Deep Learning Benchmark Studies on an Advanced AI Engineering Testbed from the Open Compass Project

Mei-Yu Wang
Pittsburgh Supercomputing Center,
Carnegie Mellon University
Pittsburgh, PA, USA
meiyuw@andrew.cmu.edu

Julian A. Uran
Pittsburgh Supercomputing Center,
Carnegie Mellon University
Pittsburgh, PA, USA
julian@psc.edu

Paola A. Buitrago
Pittsburgh Supercomputing Center,
Carnegie Mellon University
Pittsburgh, PA, USA
paola@psc.edu

ABSTRACT

We present the Open Compass project's pilot deep learning benchmark results with various AI accelerators. Those accelerators are NVIDIA V-100 and A-100, AMD MI100, as well as emerging novel accelerators such as Cerebras CS-2 and Graphcore. We evaluate their performance on various deep learning training tasks. We then discuss key insights from our experiments and share experiences about evaluating and integrating those novel AI accelerators with our supercomputing systems.

CCS CONCEPTS

• **Hardware → Hardware accelerators**; • **Computing methodologies → Machine learning**.

KEYWORDS

Artificial Intelligence, Benchmarking, BERT, UNet, Large Language Model, Neocortex, Wafer-Scale Engine, Graphics Processing Unit, Intelligence Processing Unit, Bridges-2

ACM Reference Format:
Mei-Yu Wang, Julian A. Uran, and Paola A. Buitrago. 2023. Deep Learning Benchmark Studies on an Advanced AI Engineering Testbed from the Open Compass Project. In *Practice and Experience in Advanced Research Computing (PEARC '23), July 23–27, 2023, Portland, OR, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3569951.3597596

1 INTRODUCTION

Open Compass[2] is an exploratory research project at the Pittsburgh Supercomputing Center (PSC) to conduct academic pilot studies on an advanced engineering testbed for artificial intelligence (AI). Open Compass includes the development of an ontology to describe the complex range of existing and emerging AI hardware technologies and identify benchmark problems that represent different challenges in training deep learning models. These benchmarks are then used to execute experiments in alternative advanced hardware solution architectures. One such effort for exploring alternative AI hardware solutions at PSC is the Neocortex system [3], an NSF-funded resource that targets the acceleration of AI-powered scientific discovery by vastly shortening the time required for deep learning training and fostering greater integration of deep learning with scientific workflows. Here we present some preliminary results on analyzing the effects of different accelerator types, including Cerebras CS-2, NVIDIA A-100, V-100, AMD MI100, and Graphcore IPU, for popular deep learning models applicable, such as image processing and language models. We discuss the insights from our experiments and plans for future exploration.

PEARC '23, July 23–27, 2023, Portland, OR, USA

ACM ISBN 978-1-4503-9985-2/23/07.
https://doi.org/10.1145/3569951.3597596

2 OVERVIEW OF EVALUATED AI ACCELERATORS

The diverse hardware type considered here covers main-stream devices for deep learning studies and innovative AI accelerators tailored to speed up large-scale AI workloads. In this work, we perform pilot benchmarking tests on the following accelerators: Cerebras CS-2, Graphcore BOW-IPU, AMD MI100, NVIDIA A-100, and NVIDIA V-100. We use NVIDIA A-100 and V-100 results as our baseline to evaluate other accelerators. Table 1 provides an overview of the hardware characteristics and software stack. Below, we also describe the setup of machines that host those accelerators.

- The Cerebras CS-2 is a second-generation wafer-scale engine (WSE) from Cerebras Systems. It is a single-chip processor containing 850,000 AI-optimized cores and 40 GB of on-chip SRAM. The CS-2 is designed for large-scale AI applications. It is powered by a server HPE Superdome Flex 280, which is an eight-chassis server with up to 448 cores and 24TB of memory, connected to each CS-2 machine via a path of 8 x 100 GbE individual connections, used by coordinated worker processes to stream job data into the Cerebras appliances.
- The Graphcore IPU POD-4 is a four-chip BOW-2000 IPU system from Graphcore. This IPU-POD system is designed for large-scale AI applications, and our setup is a subset of the smallest system for the Colossus Mk2 architecture. Our configuration contains 5,888 IPU cores and 3.6 GB of on-chip SRAM. The system is powered by an Exxact TS2-158632687-AES twi-socket server by Intel 8280L CPUs, for a total of 64 cores and 512GB of RAM, via a direct-attach 100GbE connection.
- The AMD MI100 is a GPU that contains 7,680 stream processors and 32GB of HBM2 memory. Our setup is powered by the same Exxact TS2-158632687-AES server driving the Graphcore equipment.
- An NVIDIA V-100 GPU contains 5,120 CUDA cores and 32 GB of HBM2 memory. It is powered by a two-socket HPE Apollo 6500 Gen10 server with two Intel 6248 CPUs for a

