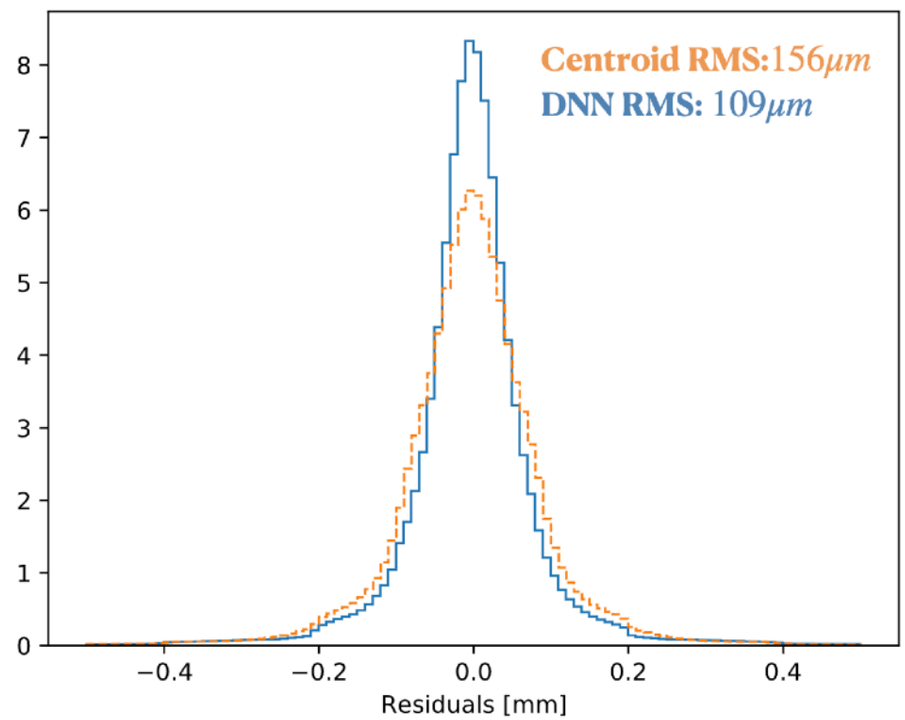
Studies on FPGA inference for tracking

Maria Carnesale, Francesco Di Bello, Francesco Giuli
Stefano Rosati, Francesco Safai Tehrani, Stefano Veneziano

