WP2 FPGA Status Report

Simone Gennai & Bernardino Spisso

ATLAS use case

CNN HLS4ML implementation studies for the ATLAS RPC Level 0 trigger system

- We are making different CNN HLS4ML implementation studies to demonstrate the feasibility of deploying machine learning algorithm on FPGA for the L0 trigger of the ATLAS experiment.
- The final aim is to understand if it is possible to optimize the design to obtain an efficient and fast inference.
- Our architecture is based on a 2D Convolutional Neural Network; we used the HLS4ML library to implement it on a Xilinx Virtex UltraScale+ xcu250-figd2104-2l-e FPGA.
- We are trying to reduce the complexity of the neural network because we had some issues related to the design routing.

ATLAS use case

Post implementation studies

We designed many CNN architectures, changing some of the neural network parameters (such as weight quantization, data type, number of kernels, number of layers), because Vivado HLS is not always able to complete the implementation.

Sometimes the design is not routable as its global congestion is high (depending on the CNN parameters).

HLS4ML library & Vitis/Vivado 2024.1

CNN input 384x9

6 Convolutional layers 2D

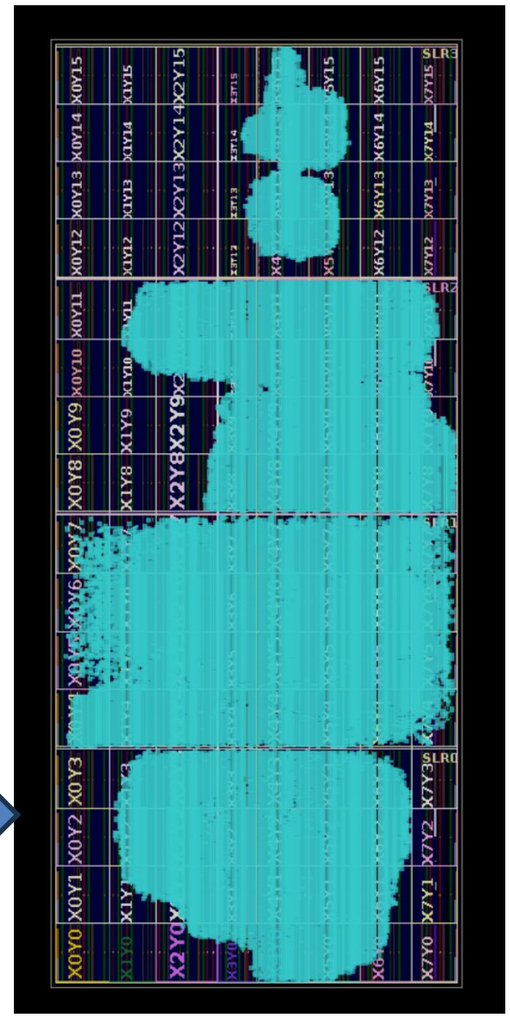
25 kernels 3x3

Weight 4-bit precision (qkeras)

FPGA occupancy

Site Type	Used	Fixed	Prohibited	Available	Util%
CLB LUTs	548546	0	0	1728000	31.74
LUT as Logic	548523	0	0	1728000	31.74
LUT as Memory	23	0	0	791040	<0.01
LUT as Distributed RAM	0	0			
LUT as Shift Register	23	0			
CLB Registers	605855	0	0	3456000	17.53
Register as Flip Flop	605855	0	0	3456000	17.53
Register as Latch	0	0	0	3456000	0.00
CARRY8	264	0	0	216000	0.12
F7 Muxes	2138	0	0	864000	0.25
F8 Muxes	42	0	0	432000	<0.01
F9 Muxes	0	0	0	216000	0.00

Site Type	Used	Fixed	Prohibited	Available	Util%
DSPs	24	0	0	12288	0.20
DSP48E2 only	24				



This is our first implementation in a Xilinx Virtex UltraScale+ xcu250-figd2104-2l-e FPGA.

We need to reduce the complexity of the Neural Network to optimize the design.

FPGA Clusters

Istituto Nazionale di Fisica Nucleare

- **We have finally received the clusters in both Naples and Milan:**

A total of three clusters equipped with 21 FPGA boards

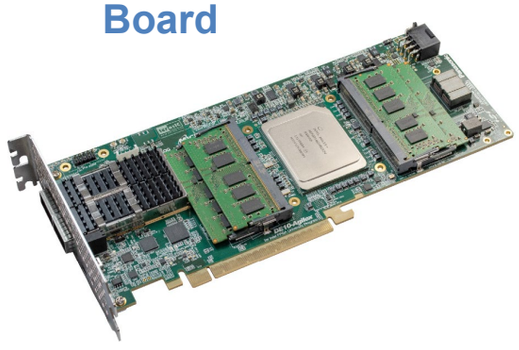
Naples:

