Fast inference on FPGA for the ATLAS Muon Trigger

156th LHCC meeting – CERN – 27th November 2023

M. Carnesale, F. A. Di Bello, F. Giuli, S. Rosati, S. Veneziano

From Simulation to FPGA Implementation

Toy model: detector with 3 stations immersed in a 1 T magnetic field
Single muon events: 2, 5, 10 and 15 kHz/cm^2 hits rate expected in the inner station of ATLAS Muon Spectrometer end-cap (New Small Wheel) at HL-LHC [1]

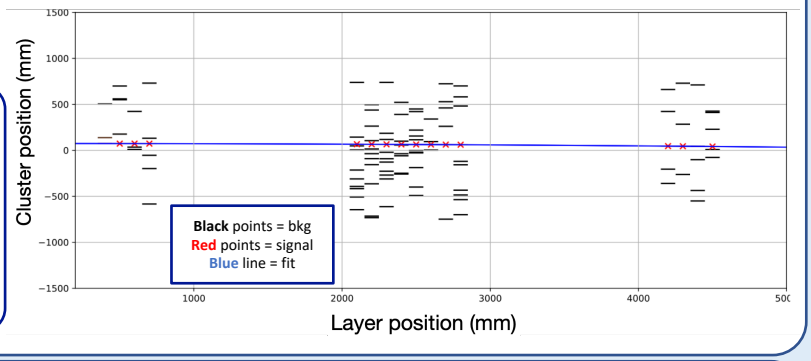
Neural networks applied to

Cluster reconstruction

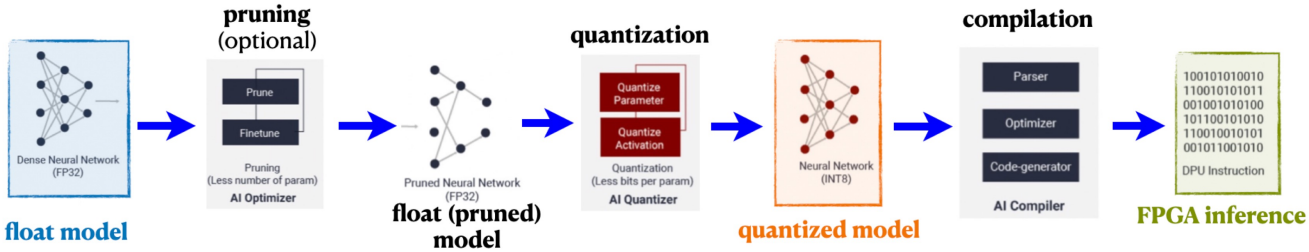
DNN trained to identify clusters produced by muons in Micromegas and sTGC detectors

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)



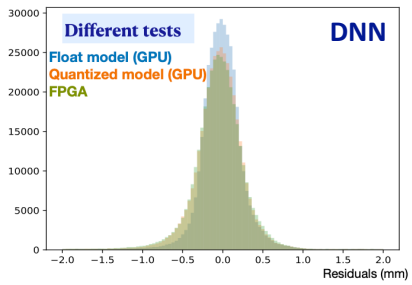
Advantages of FPGA : very fast - low energy consumption
Using Xilinx [2] FPGA architectures: U50/U250/Versal

Xilinx Vitis-AI [3]: platform provides development environment for deploying deep learning models on FPGAs

Deployment on Xilinx U50, U250 and Versal VCK5000

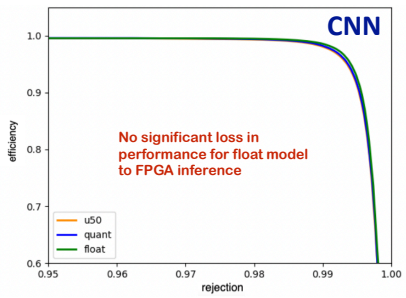
DNN for cluster reconstruction

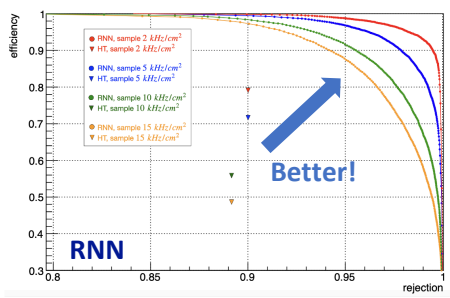
Dense neural network trained to reconstruct the hit position of the track



CNN/RNN for pattern recognition performance

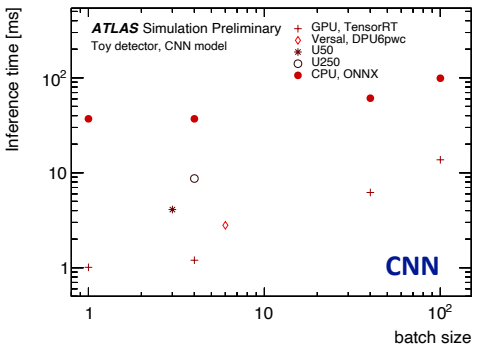
CNN not ideal for existing trigger algorithms working directly on sparse data (RNN)





Timing studies

- ● : using ONNX [4] tool
- + : NVidia RTX A5000 board with 24GB of GDDR6 memory, using TensorRT [5]
- ✳ ○ : use quantization and compilation steps in Vitis-AI workflow
- ◊ : with DPU6pwc



Conclusions

- When running on FPGAs, similar performances to the float model
- Overall CPU already meets the requirement imposed by the HLT latency – **O(1 ms)**
- CPU load to be studied, as well as the power dissipations

[1] CERN-LHCC-2013-006, [2] https://www.xilinx.com, [3] https://www.xilinx.com/products/design-tools/vitis/vitis-ai.html, [4] https://onnx.ai, [5] https://developer.nvidia.com/tensorrt