System	Cerebras CS-2	Graphcore IPU POD-4	AMD MI100	NVIDIA V-100	NVIDIA A-100
Architecture	WSE-2	Colossus Mk2 IPU	CDNA	Volta	Ampere
Cores	850.000	1.472 /IPU	120 CU (7.680 SP)	5,120c CUDA, 640c Tensor /GPU	6,912c CUDA, 432c Tensor /GPU
Memory	40GB SRAM	0.9GB /IPU	16MB+16MB L1, 60MB L2 32GB HBM2 ECC /GPU	128KB L1, 6MB L2 32GB HBM2 /GPU	192KB L1, 40MB L2 80GB HBM2 /GPU
Memory Bandwidth	20 PB/s	261 TB/s	1,2 TB/s /GPU	900 GB/s /GPU	1.555 GB/s /GPU
Performance (FP16)	5.13 PFLOPS (mix of FP32 & FP16)	1 PFLOPS	184.6 TFLOPS /GPU	15.7 TFLOPS /GPU	19.5 TFLOPS /GPU
Performance (FP32)		250 TFLOPS	23.1 TFLOPS /GPU	7.8 TFLOPS /GPU	9.7 TFLOPS /GPU
Precision support	FP16, FP32, cbfloat	FP16, FP32	FP16, FP32, FP64 BF16, INT4, INT8	FP16, FP32, BF16 INT4, INT8	FP16, FP32, BF16 TF32, INT4, INT8
Process	7nm	7nm	7nm	12nm	7nm
Framework Support	TF, PyTorch, Cerebras SDK	TF, PyTorch, ONNX, PopArt	PyTorch, ONNX, MxNET, ROCm	TF, PyTorch, ONNX, MxNET, CUDA	PyTorch, ONNX, MxNET, CUDA
Host System	HPE Superdome Flex 280 32x Intel 8280L 24TB RAM	Exxact TS2-158632687-AES 2x AMD EPYC Milan 7543 512GB RAM		HPE Apollo 6500 G10 2x Intel 6248 512GB RAM	Exxact TS4-195183185 2x AMD EPYC 7543 2TB RAM

Table 1: Features of evaluated AI accelerators

total of 80 cores and 512GB of RAM. These GPUs are hosted by the Bridges-2[1] GPU nodes.

- An NVIDIA A-100 GPU contains 6,912 CUDA cores and 80 GB of HBM2 memory. It is powered by an Exxact TS4-195183185 server with two AMD EPYC 7543 CPUs for a total of 64 cores and 2TB of RAM.

3 EVALUATED DEEP LEARNING BENCHMARKS

Below we discuss the details of our deep learning benchmark implementation, such as model architecture, dataset, and associated parameters used for our evaluation study. Here we divide our work into two parts: (1) deep learning benchmark performance across different hardware. (2) benchmarking Cerebras Model Zoo [1] performance on Neocortex.

3.1 Deep learning benchmark performance across different hardware

Here we focus on two deep learning models: UNet[10] and BERT[4], and present results by running those models across different hardware with the following setting. Those applications are run with FP16/amp mixed precision on A-100, V-100, AMD M100, FP16 on Graphcore, and mixed precision on CS-2 runs (equivalent to FP16 for A-100). The implementations of BERT and UNet models on A-100, V-100, and Graphcore use TensorFlow 2, and CS-2 uses TensorFlow Estimator Framework. We adopt Horovod for A-100/V-100/AMD MI100 multi-GPU data parallelism framework.

1) *UNet*: UNet is a convolutional neural network originally proposed for biomedical image segmentation [10]. We follow the UNet implementation of Cerebras Model Zoo Release 1.6.0 [2]. The model is a variant of the original UNet [10] model for which the order of the layers, filter size, padding, and stride setting in the repeated blocks in the contracting path and expansive path are different. We note that this version of the Cerebras UNet model is experimental, so its performance on CS-2 may not be representative. The model is trained with cross-entropy loss and Adam optimizer on the DAGM 2007 competition dataset [3], which is a synthetic dataset with 256×256 pixels greyscale images for defect detection on textured surfaces.
2) *BERT*: BERT[4] is an encoder-only transformer-based model designed for natural language understanding. The adopted implementation follows the original BERT model but is trained with an AdamW optimizer. We perform pretraining tasks with BERT Large model using the OpenWebText dataset [6, 8] with a maximum sequence length of 128.