CERN / Genova / Roma

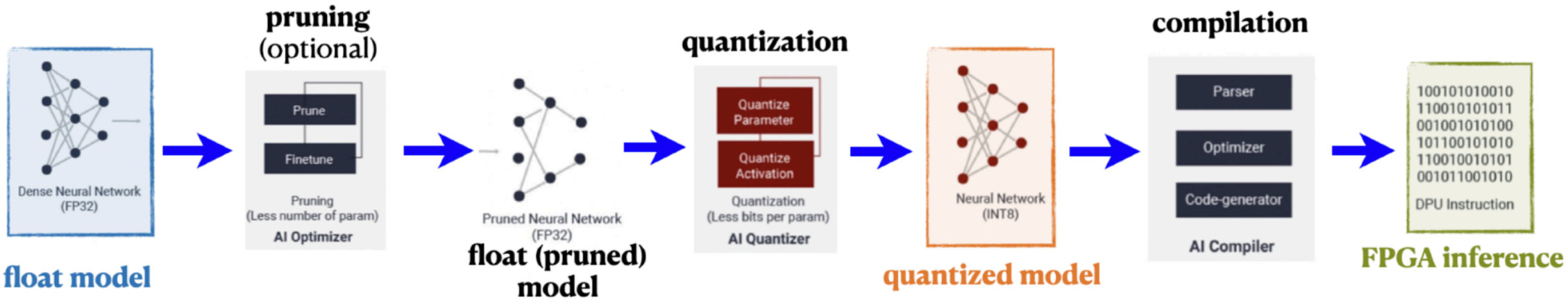
Introduction

- Work started in the framework of muon offline reconstruction and high-level trigger studies for phase-II ATLAS upgrades
- Muon track reconstruction in the high density environment of Hi-Lumi LHC
- Studying machine-learning based algorithms for:
  - Clusters position reconstruction in muon strip detectors (Micromegas, STGC)
    - DNN with strip-level quantities (position, charge, time)
  - Pattern recognition and tracking
    - RNN (LSTM) or CNN with cluster / strip level quantities
- E.g. optimize the resolution of strip detectors
  - Correct for charge sharing, pitch-dependent, effects
  - Take into account showers close to high-momentum muons and other backgrounds
- Study of ML models inference on FPGA, for application to high-level trigger tracking
- Models development in Tensorflow/Keras, inference via Tf, ONNX (CPU/GPU), TensorRT (GPU), Xilinx Vitis-AI, Mipsology (FPGA)

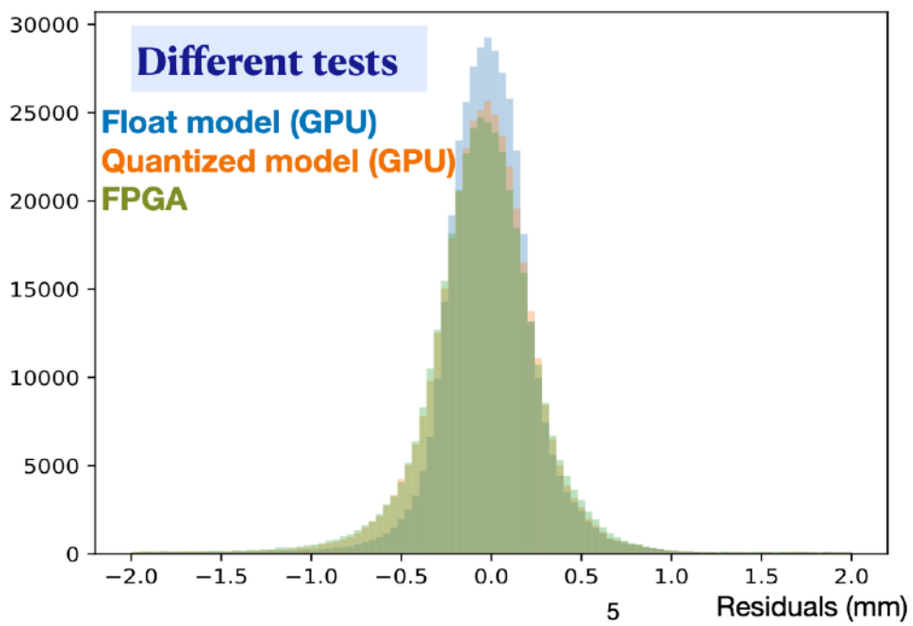


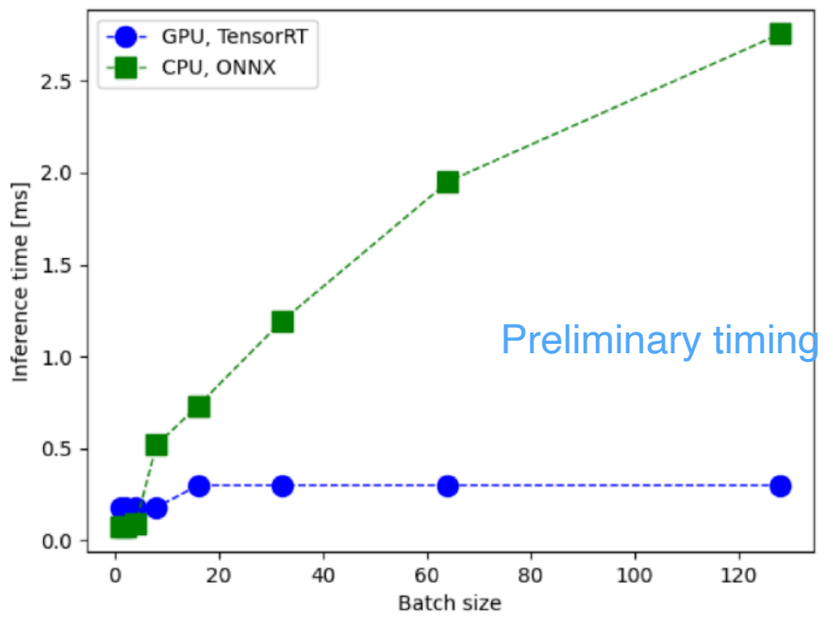
FPGA inference of clusters position

- DNN FPGA inference via Vitis-AI inference acceleration (Xilinx)
  - Model optimization (remove less important nodes and connections) -> not used yet
  - Quantization (from float to 8-bits integers)
  - Compilation specific for each Data Processing Unit (DPU) type



