The **C**entral **T**rigger **P**rocessor **b**oard for the SiPM AdvCam of the CTAO-LST

M. Molina, A. Pérez-Aguilera, J. Buces, J. A. Barrio, L. A. Tejedor, D. Nieto

25th IEEE Real Time Conference
La Biodola - Elba, Italy - 26/05/2026

Photo credit: Otger Ballester (IFAE) - CTAO Flickr

CTAO

GAE

IPARCOS

UNIVERSIDAD COMPLUTENSE MADRID

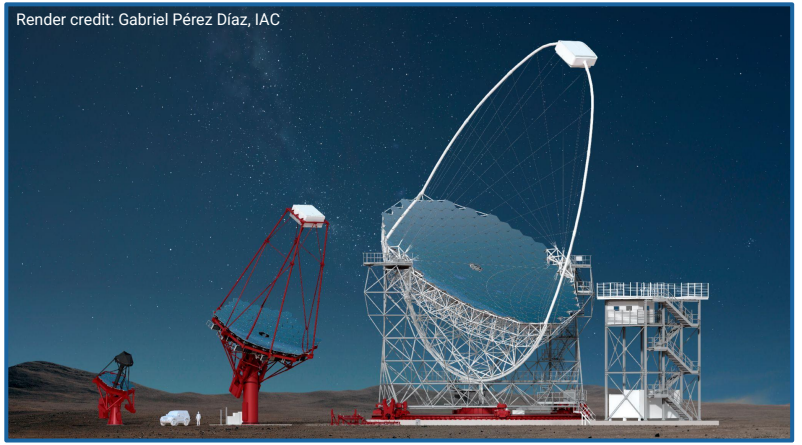
Comunidad de Madrid

QUASAR SCIENCE RESOURCES, S.L.

Artemisa
ARTificial Environment for ML and Innovation in Scientific Advanced Computing

The Cherenkov Telescope Array Observatory

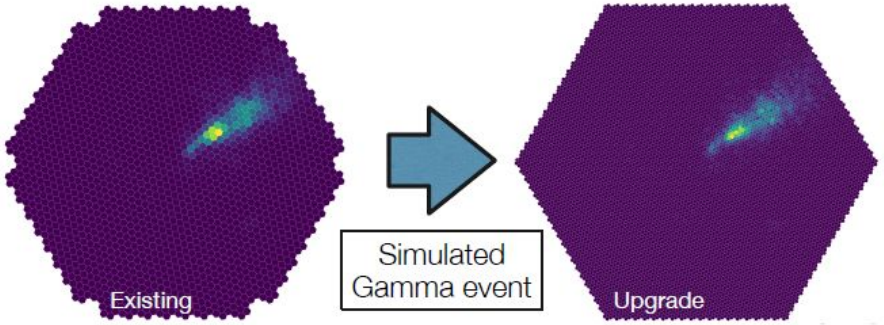
- World's largest ground-based observatory for very-high-energy gamma-ray astronomy
- Telescopes detect Cherenkov light flashes produced by gamma rays interacting with the atmosphere
- Energy range: 20 GeV to 300 TeV
  - Covered by three telescope classes (LST, MST, SST)
- LSTs target the low-energy range → the faintest Chrenkow showers
  - 4 telescopes at CTAO-North (La Palma) and 3 at CTAO-South (Chile)
- LST-1 in commissioning and producing early-science results

