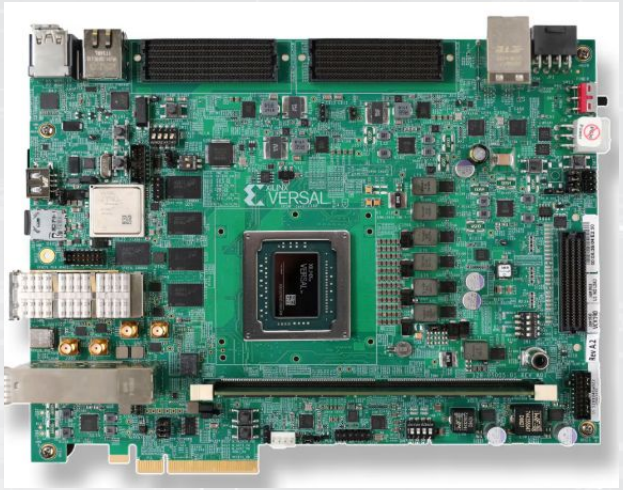
Annual PhD Review - 2024-2025

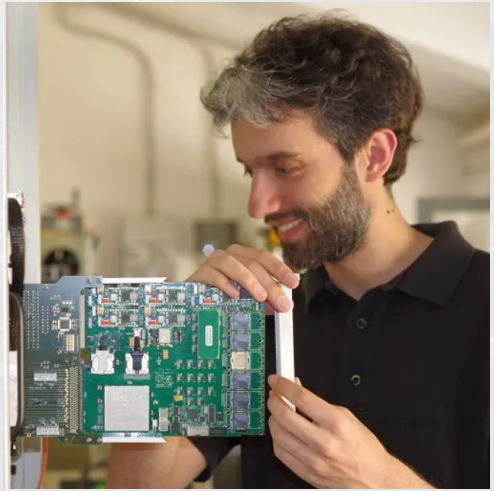
Sapkas Michail, University of Padova

About my work

- I work with FPGAs (SoC) and the AI Engines of the Versal
- I am implementing “Artificial Intelligence” algorithms such as:
  - CNNs
  - RNNs



- Big thanks to my co-Supervisor Andrea Triossi for his incredible support:



Courses and Workshops

This Year (1st) I passed 3 PhD courses:

- Neuromorphic Computing with Andrea Duggento (1.5 CFU)

- Programmable System on Chip (SoC) for data acquisition and processing with Andrea Fabbri (4 CFU)

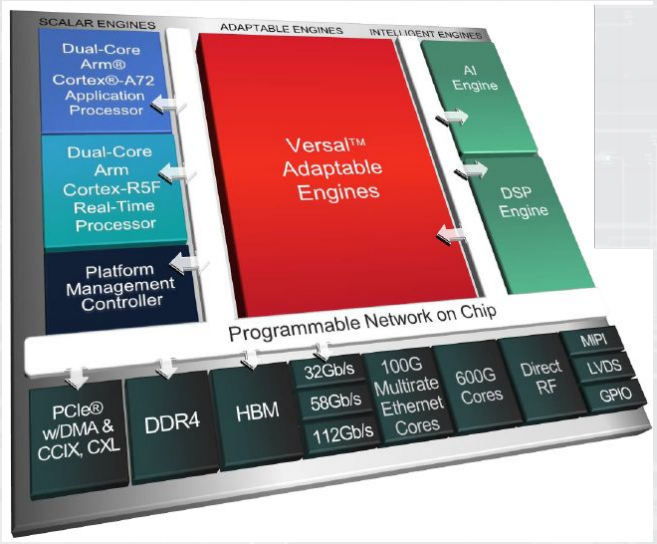
- Big Data Modelling and Learning with Ester Pantaleo (1.5 CFU)

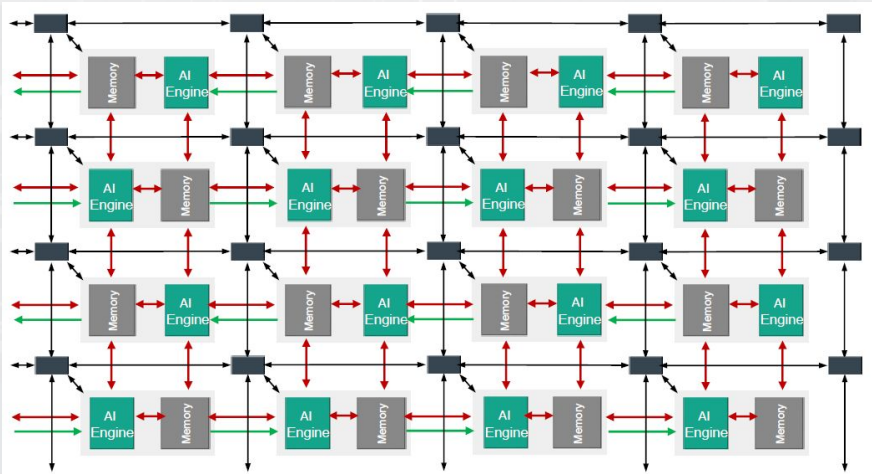
What are ACAPs and AIEs

- ACAP = Adaptive Compute Acceleration Platform = CPU + FPGA + AIE

- AIE = 400 real-time RISC-V CPU cores placed in a Tiled Array like this:

YOU are tasked to program them from scratch! From their (inter)-connectivity and functionality up to their physical placement!





Project 0: Porting the Particle Flow algo

- In the first 2 months of my PhD I was tasked with implementing the CMS Particle Flow algorithm on the Versal AI Engines
- A partial implementation showed that the algorithm - as it is written in HLS - does not really benefit in running exclusively on the AI Engines
- Even if we decide to do some numericals in the AIE and the rest of the logic in the FPGA the main problem remains:

The interface Tiles between PL and AIE are a significant bottleneck for extreme latency applications

Project 1: Deployment of Recurrent Neural Networks on the AMD Versal AI Engine

- Continuing my Master Thesis, I designed a **fully parallel AI Engines** execution of the **Gated Recurrent Unit**. The implementation is **general** and can apply to any low - latency application. The AIEs allow for **Floating Point operations** and models that scale up to **millions of FP parameters**.

- **In contrast with the previous bottleneck problem, the ability of the AI Engines FP data paths to broadcast data, works in our advantage! (everyone accepts the same input vector)**

An “exercise” that took almost 1 year to complete

- The Versal AI Engines is the “new kid” in the block (AMD documentation / bugs)
- The complexity of coordinating all the modules and data transfers between DDR Memory, FPGA and PS scales exponentially with introducing the AIE.