Resolution
of the quantized model
compatible with the
float implementation

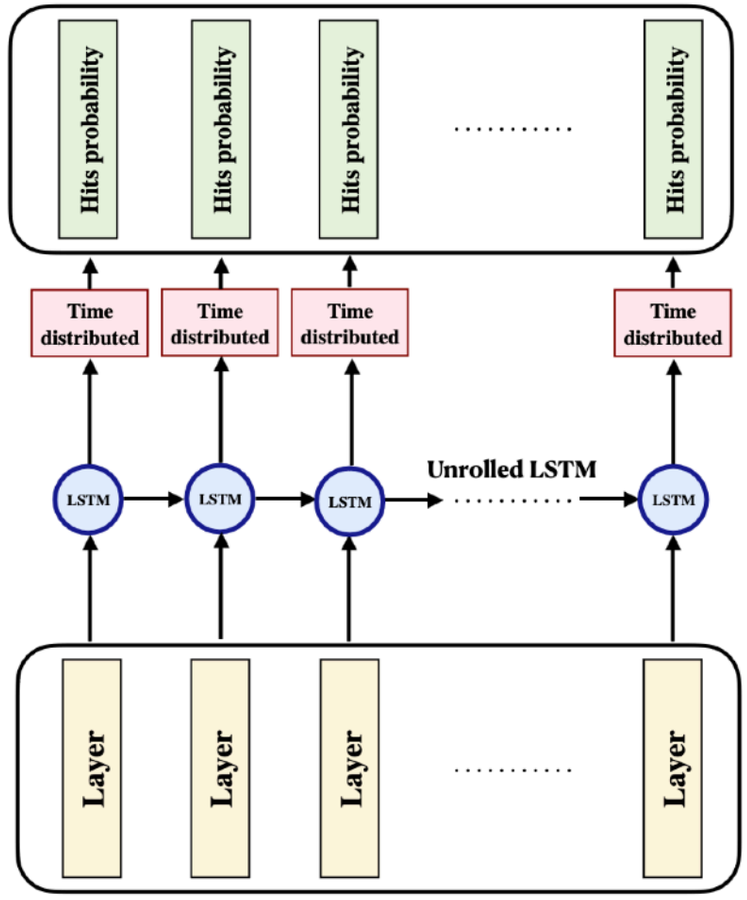


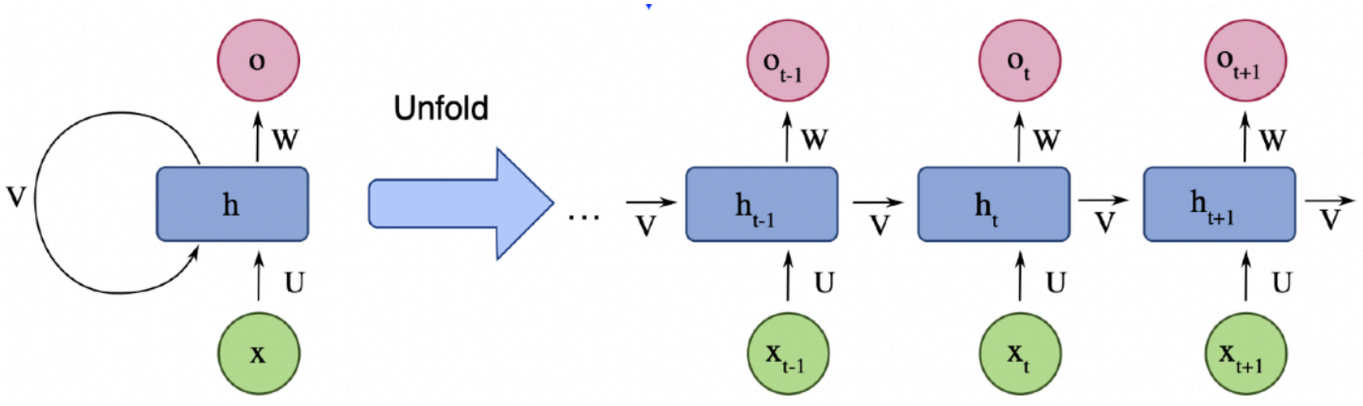


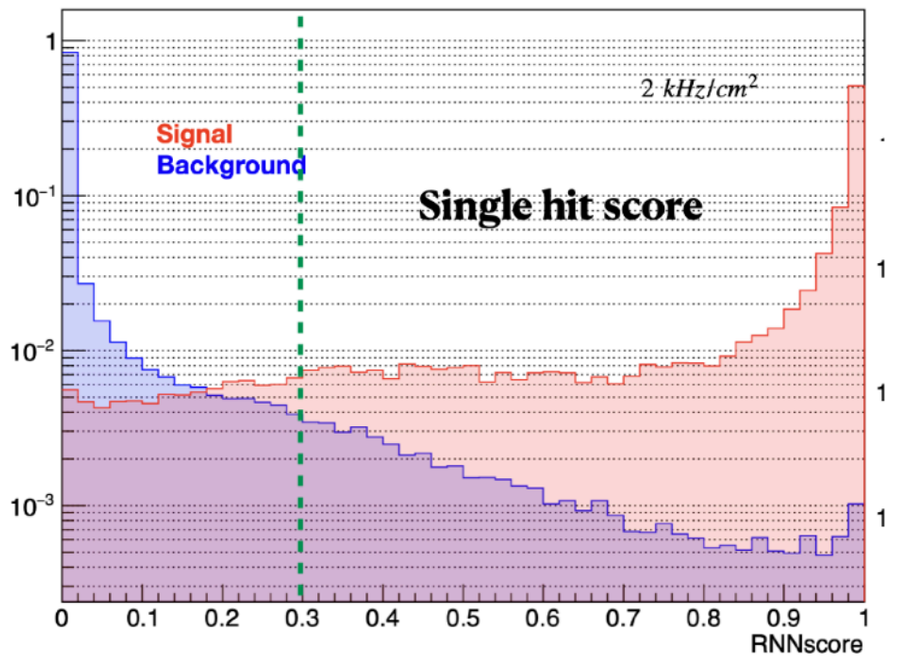
Preliminary timing tests

Pattern recognition in high-density environments

- Second step of the track reconstruction is the pattern recognition
  - Identify hits belonging to tracks, background hits and holes (inefficient layers)
- Recurrent Neural Network (RNN) based on LSTM (Long-Short Term Memory) nodes
  - Typically used for text, speech and sound recognition
  - Work directly on sparse data (no building of large images needed)
  - In the case of pattern reco, the time structure is given by the tracking layers
- As RNN are for the moment not supported by Xilinx, also tested a CNN (image processing)