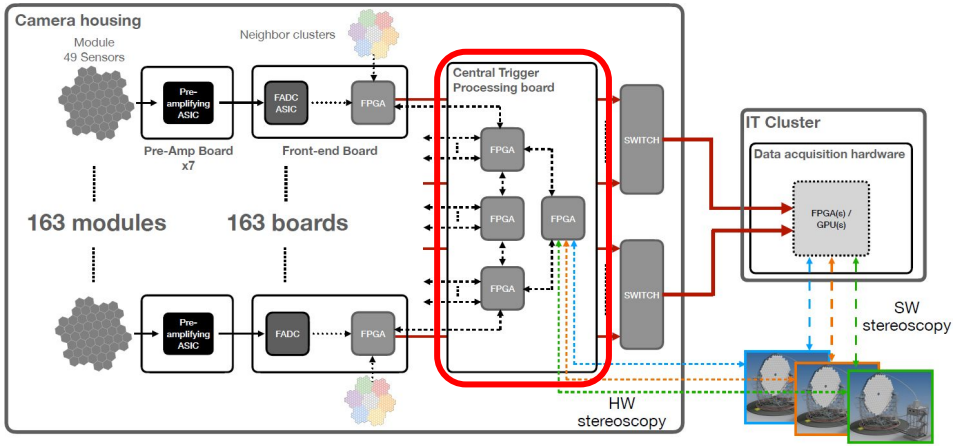


M. Molina mmolin11@ucm.es

The future LST Advanced Camera

- Mid-term upgrade for CTAO telescopes
- Transition from photomultiplier tubes (PMTs) to silicon photomultipliers (SiPMs)
- Higher granularity (7987 px, 163 FEBs)
  - 4x higher resolution with smaller pixels
- Completely digital architecture
  - Replacing analog memories with a continuous digital readout at FEBs
- The data challenge
  - Input rate: 1 GHz (1 frame/ns)
  - DAQ limit: ~40 kHz
  - **Goal:** Maximize noise rejection without losing events with a multi-stage trigger



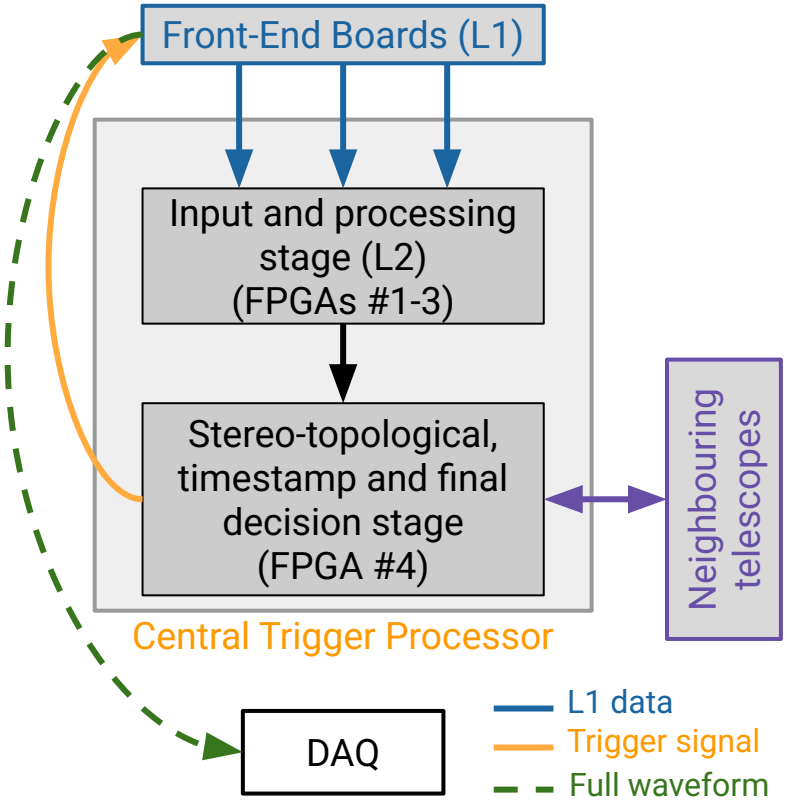


Images credit: M. Heller (UniGe)

