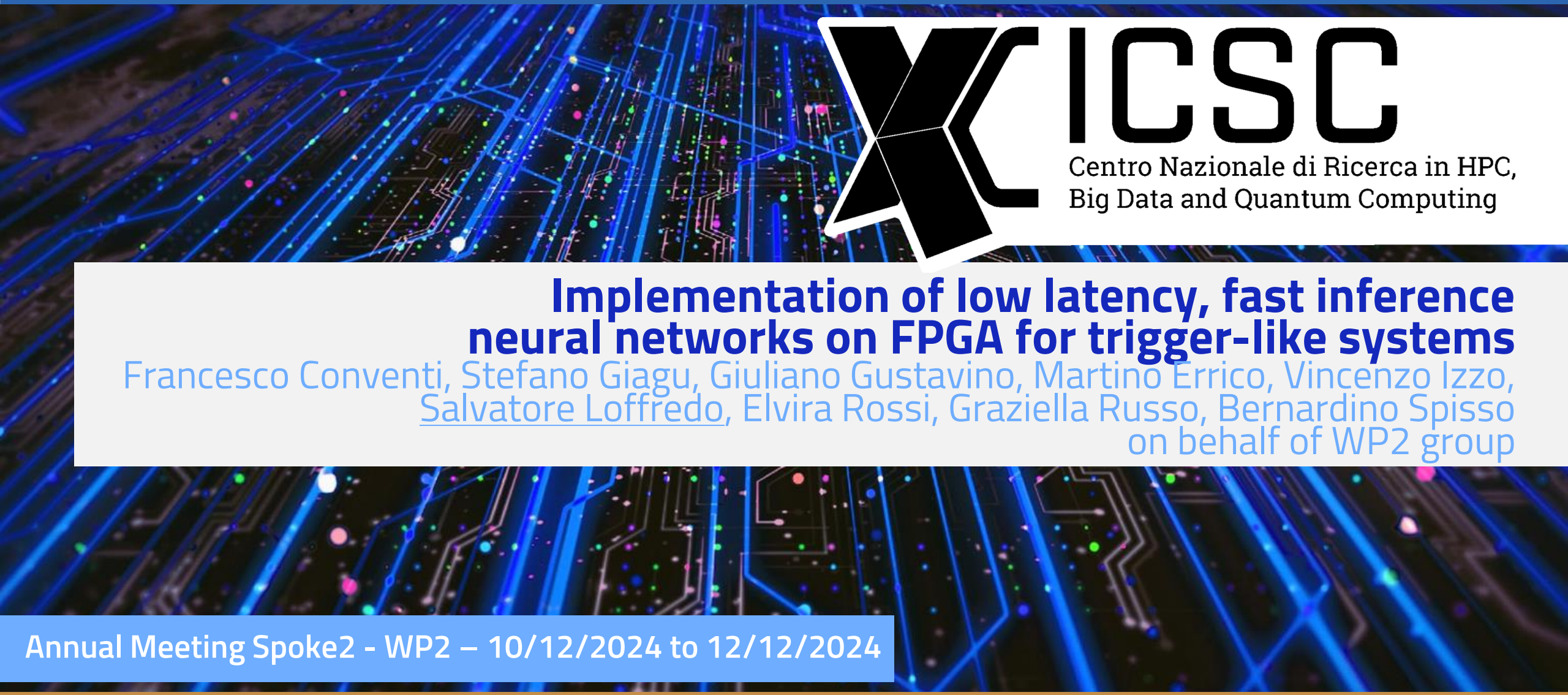
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing



Implementation of low latency, fast inference neural networks on FPGA for trigger-like systems

Francesco Conventi, Stefano Giagu, Giuliano Gustavino, Martino Errico, Vincenzo Izzo, Salvatore Loffredo, Elvira Rossi, Graziella Russo, Bernardino Spisso
on behalf of WP2 group

Annual Meeting Spoke2 - WP2 – 10/12/2024 to 12/12/2024

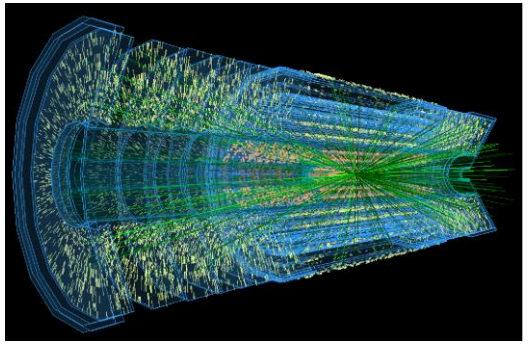
High Level and Low Level trigger systems for HEP Experiments