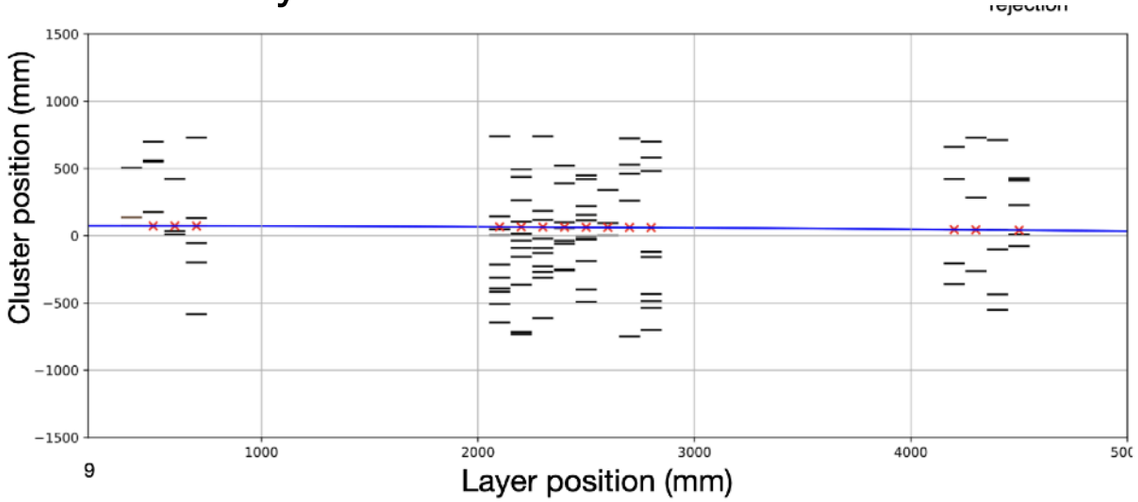


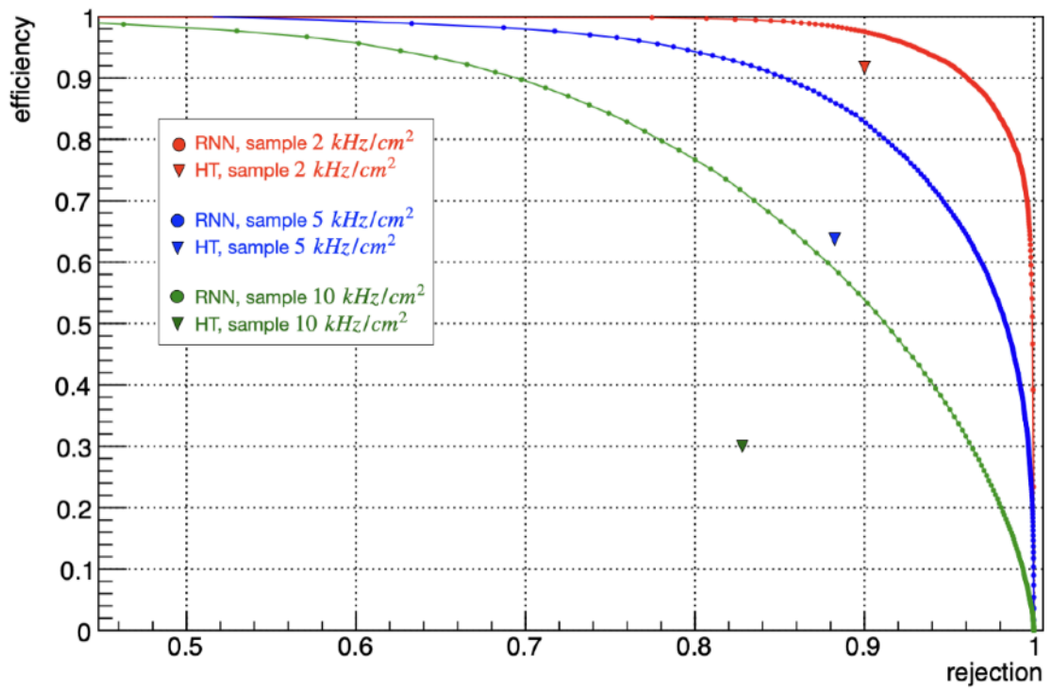




Pattern recognition

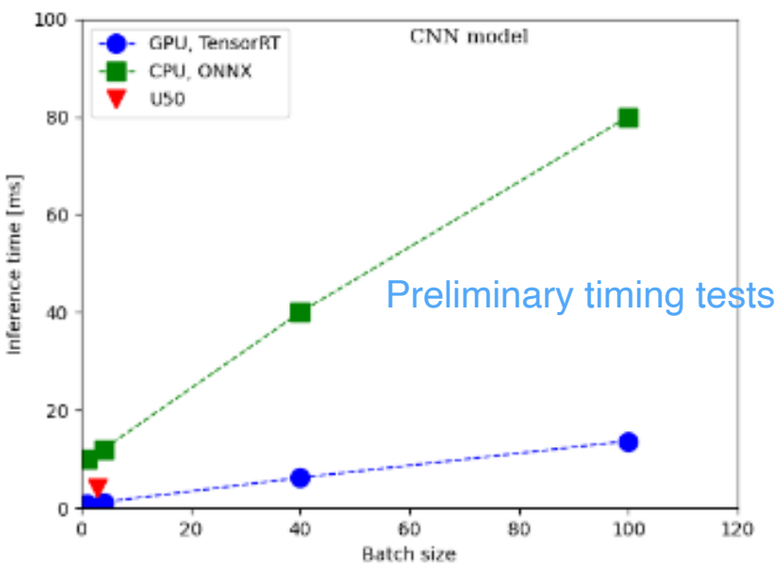
Performance depends on detector structures and background levels -> various situations under study