M. Molina mmolin11@ucm.es

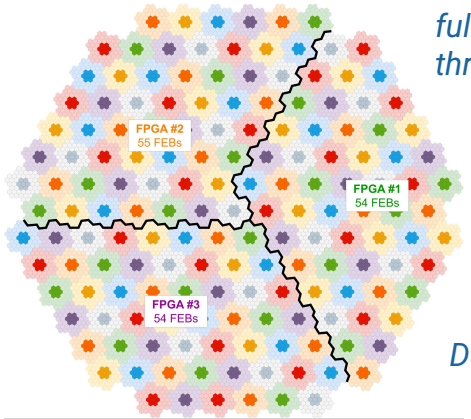
The Central Trigger Processor board (CTPb)

- **Input stage**
  - Receives L1 data @ 1GHz
  - Handles data deserialization
- **Processing stage**
  - Executes L2 algorithm in real-time
  - First rate reduction step to ~2 MHz
- **Stereo-topological trigger stage**
  - Searches for spatio-temporal coincidences with neighbour IACTs
  - Reduces mono rate to ~40 kHz
- **Final decision and readout**
  - Event timestamping
  - Trigger signal is sent to the FEBs to initiate DAQ readout



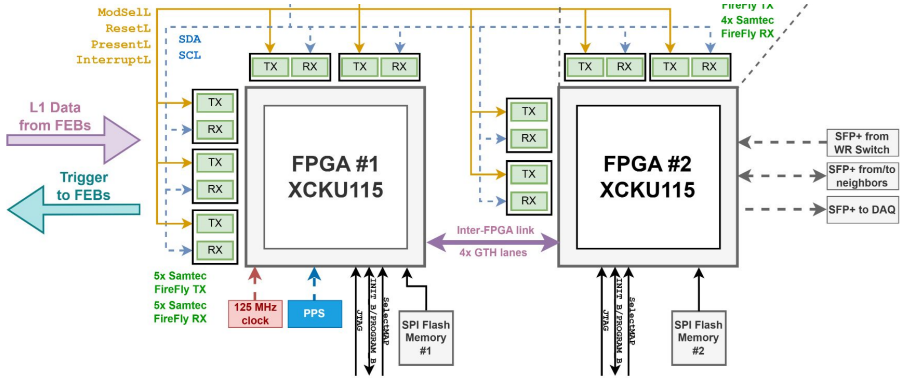
CTPb prototype: main design decisions

- The final **CTPb** design and manufacture will be very complex (and expensive)
  - 1/3-camera prototype as intermediate step
  - Prototype validation will demonstrate scalability
- Two **XCKU115** FPGAs: L2 processing (FPGA#1) and WR timing + stereo trigger (FPGA#2)
  - Selected for final design → high GTH transceiver count and enough resources for CNN inference
- FPGA#2 → second FireFly interface
  - Replicates final system input stage to test inter-sector data exchange