3.2 Cerebras Model Zoo performance on Neocortex

We evaluate the performance of Cerebras Model Zoo models running on the CS-2 servers in the Neocortex system. The runs are done using the default setting of Cerebras Model Zoo R1.6.0 (with various batch sizes)[1] and software stack R1.6.0 with pipeline mode. We can potentially further improve the performance with a newer software stack and weight-streaming mode. However, some of those features will require changing the current setting of the Neocortex system. We list the properties of tested models in Table 2, which include: UNet, MLP (fc_mnist), T5[9] base and small, Transformer[11] large and base, BERT[4] large and base, and GPT-2[8] large, medium, and small models. Please check the Cerebras Model Zoo Github[1] for more details about the model implementation and description.

4 RESULTS

We present training throughput and end-to-end execution time measurements from running deep learning models across a diverse set of AI hardware and from a selection of Cerebras Model Zoo examples running on Neocortex.

4.1 Deep learning benchmark performance across different hardware

Figure 1 shows the training throughput (samples/sec) of the BERT Large (left panel) and UNet model (right panel) for different hardware. The number of devices in Figure 1 varies due to the configuration of the systems. For example, currently we have not

[1] https://www.psc.edu/resources/bridges-2/
[2] https://github.com/Cerebras/modelzoo/tree/R_1.6.0/modelzoo/unet/tf
[3] https://www.kaggle.com/datasets/mhskjelvareid/dagm-2007-competition-dataset-optical-inspection

Model	Dataset	Fabric core utilization	Batch size per device
UNet	DAGM 2007[a] & Severstal dataset[b]	62.9%	64, 128, 256
MLP (fc_mnist)	MNIST	0.7%	128, 256, 512
T5 base & small[c]	the Colossal Clean Crawled Corpus (C4) Dataset[d]	64.4% & 18.5%	128, 256, 512
Transformer large & base	WMT-2016[e]	62.0% & 29.6%	2048, 4096
BERT large & base	OpenWebText (maximum sequence length =128, 512)	94.7% & 77.9%	256, 512, 1024
GPT-2 large, medium, & small	OpenWebText (maximum sequence length = 1024)	94.7% , 91.9%, & 78.3%	32, 64, 128

Table 2: Cerebras Model Zoo examples considered in the study.

[a] https://www.kaggle.com/datasets/mhskjelvareid/dagm-2007-competition-dataset-optical-inspection
[b] https://www.kaggle.com/c/severstal-steel-defect-detection/overview
[c] https://github.com/google-research/text-to-text-transfer-transformer/blob/main/released_checkpoints.md#t511
[d] https://commoncrawl.org
[e] https://www.statmt.org/wmt16/it-translation-task.html

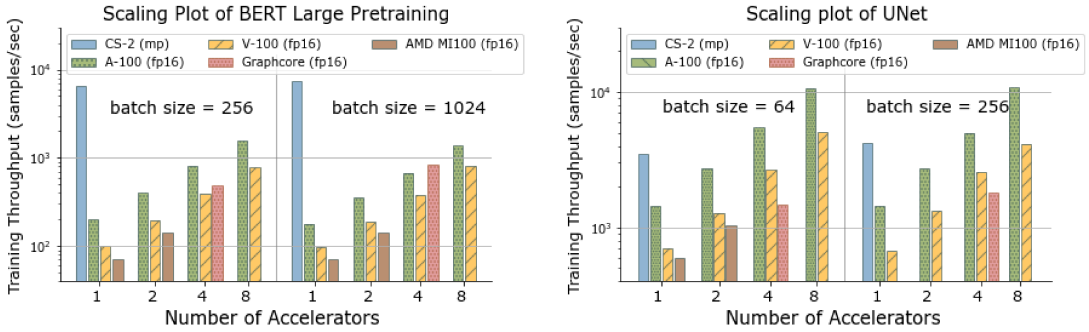

Figure 1: Training throughput for Deep Learning models, including BERT Large model pretraining and UNet, for various hardware and batch size.

