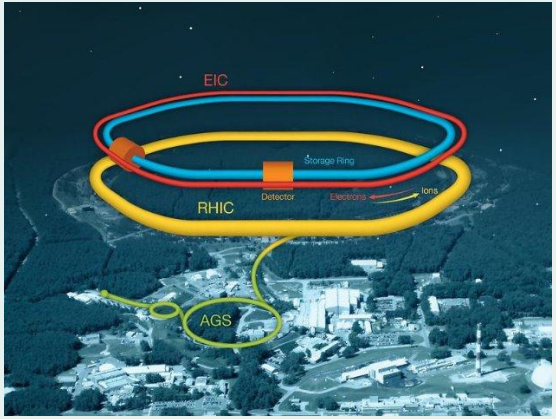
25th IEEE Real Time Conference,
28th May 2026

Online Data Reduction for the ePIC dRICH Using a Multi-FPGA Neural Network

Cristian Rossi
(INFN Sezione di Roma, APE Lab)

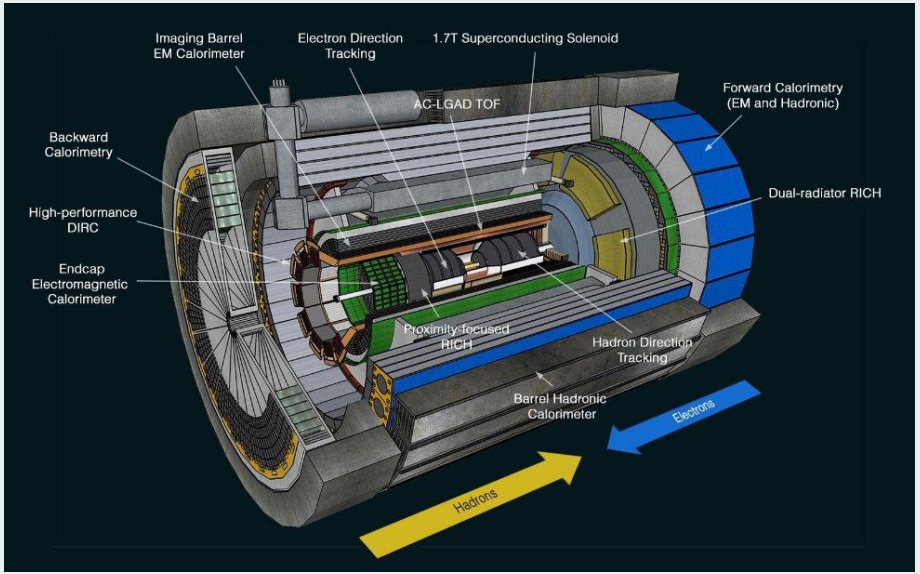
INFN
Istituto Nazionale di Fisica Nucleare

EIC ePIC: Overview



The **ePIC collaboration** currently consists of almost 500 members from 171 institutions and is working jointly with the DOE **EIC Project** to realize the ePIC detector.

ePIC will be an ~10-meter long cylindrical barrel **detector** with additional instrumentation that extends to up to 45m in each direction down the EIC beamline.



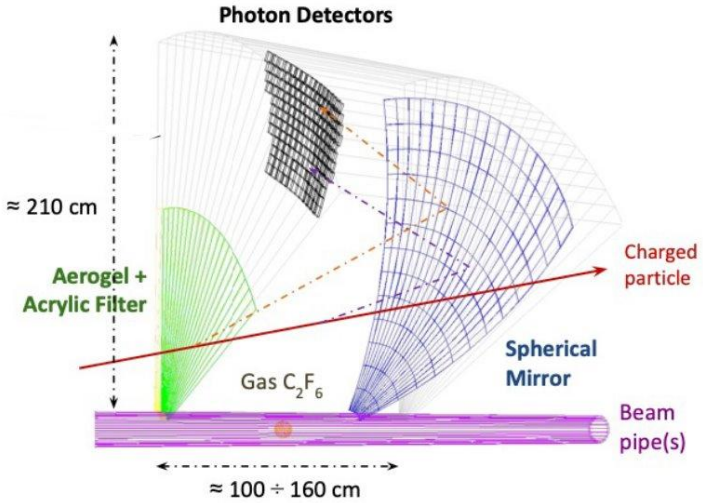
- A 1.7 Tesla superconducting magnet
- High-precision silicon detectors for particle tracking
- Precise calorimeters for measuring particles electromagnetic energy
- A suite of **particle identification (PID) detectors**
- Dense calorimetric detectors to allow the measurement of “jets“

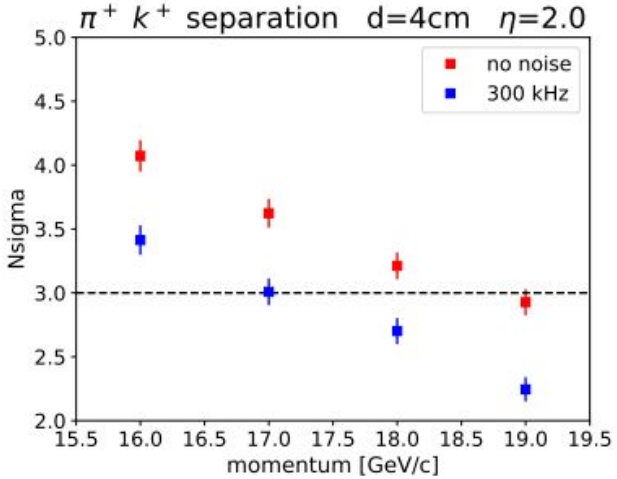
dRICH: Design and PID

- A **d**ual **R**ing **I**maging **CH**erenkov detector **(dRICH)** will be employed in the forward region **(1.5 < η < 3.5)** to provide efficient **hadron PID** from 3 GeV/c to 50 GeV/c .

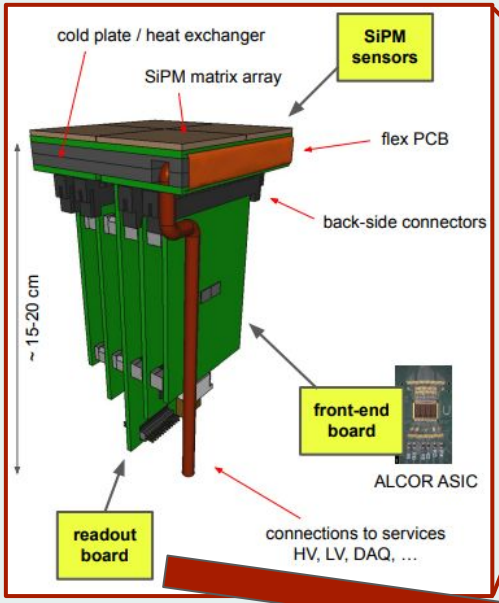
- The dRICH comprises two different radiators, aerogel and gas ($C2\ F6$), to cover the entire momentum range.

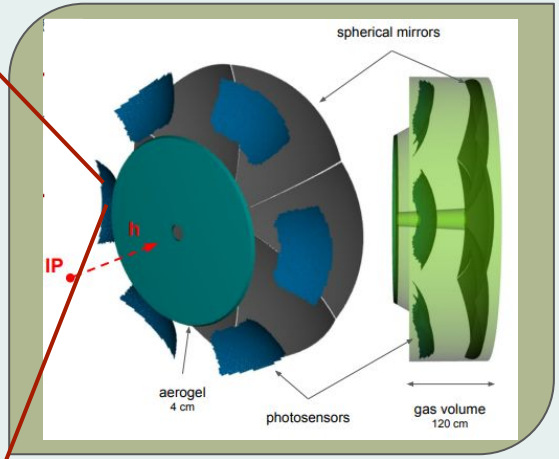
- **SiPM based photosensors** are placed in six spherical sectors to detect Cherenkov photons which are focused by six corresponding spherical mirrors

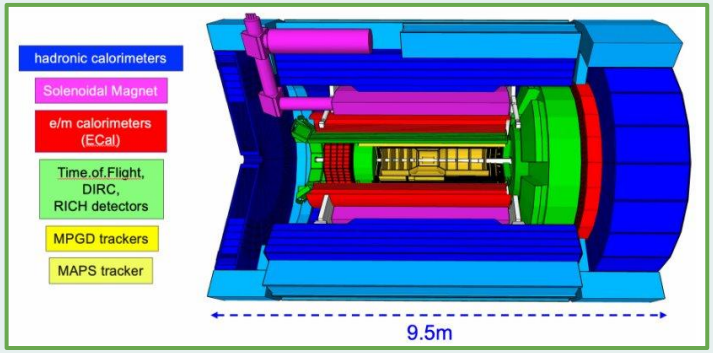




dRICH ⇒ RDO and ePIC DAQ







Forward region

- 1 photodetector unit **PDU**: 4x64 SiPM array device (**256 channels**), **4 FEBs, 1 RDO**
- **1248 PDUs for full dRICH readout**
- **319488 readout channels** divided in six sectors

FELIX

DAM

Data Aggregation Module

Assembled FLX-155

- 42 links from PDUs to Felix-155 board
- 30 Felix-155 boards in total

SERVER

ePIC processing and storage system

bandwidth/throughput issue

Analysis of Output Bandwidth

- Sensors DCR: 3-300 kHz (**increasing with radiation damage** ⇒with experiment lifetime).
- Considering planned techinques to manage SiPMs irradiation (e.g. annealing):
  - **worst DCR case: 300 kHz**

➔ Full detector throughput (FE): **6.792,19 Gbps**

➔ a **reduction is needed** to cope with 30 channel (30x100GbE) bandwidth availability

➔ Single DAM output bandwidth : **226,41 Gbps**

dRICH DAQ parameters	
RDO boards	1248
ALCOR64 x RDO	4
dRICH channels (total)	319488
Number of DAM	30
Input link in DAM	42
Output links from DAM to TP	1
Number of DAM Trigger Processor	1
Input link to DAM Trigger Processor	30
RDO-DAM Link Bandwidth (VTRX+) [Gb/s]	10
DAM to Echelon-0 Switch Bandwidth [Gb/s]	100
dRICH Interaction tagger reduction factor	1
Interaction tagger latency [s]	1,00E-04
EIC parameters	
EIC Clock [MHz]	98,522
Orbit efficiency (takes into account gap)	0,92

dRICH data stream analysis		Limit
Sensor rate per channel [kHz]	300,00	4.000,00
Rate post-shutter [kHz]	276,00	800,00
Throughput to serializer [Mb/s]	172,50	788,16
Throughput from ALCOR64 [Mb/s]	1.380,00	
Throughput from RDO [Gb/s]	5,39	10,00
Input at each DAM [Gbps]	226,41	420,00
Buffering capacity at DAM [Mb]	23,18	
Output from each DAM [Gbps]	226,41	100,00
Aggregated dRICH data throughput		
Total input at DAM [Gb/s]	6.792,19	
Total output from DAM [Gb/s] to Echelon	6.792,19	

data reduction system

Analysis of Output Bandwidth

- Sensors DCR: 3-300 kHz (**increasing with radiation damage** ⇒with experiment lifetime).
- Considering planned techinques to manage SiPMs irradiation (e.g. annealing):
  - **worst DCR case: 300 kHz**

➔ Full detector throughput (FE): **6.792,19 Gbps**

➔ a **reduction is needed** to cope with 30 channel (30x100GbE) bandwidth availability

- EIC beams bunch spacing: ~10 ns ⇒ bunch crossing rate of 100 MHz
- For the low interaction cross-section (DIS) ⇒ one interaction every ~100 bunches ⇒ interaction rate of ~1MHz.