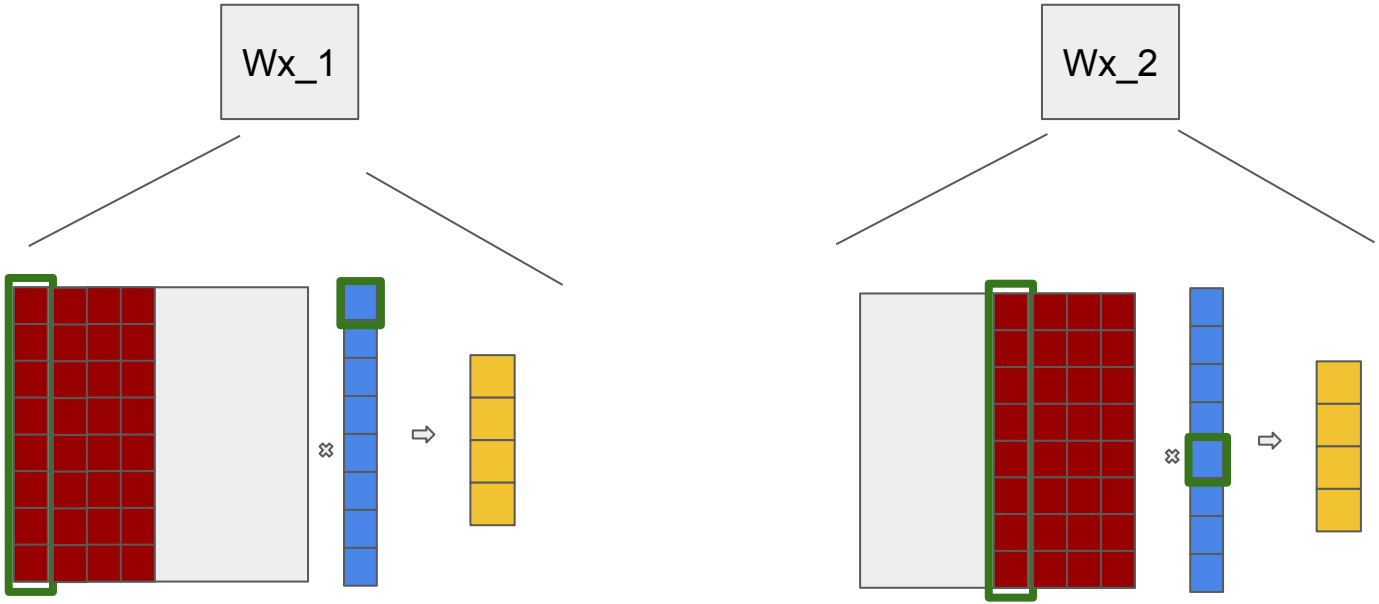
- Invaluable experience gained by programming the AIE
- Most EE / CA PhDs spend their time in trying to get closer to the theoretical throughput of the device in a mock-up GEMM (Matrix-Matrix Mult) scenario. That's because the device is being marketed as a GPU alternative.

- **We are one of the very few that are trying to use the AIE for Real Time - Ultra Low Latency Applications. Targeting low-latency also changes the way you compute!**

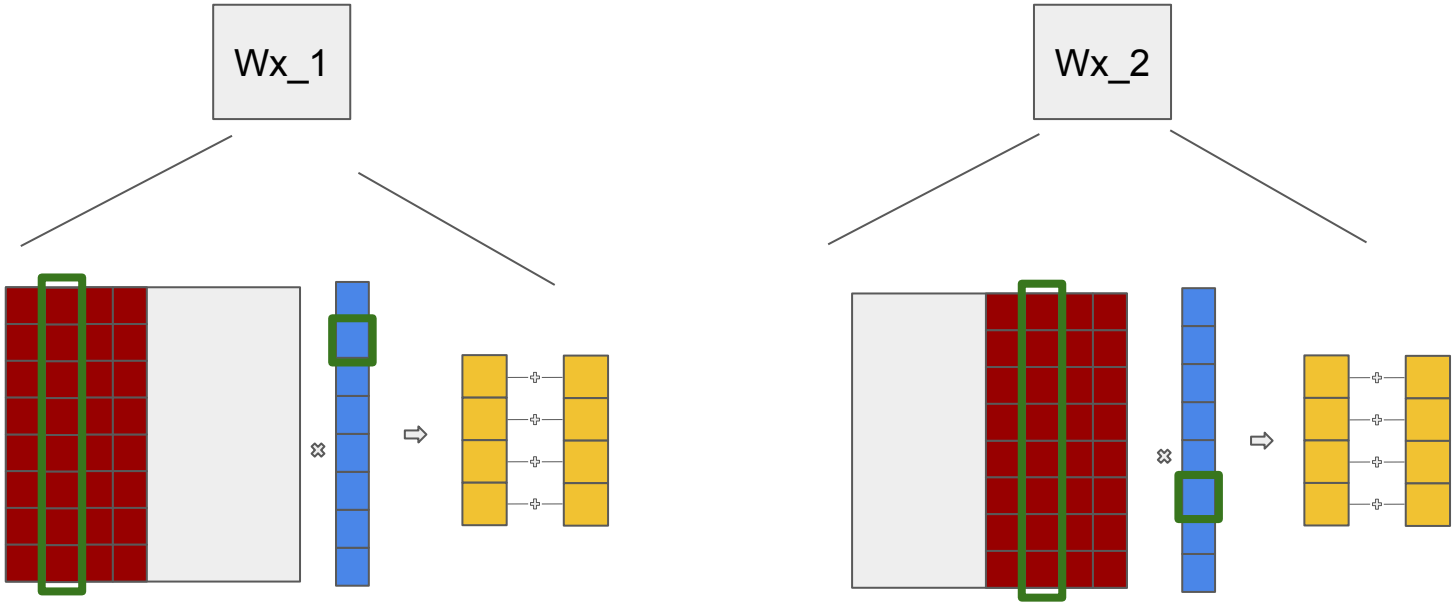
Project 1: Deployment of Recurrent Neural Networks on the AMD Versal AI Engine

$$
\begin{aligned}
z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\
r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\
\hat{h}_t &= \phi(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t
\end{aligned}
$$

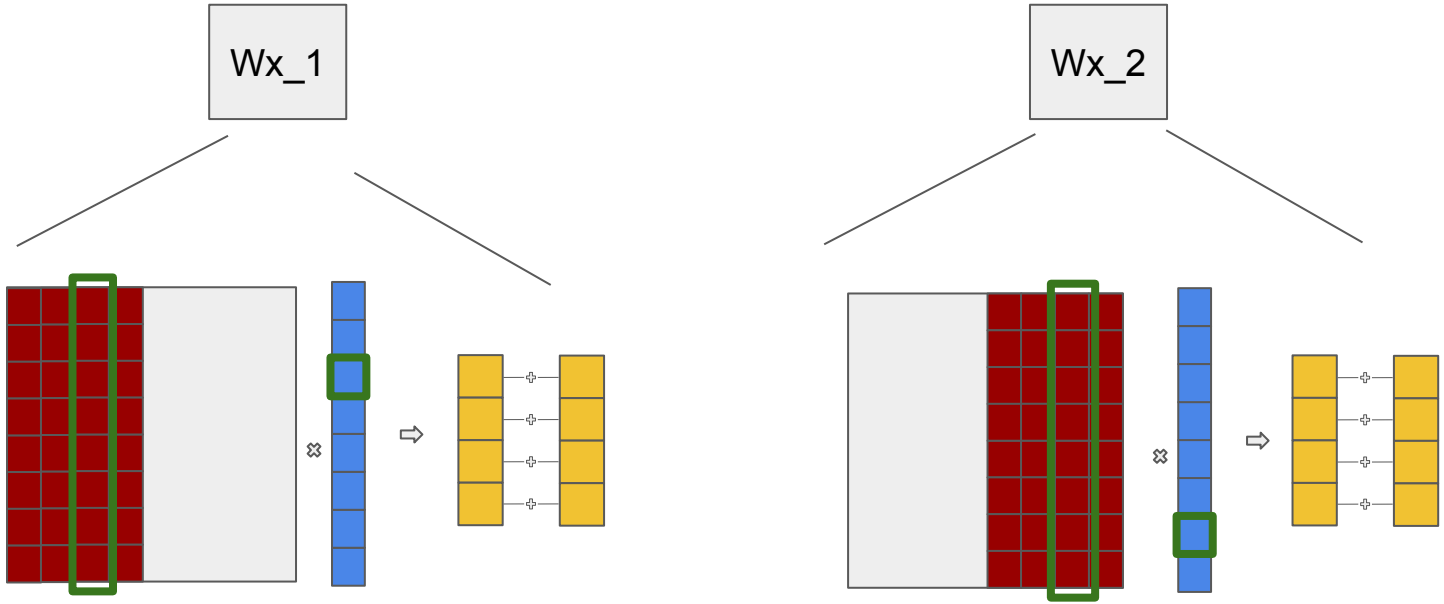
Split the columns into multiple tiles - cascade the result