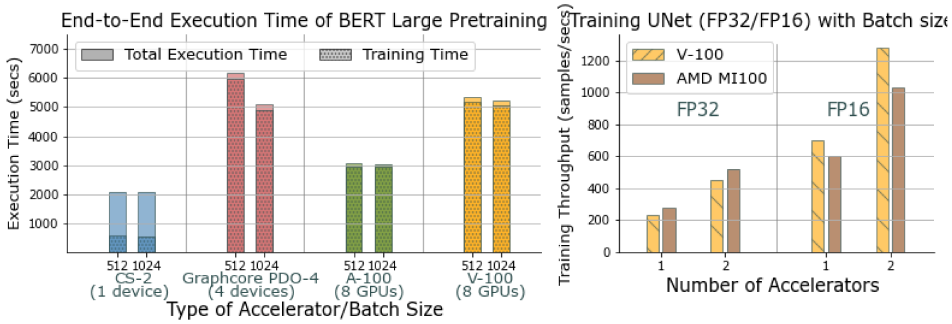

Figure 2: *Left Panel*: End-to-end execution time (solid color histograms), including compilation, I/O, and data pre-processing, versus training time (dotted histograms) for training BERT large model with 4,096,000 samples. *Right Panel*: Training throughput of training UNet with FP32 and FP16 for V-100 and AMD MI100 GPUs.

set up training across multiple CS-2 accelerators. We currently mainly explore settings that running with all four Graphcore IPUs in our system. We list the performance of V-100 and A-100 up to the number of GPUs hosted on a single node (eight). Regardless the heterogeneous setting and nature of the systems, we list the performance for a given number of accelerators for comparison. For BERT pretraining task, the performance of CS-2 outperforms other accelerators significantly, either with single or multiple devices. Graphcore outperforms main-stream GPUs (same number of devices) when training with large batch size $\geq$ 1024. For the UNet model, the performance of CS-2 is the highest among the considered hardware with a device number less or equal to two. However, the performance of A-100s and V-100s starts to catch up or even outperform when training with device numbers greater than four. For BERT with a batch size of 256/1024, throughput improvements observed for one CS-2 against eight A-100s and eight V-100s are 4.1×/5.2× and 8.3×/9.1×, respectively. For Graphcore IPU POD-4 with a batch size of 256/1024, the throughput is 0.6×/1.3× and 1.3×/2.2× compared to four A-100s and four V-100s.

When training with FP16/mixed-precision, the AMD GPUs have lower training throughput than V-100s. However, as shown in the right panel of Figure 2, the training throughput of AMD MI100

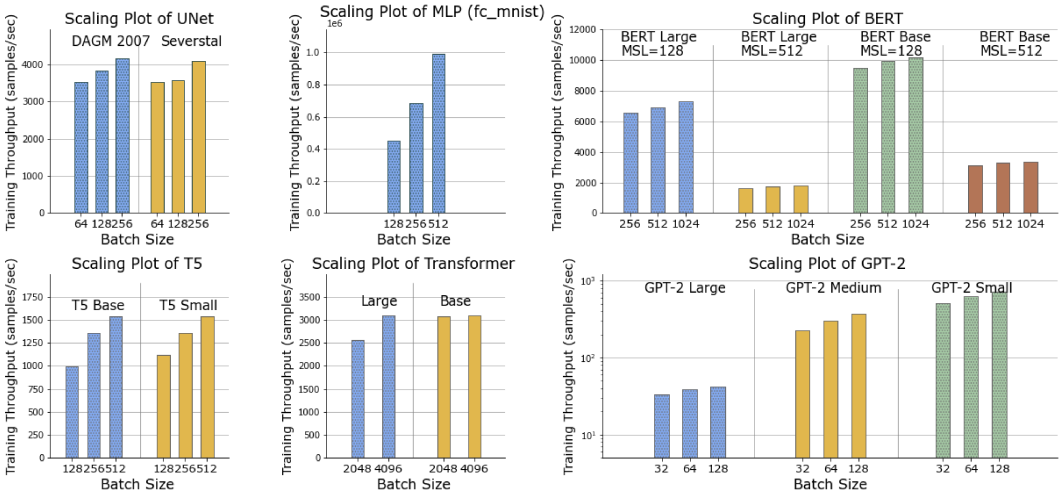

Figure 3: **Training throughput for Cerebras Model Zoo Deep Learning models trained with CS-2.**

with FP32 precision is comparable or slightly higher (1.1x/1.2x for one/two GPUs) compared to V-100 GPUs for the UNet test we performed. Herefore, the current AMD MI100 mixed precision configuration may need to be optimized for some of the deep learning tasks.

The training throughput for A-100s, V-100s, and AMD MI100s, in general, is not sensitive to varying batch sizes in the range we are testing, and in some cases, the performance may drop slightly with a larger batch size. For CS-2 and Graphcore, the performance gain is more prominent with increasing batch size. The CS-2 performance gain is 1.1× from batch size 256 to 1024 for BERT and 1.2× from batch size 64 to 256 for UNet, and for Graphcore, it is 1.7× for BERT and 1.2× for UNet.