➔ A system tagging dark current noise-only events can solve the throughput issue (reducing down to 1/5 the data throughput)

➔ Single DAM output bandwidth : **45,64 Gbps**

dRICH DAQ parameters	
RDO boards	1248
ALCOR64 x RDO	4
dRICH channels (total)	319488
Number of DAM	30
Input link in DAM	42
Output links from DAM to TP	1
Number of DAM Trigger Processor	1
Input link to DAM Trigger Processor	30
RDO-DAM Link Bandwidth (VTRX+) [Gb/s]	10
DAM to Echelon-0 Switch Bandwidth [Gb/s]	100
dRICH Interaction tagger reduction factor	5
Interaction tagger latency [s]	1,00E-04
EIC parameters	
EIC Clock [MHz]	98,522
Orbit efficiency (takes into account gap)	0,92

dRICH data stream analysis		Limit
Sensor rate per channel [kHz]	300,00	4.000,00
Rate post-shutter [kHz]	276,00	800,00
Throughput to serializer [Mb/s]	172,50	788,16
Throughput from ALCOR64 [Mb/s]	1.380,00	
Throughput from RDO [Gb/s]	5,39	10,00
Input at each DAM [Gbps]	226,41	420,00
Buffering capacity at DAM [Mb]	23,18	
Output from each DAM [Gbps]	45,28	100,00
Aggregated dRICH data throughput		
Total input at DAM [Gb/s]	6.792,19	
Total output from DAM [Gb/s] to Echelon	1.358,44	

Data reduction with ML Classifier

Online **Signal/Noise discrimination** using ML

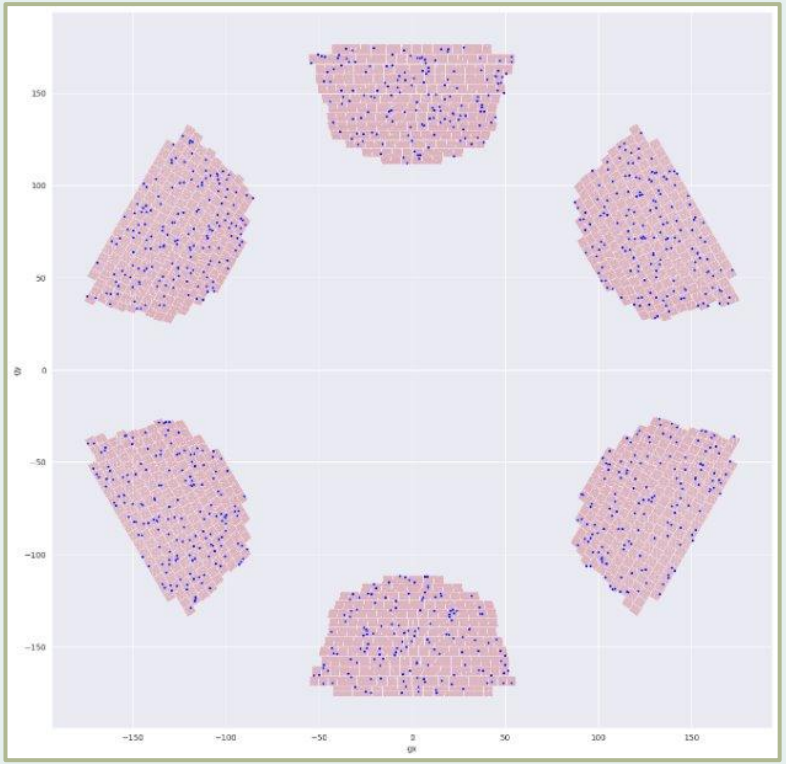
- **Signal**
  (i.e Merged Phys Signal+ Bkg)
  - **Physics Signal:**
    - ➔ e.g. DIS, SIDIS, ...
  - **Physics Background:**
    - ➔ e/p interaction with beam pipe
    - ➔ synchrotron radiation
- **SiPM Noise:**
  - Dark Count Rate (DCR) modeled on the reconstructed dataset

Discriminate between **Noise-Only** and **Signal+Noise** events

dRICH Data Reduction: Classes Definition

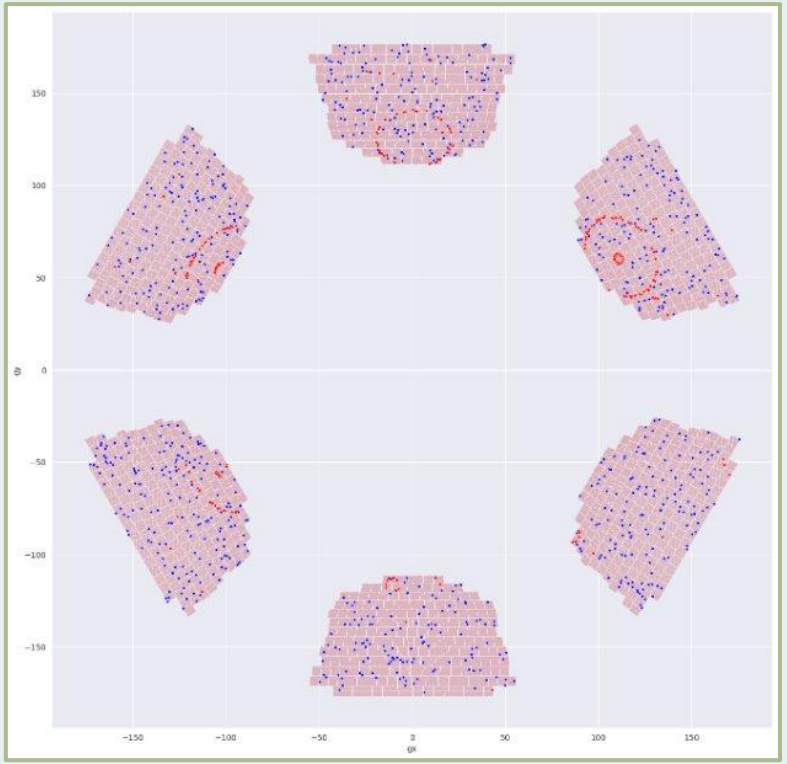
(Negative)

Noise-Only



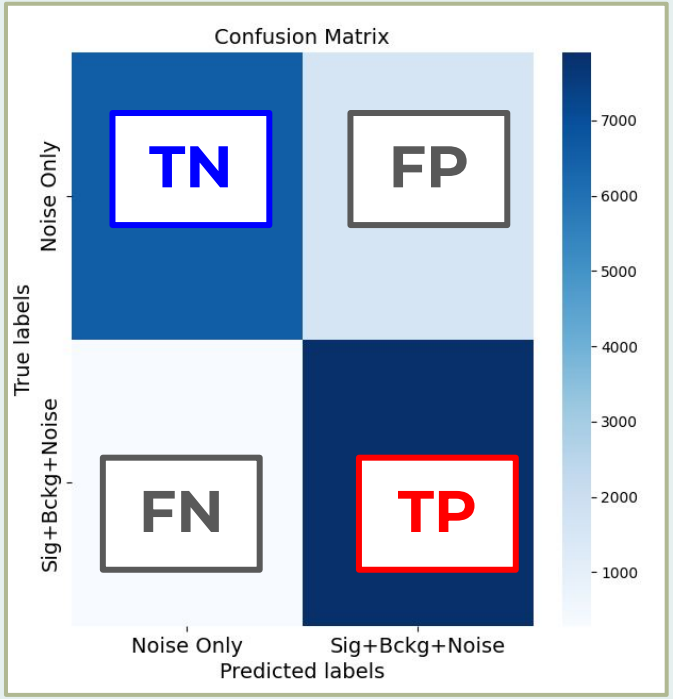
(Positive)

Signal+Background+Noise



dRICH Data Reduction: Performance Evaluation and Targets

- To evaluate the model performance, we introduce:
  - **True Positive Rate (TPR, or *Sensitivity*):** percentage of correctly classified true-labeled Signal+Background+Noise events
  - **True Negative Rate(TNR, or *Specificity*):** percentage of correctly classified true-labeled Noise-Only events

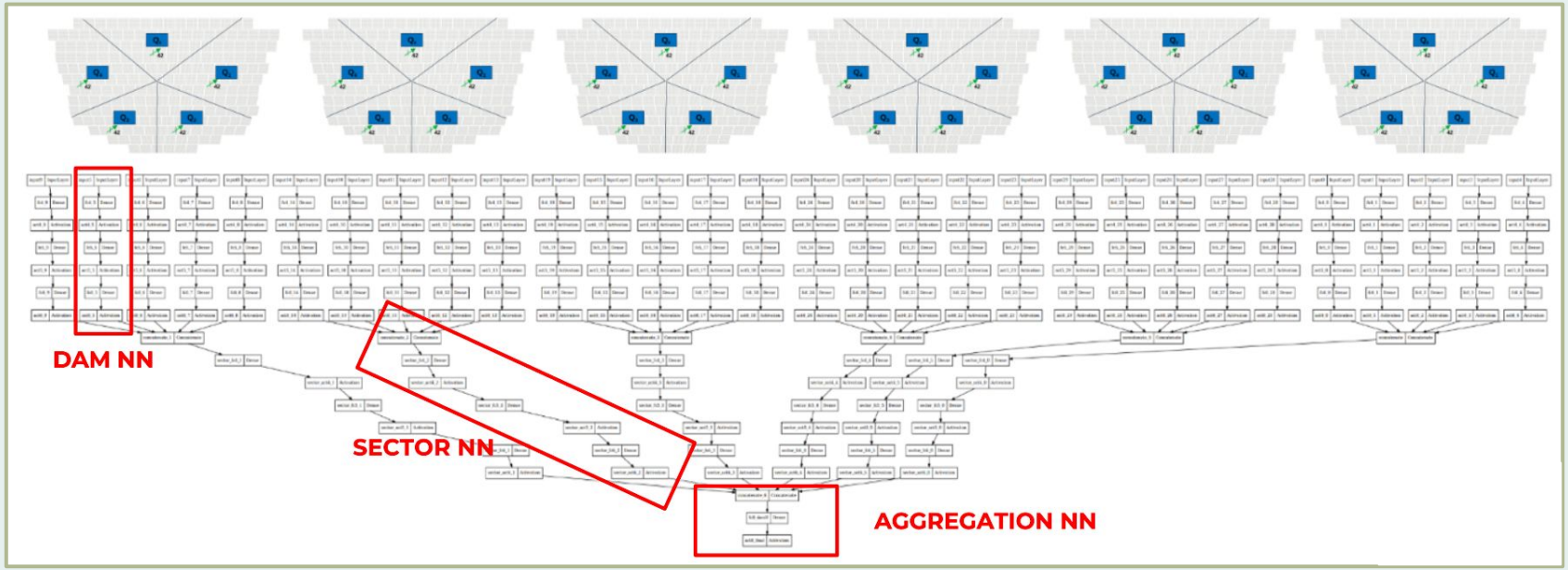


$$TPR = \frac{TP}{TP + FN}$$ as high as possible

$$TNR = \frac{TN}{TN + FP}$$ >= 80% [⇒ reduction factor >=5]

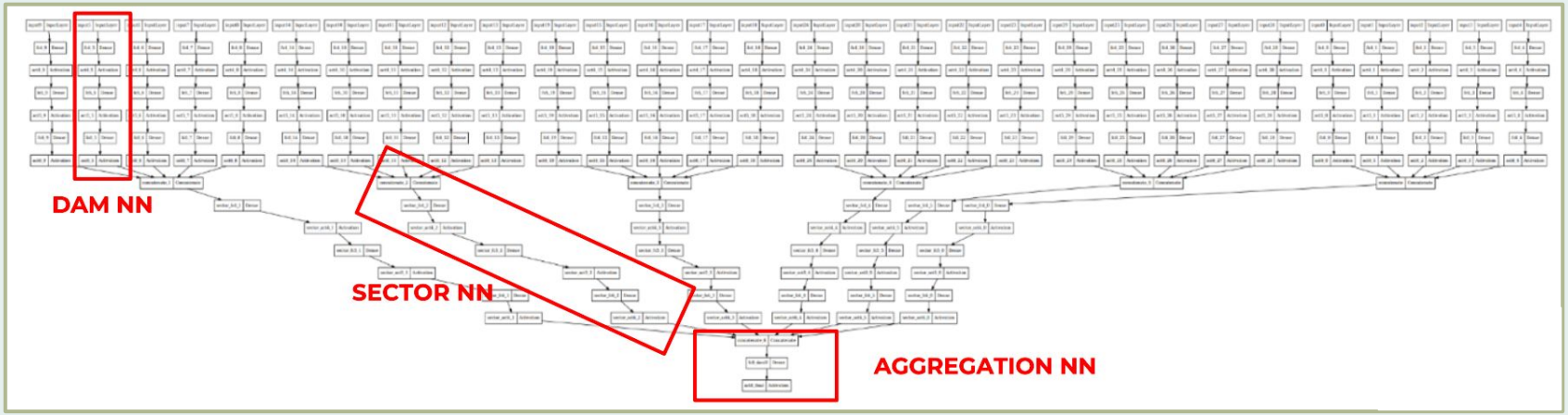
dRICH Data Reduction: Tensorflow-Keras Model Definition

- The NN model mimics the DAQ system architecture
  - **30** (# of subsectors x # of sectors) **DAM MLP networks** deployed on 30 **DAM FPGAs**.
  - **6 sector MLP networks + 1 aggregation network** deployed on the **TP FPGA**.