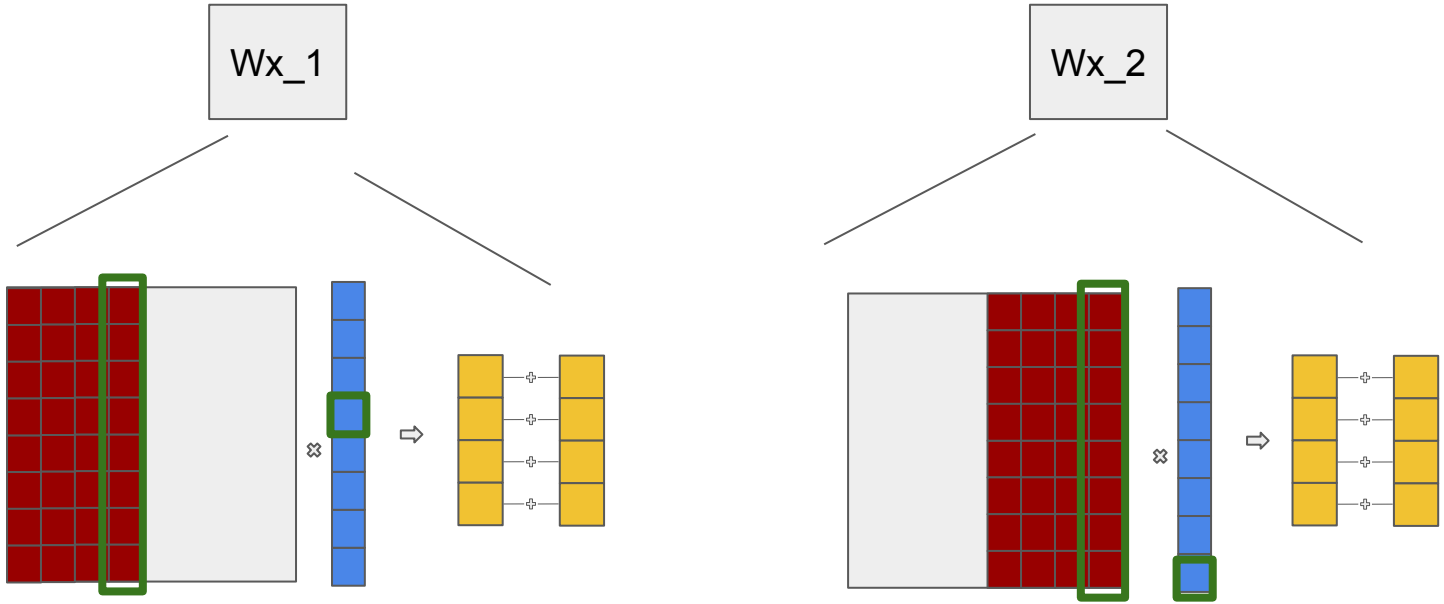
Split the columns into multiple tiles - cascade the result



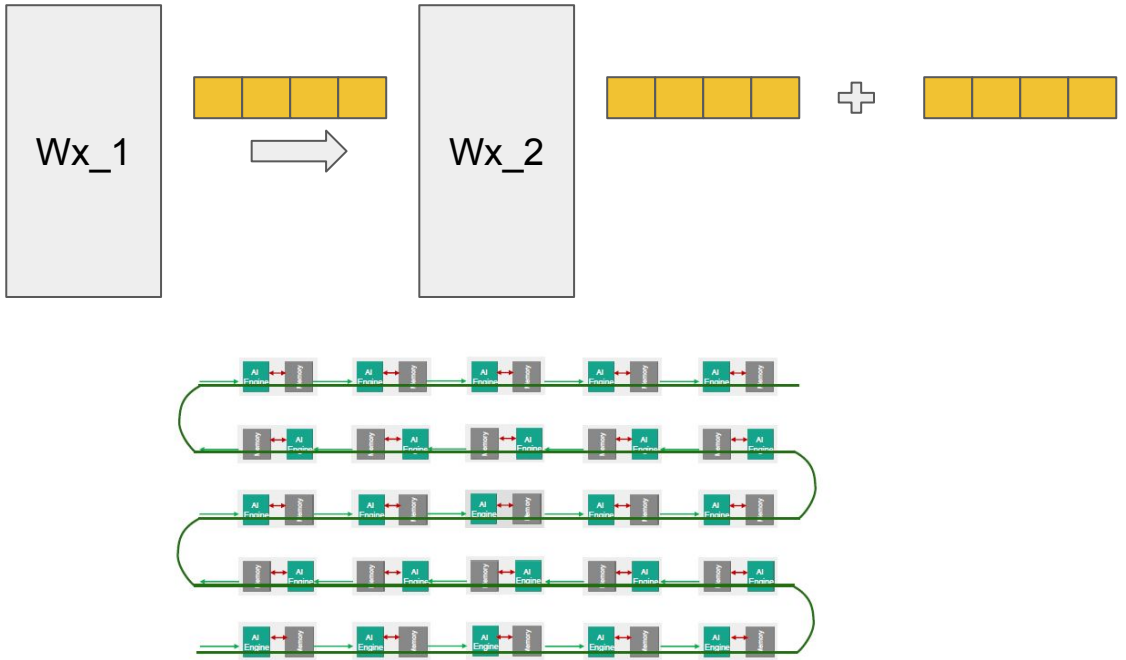
Split the columns into multiple tiles - Cascade the result



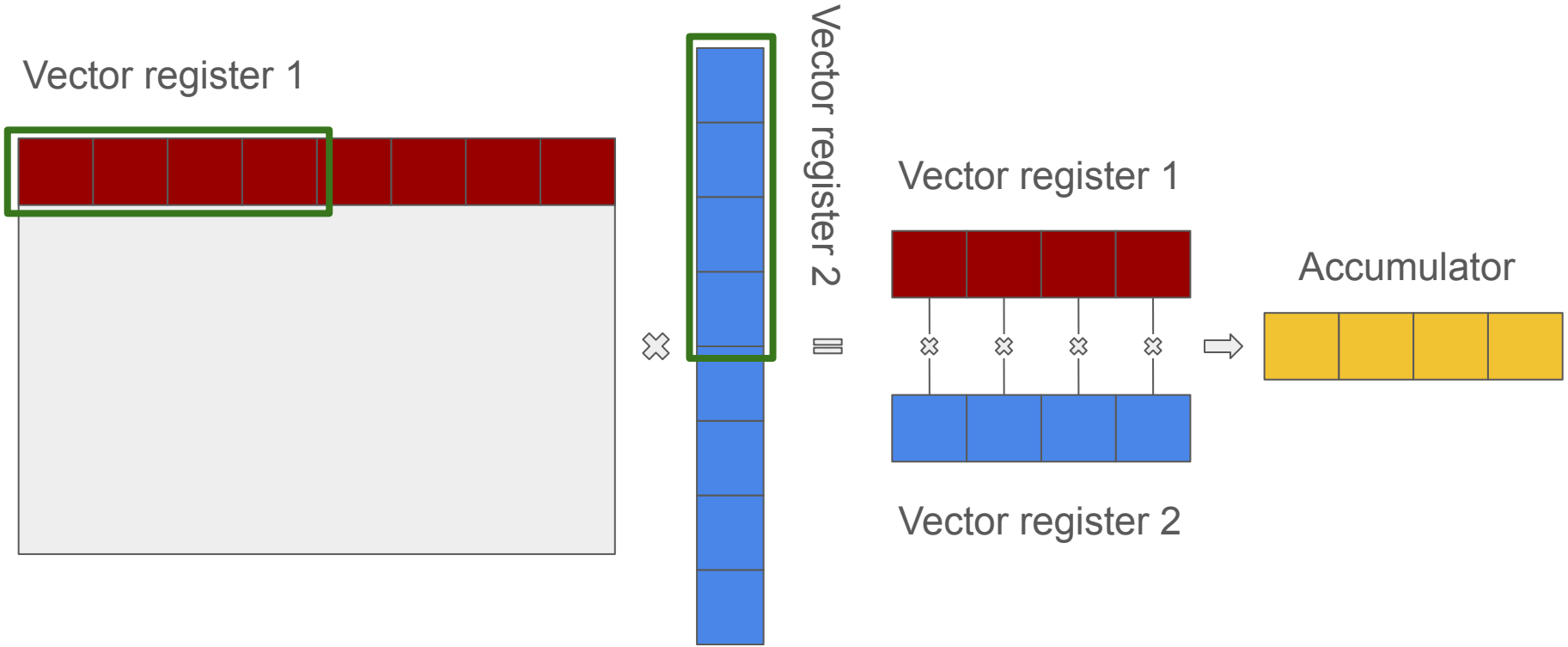
Split the columns into multiple tiles - Cascade the result



