Machine Learning acceleration in the Global Event Processor of the ATLAS Trigger Update

Zhixing “Ethan” Jiang (University of Washington) on behalf of ATLAS global event trigger group

BACKGROUND

- ATLAS HL-LHC undergoing trigger system upgrade
- The upgrading Global trigger subsystem is FPGA based
- Two tools, HLS4ML and fwX are being used to generate new algorithm with super low latency

HLS4ML & fwX

- HLS4ML and fwX are tools for generating ultra-low-latency models
- HLS4ML and fwX use high-level synthesis (HLS) to convert the model
- HLS4ML support CNN, DNN, RNN, Transformer, GNN;
- fwX support BDT

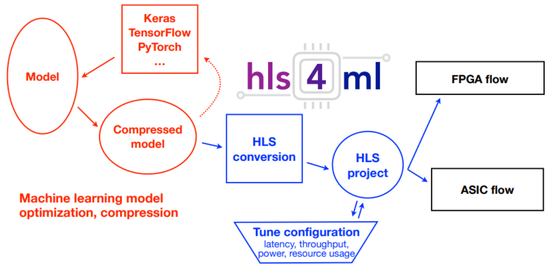


Fig1. The workflow of hls4ml

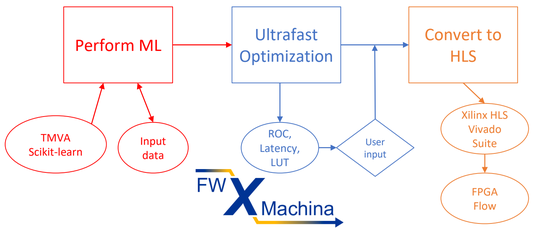


Fig2. The workflow of fwX

Architecture of the Framework

Framework of LHC upgrade:

- Trigger algorithms deployed in Global Event Processors (GEP)
- Data sent to GEP at each Bunch Crossing
- GEP processes data using ML algorithm in pipelining
- Hard requirement for latency < 1.5 us

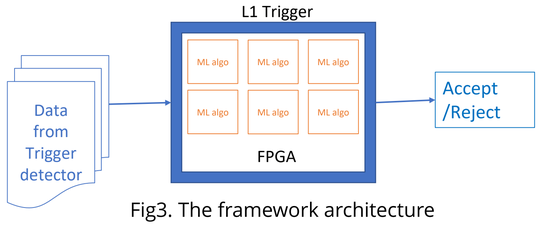


Fig3. The framework architecture

Frame Testing:

- Build a test vehicle for testing
- Similar structure as the Framework but smaller
- Deployed on the physical board for testing
- The Design has been tested on the FPGA VCU118

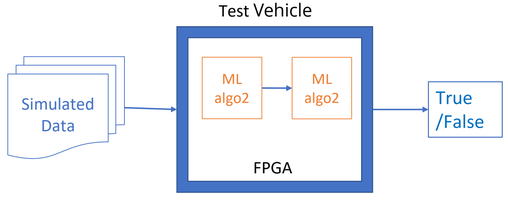


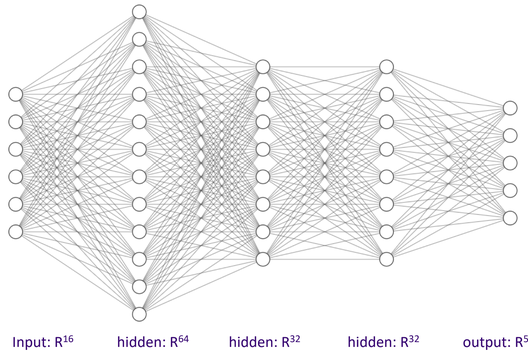
Fig4. The architecture of the test Vehicle

Deployment of ML models into the framework

Dense model using HLS4ML

A Dense Layer model used for solving the Jet-tagging classification problem

- Model structure:



Input: R^{16} hidden: R^{64} hidden: R^{32} hidden: R^{32} output: R^{5}

- Model resource / latency:

Resource	Utilization	Utilization %
DSP	2784	22
FF	13375	~0
LUT	109060	6
BRAM	8	~0
Latency	12 cycles	60.00 ns

BDT model using fwX

A Boost Decision Tree model to discriminate between VBF Higgs and multi-jet events

- Model Hyperparameter:

Configuration	Optimized
Bin Engine	LUBE
# of input variables	5
Input bit width	8
Cut threshold bit width	8
Output score bit width	16
Max. depth	4
# of training trees	100
# of final trees	100
Cut eraser, threshold	Yes, 5%
# of bins	40k

Fig5: The FPGA for testing design

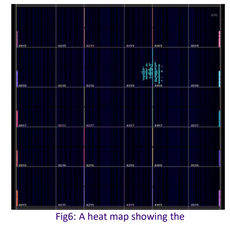


Fig6: A heat map showing the resource of the FPGA

- Model resource / latency:

Resource	Utilization	Utilization %
DSP	2	~0
FF	35748	1
LUT	71421	4
BRAM	6	~0
Latency	7 cycles	21.875 ns

Summary

In this research, we proofed using the machine learning model generated by hls4ml to design the algorithm in APU would be practical and suitable. We proofed it by:

1. Discover the structure of the APU
2. Discover the structure of the hls4ml & FwX model
3. Deploy the model into APU
4. Testing the result of the ML APU.

Reference

J. Duarte and S. Han and P. Harris and S. Jindariani and E. Kreinar and B. Kreis and J. Ngadiuba and M. Pierini and R. Rivera and N. Tran and Z. Wu, Fast inference of deep neural networks in FPGAs for particle physics, Journal of Instrumentation, doi: 10.1088/1748-0221/13/07/p07027

T.M. Hong and B.T. Carlson and B.R. Eubanks and S.T. Racz and S.T. Roche and J. Stelzer and D.C. Stumpp, Nanosecond machine learning event classification with boosted decision trees in FPGA for high energy physics, Journal of Instrumentation doi:10.1088/1748-0221/16/08/p08016

Global Trigger Community (2021) ATLAS TDAQ Phase-II Upgrade: Firmware Specifications for the Global Trigger. ATL-COM-DAQ-2021-098