We also measure the end-to-end execution time of performing BERT large model pretraining with 4,096,000 samples. The results are shown in the left panel of Fig 2. Please note that the number of devices used here differs for different hardware: one for CS-2, four IPUs for Graphcore, and 8 GPUs for A-100 and V-100. Note the number of devices for each hardware is different, which reflect the available devices for a given node/server in our system. For most of the hardware, the compilation and data pre-processing time is much less significant compared to the actual training time. However, for CS-2, the compilation time is generally much longer than for other accelerators. Nevertheless, the high training throughput for CS-2 compensates for the high compilation time, so CS-2 measures the shortest total execution time among the considered hardware and demonstrates its capable of performing efficient large model training tasks.

4.2 Cerebras Model Zoo perfomance on Neocortex

Figure 3 shows the training throughput (samples/sec) of various Cerebras Model Zoo examples. In those examples, the CS-2 server displays its capability of performing training with large batch sizes with high efficiency. As discussed in Section 4.1, CS-2 demonstrates superb performance on large deep learning tasks such as training BERT and GPT-2. The training throughput benefits from increasing batch size and shows a scaling relation between performance and batch size. For example, the training throughput for T5 increases by 1.4× and 1.5× when increasing the batch size from 125 to 256 and 512, and for BERT large model, it increases by 1.1× and 1.2× when increasing the batch size from 256 to 512 and 1024. In this work, we report the result of measuring training throughput only, but we will present results from other ongoing tests and different performance metrics in future work.

5 DISCUSSION

In this work, we present the pilot deep learning benchmark results of the Open Compass project. We evaluate the performance of several main-stream GPU devices and novel AI accelerators. Their performances vary for different deep learning tasks and training settings. For example, the CS-2 shows promising results in providing efficient training solutions for large language models like BERT and GPT-2. The performance of Graphcore is most optimized with a large model size and batch size, which outperforms the results of training the BERT Large model with the same number of A-100 devices. For AMD MI100, the training performance with FP32 precision is higher than V-100s, but for FP16, it is not. Given that the theoretical performance of AMD MI100 in FP16 (see Table 1) is actually higher than V-100s, there might be room for improvement for the FP16 training configurations.

We also note that there are several limitations in our current hardware settings. For example, due to the system architecture, we are unable to upgrade our CS-2 servers to the latest software stack yet (R1.8.0 as of April 2023) or perform the weight-streaming mode training robustly. Therefore the results we present here may not reflect the best performance of what CS-2 can achieve. We also note that the Graphcore IPU POD-4 we tested is a subset of the system, which usually consists of 16 or 64 IPUs (POD-16 or POD-64) for the complete system. Nevertheless, these novel accelerators still perform well in some experiments. We also find that AMD MI100 is a good alternative to NVIDIA GPUs, given its competitive computing power compared to V-100s. However, the mix-precision

training configuration can be further optimized to achieve better performance.

Several ongoing efforts are working on evaluating novel AI accelerators for open science deep learning workloads (e.g., [5, 7]). Some tasks and experiment settings are similar to ours, but there has yet to be a consensus on benchmarking diverse AI accelerator types. We plan to continue conducting experiments on our AI testbed to test more diverse deep learning model types and provide insights to optimize deep learning workloads for the open science community.

REFERENCES

[1] Cerebras Model Zoo R1.6.0. https://github.com/Cerebras/modelzoo/tree/R_1.6.0.

[2] Paola A. Buitrago and Nicholas A. Nystrom. 2019. Open Compass: Accelerating the Adoption of AI in Open Research. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)* (Chicago, IL, USA) *(PEARC '19)*. Article 72, 9 pages.

[3] Paola A. Buitrago and Nicholas A. Nystrom. 2021. Neocortex and Bridges-2: A High Performance AI+HPC Ecosystem for Science, Discovery, and Societal Good. In *High Performance Computing*. Springer International Publishing, 205–219.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. 4171–4186. https://doi.org/10.18653/v1/n19-1423

[5] Emani and et al. 2022. A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads. In *2022 IEEE/ACM International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*. 13–25. https://doi.org/10.1109/PMBS56514.2022.00007

[6] Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus

[7] Abhinand Nasari and et al. 2022. Benchmarking the Performance of Accelerators on National Cyberinfrastructure Resources for Artificial Intelligence / Machine Learning Workloads. In *Practice and Experience in Advanced Research Computing* (Boston, MA, USA) *(PEARC '22)*. Article 19.

[8] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[9] Colin Raffel and et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, Cham, 234–241.

[11] Ashish Vaswani and et al. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.