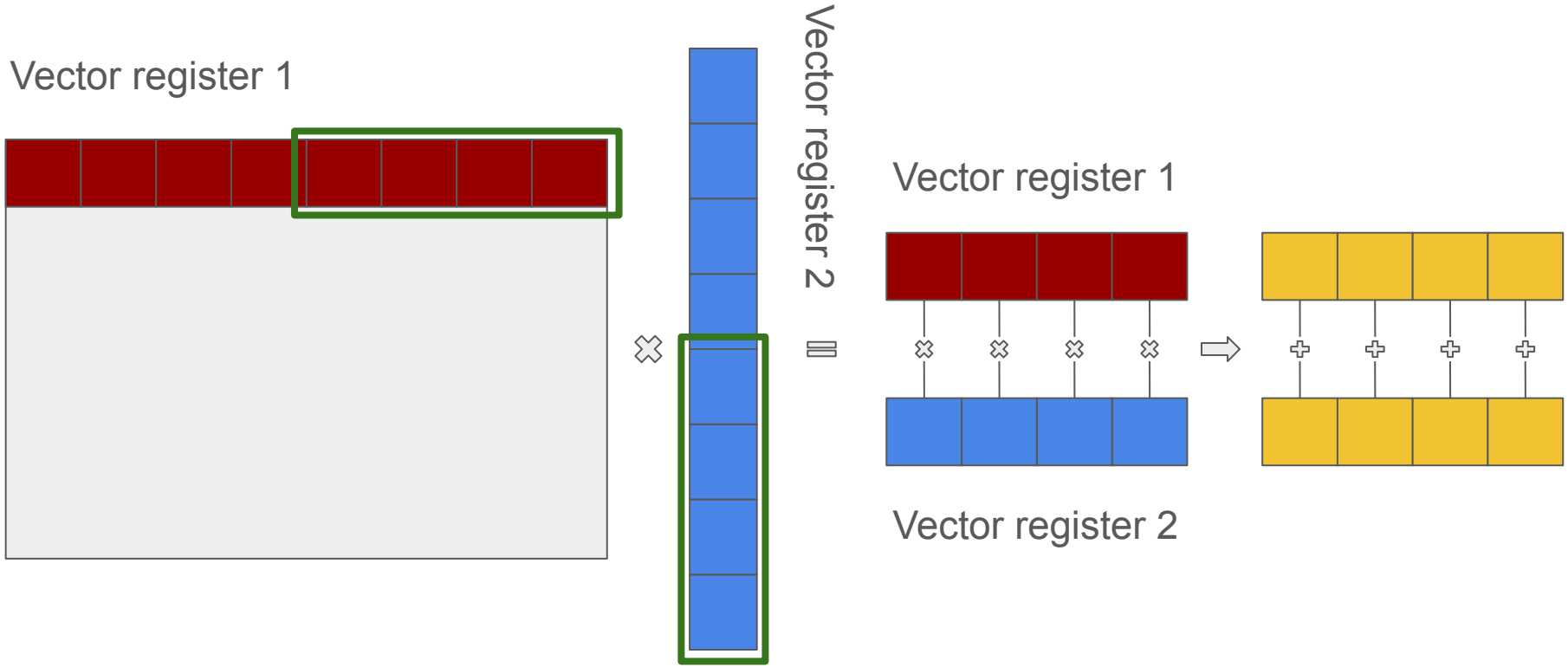
And use Cascade



Multiply Accumulate: Rows

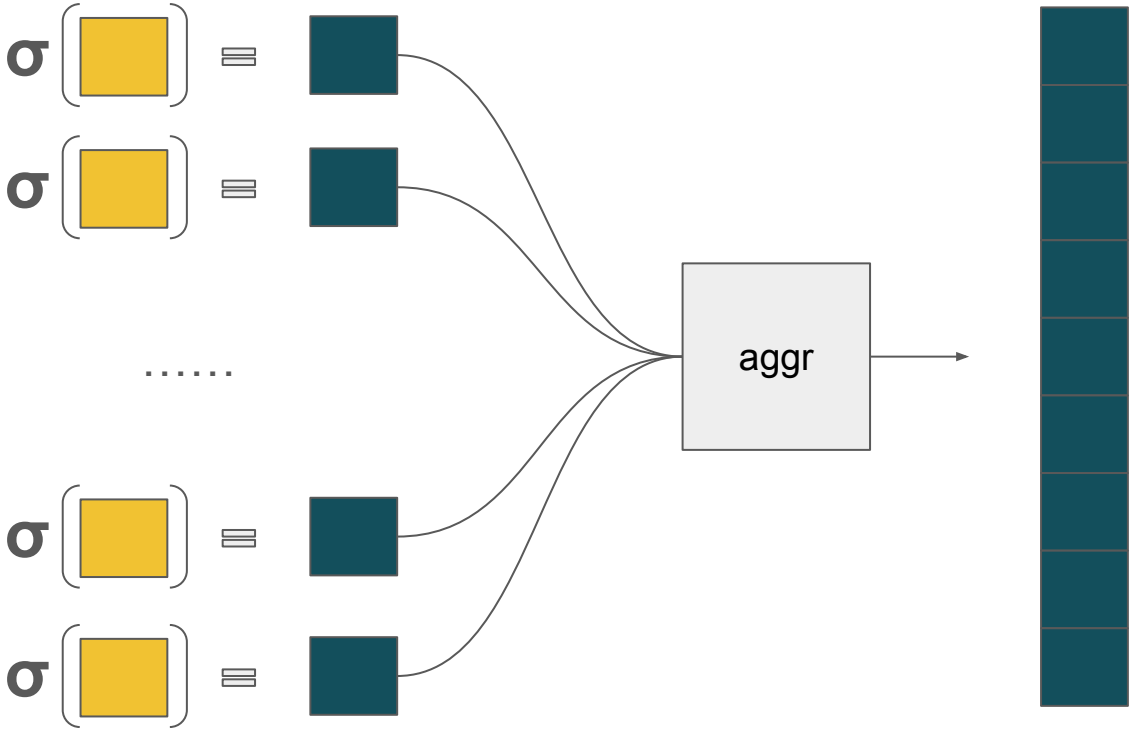


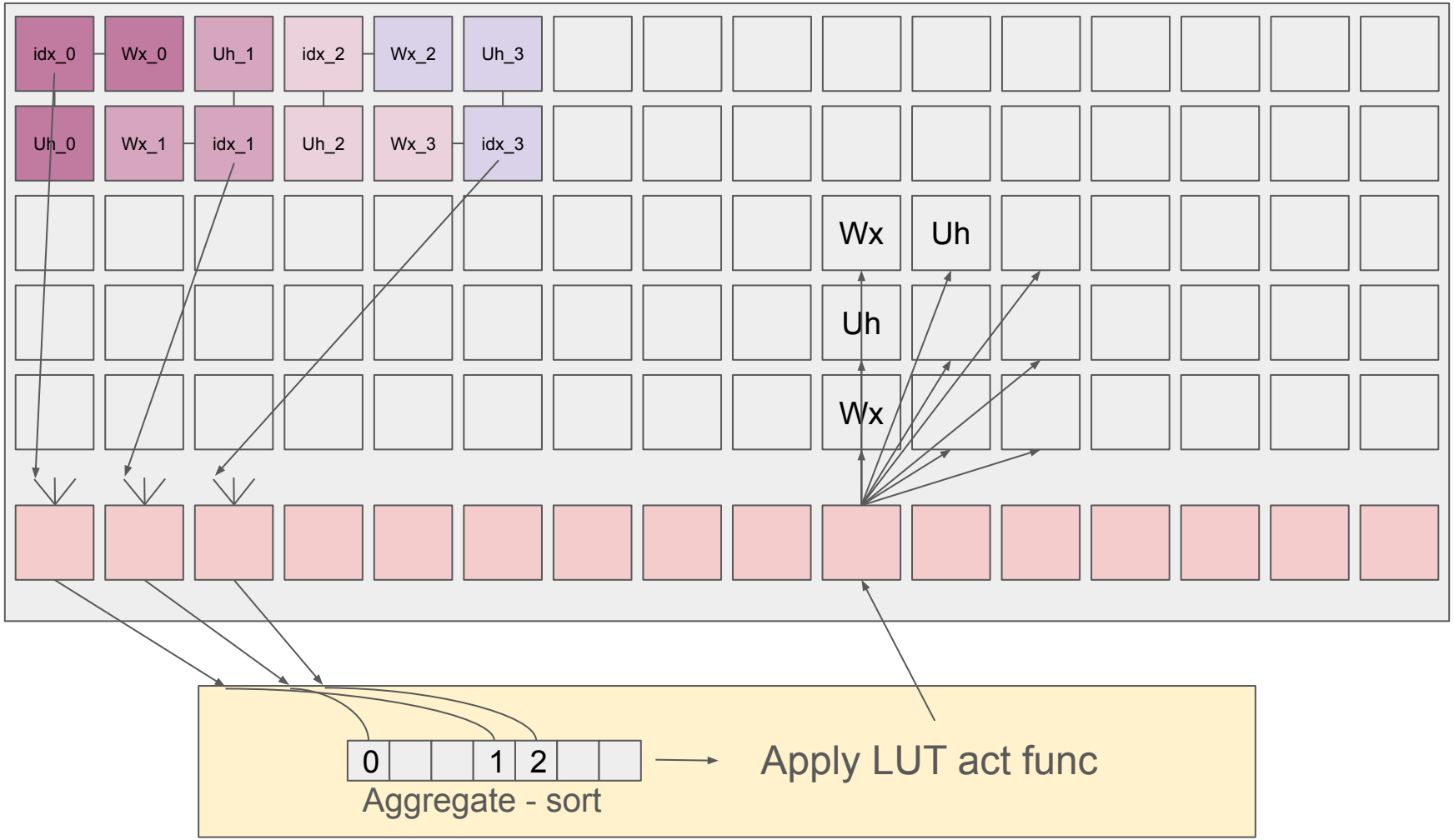
Multiply Accumulate: Rows



Multiply Accumulate: Rows