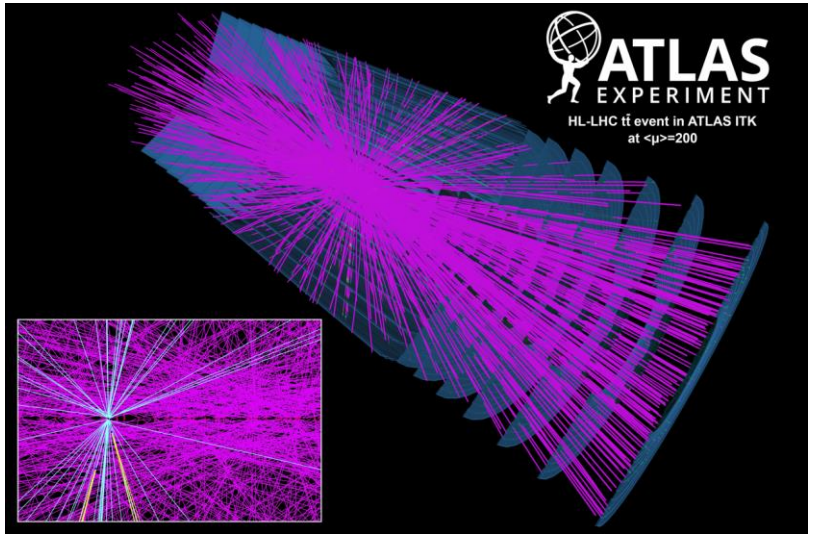
Goal: to provide, by means of Neural Networks, the High Energy Physics Experiments, such as ATLAS (A Toroidal LHC Apparatus), High Level and Level-0 trigger systems capable to sustain high rate and predictable, fixed, low latency.

Solution being tested: the use of fast Convolutional and Graph Neural Networks on FPGA (Field Programmable Gate Array).

Outline of the talk

- ✓ Neural Networks in HEP trigger
- ✓ Lightweight Neural Networks
- ✓ Use case: the L0 Muon Barrel Trigger of the ATLAS experiment for HL-LHC
- ✓ Development pipeline
- ✓ Conclusions and outlook

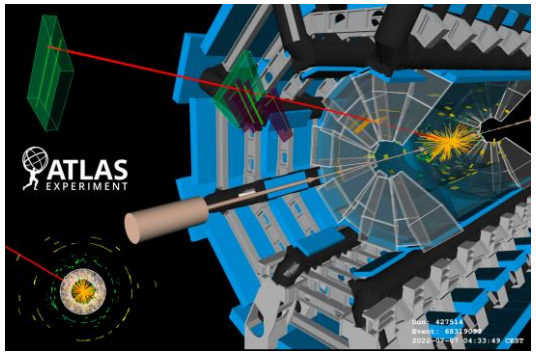




Neural networks for the trigger of the HEP experiments

HEP trigger systems have stringent requirements in terms of rate and latency:

Detector collisions



40 MHz

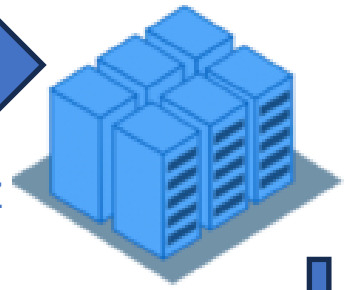
L1 trigger

100ns -1us

Low latency **AI inference**

100 kHz

High-Level Trigger

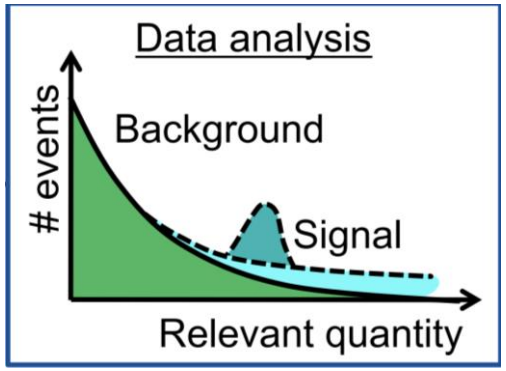


1ms -1s

High throughput **AI inference**

1 kHz

Offline



To fulfil the needs of the triggers that are in development for the upgraded LHC detectors, such as ATLAS, we are currently developing different architectures based on Convolutional and Graph Neural Networks on the latest generation of hardware accelerators (FPGAs, GPUs).

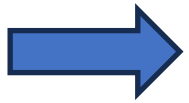
Lightweight Neural Networks

The target is to build fast inference, low latency NNs.

The **NN networks** dimensions represent a bottleneck for the trigger requirements; so, the solution is to use multi-stage compression approach (pruning and quantization) to reduce the NNs resource utilization.

Optimized hardware for efficient and fast inference is needed.

Furthermore, NNs architecture needs to have good performance in terms of power budget respect to GPU implementations.