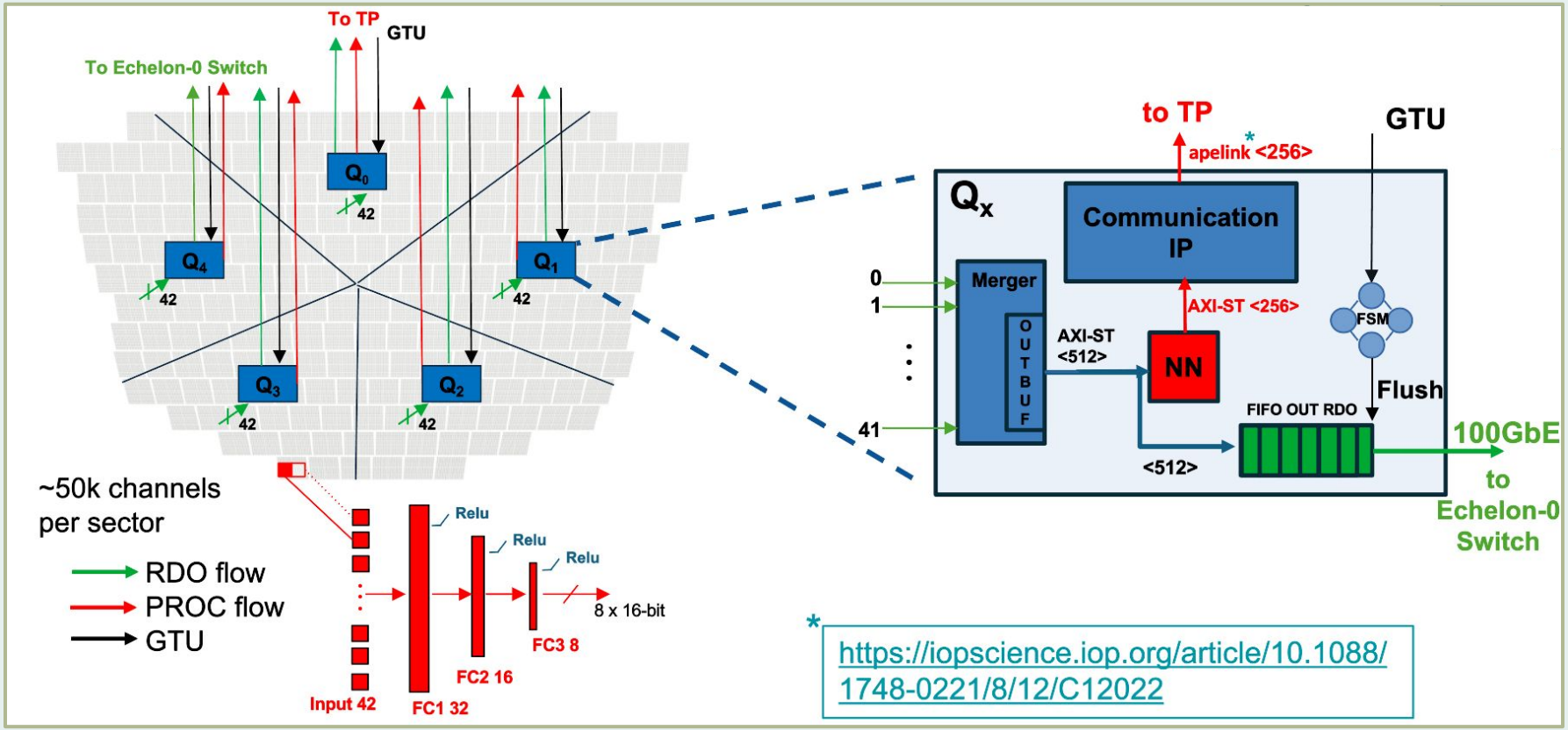


dRICH Data Reduction: Tensorflow-Keras Model Definition

- The 30 DAM networks are concatenated to feed 6 intermediate model (called **Sector NN**) to be deployed on an additional **Trigger Processor (TP) FPGA**.
- Each Sector NN work on the aggregated information of a single sector (5 DAMs).
- The 6 outputs from Sector NNs are then aggregated and processed in a **lightweight TP NN** (single layer, 5 input neurons), deployed on the same TP FPGA

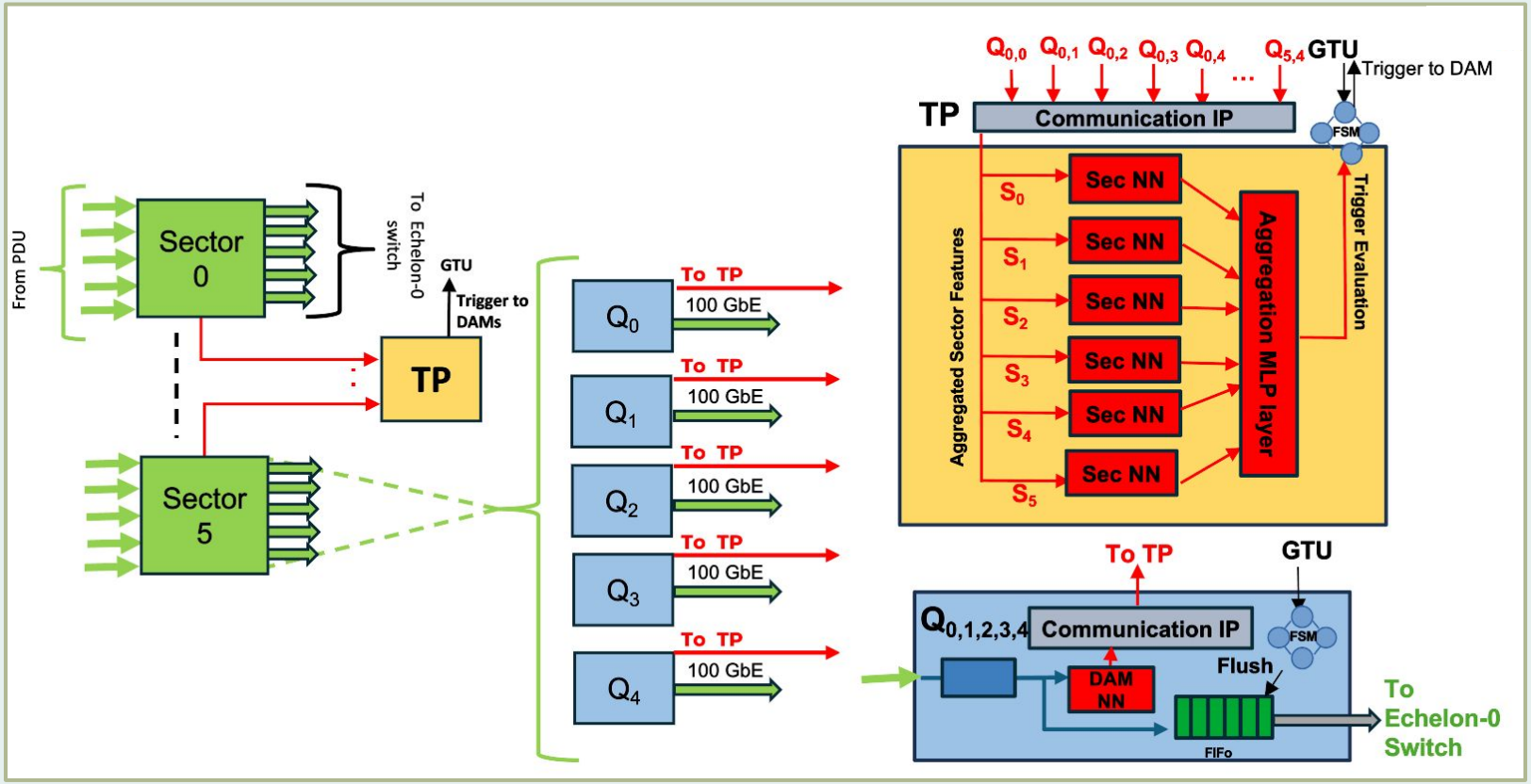


dRICH Data Reduction Stage on FPGA: Subsectors' Design

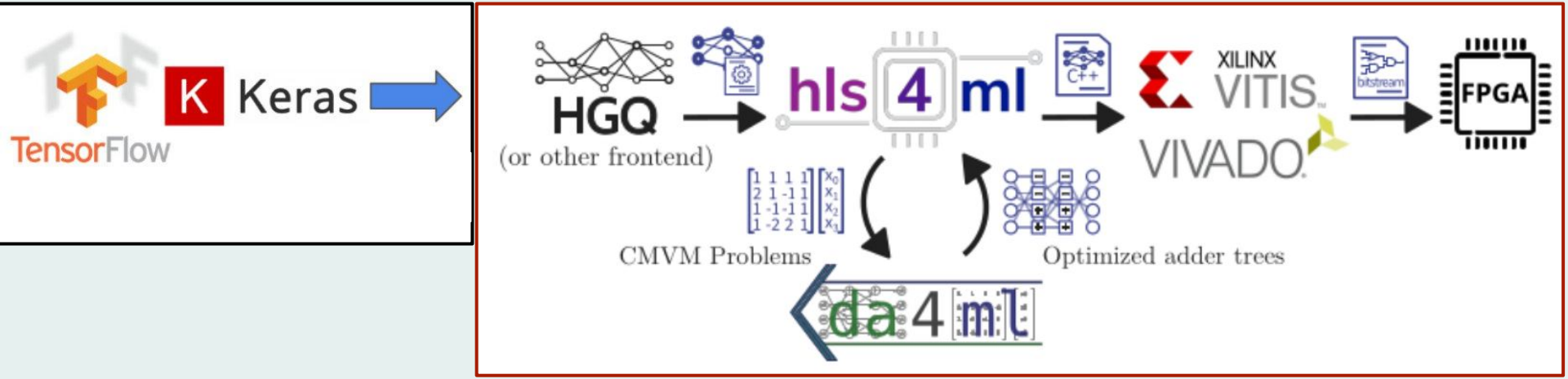


* https://iopscience.iop.org/article/10.1088/1748-0221/8/12/C12022

dRICH Data Reduction Stage on FPGA: Subsectors' Design



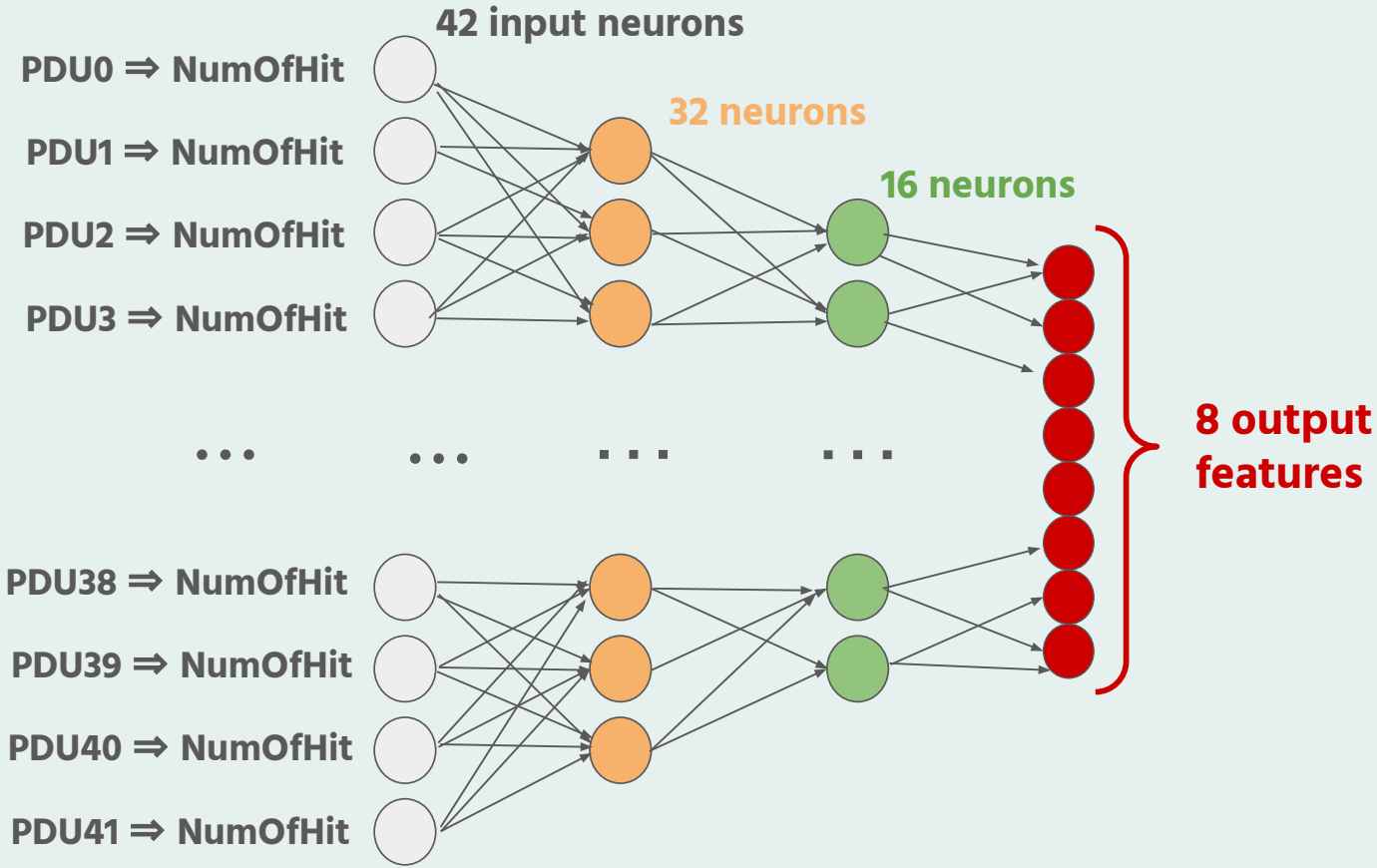
dRICH Data Reduction: How? ⇒ Design and Implementation