Aggregate partial results w Packet Stream: Merge



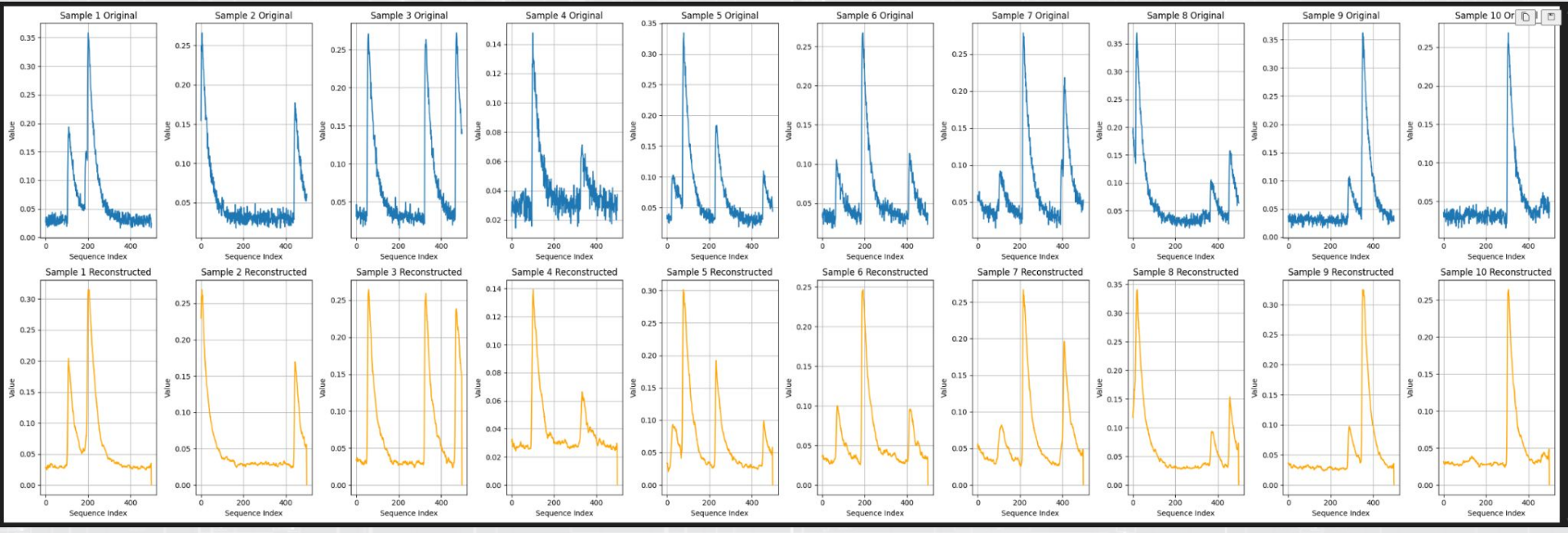

idx_0
Wx_0
Uh_1
idx_2
Wx_2
Uh_3
Uh_0
Wx_1
idx_1
Uh_2
Wx_3
idx_3
Wx
Uh
Uh
Wx
0
1
2
Aggregate - sort
Apply LUT act func