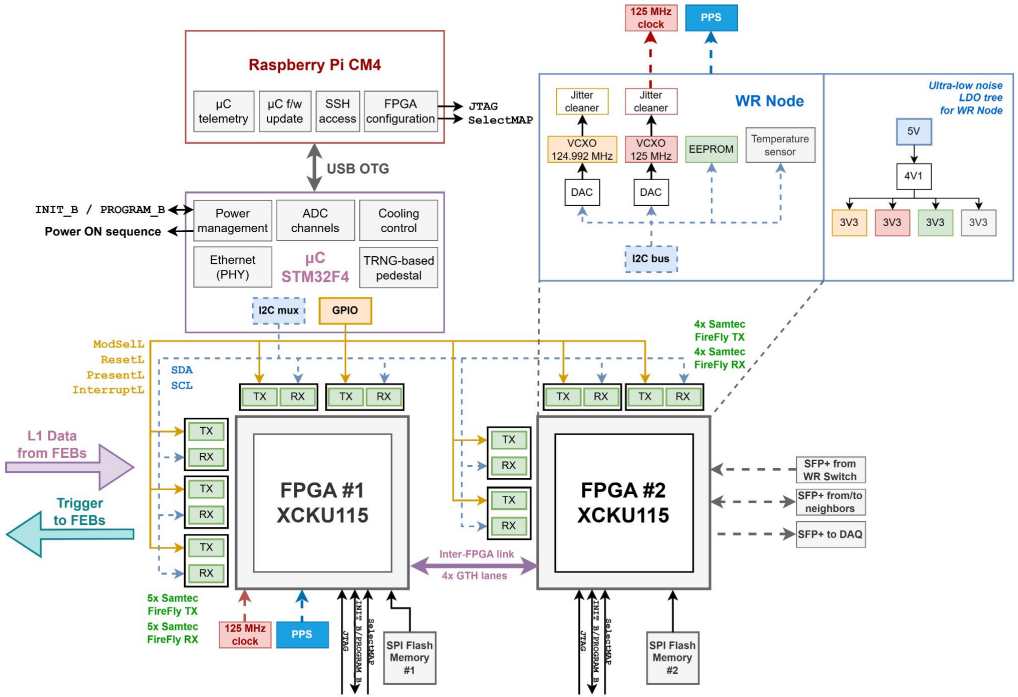
full camera three-sector splitting

Dual-FPGA prototype design (detail)



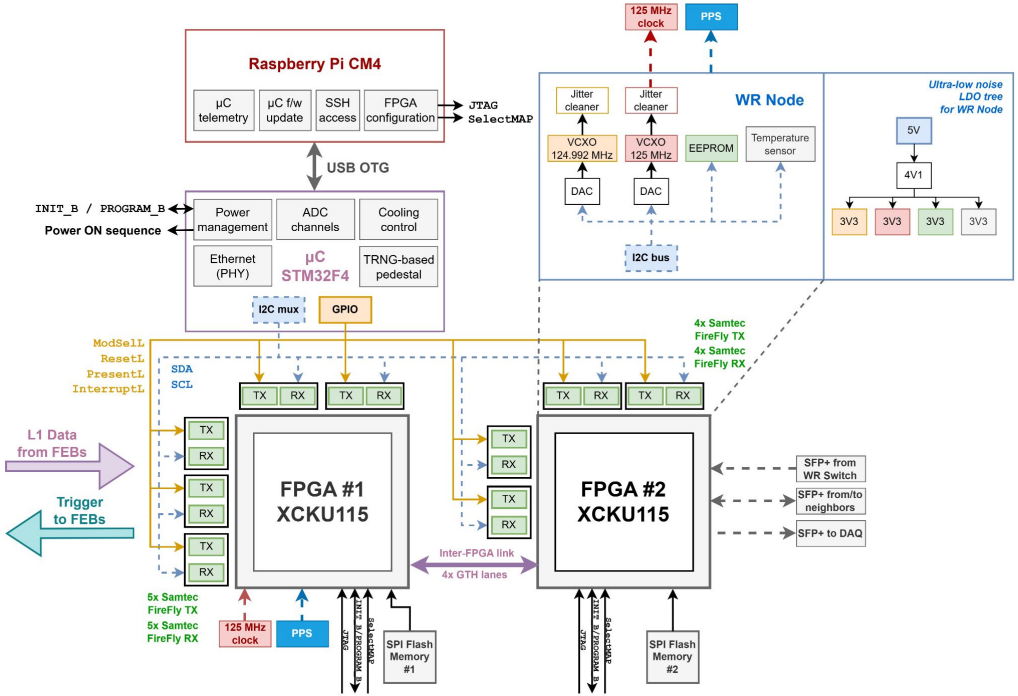
CTPb prototype: data processing architecture

- ➔ FPGA#1: L1 reception and L2 trigger
  - ◆ Receives L1 from up to 60 FEBs
  - ◆ Returns trigger decision
- ➔ FPGA#2: WR timing and stereo trigger
  - ◆ Runs stereo-topo trigger algorithm and timestamping
  - ◆ Additional SFP+ ports
  - ◆ Up to 48 additional FEBs to replicate final CTPb input stage
- ➔ Inter-FPGA link
  - ◆ 4x GTH lanes at 16 Gbps
- ➔ White Rabbit node
  - ◆ Dedicated hardware for low-jitter and stable synchronization