- **Design targets** (sensitivity, specificity, throughput, latency) and **hardware constraints** (mainly FPGA resource usage) must be taken into account and verified at any stage.
- For the **quantization step**, we aligned our system with the new state-of-the-art libraries for FPGA ML model implementation

  ⇒ **HGQ2** (High Granularity Quantization), **hls4ml**, **da4ml**

DAM Input ⇒ Subsector Input and Feature Extraction



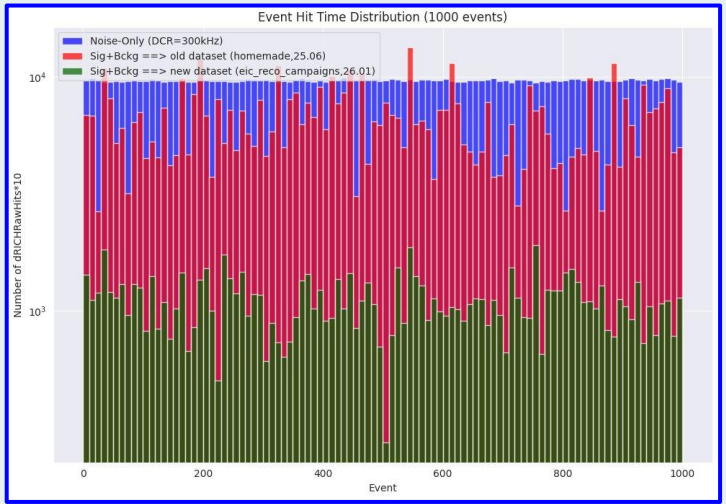
dRICH Data Reduction: Dataset Generation

- Dataset "homemade" generated using **EICrecon v25.06 ⇒ potential bias**
- To assess system performance and minimize potential dataset bias:
  ⇒ model was trained and validated on data from **EIC simulation/reconstruction campaigns**
- This ensures consistency with more recent software versions (e.g. **v26.01**)

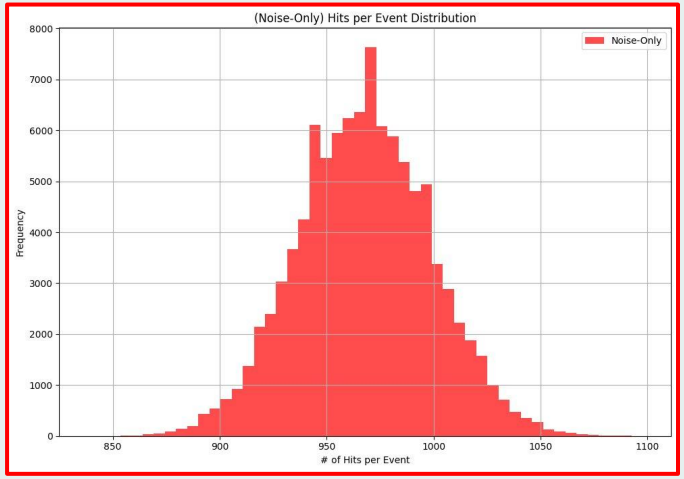
However:
⇒ average number of hits of v26.01 physics-related events seems 1 order of magnitude lower wrt v25.06
⇒ **where does it lead?**

Sig+Bckg Dataset
(generated within ePIC software framework workflow)

- **Montecarlo Events** *(Physics Sig + Physics Bg)*
- **(GEANT4) Simulation** *(ePIC detectors output)*
- **Reconstruction** *(digitization, quantum efficiency)*



dRICH Data Reduction: Dataset and Noise Generation



Sig+Bckg Dataset

(generated within ePIC software framework workflow)

Montecarlo Events

(Physics Sig + Physics Bg)

(GEANT4) Simulation

(ePIC detectors output)

Reconstruction

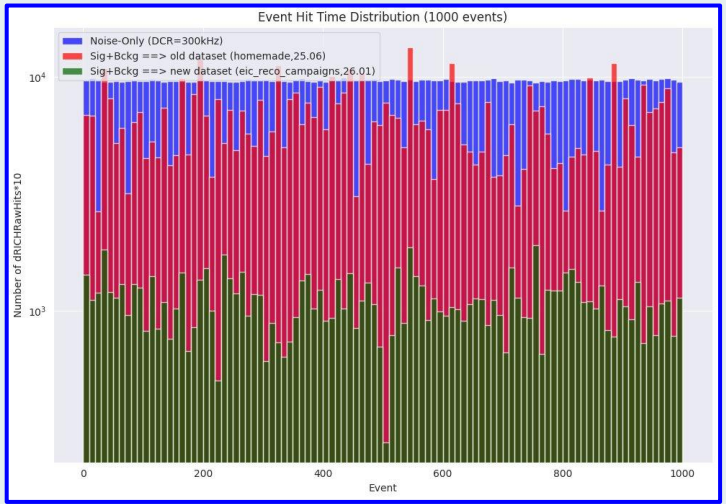
(digitization, quantum efficiency)

(Python) Noise Generation

(dRICH SiPM Dark Count)

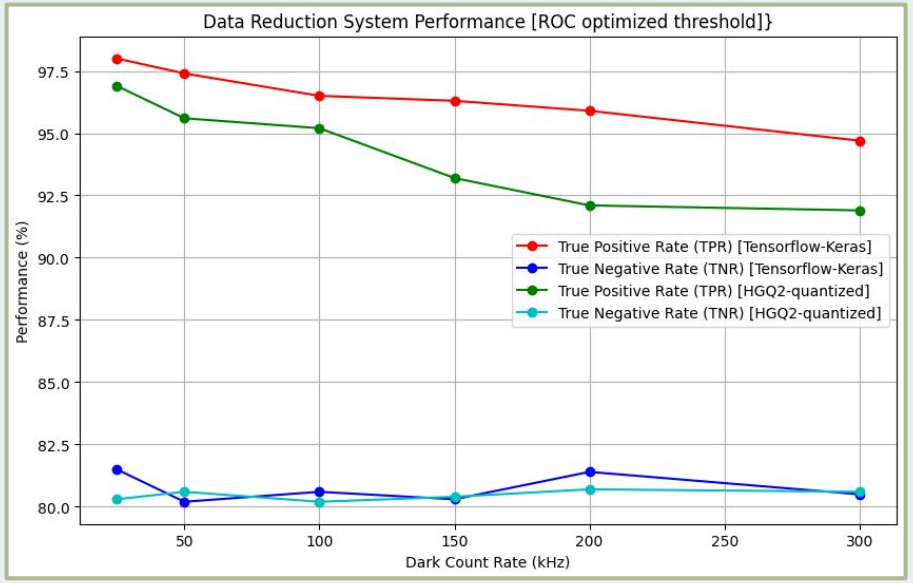
Noise-Only Dataset

Sig+Bckg+Noise Dataset



dRICH Data Reduction: Training and Validation (EICrecon v25.06)

- We trained the 30 MLP DAM models concatenated to the single MLP TP model by using 100k Signal+Background+Noise and 100k Noise-Only events.
- **200k balanced dataset** (90% training set, 8% testing set, 2% validation set) varying the Dark Count Rate parameter



➔ **Tensorflow-Keras (floating-point) model:**

⇒ drop of classification performance with increasing DCR (e.g. increasing number of noise hits per event)

⇒ TPR ~ 95% for noisiest case (DCR = 300 kHz).

➔ **(HQ2) Quantized model:**

⇒ prediction performance drop after quantization step

⇒ TPR>90% for noisiest case (DCR = 300 kHz)

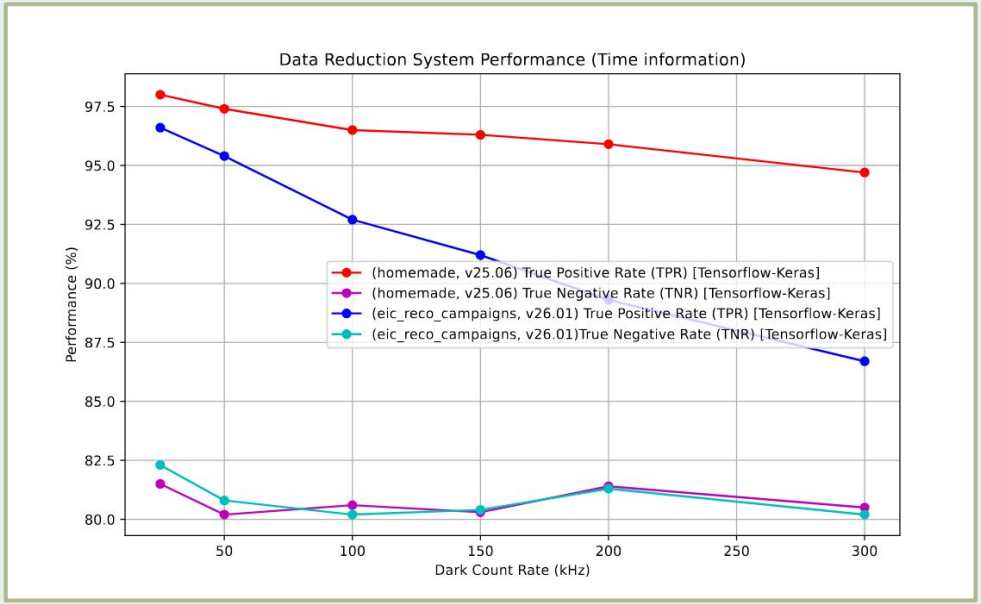
dRICH Data Reduction: Training and Validation (versions comparison)

- We made a comparison between model **TPR performance** coming from the different training with the two dataset generated with different EICrecon version

➔ **Tensorflow-Keras (floating-point) model:**

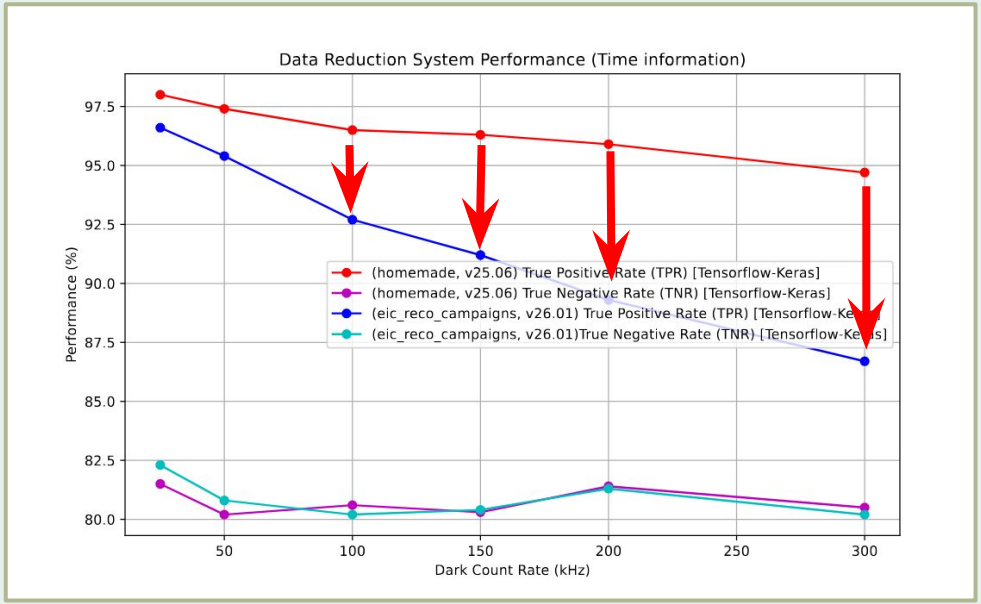
⇒ drop of classification performance with increasing DCR (e.g. increasing number of noise hits per event)

⇒ TPR ~ 87% for noisiest case (DCR = 300 kHz) for the v26.01.



dRICH Data Reduction: Training and Validation (versions comparison)

- We made a comparison between model **TPR performance** coming from the different training with the two dataset generated with different EICrecon version



➔ **Tensorflow-Keras (floating-point) model:**

⇒ drop of classification performance with increasing DCR (e.g. increasing number of noise hits per event)

⇒ TPR ~ 87% for noisiest case (DCR = 300 kHz) for the v26.01.

