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

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

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

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:

Model resource / latency:

Resource Utilization Utilization %
DSP 2784 22
FF 13375 ~0
LUT 109560 6
BRAM 8 ~0

Latency 12 cycles 60.00 ns

Summary
In this research, we proved using the machine learning model generated by HLS4ML to design the algorithm in APU would be practical and suitable. We proved 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
Global Trigger Community (2021) ATLAS TDAQ Phase-II Upgrade: Firmware Specifications for the Global Trigger. ATLAS-TDAQ-2021-018