CTPb prototype: control, configuration and power

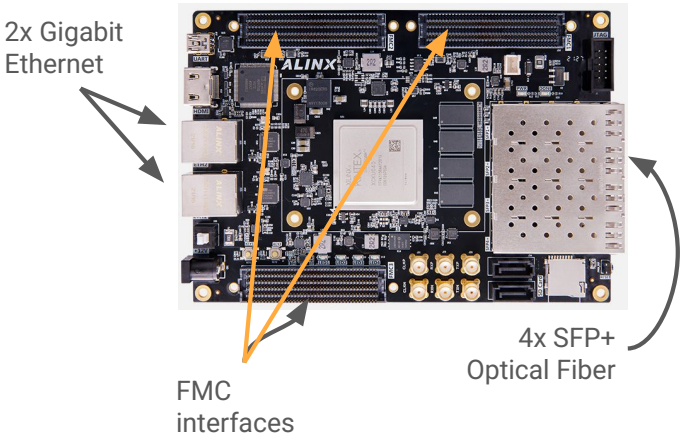
- **Raspberry Pi CM4**
  - Remote access via SSH
  - STM32F4 firmware updates
- **STM32F4 (BMC)**
  - Power sequencing and monitoring
  - FireFly diagnostics via I2C
  - Temperature check + fan control
- **Boot modes**
  - Autonomous: QSPI Flash
  - Remote: SelectMAP (via CM4)
  - JTAG daisy chain for debugging
- **Power budget → ~192 W**
  - XPE with 80% utilization, all GTH lanes active → ~159 W
  - 18 FireFly modules → ~33 W



M. Molina mmolin11@ucm.es

Validating the design: testbenches summary

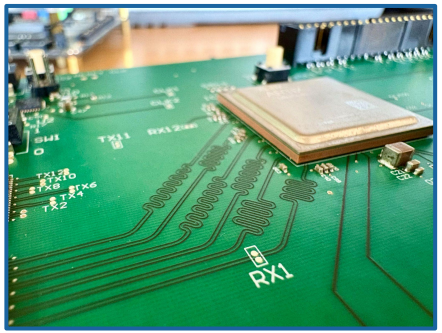
#1 Machine Learning @ FPGAs

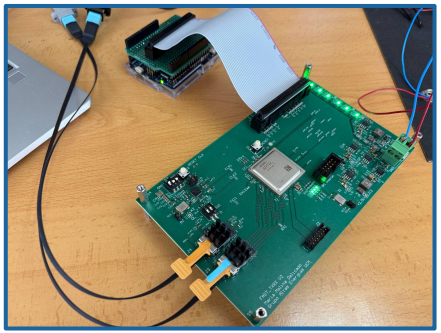


➔ 2x ALINX AMD Xilinx Kintex US **XCKU040**

- 20 gigabit transceivers @ 16.3 Gbps
- 4GB high-speed DDR4 RAM

#2 High-speed lines testing (FastTXRX)





➔ Main components

- Xilinx Artix US+ with 12 gigabit transceivers
- Two 12-channel Samtec FireFly optical connectors (TX and RX)

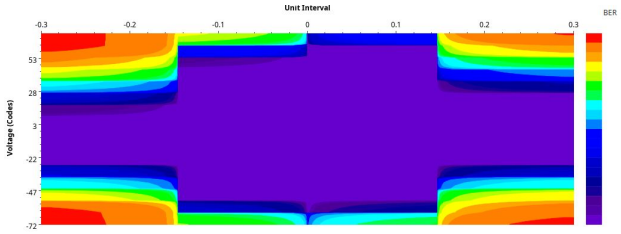
➔ Design and manufacturing

- 12 layer PCB design and high-speed differential pair routing
- Validation of the substrate and the PCB manufacturer

