AIgean: An Open Framework for Machine Learning on Heterogeneous Clusters

Naif Tarafdar*, Giuseppe Di Guglielmo†, Philip C Harris‡, Jeffrey D Krupa‡,
Vladimir Loncar§, Dylan S Rankin‡, Nhan Tran¶, Zhenbin Wu||, Qianfeng Shen*, Paul Chow*
* University of Toronto
† Columbia University
‡ Massachusetts Institute of Technology
§ CERN
¶ Fermilab
|| University of Illinois
Email: {naif.tarafdar, qianfeng.shen}@mail.utoronto.ca, giuseppe.diguglielmo@columbia.edu,
{pcharris, jkrupa, drankin}@mit.edu
{vladimir.loncar, zhenbin.wu}@cern.ch, ntran@fnal.gov, pc@eecg.toronto.edu

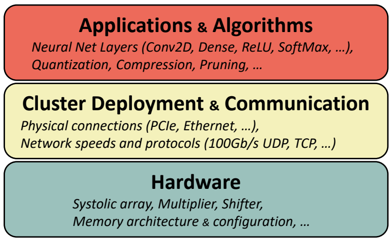
Applications & Algorithms
Neural Net Layers (Conv2D, Dense, ReLU, SoftMax, …),
Quantization, Compression, Pruning, …
Cluster Deployment & Communication
Physical connections (PCIe, Ethernet, …),
Network speeds and protocols (100Gb/s UDP, TCP, …)
Hardware
Systolic array, Multiplier, Shifter,
Memory architecture & configuration, …

Fig. 1. Abstraction Stack for common Machine Learning Frameworks.

Machine learning (ML) in the past decade has been one of the most popular topics of research within the computing community. Interest within the computing field ranges across all levels of the computation stack. We show this stack in Figure 1. This work introduces an open framework, called AIgean, to build and deploy machine learning (ML) algorithms on a heterogeneous cluster of devices (CPUs and FPGAs). Users can flexibly modify any layer of the machine learning stack in Figure 1 to suit their need. This allows both machine learning domain experts to focus on higher algorithmic layers, and distributed systems experts to create the communication layers below.

We leverage two open-source projects: Galapagos [3], for multi-FPGA deployment and `hls4ml` [1], for generating machine learning kernels synthesizable using Vivado HLS. We use particle detection in the physics domain to provide the first driving applications that help us to characterize the framework. To use AIgean, the user provides a machine learning algorithm and the resources of their cluster. Then AIgean converts the algorithm into appropriate IP cores and provides the off-chip communication between devices. `HLS4ml` was adapted to provide streaming interfaces. This fits the Galapagos model and works well with single inference. We designed a bridge from `hls4ml` to Galapagos to convert fixed-point streams into Galapagos streams that can be received from any compute kernel within our cluster.

We demonstrate the effectiveness of AIgean with two use cases: a small network running on a single network-connected FPGA and an autoencoder running on three FPGAs, and compare to SDAccel [4]. Our small neural-network single-FPGA implementation can implement a single inference in 0.08 ms as opposed to 2.9 ms in SDAccel, highlighting the efficacy of a network-connected accelerator for a single inference case. Our 3-FPGA autoencoder implementation performs a batch-size of 2400 inferences in 0.08 ms as opposed to 0.26 ms on a single FPGA in SDAccel, showing the need for multi-FPGA fabrics as it allows users to target large implementations of their machine learning circuitry, these implementations can perform better than smaller implementations. Multi-FPGA fabrics also make it possible to implement large networks such as ResNet-50, which is work in progress. Preliminary results before any optimizations have been applied shows that we can achieve a throughput of 200 images/s using 5 FPGAs, which can be compared to Brainwave, which has a throughput of 559 images/s [2]. We expect our results to improve significantly once we apply optimizations.

References

[1] J. Duarte, P. Harris, S. Hauck, B. Holzman, S.-C. Hsu, S. Jindariani, S. Kha, B. Krei, B. Le, M. Liu *et al.*, "FPGA-accelerated machine learning inference as a service for particle physics computing," *arXiv preprint arXiv:1904.08986*, 2019.

[2] J. Fowers *et al.*, "A Configurable Cloud-scale DNN Processor for Real-time AI," in *Proceedings of the 45th Annual International Symposium on Computer Architecture*, ser. ISCA '18. Piscataway, NJ, USA: IEEE Press, 2018, pp. 1–14. [Online]. Available: https://doi.org/10.1109/ISCA.2018.00012

[3] N. Tarafdar, N. Eskandari, V. Sharma, C. Lo, and P. Chow, "Galapagos: A full stack approach to FPGA integration in the cloud," *IEEE Micro*, vol. 38, no. 6, pp. 18–24, 2018.

[4] Xilinx Inc., "SDAccel Development Environment," https://www.xilinx.com/products/design-tools/software-zone/sdaccel.html, 2020.

2576-2621/20/$31.00 ©2020 IEEE
DOI 10.1109/FCCM48280.2020.00072