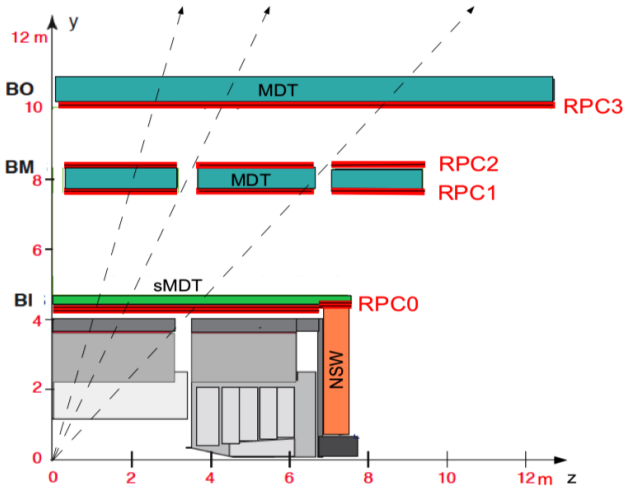


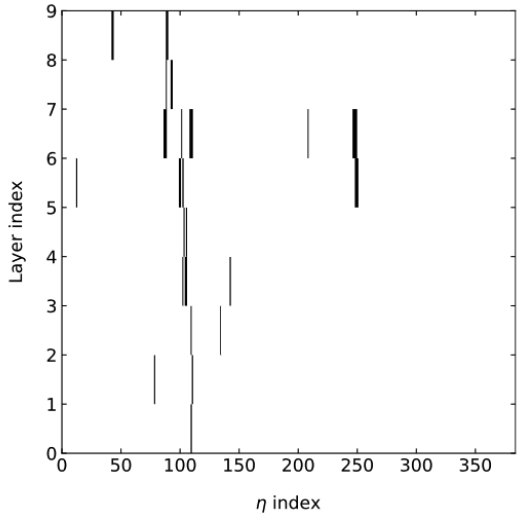
FPGA architecture

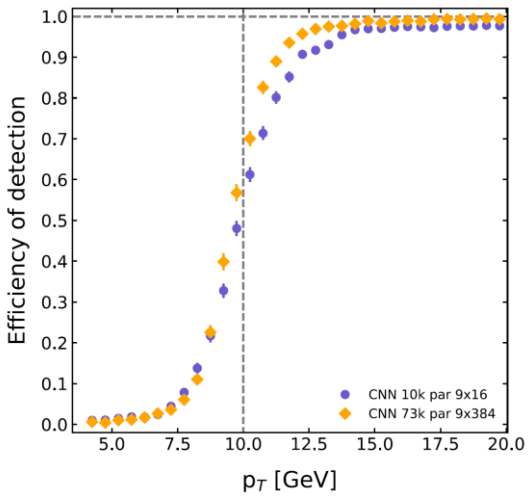
The final aim is to show that it possible to build Lightweight Neural Network to be implemented on FPGA for HLT and L0 triggers.

CNN and GNN FPGA implementations are currently under study.

L0 Muon Barrel Trigger of the ATLAS Experiment for HL-LHC





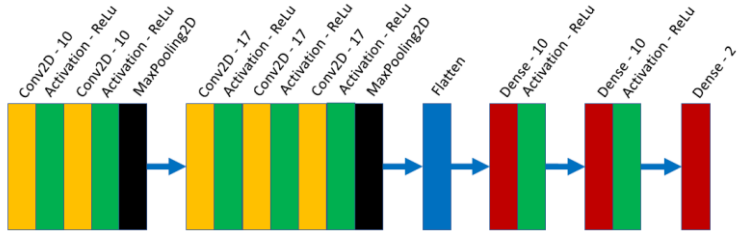


Model (9 × 16)	BRAM	DSPs	FF	LUT	Latency (cycles)
Teacher	1123	31.7 k	2.4 M	265.6 k	640
Student 32 bit	171	3.8 k	290 k	46 k	249
QStudent 4 bit	11	6	15.0 k	30.0 k	184
QStudent 3 bit	11	0	11.1 k	23.3 k	182

FPGA Xilinx Ultrascale+ XCV13P
occupancy and inference time per event

Model (9 × 16)	BRAM	DSPs	FF	LUT
Teacher (%)	20.9	258.0	69.4	15.3
Student 32 bit (%)	3.2	31.0	8.4	2.7
QStudent 4 bit (%)	0.2	0.05	0.4	1.7
QStudent 3 bit (%)	0.2	0	0.3	1.3

A use case is the trigger candidate identification algorithm for the L0 muon trigger in the barrel region of ATLAS.