Signal integrity tests results

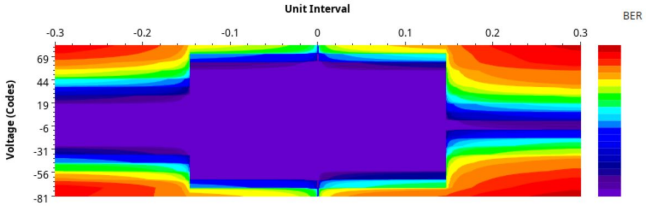
➔ Raw PRBS data at 10.3125 Gbps

- Best case: 2 m MTP cable, 2 connectors
- Worst case: 3x 30 m segments, 2 breakout boxes, 4 connectors
- BER ≤ 10^{-13} in both cases



Top: best case
Bottom: worst case
Outermost BER contour 10^{-12}

Eye-diagrams (IBERT)

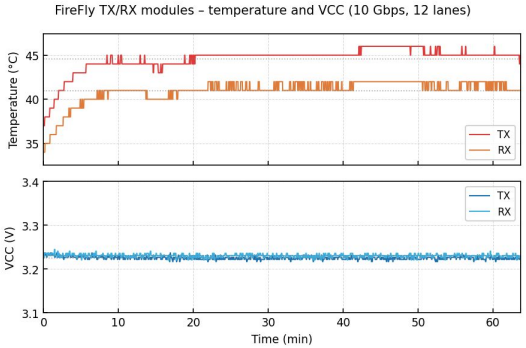


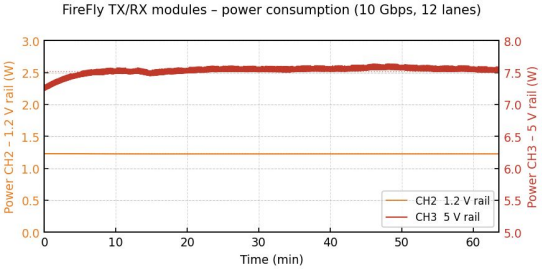
➔ 300 m coil test

- Minimum achievable BER 10^{-10}
- Limited by OM3 modal dispersion at this distance and data rate

➔ Optical power budget

- OM3 fiber attenuation: 3.5 dB/km at 850 nm, 0.75 dB per connector
- Worst-case loss: 6.315 dB, maximum reach: 290 m





M. Molina mmolin11@ucm.es

CTPb ↔ FEBs link protocol evaluation

- ➔ Candidates evaluated at 10.3125 Gbps, 64B/66B encoding, no FEC
  - Raw PRBS (baseline), Aurora 64b/66b, JESD204C
  - BER ≤ 10^{-13} in all cases
- ➔ Aurora 64b/66b
  - Up to 2-cycle latency uncertainty
- ➔ JESD204C Subclass 1
  - Deterministic latency confirmed across all fiber configurations
  - Pipeline latency constant at 6 clock cycles for all fiber lengths
  - Buffer level increases monotonically with fiber length

BER test setup

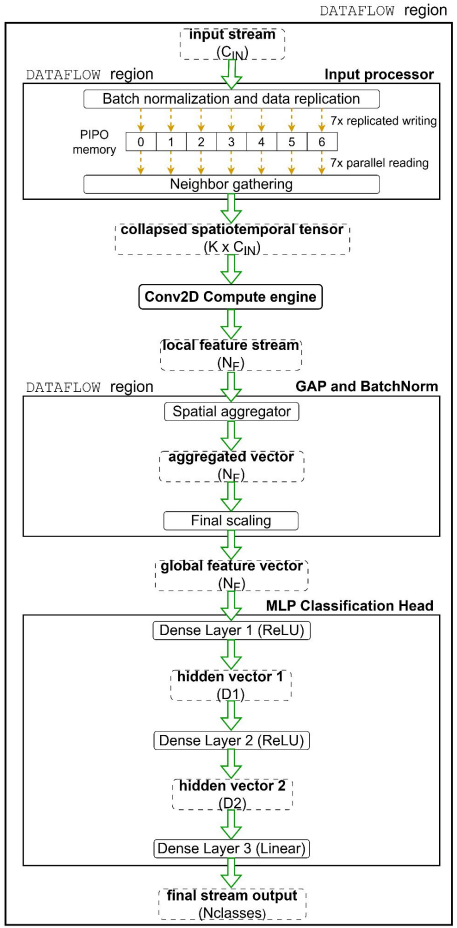
Optical fibers (left), breakout boxes (center) and FastTXRX custom board (right)