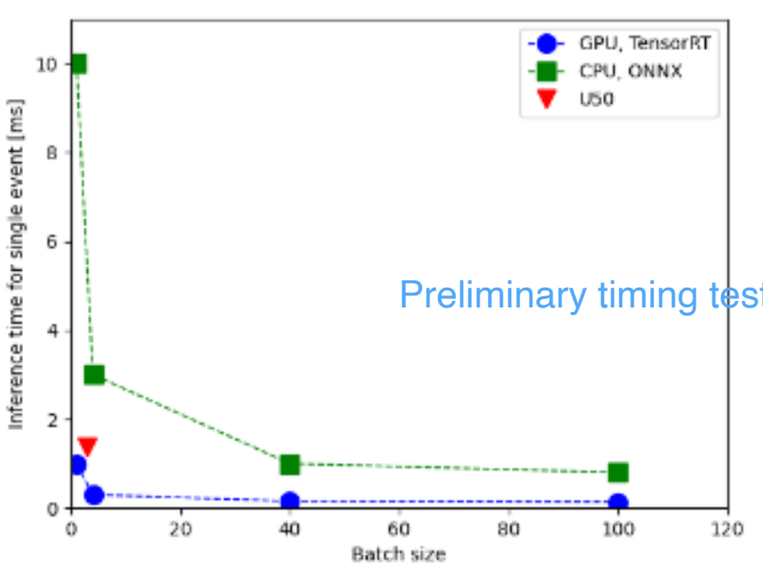


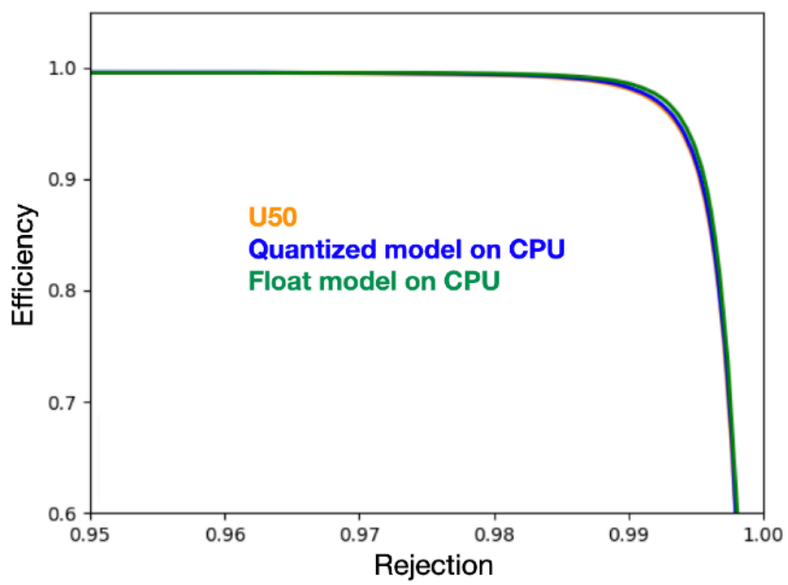


We'll start testing the pattern reco RNN (on CPU) during the ongoing LHC run-3

- CNN (tested with lower backgrounds) inference tests (U50 ST, CPU/GPU MT)







Testbeds

- At CERN (ATLAS TDAQ):
  - Server with AMD Epyc 7302 processor (2.9 GHz, 128 GB sys memory), with GPU NVidia RTX A5000 (24 GB GDDR6 memory)
  - DPU: Xilinx Alveo U50, Xilinx Alveo U250, Versal VCK5000

- In Rome (recently installed in the ATLAS lab 2nd floor):
  - Dual CPU Intel Xeon Gold 6346 (3.10 GHz, 128 GB sys memory), with GPU NVidia A100 (80 GB)
  - DPU's to be added