- **2 Servers with:**
  - **8 Xilinx U55C Alveo boards**

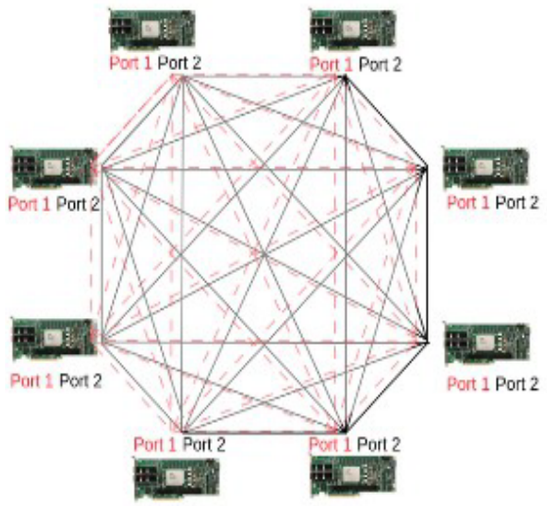
Milano-Bicocca:

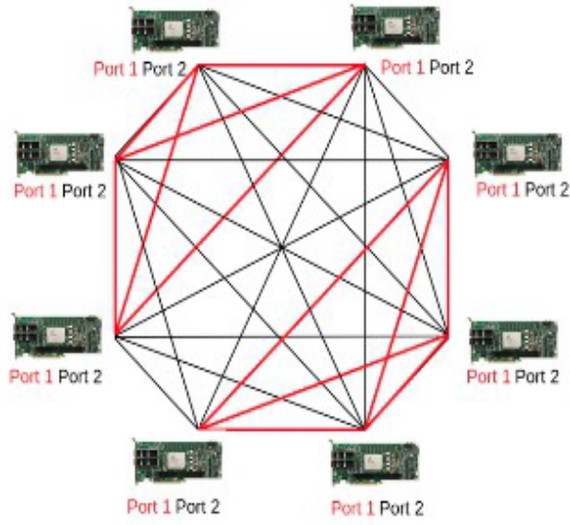
- **2 Servers with:**
  - **8 Xilinx U55C Alveo boards**
- **2 Servers with:**
  - **8 Intel Agilex Terasic boards**

Intel Agilex Terasic Board



Xilinx U55C Alveo Board





Unfortunately for the Intel cluster, the optical modules for Intel are not compatible with FPGAs something that had been assured to us by the seller, so their use is now limited.

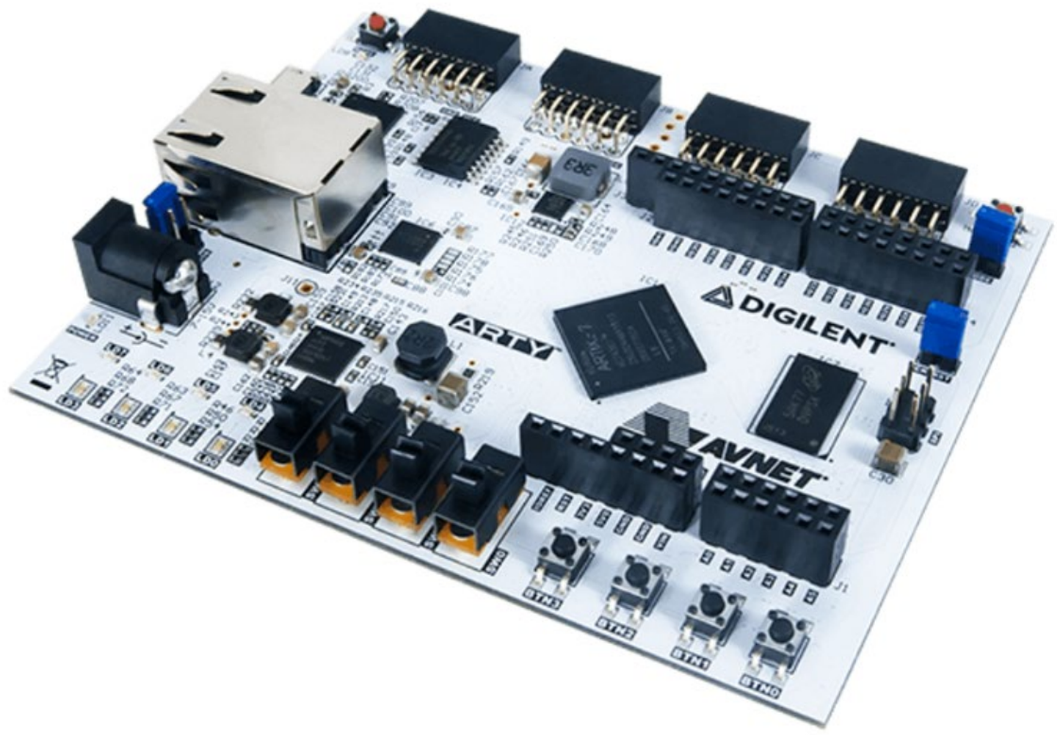
For Xilinx, Mirko Mariotti will soon test the full mesh configuration

FPGA Introductory Course

Istituto Nazionale di Fisica Nucleare

- We have re-proposed an introductory course for FPGA programming in Milan:
- https://agenda.infn.it/event/45908/
- Professors: M. Mariotti, A. Triossi, S. Summers
- The course used servers with FPGAs to give each user a
- Complete system for compiling firmware.