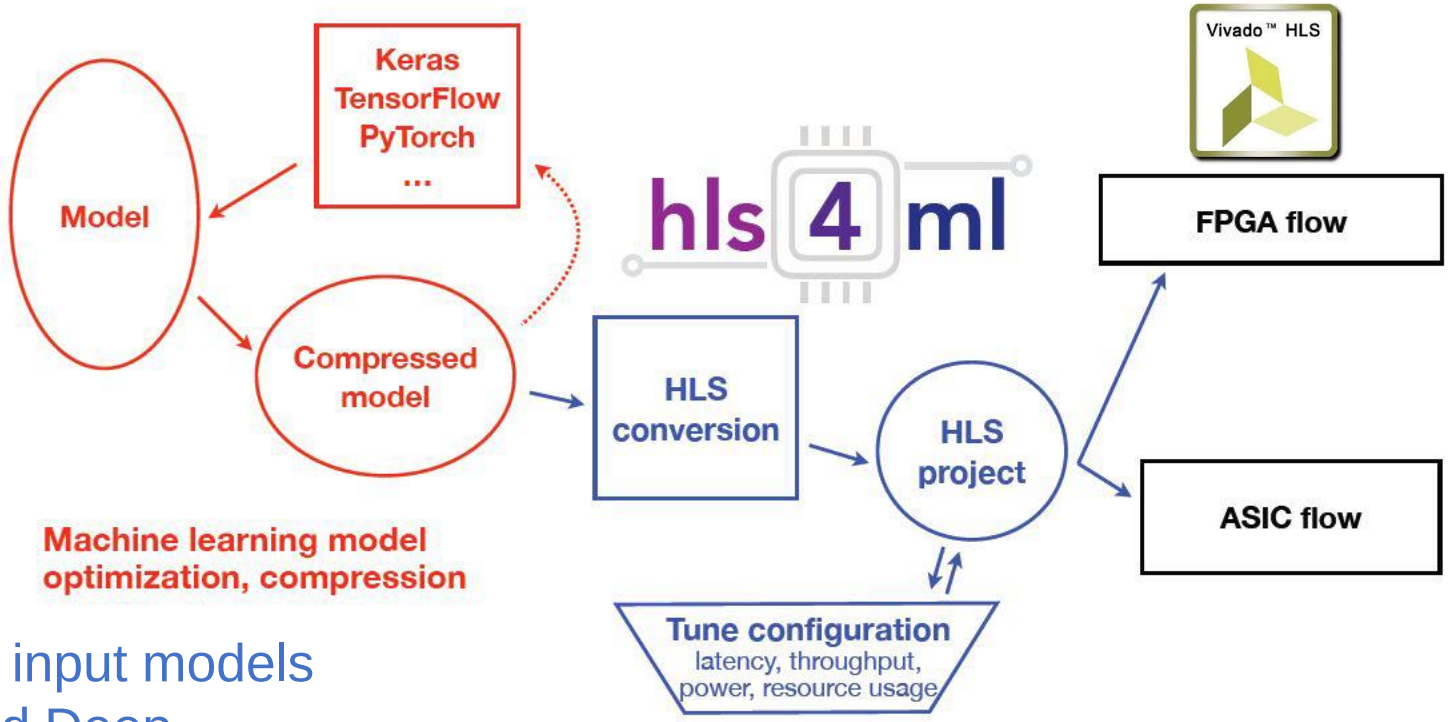
Different CNN implementations are tested, such as in this work:

Model compression and simplification pipelines for fast deep neural network inference in FPGAs in HEP

The latest FPGA implementations show promising performance, corresponding to a **latency of about 100 ns** using HLS4ML library and to a **latency of 82 ns** using a custom made VHDL CNN. These values are within the latency budget of the L0 trigger foreseen for the RPC.

Development pipeline

The implementation of NN models has been performed using the HLS4ML library, which in combination with Vivado HLS package translate a Tensorflow model into VHDL code.



HLS4ML reads as input models trained on standard Deep Learning libraries.

CNN and GNN are supported.

https://fastmachinelearning.org/hls4ml/
https://arxiv.org/abs/1804.06913
https://github.com/fastmachinelearning/hls4ml

Conclusions and Outlooks

The study demonstrates the feasibility of deploying machine learning algorithms on FPGA accelerator cards for trigger selection in HEP experiment.

FPGA-based CNN inference shows good performance in terms of maximum operation frequency, latency and logic resource occupation.

FPGA-based GNN, such as Garnet, is under study.

We are working on an HLT anomaly detection algorithm to be implemented in FPGA; it is a track-based jet anomaly tagger for BSM physics, such as dark-QCD, displaced jets, etc...

Next steps:

- An AMD Alveo cluster is currently being constructed at the INFN-Naples; it will allow us to investigate the performance and scalability of the Atlas L0 and HLT algorithms on multi-FPGA systems.
- An AMD Alveo cluster and an Intel Agilex cluster are under construction in Milano Bicocca site.
- The goal of the activity is to demonstrate that NN models can be used to fulfil the needs of the triggers that are in development for the upgraded LHC detectors.

Thank you for your attention