FPGA-accelerated Machine Learning Inference for LHC Trigger and Computing at SC 19

Philip Harris*[1] and Javier Duarte†[2]

[1]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA
[2]Department of Physics, University of California San Diego, La Jolla, CA

November 18, 2019

Abstract

UCSD and MIT will lead a group of collaborators demonstrating real-time FPGA-accelerated machine learning inference. Machine learning is used in many facets of LHC data processing including the reconstruction of energy deposited by particles in the detector. The training of a neural network for this purpose with real LHC data will be demonstrated. The model deployment and acceleration on a Xilinx Alveo card using a custom compiler called `hls4ml` will also be shown. An equivalent setup utilizing an NVIDIA GPU will also be presented allowing for direct comparison. This demonstration will serve to illustrate a first prototype of a new approach to the real-time triggering and event selection with LHC data aimed at meeting the challenges of the second phase of the LHC program, the High Luminosity LHC, which is planned to run from 2026-2037, following the development of further prototypes following this approach during the upcoming LHC data taking runs in 2021-2023.

The Large Hadron Collider (LHC) at the European Center for Nuclear Research (CERN) is the definitive instrument to perform high energy physics measurements. It represents both an engineering marvel and, with the discovery of the Higgs boson, an unrivaled scientific instrument [1, 2]. Over the past six years, the LHC has continued to reach new heights in our understanding of particle physics. At the time of the Higgs boson discovery, little was known about the properties of this new particle. With more data we have continued to improve our understanding. In fact, it was only recently that we were able to establish the Higgs boson's interaction with fermions [3, 4]. The LHC will continue to run for the next 15 years and will bring a wealth of particle physics data that will cover many areas including the Higgs boson and dark matter [5, 6, 7, 8, 9, 10, 11]. However, there is concern amongst the research community that continual running will yield diminishing returns. There are no energy increases expected for the LHC's future, and the scheduled upgrades of the LHC detectors aim to maintain, but not improve performance.

To enhance the sensitivity of the acquired data to new physics, we aim to improve the data acquisition chain. At a data rate of over 50 terabytes per second, the LHC has the largest data rate in the world. Processing of this data is done in a set of tiers. With each successive tier, data is filtered, selecting only a fraction of the initial collision rate, so that more information can be used to process the event and determine if the event is sufficiently interesting to save to disk. In each tier, we know that many high quality events that could be used to enhance physics measurements are being thrown out. Here, we aim to recover these events by targeting the second tier, known as the high level trigger (HLT) to investigate the level of speed up of reconstruction by assessing the reconstruction improvement of heterogeneous computing. This system currently consists of a CPU based system consisting of order 10,000 cores located on site at the LHC. As a proof of concept, P. Harris, J. Duarte, and collaborators have shown a speed-up of $\times 200$ in latency for inference of ResNet-50 with a Microsoft Brainwave FPGA service over a CPU [12].

*pcharris@mit.edu
†jduarte@ucsd.edu

We propose to translate algorithms involved in reconstruction at the HLT to ML-based algorithms and then to port these algorithms to FPGAs, GPUs. ML algorithms have long been a mainstay in high energy physics. However, new algorithms are not adopted during the reconstruction chain due to the computing limitations and the relative difficulty of writing specialized time limited algorithms. Due to recent developments, the porting of NN based algorithms to GPU and FPGAs has become relatively simple leading to significant speeds due to the large parallelization of these algorithms. The quick translations allow for testing and deployment in heterogeneous (GPU/FPGA/ASIC) systems on short timescales, which in turn could lead to a more significant proposal for alternative computing model at the LHC.