Project 2: AI-Enhanced Data Processing on the CAEN 2745 Digitizer

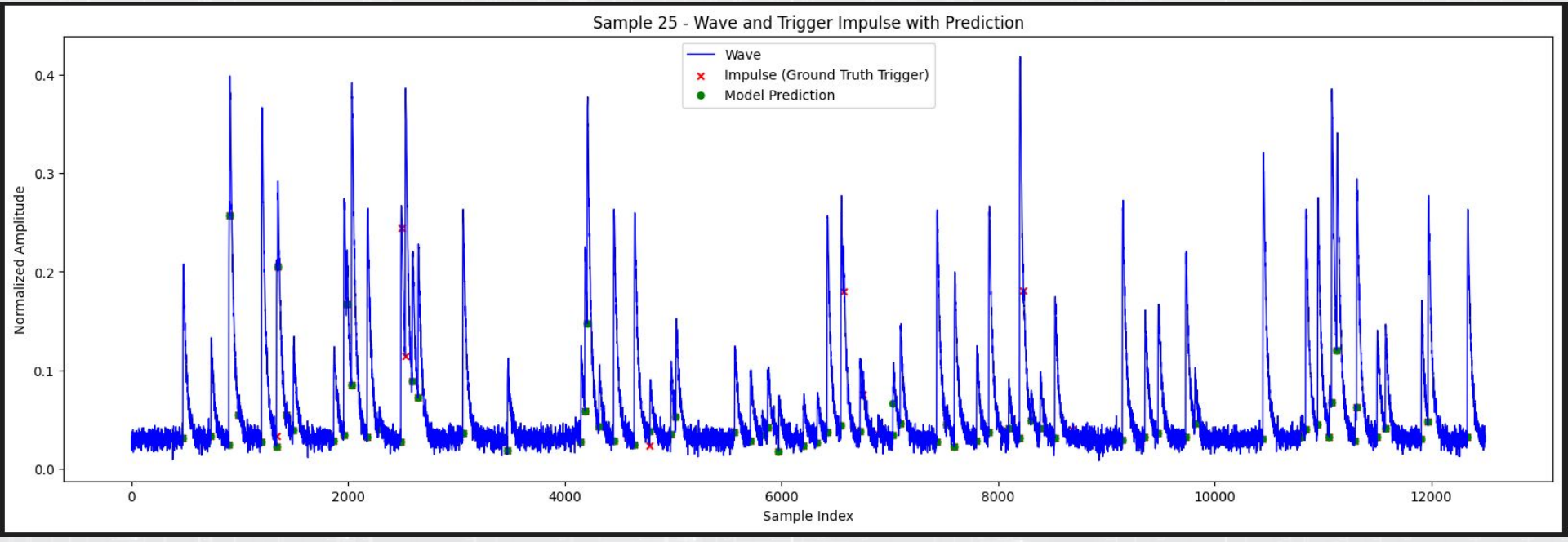
- The objective of this project is to enhance the CAEN 2745 digitizer's performance by deploying AI algorithms on the Xilinx Zynq-7000 SoC. Two primary use cases are being explored:

1. A convolutional neural network (CNN) peak detector to function as a trigger mechanism, transferring data only when significant peaks are detected.
2. A CNN-based autoencoder that continuously encodes fixed time windows of digitized data, significantly improving the throughput of the device.

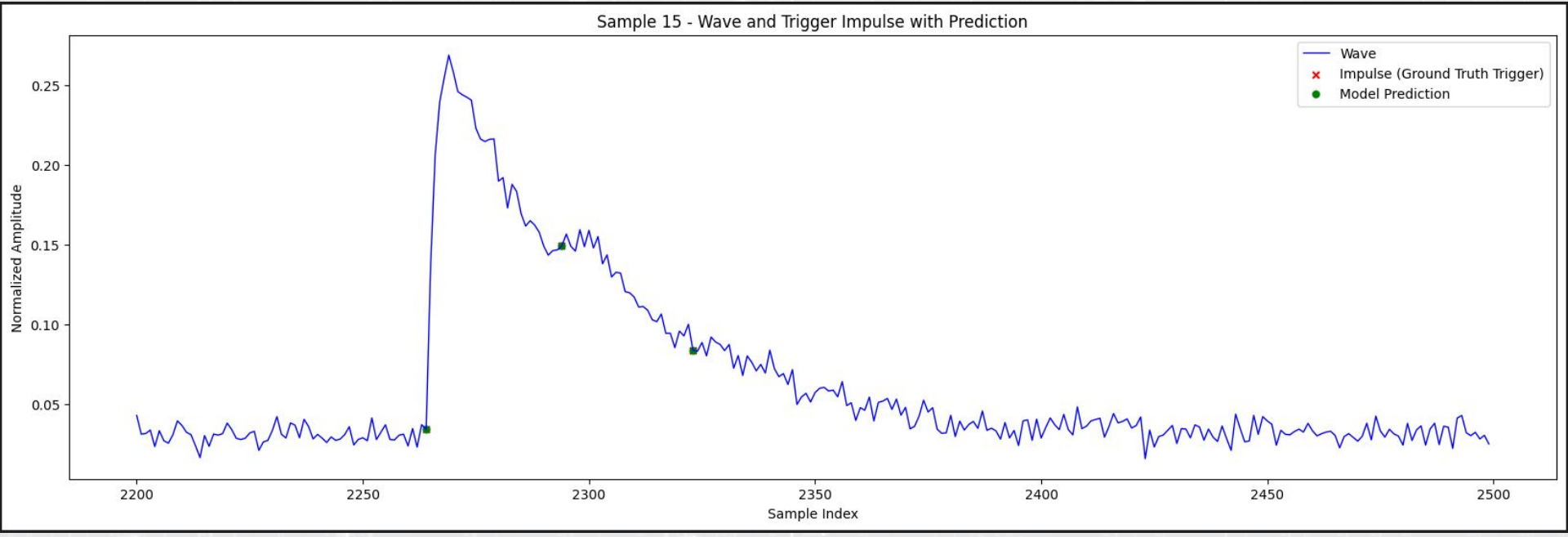
Project 2: AI-Enhanced Data Processing on the CAEN 2745 Digitizer



Project 2: AI-Enhanced Data Processing on the CAEN 2745 Digitizer



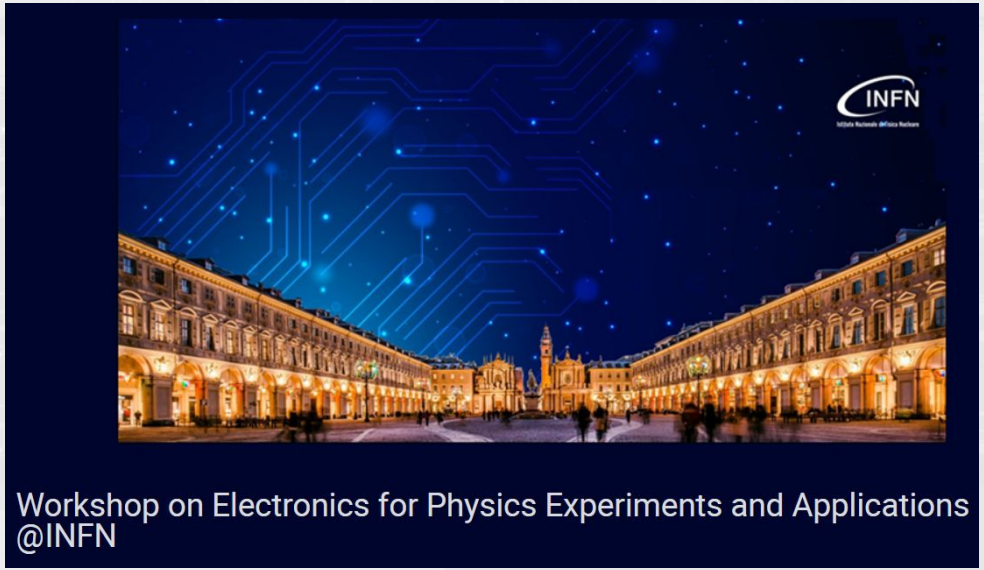
Project 2: AI-Enhanced Data Processing on the CAEN 2745 Digitizer



