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 training activity for HPC environments.

MLperf v1.0 edge inferencing benchmark results

The MLperf v1.0 HPC training series of benchmarks takes standard HPC deep learning models and runs them multiple times on a given hardware configuration and measures the time it takes to train to a given accuracy level. The main metric is the measured time to train the model to a specified level of accuracy, in minutes.

There are 2 DNN models represented in the MLperf v0.7 HPC training benchmark: the **CosmoFlow** and **DeepCAM** models. DeepCAM is a climate segmentation model but at press time, there were only 3 submissions, so we will save discussing DeepCAM results for a future report.

CosmoFlow is a 3D CNN cosmology parameter prediction model from LBNL which takes as input 3D segments of the universe (with 4 redshift buckets) and predicts OmegaM, Sigma8 and Ns cosmological parameters for that universe, at a mean average error of 0.124.

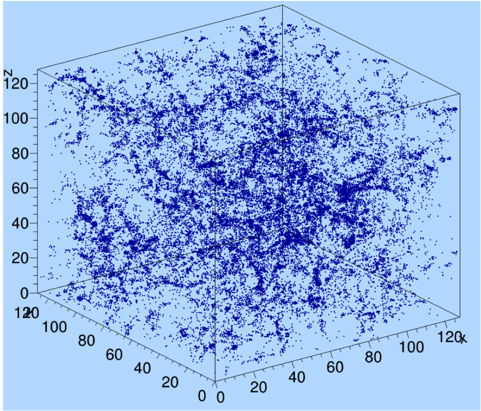

Fig. 1. Example simulation $256Mpc/h^3$, 128^3 voxel sub-volume, used as input to the CosmoFlow network. This sub-volume is taken from the full $512Mpc/h^3$ simulation of dark matter in the universe, evolved over 3 billion years to a redshift of 0 (i.e. today).

There's a paper[2] which describes the CosmoFlow CNN model[3] in more detail. Figure 1 is from the paper on CosmoFlow.

In Figure 2, we report on the top 10 MLperf v0.7 HPC CosmoFlow training results.

[1] All MLperf inferencing and training results are available at https://mlcommons.org/en/ as of 05/26/2021

[2] Please see ***CosmoFlow using DL to learn the universe at scale*** paper (https://arxiv.org/pdf/1808.04728.pdf)

[3] Reference version of the CosmoFlow model can be found at https://github.com/sparticlesteve/cosmoflow-benchmark/