➔ **JESD204C** selected as preferred protocol

M. Molina mmolin11@ucm.es

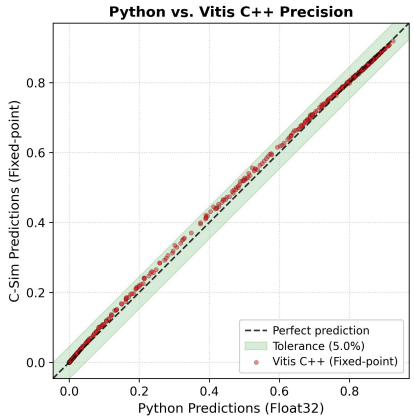
CNNs @ **CTPb**: custom Vitis HLS implementation

- **Small and custom models**
  - Required for hexagonal geometry
- **2D CNN on L1 trigger data**
  - SW ↔ HW co-design
- **Custom Vitis HLS DATAFLOW pipeline**
  - Input processor stage normalize each pixel and gather its 7 neighbors
  - Conv2D engine processes each pixel's neighborhood (one per cycle)
  - GAP accumulates all 163 pixels into a single feature vector
- **Fixed-point quantization (PTQ)**
  - C-simulation vs float32 reference



HLS Vitis DATAFLOW diagram

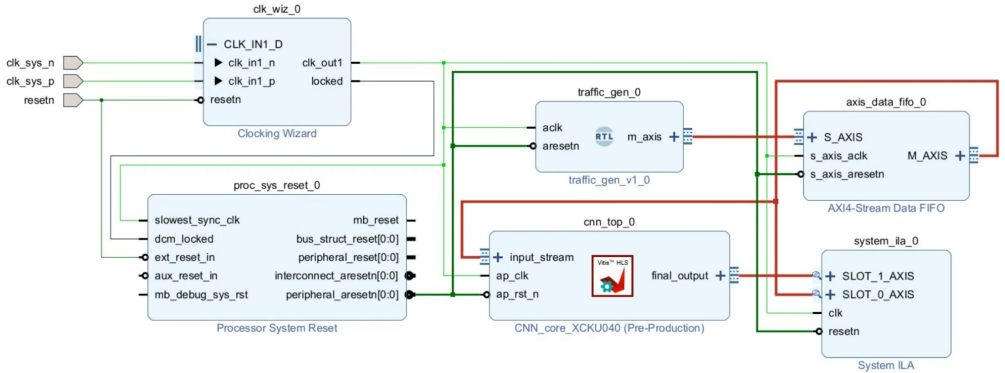
PTQ precision loss
Fixed point vs floating point



M. Molina mmolin11@ucm.es

CNNs @ **CTPb**: Hardware results (XCKU040)

- **Pipeline latency: 630–634 clock cycles**
  - 4-cycle variation from PIPO buffer
  - 1.91–1.92 µs at 330 MHz
- **Sustained throughput: 1.94 Mevents/s**
  - Initiation Interval II = 170 cycles
  - L = 163 FEB inputs + 7-cycle gap
- **Resource utilization**
  - DSP is the limiting resource
  - Conv2D → most DSP demanding
- **Power consumption**
  - CNN IP core: 3.846 W
  - 1.981 µJ per inference



