn to MLperf edge inferencing results for natural language processing in Figure 3.

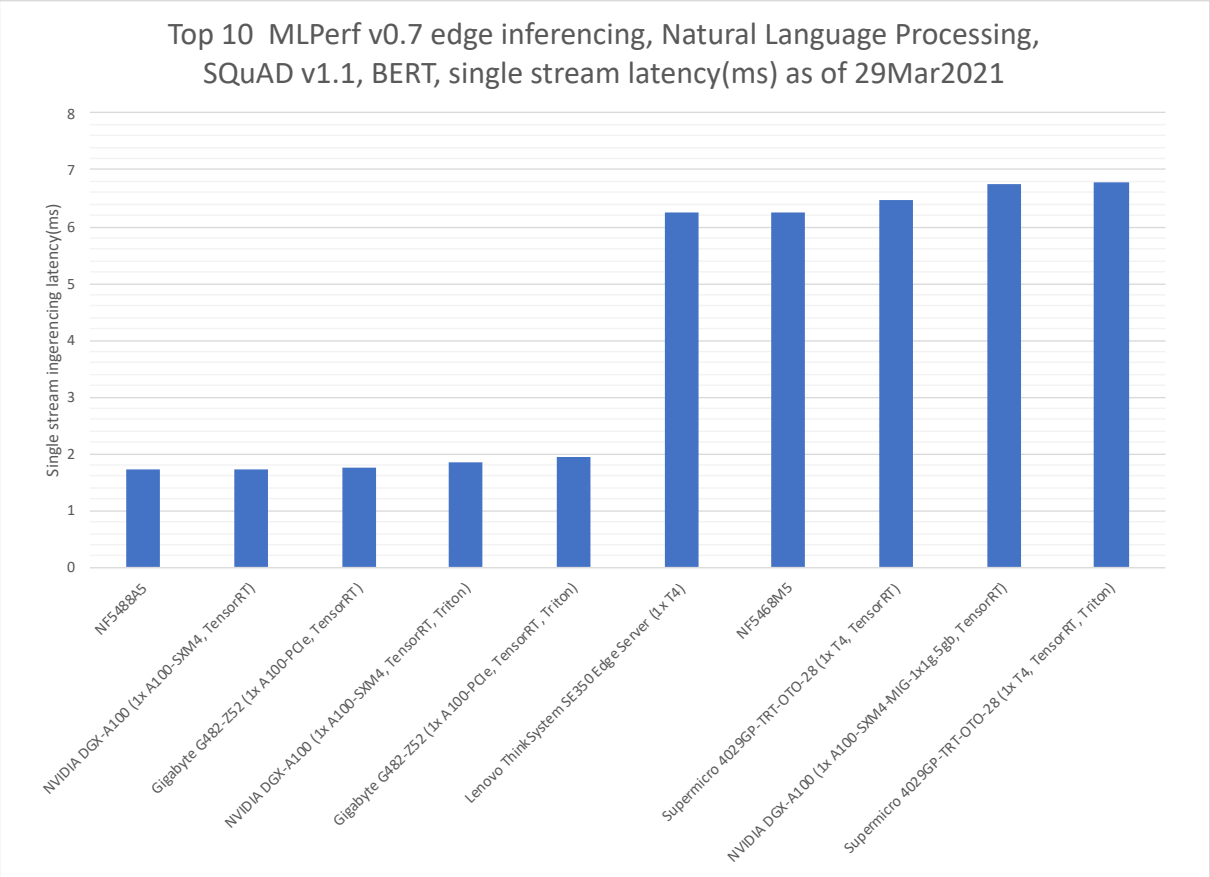

Figure 3 Top 10 MLperf v0.7 data center image classification performance results

In Figure 3, once again the A100 based systems took the top 5 slots. Latencies on natural language processing inferencing ran from 1.72 msec to 6.78 msec. There were 4 T4 systems (#6,7,8, & 10). One can see a definite bifurcation in latencies with the A100s all under 2msec and all the other systems over 6msec. And again the T4 systems all used Intel CPUs while the A100 systems all use AMD EPYC CPUs.

Again the MIG A100 result was again slower with a latency ~3.8X worse than the average of the A100 SXM4 systems.

Significance

AI-ML-DL training and inferencing activities are becoming more mainstream. As the world moves to IoT, a lot of inferencing will be done at the edge. At ~1-2msec average latencies, this should be adequate for most edge activities. However, we are not sure this speed would adequately support autonomous car driving or autonomous plane piloting, with 20-60 active sensors at any one time. It's unlikely that these systems will support multiple GPUs but that would be one way to deal with this.

© 2014-2021 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

Unclear what's really going on with the MIG submissions. It would have been great to have multiple MIG submissions at different memory levels for each workload to see how memory size impacts latencies.

This report shows that NVIDIA GPUs are the go to hardware accelerator for inferencing work at the edge. Our prior report showed that they were also dominant in data center training and inferencing activities. In our view, having one vendor so dominant in an application domain so important to the world's future, can become a problem

This is only our second performance report analyzing MLperf. We plan to discuss different training and inferencing workloads in our next report. 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.

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

© 2014-2021 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED