Studies on track finding algorithms based on machine learning with GPU and FPGA

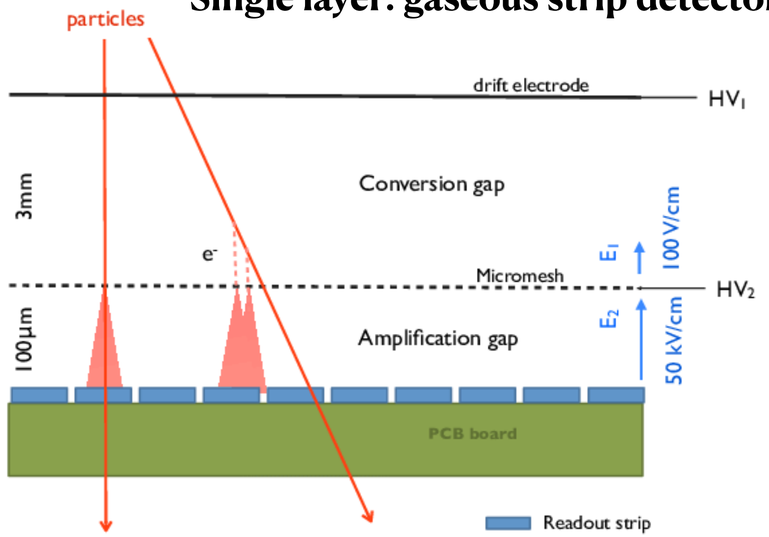
Maria Carnesale

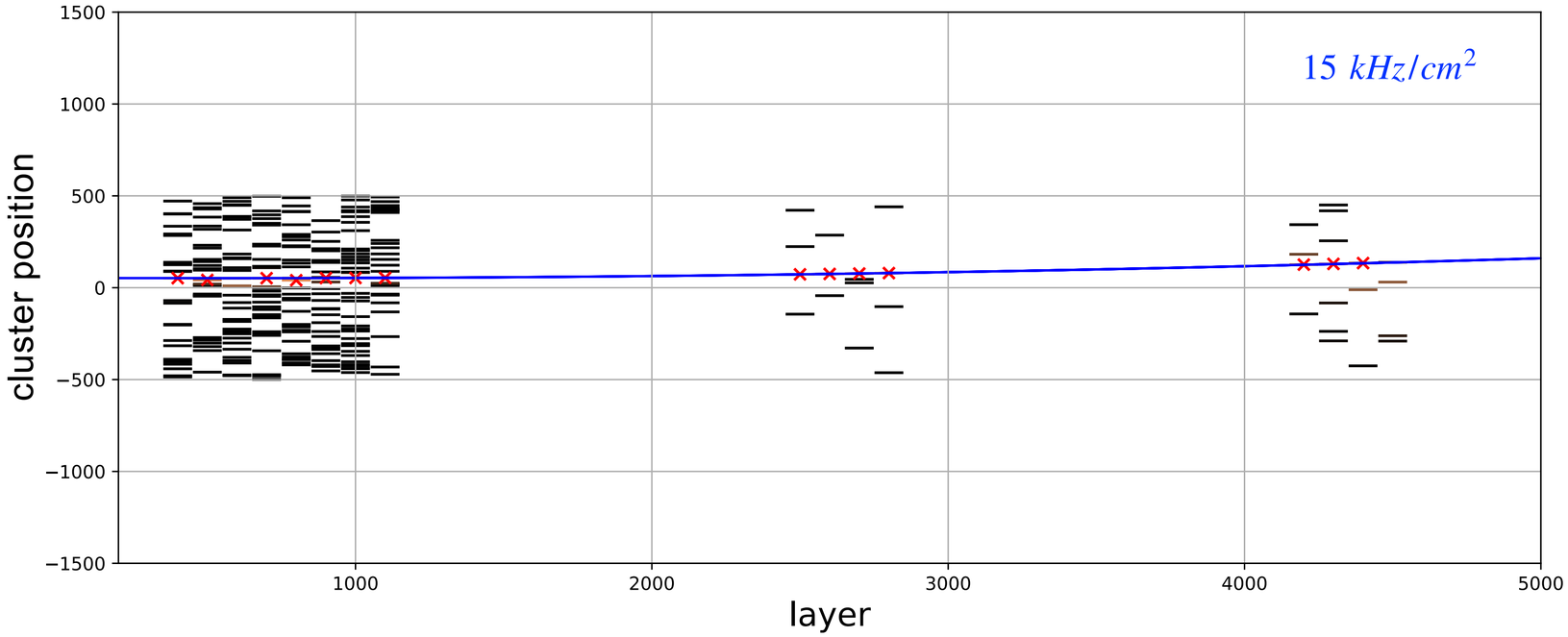
EuCAIFCon 2024 – Amsterdam – 30 Apr- 3 May 2024

ML algorithms for muon pattern recognition

- **Algorithms for cluster reconstruction and pattern recognition in gaseous strip detectors**
- Models tested are **Dense NN (DNN) and Convolutional NN (CNN)**
  - DNN trained to identify clusters produced by muons in gaseous strip detectors
  - RNN/CNN trained to identify tracks in events with high occupancy

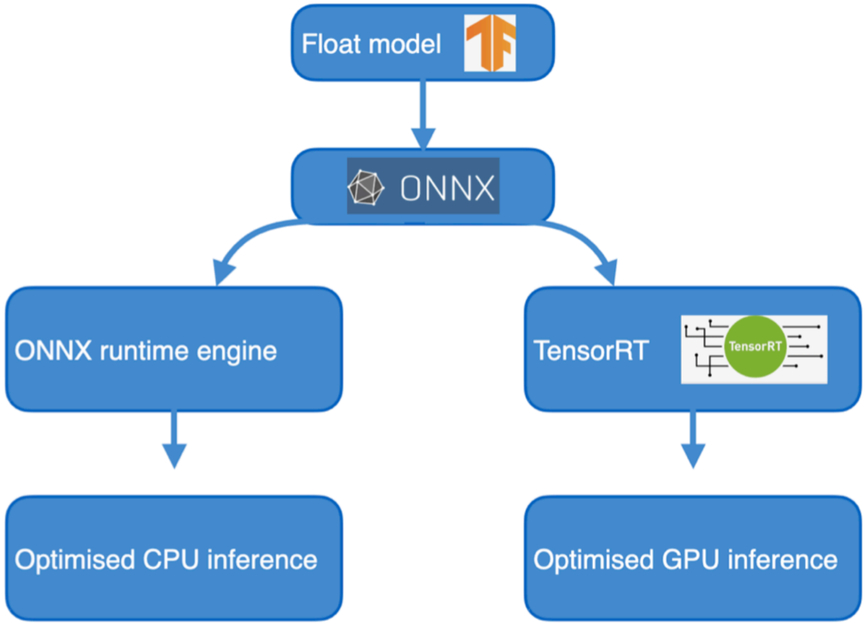
Single layer: gaseous strip detector



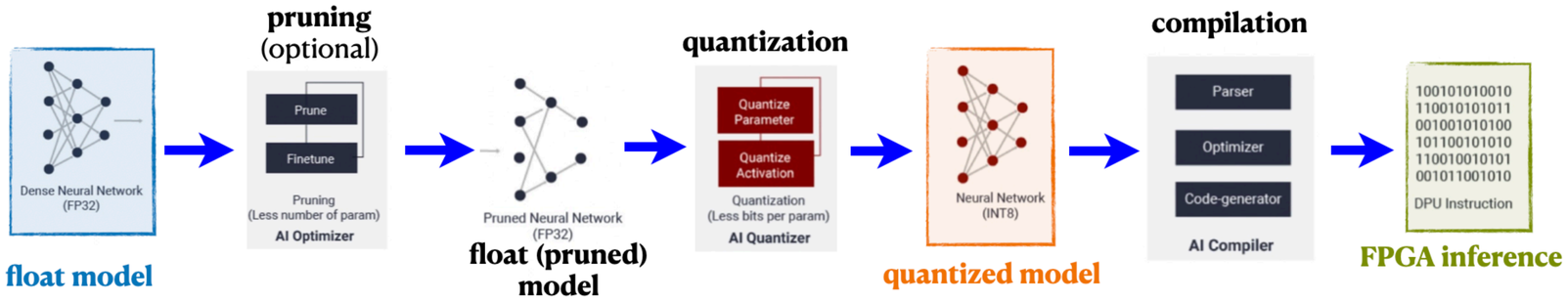


ML algorithms tested on CPU/GPU/FPGA

- **Study of inference time and performance on different architectures:**
  - **CPU**: using ONNX
    - Open Neural Network Exchange: open source framework that optimizes the usage of CPU resources
  - **GPU**: using tensor flow and tensorRT
    - Framework produced by NVIDIA to run optimized inference on GPU

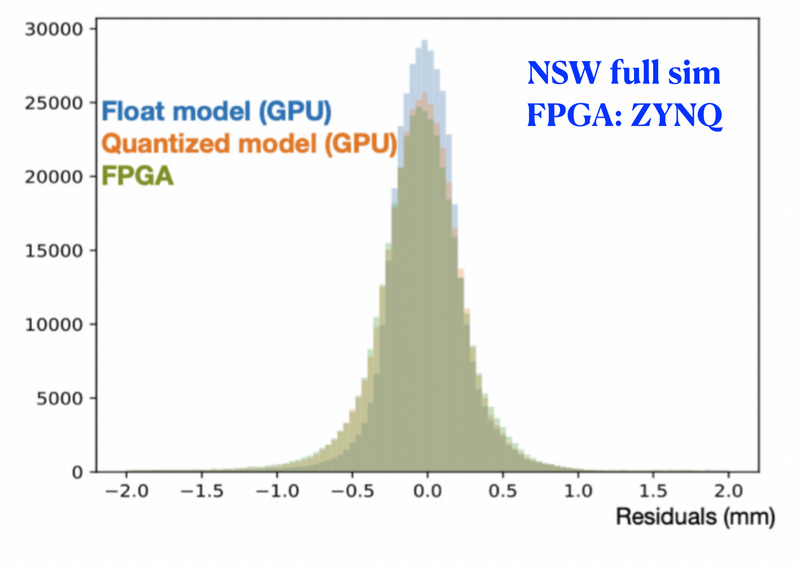


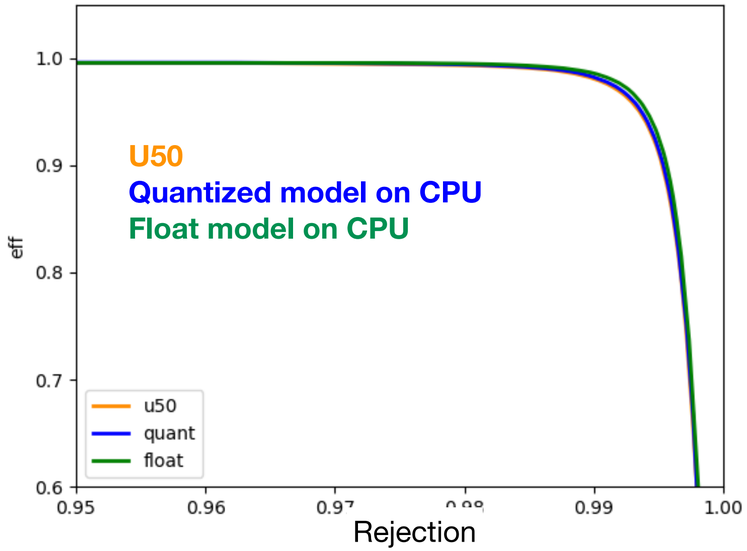
- **FPGA**: using Vitis-AI workflow provided by Xilinx for inference acceleration or HLS4ML and vivado