Attended Conferences, Schools and Workshops

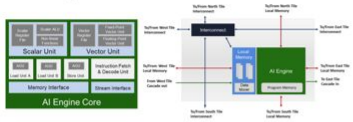
This Year (1st) I attended 1 Workshop, 1 School and 3 Conferences:

- Workshop on Electronics for Physics Experiments and Applications @INFN Mar 5 – 7, 2025 Torino



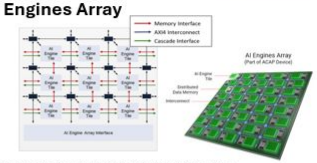
Exploring the AMD Xilinx Versal AI Engine (AIE) for low latency applications

AI Engine Tile



- 32 KB memory but can use up to 4 memory modules in all four directions as **one 128KB contiguous block** of memory
- **Scalar** RISC processor (called Scalar Unit)
- **Vector** unit featuring a Vector **fixed-point/integer** unit and a Single-precision **floating-point** (SPFP) vector unit
- Three data memory ports allow for **two loads and one store**
- Connections of **two input streams** and **two output streams**
- **One cascade stream**

AI Engines Array



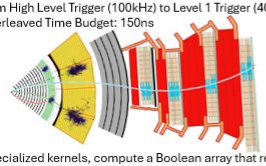
- A two-dimensional array of **400 AI Engine Tiles**
- Connected by **Memory**, **Stream** and **Cascade** Interfaces

Graphs and Kernels

- The user programs **kernels** which operate on input and output data and they represent physical AI Tiles
- Graphs are collections of connected kernels
- The top level graph represents the operations of the whole AI Engine Array

CMS Particle Flow

- The next upgrade of CMS will bring the Particle Flow Algorithm from High Level Trigger (100kHz) to Level 1 Trigger (40MHz)
- Interleaved Time Budget: 150ns



- Specialized kernels, compute a Boolean array that represents:
  - A **Muon** to a **Track** linking
  - A **Track** to a **Calorimeter Cluster**

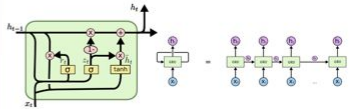
- Preliminary results show extremely fast Muon-Track Linking **but** Track-Calo Linking remains challenging due to more complex calculations

Objects / Kernel	Muon – Track [ns]	Track – Calo [ns]
1	80.0	260.8
2	115.2	350.0
3	169.6	389.6
4	220.5	482.4

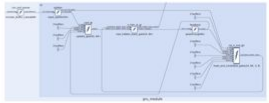
- **Main Challenges:**
  - Data transfers through the PL-AIE **Interface Tiles** creates the main bottleneck of the problem
  - Unlike FPGA fabric, the AIE processors cannot accept data while computing, which critically reduces the **Iteration Interval**
  - The need to manipulate single elements in the vectors breaks the pipelining and introduces latency

Gated Recurrent Unit

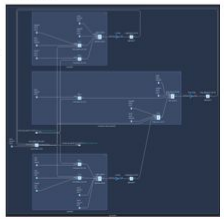
- Recurrent Neural Networks are a class of neural networks specifically designed for processing sequential data
- Initially designed to be used for **real time control** and trained using **reinforcement learning** in a synchrotron light source



- Physics experiments could benefit by achieving **~ µs inference**

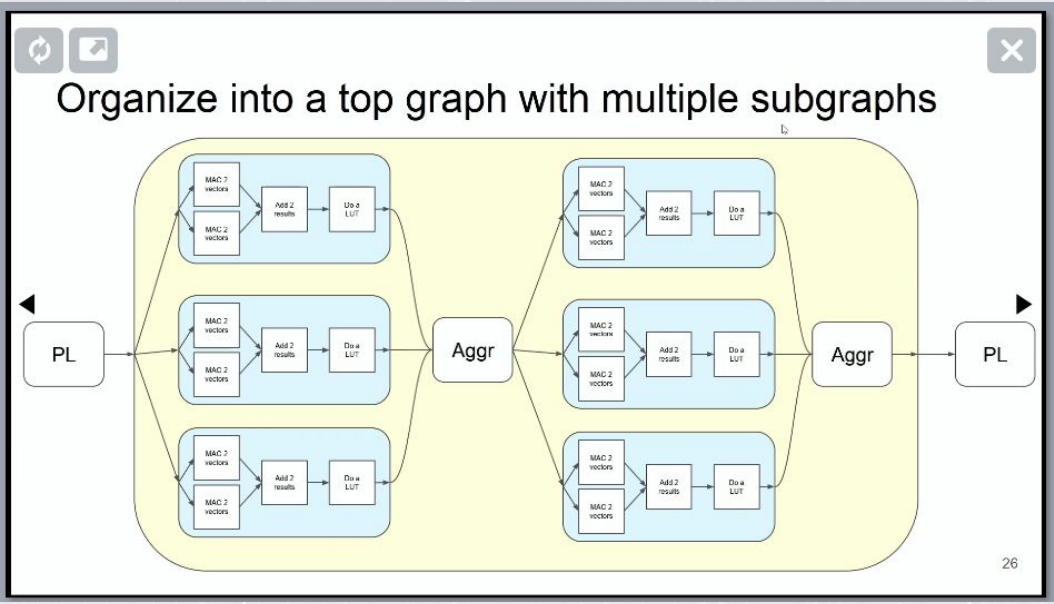


- A Preliminary implementation used a **compute** approach for the **activation functions** and only three kernels
- **Main Challenges:**
  - Limited memory to store the parameters of the model
  - The sequential nature of the model introduces latency
  - Multiple connectivity choices and distribution/scheduling strategies create a huge design space to explore
- **in Progress:**
  - Create a framework in which a fully distributed version can be deployed
  - Schedule correctly the workload



Attended Conferences, Schools and Workshops

- **Conference -** 2nd FPGA Developers' Forum (FDF) meeting May 20 – 23, 2025 **CERN**