➔ huge drop wrt v25.06!!

➔ **how can we improve performance?**

Time-Information Input for Multi MLP model

- We start to evaluate how to implement the **timing information** to enhance the performance of the data reduction system.
- dRICH hits **timing distribution**

  ⇒ **binwidth connected** to the bit resolution available from Coarse Counter and Fine Counter

- **Time normalized to first hit in each event**

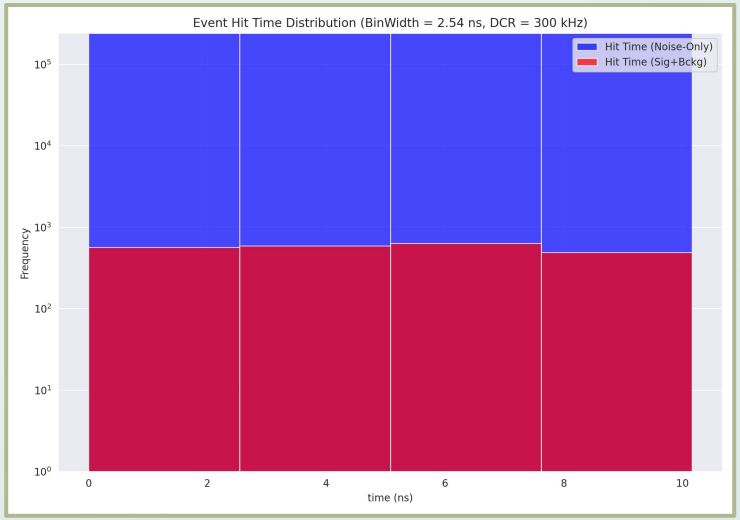
➔ **binwidth = 2.54 ns**

394 MHz → LSB = 2.54 ns

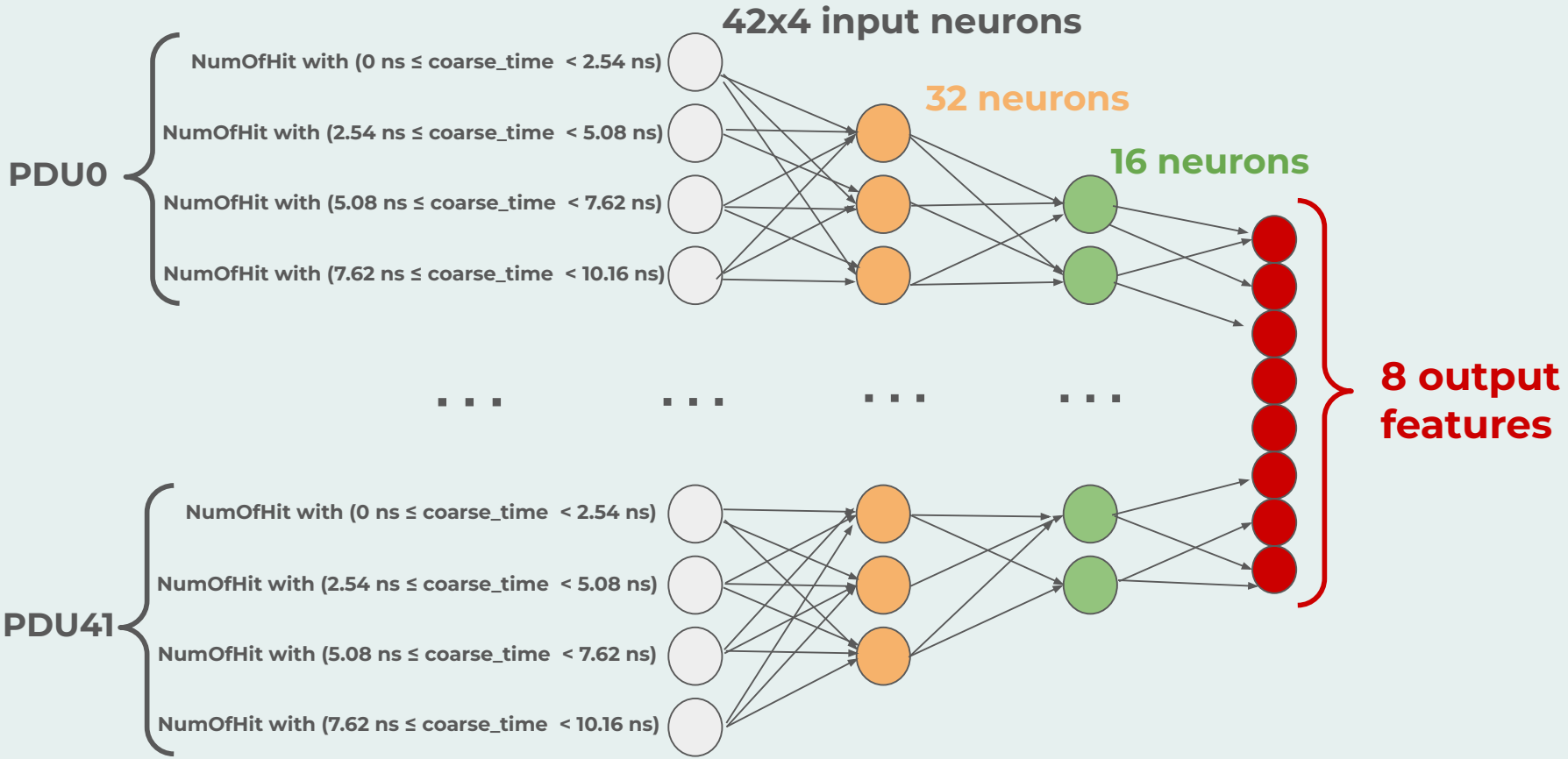
Calibration needed to use it at DAM level

31 29	28 26	25 24	23 9	8 0
Col ID	Pix ID	TDC ID	Coarse Counter	Fine Counter

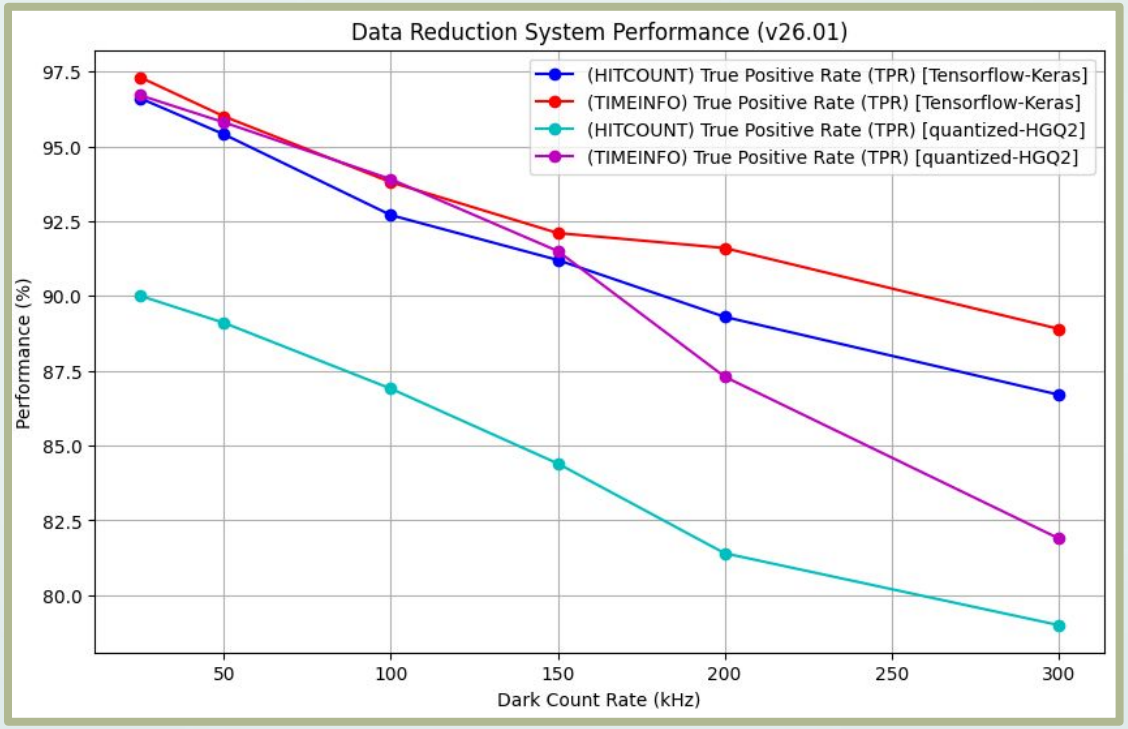
TIME



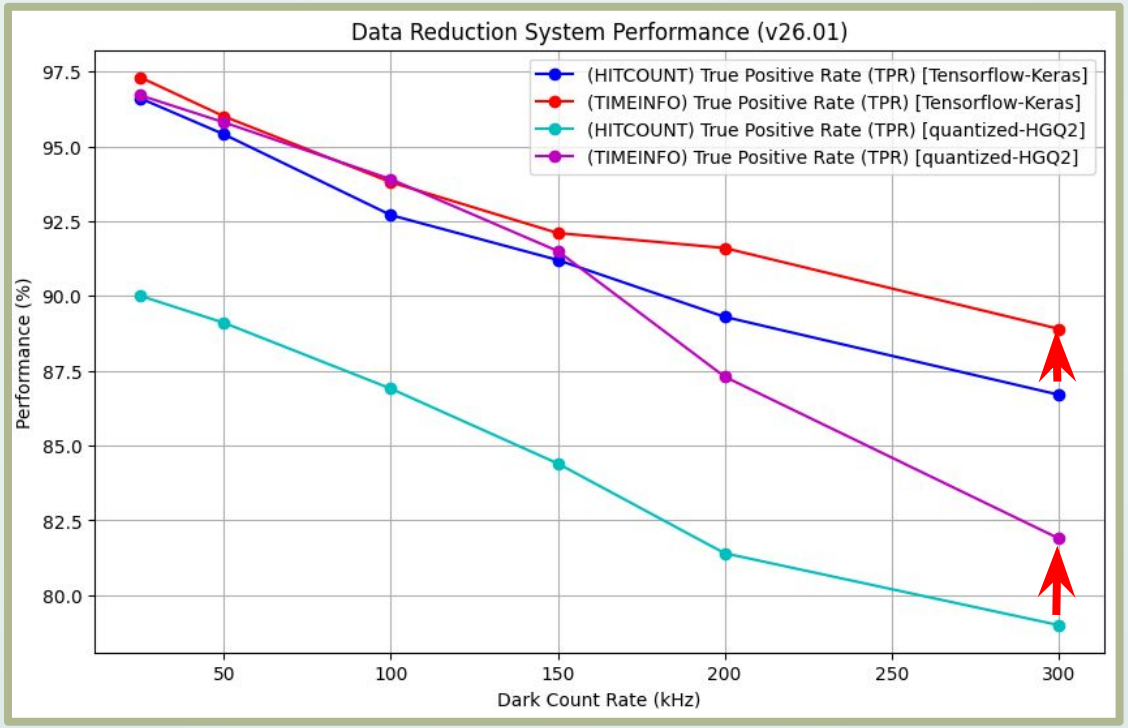
DAM Input ⇒ Subsector Input for Time Information