To illustrate how effective an ML accelerator is for low latency processes, we have identified an algorithm that is projected to be one of the leading limitations in the reconstruction of the high-level trigger (the second stage of filtering): the reconstruction of signals in the hadron calorimeter of the Compact Muon Solenoid (CMS) detector [13]. We first rewrote the existing algorithm, which is non-ML based, using a multilayer perceptron regression neural network. Then we ported this algorithm to an FPGA using `hls4ml` [14], a translation tool to convert machine learning models in common frameworks into firmware using High-Level Synthesis, developed by P. Harris (MIT), J. Duarte (UCSD), and collaborators at Fermilab, CERN, and UIC. With the current algorithm, we find a throughput of roughly 3 ms for approximately $16,000$ inferences (corresponding to 1 full LHC event) using a GPU, and 2 ms using an FPGA. Additionally, we have executed this algorithm on a GPU using the existing tools within `TensorFlow`. When going from another machine in the same cluster, the overall latency increases by 10 ms without a change in throughput. On a Xilinx Alveo U250 FPGA, we are still characterizing the ultimate achievable throughput and latency. However, we have observed in previous tests that similar latency and throughput rates can be obtained.

The GPU-based algorithm is now being run as a service (aaS) at 3 locations across the US: at UCSD with a 7 GPU machine, at MIT with a 1 GPU machine, and at Fermilab with a 3 GPU machine. Additionally, local service is being run at SC19. Two FPGA based services are also operational utilizing Xilinx Alveo cards, one at Fermilab and one at UCSD. Both the GPU and FPGA algorithms are served by the same machine.

In this demonstration, we will run an emulation of real LHC data taking by sending roughly $16,000$ inference requests at a rate of approximately 1 kHz. The maximum permissible average latency for the high-level trigger is 200 ms for the full system. As such, we target latency of less than about 10 ms for each set of $16,000$ inferences.

References

[1] ATLAS collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett. B* **716** (2012) 1 [`1207.7214`].

[2] CMS collaboration, *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC*, *Phys. Lett. B* **716** (2012) 30 [`1207.7235`].

[3] CMS collaboration, *Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector*, *Phys. Lett. B* **779** (2018) 283 [`1708.00373`].

[4] CMS collaboration, *Observation of Higgs boson decay to bottom quarks*, *Phys. Rev. Lett.* **121** (2018) 121801 [`1808.08242`].

[5] M. R. Buckley, D. Feld and D. Goncalves, *Scalar Simplified Models for Dark Matter*, *Phys. Rev.* **D91** (2015) 015017 [`1410.6497`].

[6] P. Harris, V. V. Khoze, M. Spannowsky and C. Williams, *Constraining Dark Sectors at Colliders: Beyond the Effective Theory Approach*, *Phys. Rev. D* **91** (2015) 055009 [`1411.0535`].

[7] U. Haisch and E. Re, *Simplified dark matter top-quark interactions at the LHC*, *JHEP* **06** (2015) 078 [`1503.00691`].

[8] CMS collaboration, *Search for dijet resonances in proton–proton collisions at $\sqrt{s} = 13$ TeV and constraints on dark matter and other models*, *Phys. Lett. B* **769** (2017) 520 [`1611.03568`].

[9] CMS collaboration, *Search for narrow and broad dijet resonances in proton-proton collisions at $\sqrt{s} = 13$ TeV and constraints on dark matter mediators and other new particles*, *JHEP* **08** (2018) 130 [1806.00843].

[10] CMS collaboration, *Search for dark matter produced with an energetic jet or a hadronically decaying W or Z boson at $\sqrt{s} = 13$ TeV*, *JHEP* **07** (2017) 014 [1703.01651].

[11] CMS collaboration, *Search for new physics in final states with an energetic jet or a hadronically decaying W or Z boson and transverse momentum imbalance at $\sqrt{s} = 13$ TeV*, *Phys. Rev. D* **97** (2018) 092005 [1712.02345].

[12] J. Duarte et al., *FPGA-accelerated machine learning inference as a service for particle physics computing*, *Comput. Softw. Big Sci.* **3** (2019) 13 [1904.08986].

[13] CMS collaboration, *HCAL Out Of Time Pileup Subtraction and Energy Reconstruction*, .

[14] J. Duarte et al., *Fast inference of deep neural networks in FPGAs for particle physics*, *JINST* **13** (2018) P07027 [1804.06913].