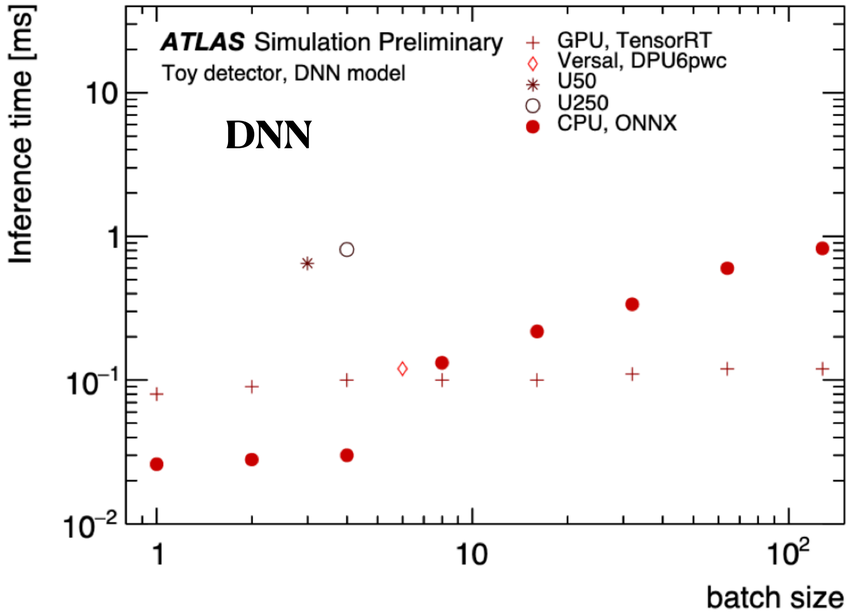


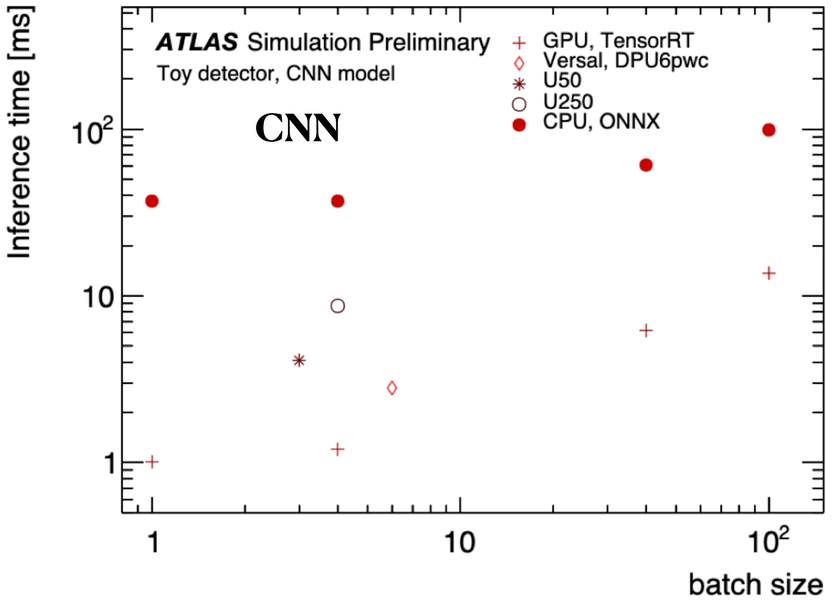
Timing and performance on CPU/GPU/FPGA

- Comparing CPU (ONNX) / GPU (TensorRT) / FPGA (Vitis AI) for DNN and CNN inference
- Small resolution degradation after quantisation
- Same performance of quantised model on CPU/GPU and FPGA









- Batch size (number of events processed in parallel) is fixed in the case of FPGA, free for CPU/GPU
- FPGA times are not a simulation