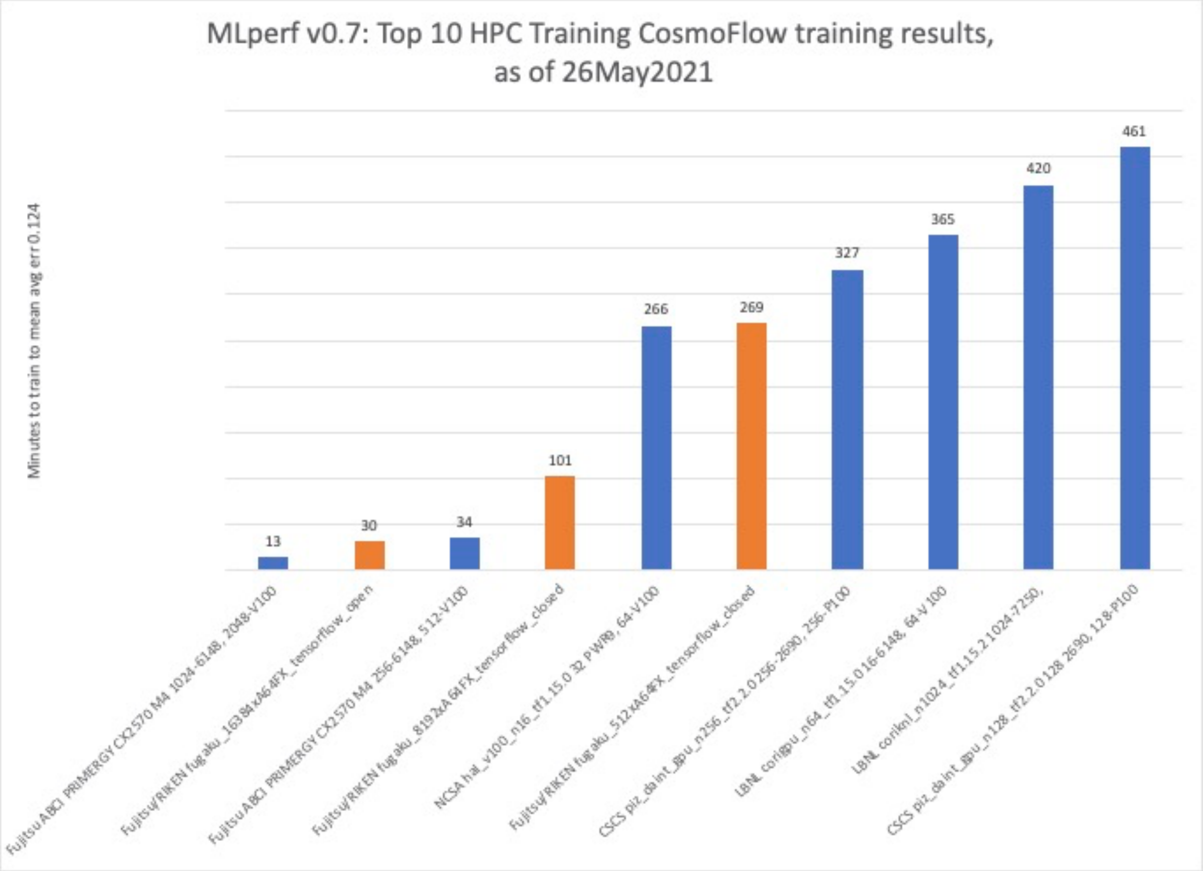

Figure 2 Top 10 MLperf v0.7 HPC training results

In Figure 2, the surprising results all came from **Fujitsu/RIKEN** submissions (#2, 4, & 6) which used no GPUs and only multiple **Fujitsu A64FX CPUs**[4]. The results achieved with this CPU alone rivaled other submissions that used CPUs and NVIDIA GPUs.

For example, the #2 submission used **16K A64FX CPUs** and 0 GPUS vs the #3 submission which used **256 Intel 6148 CPUs & 512 NVIDIA V100 GPUs** with a modest (~10%) improvement in training time. Similarly, the #6 submission used **512 A64FX CPUs** and 0 GPUs while the #5 submission used **32 IBM POWER9 CPUs & 64 NVIDIA V100 GPUs** with only a slight increase in performance time (~1%).

The Fujitsu A64FX processor has 48 ARMv8 compute cores that provide special acceleration for HPC workloads. These include ARMv8-A architecture with **Scalable Vector Extension**, **HBM2** (high bandwidth memory), **Tofu-D** (Fujitsu torus fusion proprietary) interconnect controller and PCIe gen3 support. The processor also adds **FP16, INT16 and INT8 dot product**, **scatter/gather vectorization**, along with other HPC oriented speed-ups.

In addition, all the Fujitsu/RIKEN A64FX processor submissions used **Mesh TensorFlow** with TensorFlow for their model processing software. Mesh TensorFlow was specifically designed

[4] From https://www.fujitsu.com/downloads/SUPER/a64fx/a64fx_datasheet.pdf as of 27May2021

to support SPMD (single program [instruction], multi-data) programming model for model training parallelism. We suppose that standard GPU TensorFlow provides similar capabilities for GPU model training parallelism

Significance

When we first heard of HPC making use of AL ML DL techniques for weather forecasting at SC19, we couldn't believe it. But it makes sense, there's plenty of (weather sensor) data with subsequent weather measurements that could be associated with it. So, the data's available and why not use DL models to bypass the significant weather model computations to obtain a forecast.

CosmoFlow wasn't mention then but the paper goes into significant length on how the data is obtained (through simulations) and how the model is configured and operates.

This is our fourth performance report analyzing MLperf. And at this point we have examined all the training and inferencing submissions where it makes the most sense. 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