Vivado Block Design of the hardware-in-the-loop system

Layer	LUT (%)	FF (%)	BRAM (%)	DSP (%)
Input proc.	1273 (0.5)	1568 (0.3)	73.5 (12.2)	20 (1.0)
Conv2D	9901 (4.1)	7743 (1.6)	2.5 (0.4)	1072 (55.8)
Pool. & BN	956 (0.4)	1394 (0.3)	1.0 (0.2)	9 (0.5)
Dense	2328 (1.0)	6997 (1.4)	0 (<0.1)	104 (5.4)
Others	1662 (0.7)	2762 (0.6)	0 (<0.1)	0 (<0.1)
TOTAL	**16120 (6.7)**	**20464 (4.2)**	**77 (12.8)**	**1205 (62.8)**

XCKU040 Resource utilization from the CNN implementation

Disclaimer: *Each FPGA of the final CTPb will process ⅓ of camera (3x throughput) and will have much more resources (XCKU115) → enough headroom for increase parallelization*

M. Molina mmolin11@ucm.es

CNN benchmarking (FPGA/CPU/GPU) and hls4ml custom layers implementation

- **FPGA** obtains the best results in latency, throughput and power consumption
- **CPU** → batch of 10,000 samples
- **GPU** → 1,000,000 samples in chunks of 100,000 (OOM above) to simulate streaming
- Direct comparison is limited
  - CPU/GPU operate in batch mode, FPGA as a continuous pipeline
  - Model too small to saturate GPU resources

Metric	CPU i7-1255U	GPU RTX 3090	FPGA XCKU040
Median latency	0.366 ms	0.376 ms	1.91–1.92 μs
Energy/inf (sample)	3.12 mJ	N/M[a]	N/A[b]
Throughput	42 474 inf/s	37 270 inf/s	1.94 Minf/s
Time per sample	0.024 ms	0.027 ms	0.52 μs
Energy/inf (batch)	330 μJ	134 μJ	1.98 μJ
Active power	~14 W	~5 W	3.846 W
System baseline	5.1 W	135.6 W	0.523 W
Total board power	~19 W	~141 W	5.348 W
Throughput speedup	1×	0.88×[c]	~46×
Energy efficiency gain	1×	0.41×[c]	~167×

- Three custom Keras layers implemented via **hls4ml Extension API**
  - *NeighborGatherLayer*, *IndexedConvolutionLayer* and *IndexedPoolingLayer*
- Automatic `ap_fixed` precision selection, no manual tuning required
- Work in progress
  - `io_stream` backend with DATAFLOW pipeline

M. Molina mmolin11@ucm.es

Conclusion

- ➔ 1/3-camera CTPb prototype designed as intermediate validation step towards the full system
- ➔ Optical links validated at 10 Gbps with BER ≤ 10^{-13} over fiber lengths representative of the final installation
- ➔ JESD204C identified as preferred protocol for deterministic latency
- ➔ CNN implemented on FPGA with deterministic latency of 1.91–1.92 µs and sustained throughput of 1.94 Mevents/s
- ➔ Custom layers for hexagonal geometry implemented in hls4ml *(ongoing)*

Outlook

- ➔ System demonstrator → full trigger data path validation with real FEB data (FastTXRX + AXKU040)
  - ◆ New FastTXRX board (*ongoing*) integrating WR node hardware
- ➔ 1/3-camera prototype
  - ◆ Schematics and PCB layout under development
  - ◆ Signal integrity and thermal simulations before manufacturing
- ➔ Current implementation processes one pixel per clock cycle (II governed by L)
  - ◆ Next step: increase CNN throughput for XCKU115 processing several pixels in parallel

M. Molina mmolin11@ucm.es

Acknowledgements

The research here presented has been partially supported by the MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, under grant PDC2023-145839-I00.
Also by grant *PID2022-138172NB-C42* funded by MICIU/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe",

This research is co-funded by the Community of Madrid under the call for grants for the implementation of industrial PhD programmes, awarded by Order 2637/2025 of 17 July, issued by the Regional Minister for Education, Science and Universities, reference IND2024/TIC-34250

The author(s) gratefully acknowledges the computer resources at Artemisa, funded by the European Union ERDF and Comunitat Valenciana as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV).
This work was supported in part by the AMD University Program.