- A further advanced course is being organized for October in Perugia



Updated list of papers and contributions

Istituto Nazionale di Fisica Nucleare

Tabella 1

Titolo	Autori	Link a paper/conference	Accepted for PUB/PRoceeding
Fast Neural Network Inference on FPGAs for Triggering on Long-Lived Particles at Colliders	Andrea Coccaro Francesco Armando Di Bello Stefano Giagu Lucrezia Rambelli and Nicola Stocchetti	https://arxiv.org/pdf/2307.05152.pdf	Andrea Coccaro et al 2023 Mach. Learn.: Sci. Technol. 4 045040
Sviluppo di acceleratori per il Machine Learning e sistemi di Inference as a Service su FPGA	Daniele Spiga, Diego Ciangottini , Giacomo Surace, Giulio Bianchini, Loriano Storchi , Mirko Mariotti	Workshop Loano	
KServe inference extension for a FPGA vendor-free ecosystem	Daniele Spiga, Diego Ciangottini , Giacomo Surace, Giulio Bianchini, Loriano Storchi , Mirko Mariotti	CHEP 2023	EPJ Web of Conferences 295, 11012 (2024)
Deep Learning techniques for reconstruction on ASTRI Mini-Array Monte Carlo data	Saverio Lombardi, Francesco Visconti, Michele Mastropietro	https://pos.sissa.it/444/713/pdf	PoS(ICRC2023)713
A novel explainable approach in radiomics pipeline for local recurrence prediction of lung cancer: a feasibility study exploiting high energy physics potential to evaluate the model	Mariagrazia Monteleone, Simone Gennai, Pietro Govoni, Chiara Paganelli	ACM ISBN 979-8-4007-0815-2/23/09. https://doi.org/10.1145/3632047.3632074	ACM ISBN 979-8-4007-0815-2/23/09. https://doi.org/10.1145/3632047.3632074
Triggerless data acquisition pipeline for Machine Learning based statistical anomaly detection	Gaia Grosso, Nicolò Lai, Matteo Migliorini, Jacopo Pazzini, Andrea Triossi, Marco Zanetti, Alberto Zucchetta	CHEP 2023	G. Grosso et al EPJ Web of Conf., 295 (2024) 02033
40MHz Triggerless Readout of the CMS Drift Tube Muon Detector	Matteo Migliorini, Jacopo Pazzini, Andrea Triossi, Marco Zanetti	TWEPP 2023	M. Migliorini et al 2024 JINST 19 C02050
Front-End RDMA Over Converged Ethernet, real-time firmware simulation	Gabriele Bortolato, Antonio Bergnoli, Damiano Bortolato, Daniele Mengoni, Matteo Migliorini, Fabio Montecassiano, Jacopo Pazzini, Sandro Ventura, Andrea Triossi, Marco Zanetti	TWEPP 2023	G. Bortolato et al 2024 JINST 19 C03038
Front-End Rdma Over Converged Ethernet, real-time firmware simulation	Gabriele Bortolato, Antonio Bergnoli, Damiano Bortolato, Daniele Mengoni, Matteo Migliorini, Fabio Montecassiano, Jacopo Pazzini, Sandro Ventura, Andrea Triossi, Marco Zanetti	TIPP 2023	
The CMS Level-1 trigger data scouting for LHC run 3 and the CMS phase-2 upgrade	Sabrina Giorgetti (Matteo Migliorini, Rocco Ardino, Jacopo Pazzini, Andrea Triossi, Marco Zanetti) on behalf of the CMS Collaboration	La Thuile 2024 - YSF	
Hardware implementation of quantum machine learning predictors for ultra-low latency applications	Lorenzo Borella, Alberto Coppi, Jacopo Pazzini, Andrea Stanco, Andrea Triossi, Marco Zanetti	EuCAIFCon 2024	
Quantum machine learning classifiers implemented on FPGA for ultra-low latency applications	Lorenzo Borella, Alberto Coppi, Jacopo Pazzini, Andrea Stanco, Andrea Triossi, Marco Zanetti	ICHEP 2024	

KPIs Status

Istituto Nazionale di Fisica Nucleare

Link to the flagship document

KPI ID	Description	Acceptance threshold	Status up to today
KPI2.2.3.1	Development of triggering algorithms, on-line analyses, data acquisition on FPGA	Submission of 1 paper to a peer-reviewed journal	1 paper already accepted
KPI2.2.3.2	Online scouting	Submission of 1 paper to a peer-reviewed journal	paper published in: *PoS* ICHEP2024 about scouting
KPI2.2.3.3	Development of tools to integrate several FPGAs together	Submission of 1 paper to a peer-reviewed journal	G. Bortolato et al 2024 JINST 19 C03038
KPI2.2.3.4	Organizing courses about FPGA programming on low and high level	At least two courses organized	1 course done at the end of 2023 1 VHDL course done in February 2024 1 course done in June 2025