https://cds.cern.ch/record/2932736

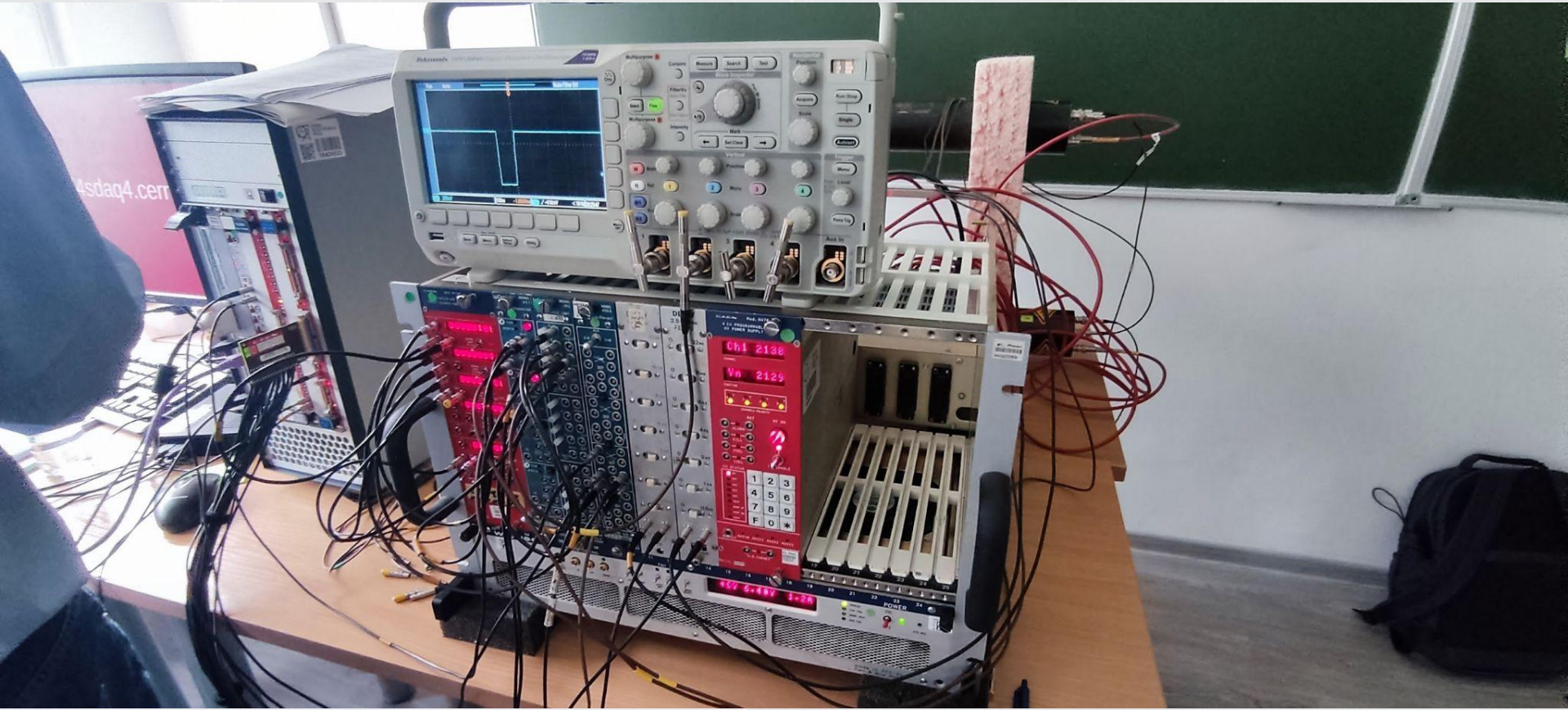
OR

https://www.youtube.com/watch?v=0G2Np1h4uVM

Attended Conferences, Schools and Workshops

- **Attended School -** ISOTDAQ 2025 - International School of Trigger and Data AcQuisition Jun 17 – 26, 2025 **Vilnius, Lithuania**

ISOTDAQ 2025 - International School of Trigger and Data AcQuisition

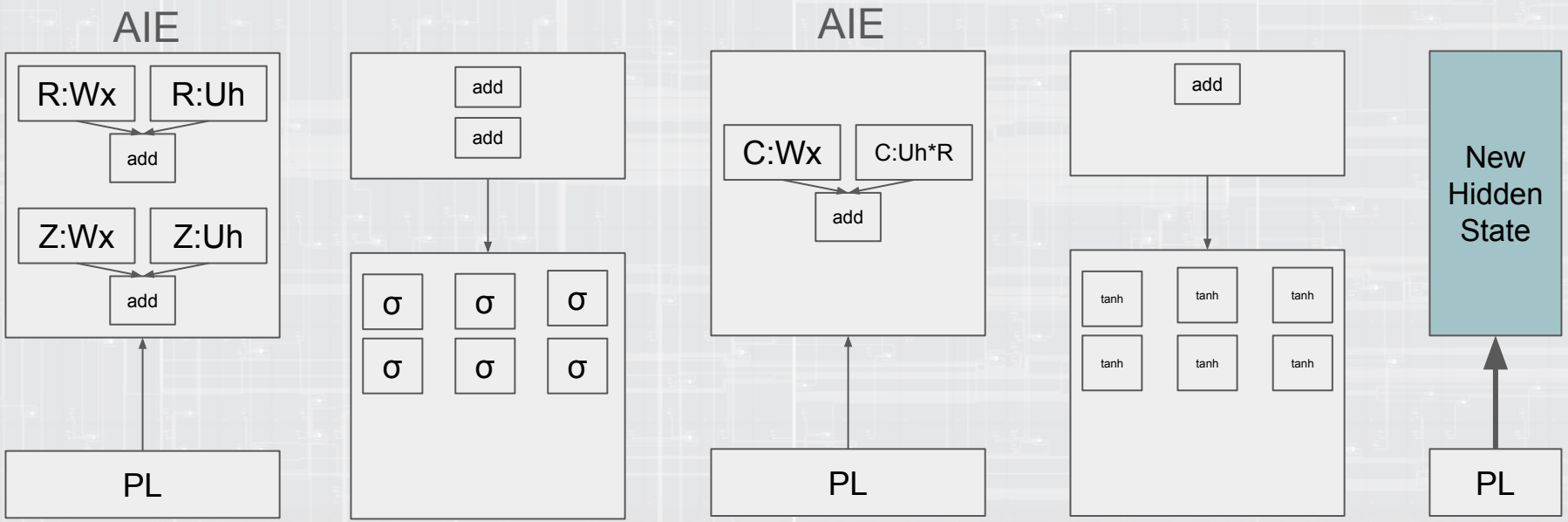

Ch1 2138
Vm 2129
POWER

Attended Conferences, Schools and Workshops

- **Conference -** TWEPP 2025 Topical Workshop on Electronics for Particle Physics Oct 6 – 10, 2025 **Rethymno, Crete, Greece**

TWEPP 2025 Topical Workshop on Electronics for Particle Physics

Attended Conferences, Schools and Workshops



For the Second Year

I need to do the CAEN project of running CNNs on the ZYNQ 7000:

- Hyperparameter tuning of the model with resource constraints
- Kernel Pruning to minimize resource consumption (there is also a method of imposing orthogonal kernels)
- Deploy with HLS4ML
- Quantize efficiently
- Test

- Estimated Time for delivery ~ 4-5 months

Future Conferences, Schools and Workshops

Advanced course to FPGA programming

Oct 27 – 29, 2025
INFN Perugia
Europe/Rome timezone

Enter your search term

- Overview
- Registration
- Participant List
- Travel & Accomodation

Starts Oct 27, 2025, 2:00 PM
Ends Oct 29, 2025, 5:00 PM
Europe/Rome

INFN Perugia
Dipartimento di Fisica e Geologia - FISGEO UNIPG
Via Alessandro Pascoli, 06123 Perugia PG
Go to map

Daniele Spiga
Pasquale Lubrano
Simone Gennai
Stefano Giagu

Bulletin Advanced Course on FPGA programming....

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Advanced course to FPGA programming

Image taken from:
https://www.amd.com/en/products/accelerators/alveo.html

The course is mainly thought for the new personnel hired by means of the ICSC project to work on projects and research activities in the framework of Spoke 2 ("Fundamental Research And Space Economy") and Spoke3 (Astrophysics & Cosmos Observations), but open to anyone interested within the fundamental physics community.

The lessons will be held in Peurgia.

Lecturers:

- Piero Vicini
- Francesco Simula
- Francesca Locicero
- Ottorino Frezza
- Andrea Biagioni

3rd year Internship

- I have been contacted by my former thesis supervisor
- He is working on machine Learning models at SLAC
- The project idea is to develop a tools that deploys AI models in the AI Engines just like HLS4ML
- Scheduled interview on the 24th of October

Physics > Instrumentation and Detectors

[Submitted on 30 May 2023]

Implementation of a framework for deploying AI inference engines in FPGAs

Ryan Herbst, Ryan Coffee, Nathan Fronk, Kukhee Kim, Kuktae Kim, Larry Ruckman, J.J. Russell

Thank you for your time!

Hit me up with any curiosities you may have.

More about my projects here: