Fast inference on FPGA for the ATLAS Muon Trigger

CHEP 2023 – Norfolk Waterside Marriott – 8-12 May 2023

Maria Carnesale (Sapienza Università di Roma)

From Simulation to FPGA Implementation

Toy model: detector with 3 station immersed in a 1 T magnetic field
Single muon events: hit rate expected in the inner station of ATLAS Muon Spectrometer end-cap (New Small Wheel) at HL-LHC

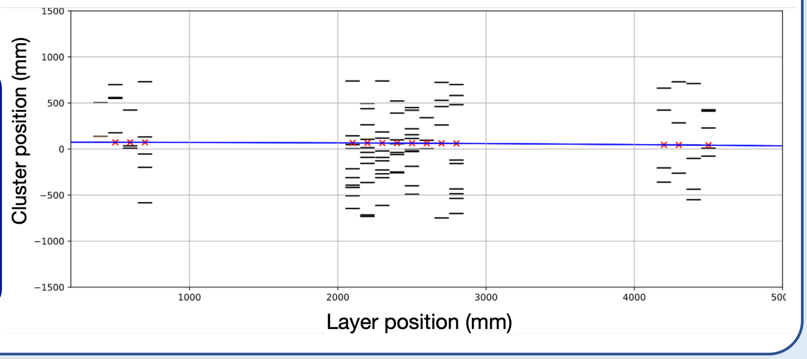
Neural networks applied to

Cluster reconstruction

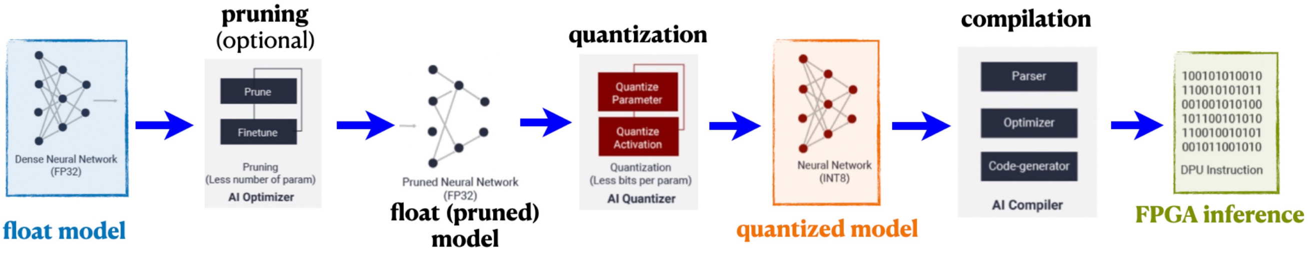
DNN trained to identify cluster produced by muons in Micromegas and sTGC detectors

Pattern recognition

RNN/CNN trained to identify tracks in events with high occupancy
RNN models not supported for FPGA inference, only CNN are tested



Model inferred in FPGA using Vitis-AI Flow (Xilinx)


pruning (optional)

quantization

compilation

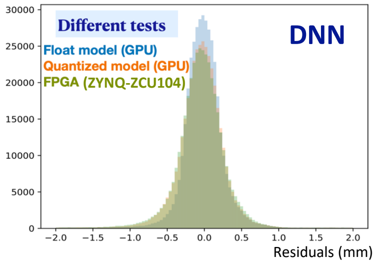
Advantages of FPGA : very fast - low energy consumption
Using Xilinx [1] FPGA architectures: U50/U250/ZYNQ

Xilinx Vitis-AI [2]: platform provides development environment for deploying deep learning models on FPGAs

Deployment on Xilinx U50 - U250 – ZYNQ ZCU104

DNN for cluster reconstruction

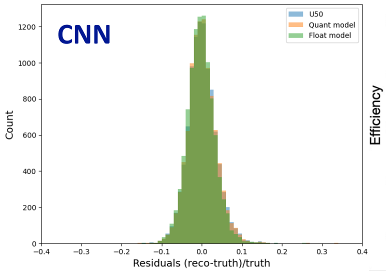
Dense neural network trained to reconstruct the hit position of the track

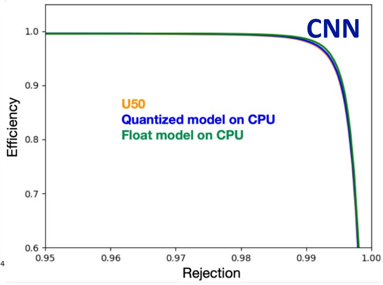


Pattern recognition performance

CNN pre-processing step to built images (to be estimated)

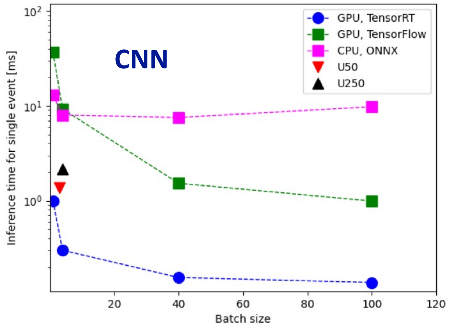
Not ideal for trigger existing algs working directly on sparse data (RNN)





Timing studies

CPU: using ONNX [3] tool
GPU: NVidia RTX A5000 board with 24GB of GDDR6 memory, using TensorRT [4]
U50 - **U250**: after quantization and compilation steps in Vitis-AI workflow



Conclusions

- When running on FPGAs, performance are similar to the float model
- FPGA timing comparable to GPU with TensorRT software

[1] https://www.xilinx.com, [2] https://www.xilinx.com/products/design-tools/vitis/vitis-ai.html, [3] https://onnx.ai, [4] https://developer.nvidia.com/tensorrt