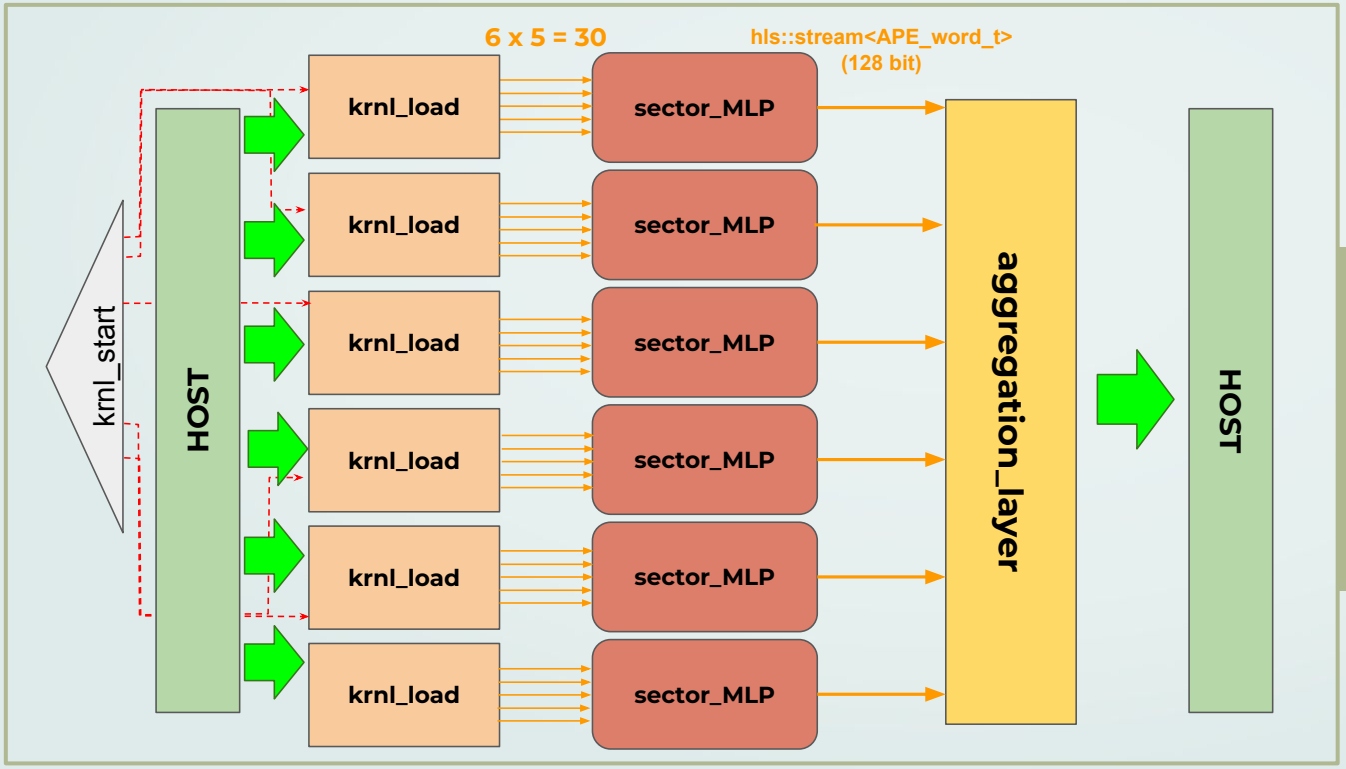
Time-Information Input for Multi MLP model



Time-Information Input for Multi MLP model

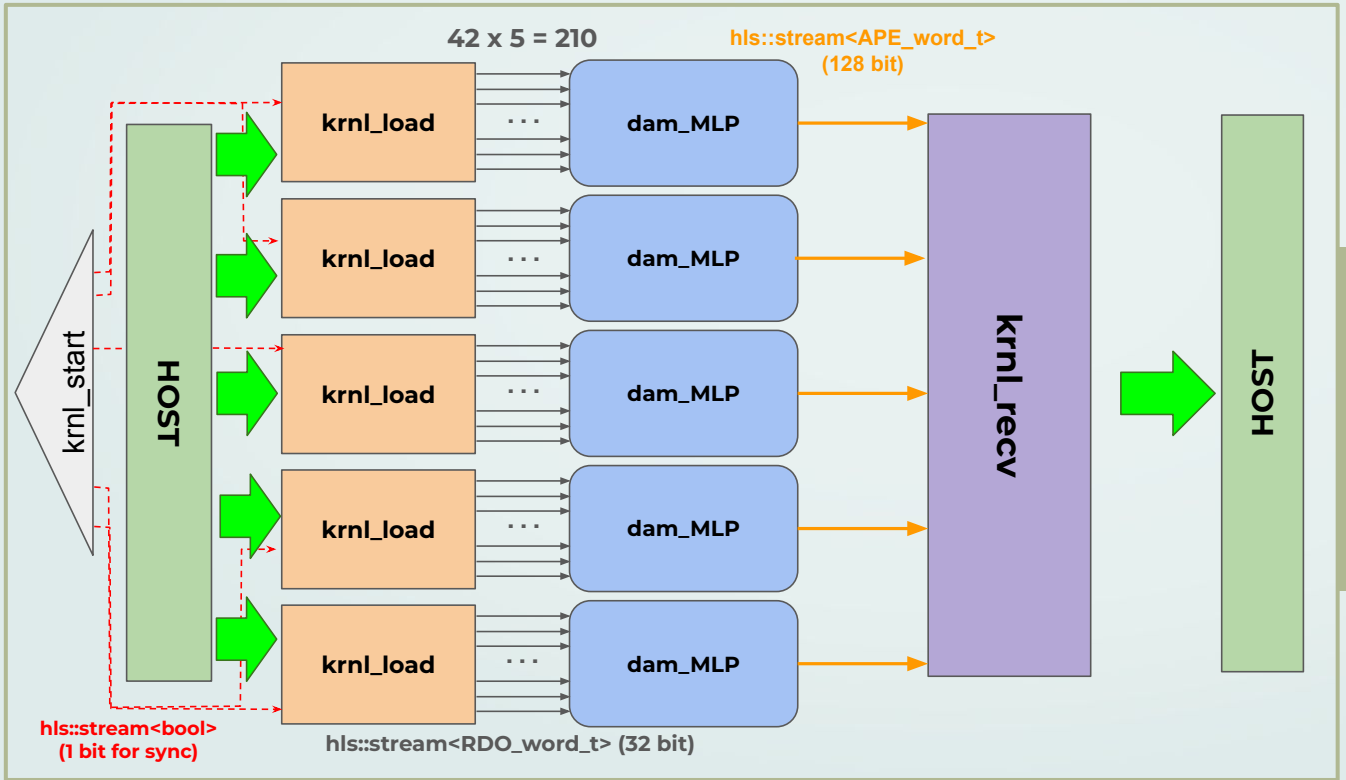


dRICH Data Reduction Stage on FPGA:
⇒ TP HW implementation



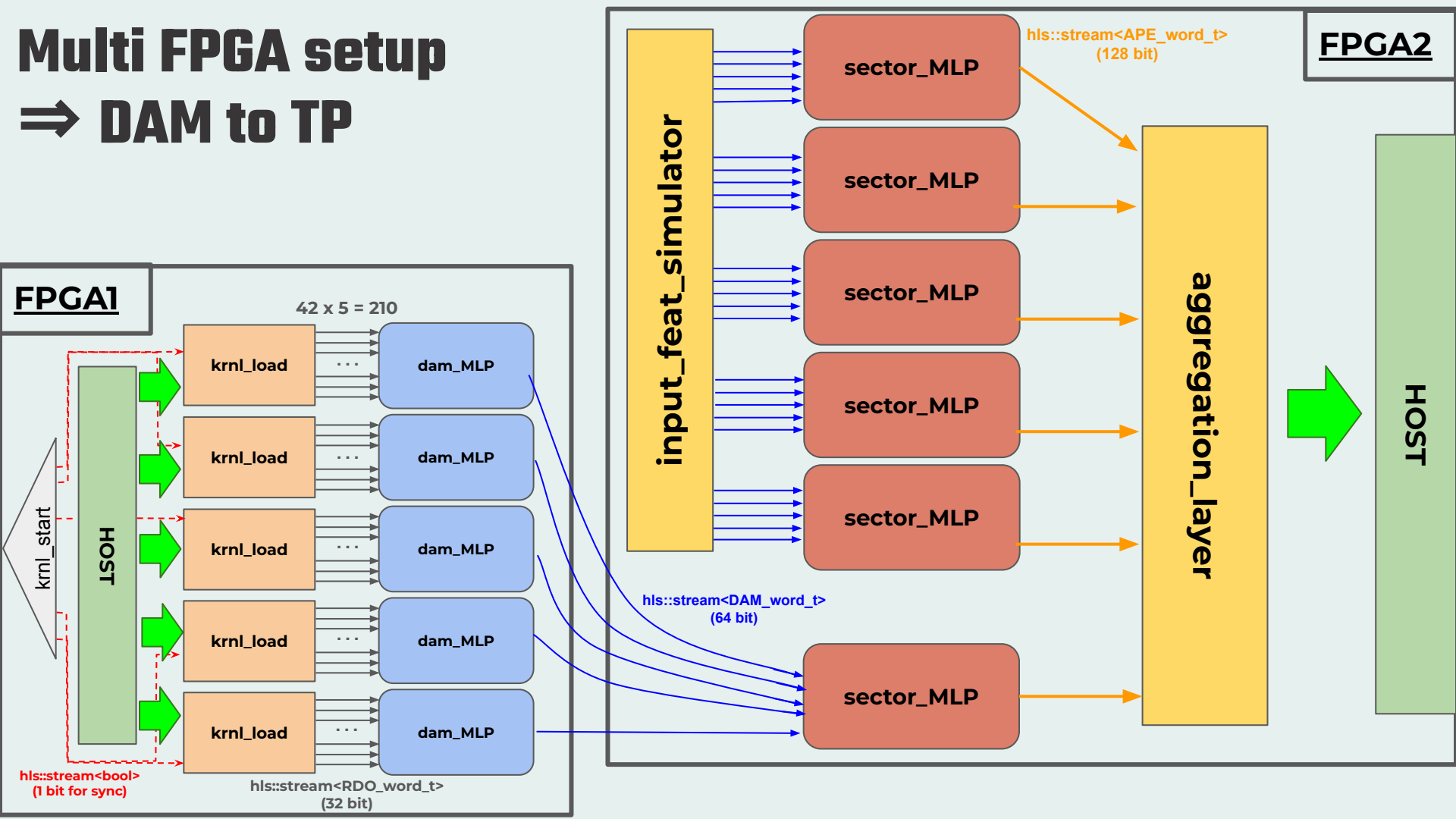
Stripped-down **HW design** (on AMD Versal™ Premium Series VPK180 Evaluation Kit) used to validate the **TP NN model (6 Sector MLP + Aggregate Layer)** implementation and to assess system performance.

dRICH Data Reduction Stage on FPGA:
⇒ 5 DAMs HW implementation



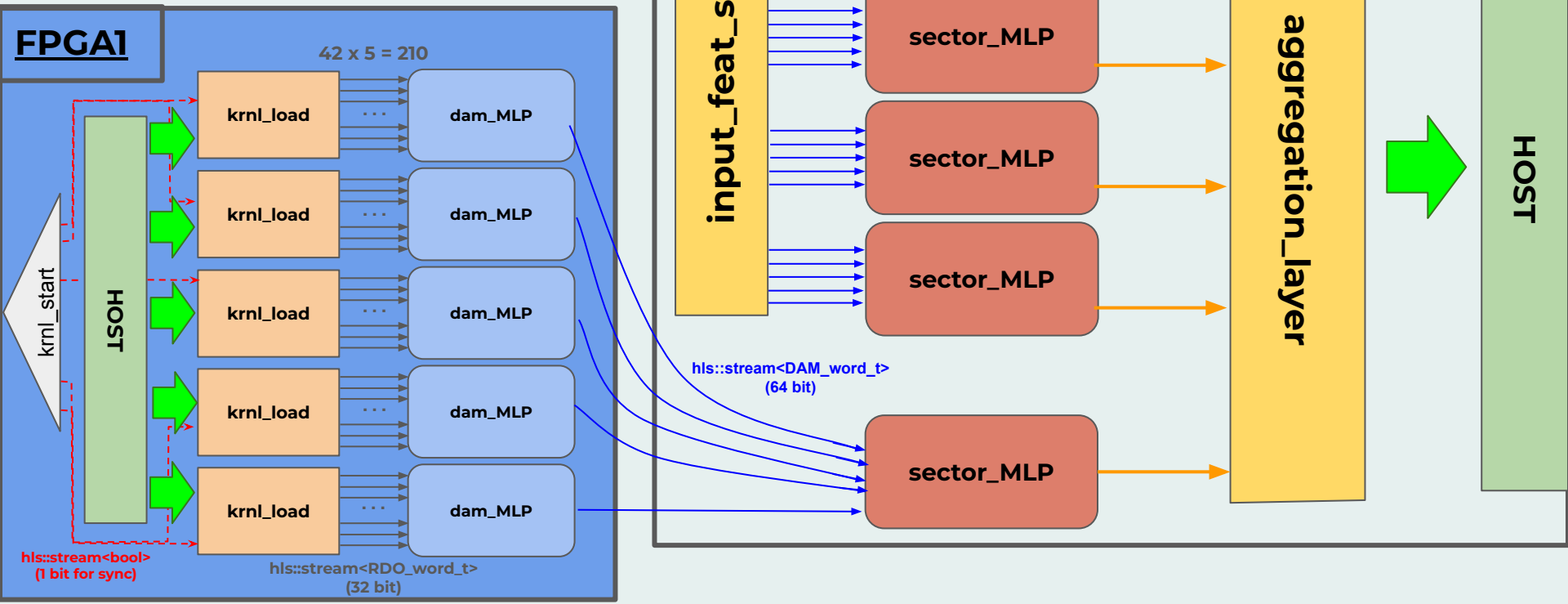
Stripped-down **HW design** (on AMD Versal™ Premium Series VPK180 Evaluation Kit) used to validate the **6 Subsector DAM NN models** implementation and to assess system performance.

Multi FPGA setup
⇒ DAM to TP



Multi FPGA setup

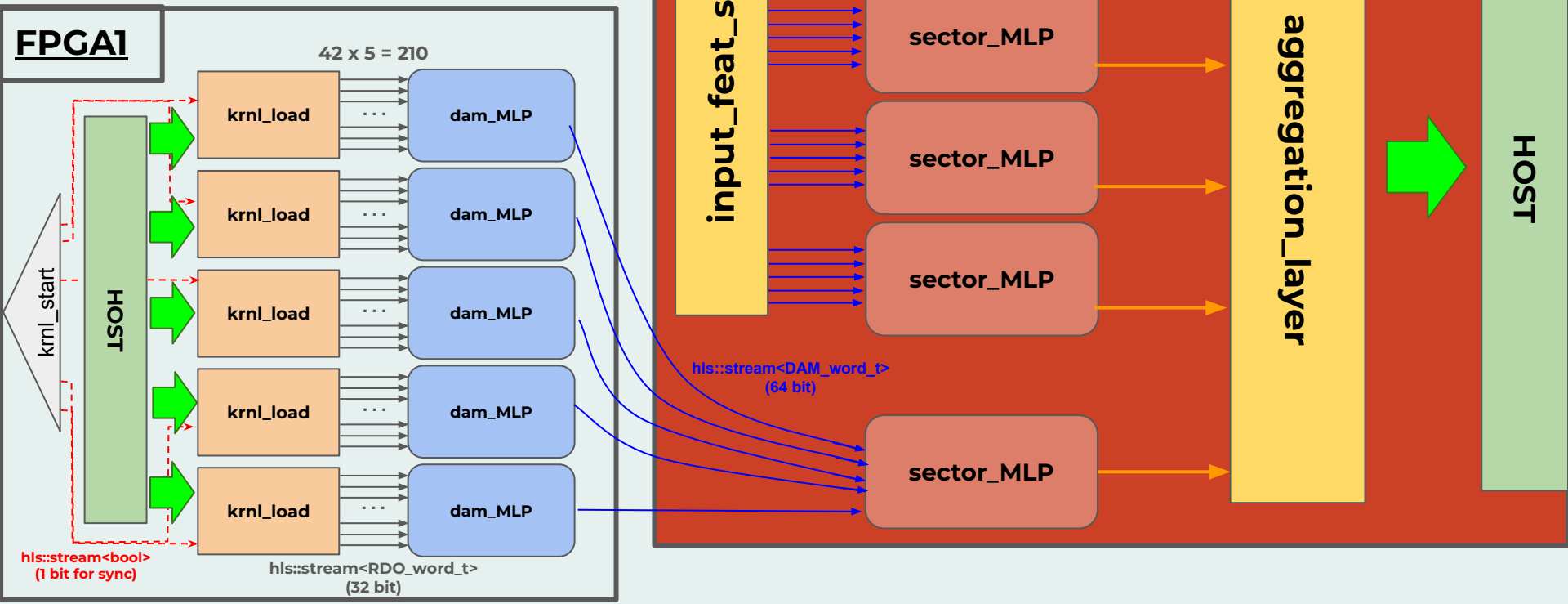
FPGA1
⇒ datastream simulation of an entire sector (5 DAMs)



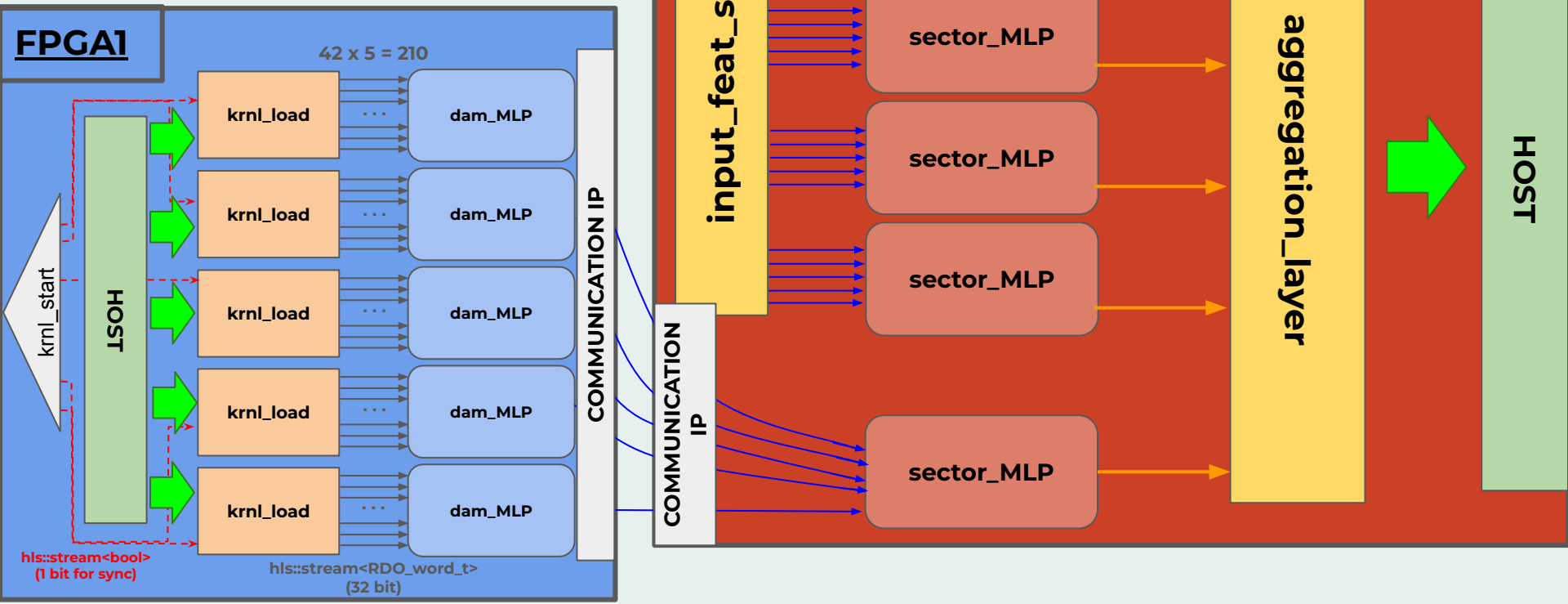
Multi FPGA setup

FPGA2

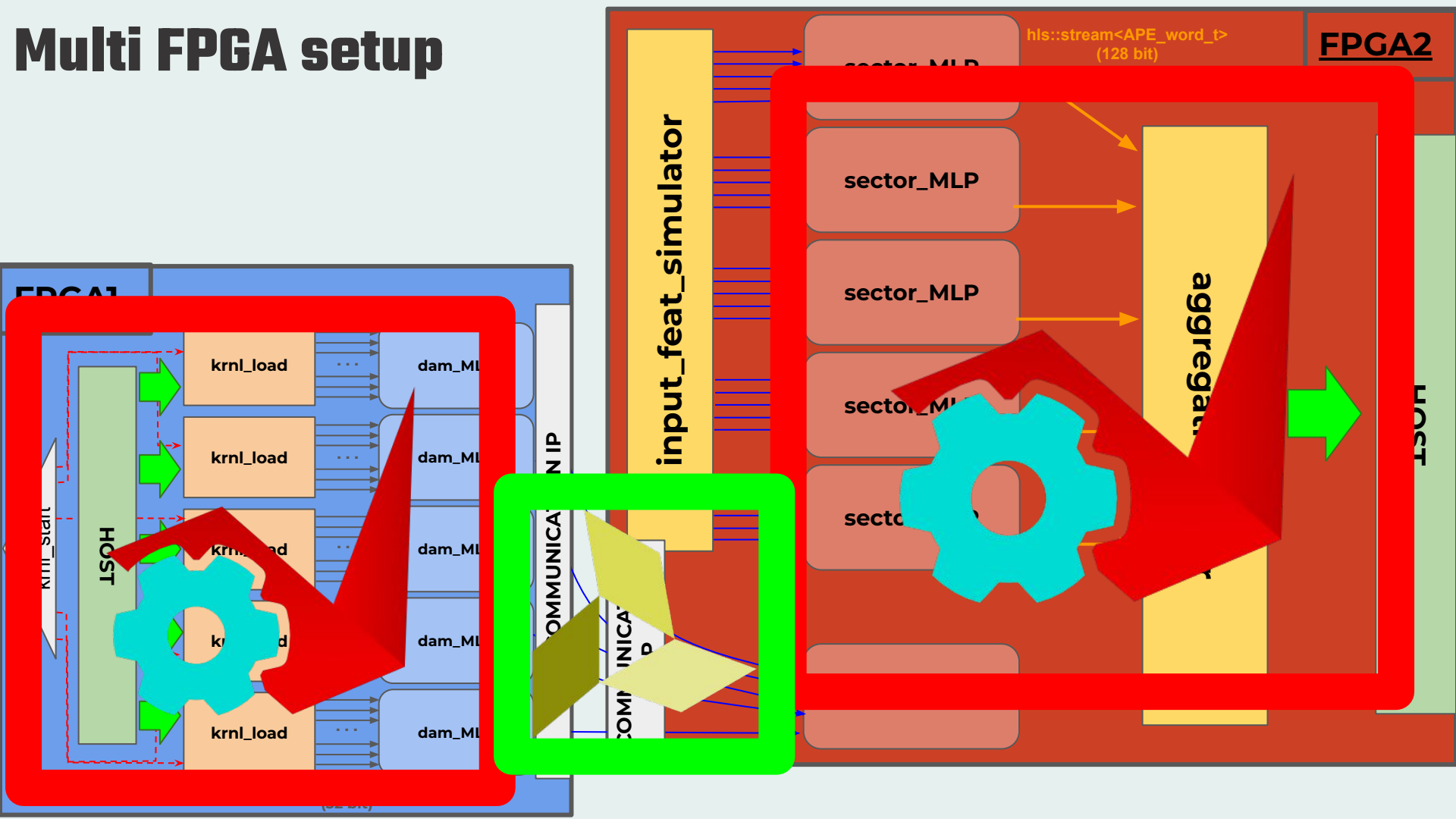
⇒ 5 input links coming aggregated as input of a single Sector_MLP
(simulated datastream for the remaining 5 ones)



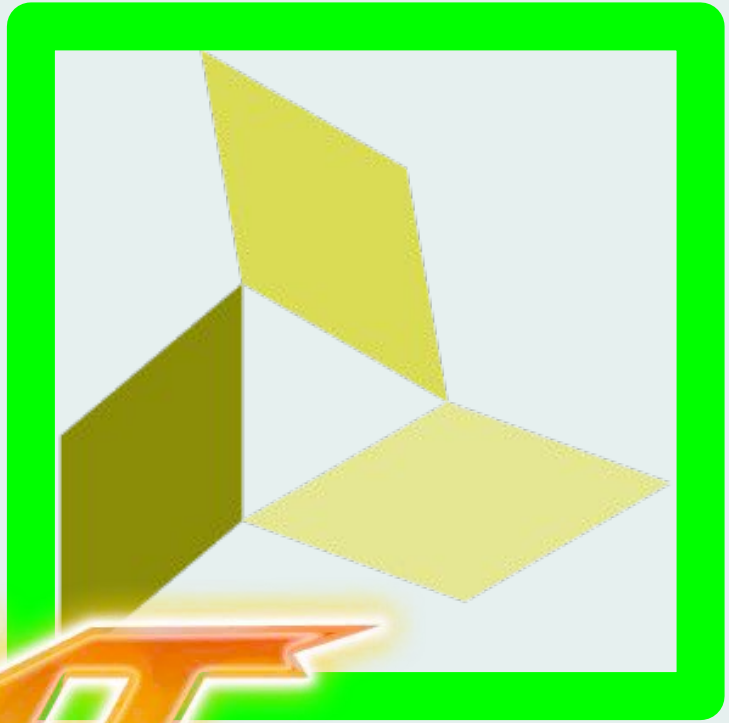
Multi FPGA setup



Multi FPGA setup



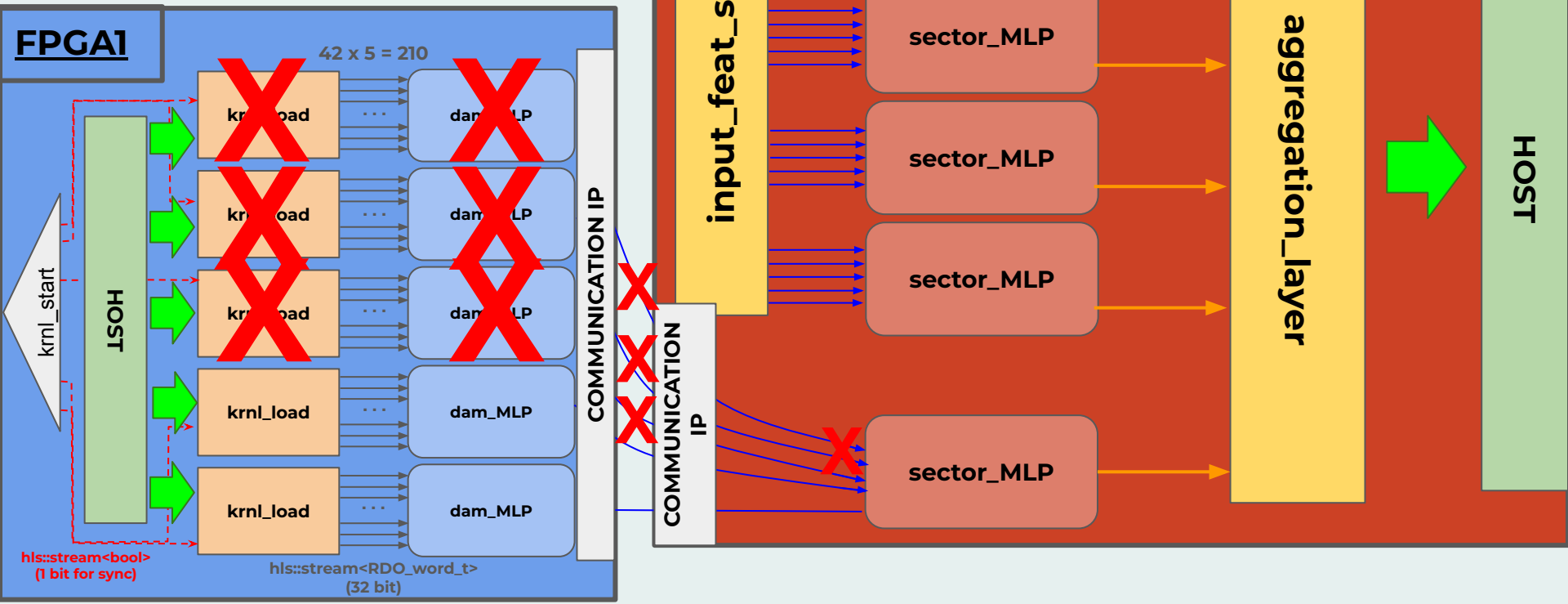
Implementation Tools ⇒ development compatibility



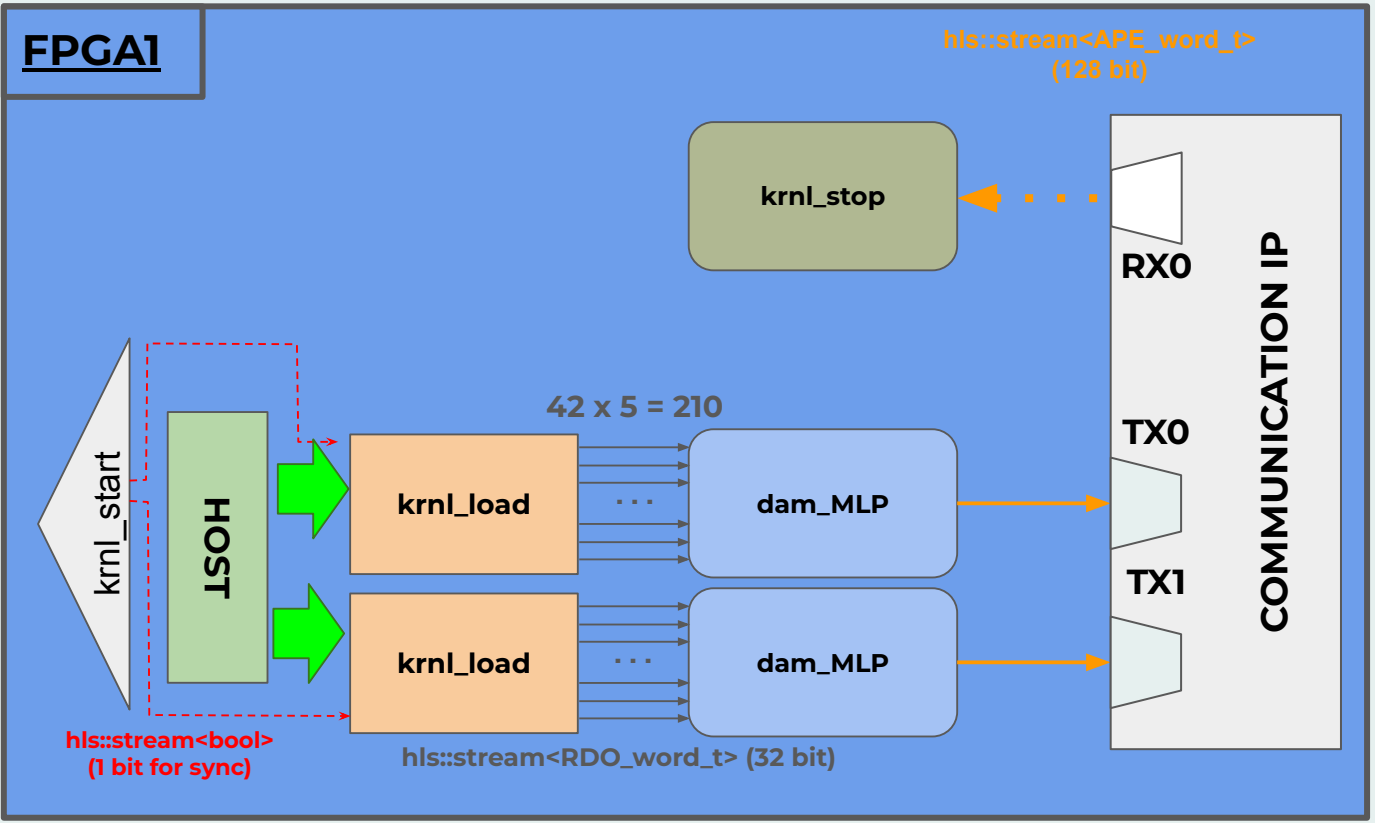
Multi FPGA setup

Unfortunately, at this stage of development we have only a working communicationIP with 2 Links

⇒ 5 Links version ongoing



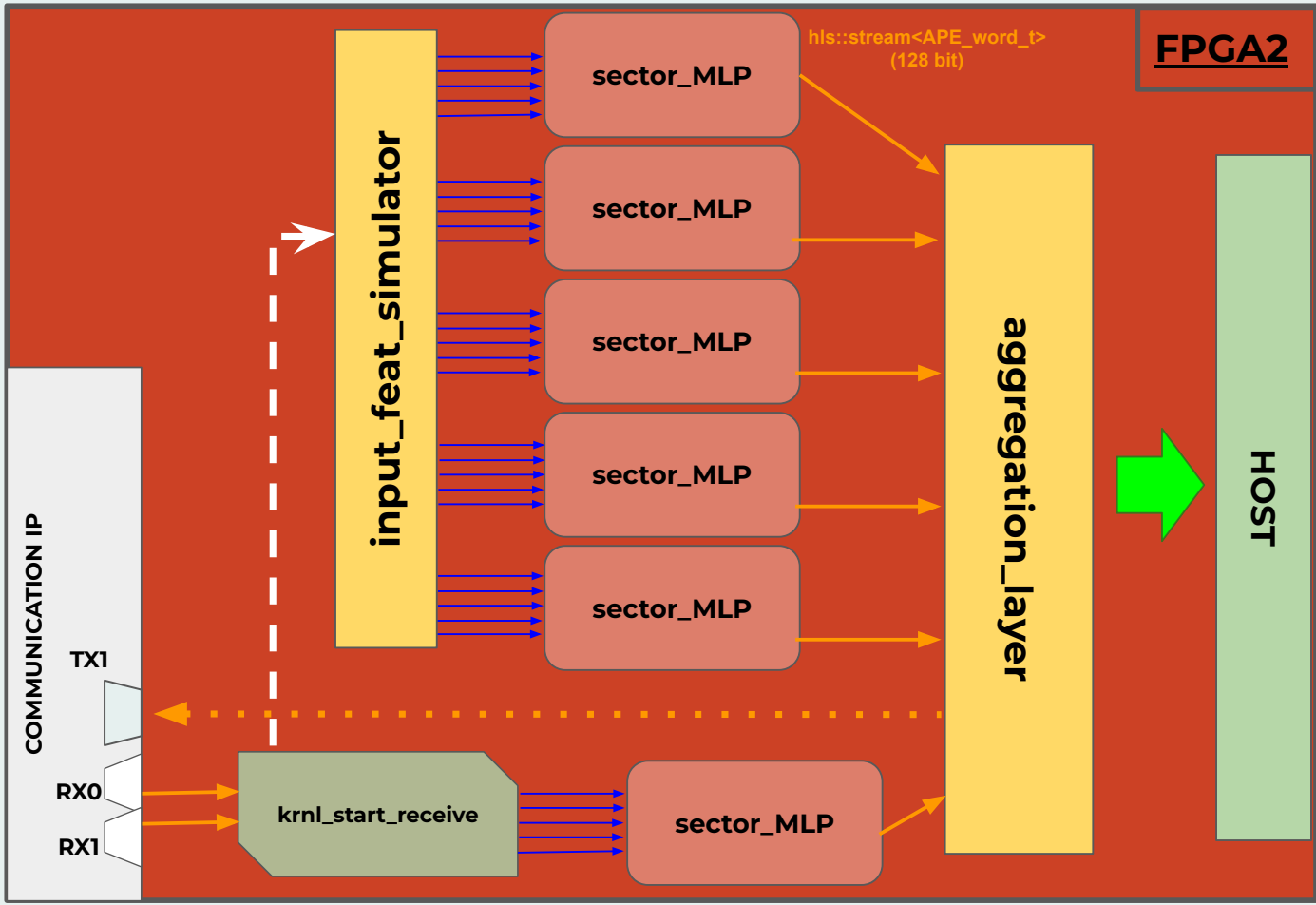
Multi FPGA testbed solution: 2DAM FPGA1



To monitor the multi FPGA system execution and correctly measuring its **throughput**:
⇒ **krnl_stop** implementation: it receives a single data packet at the end of the TP FPGA processing
⇒ allows us to measure the difference between the time of the first packet sent by DAM FPGA and the time of the last result computed TP FPGA

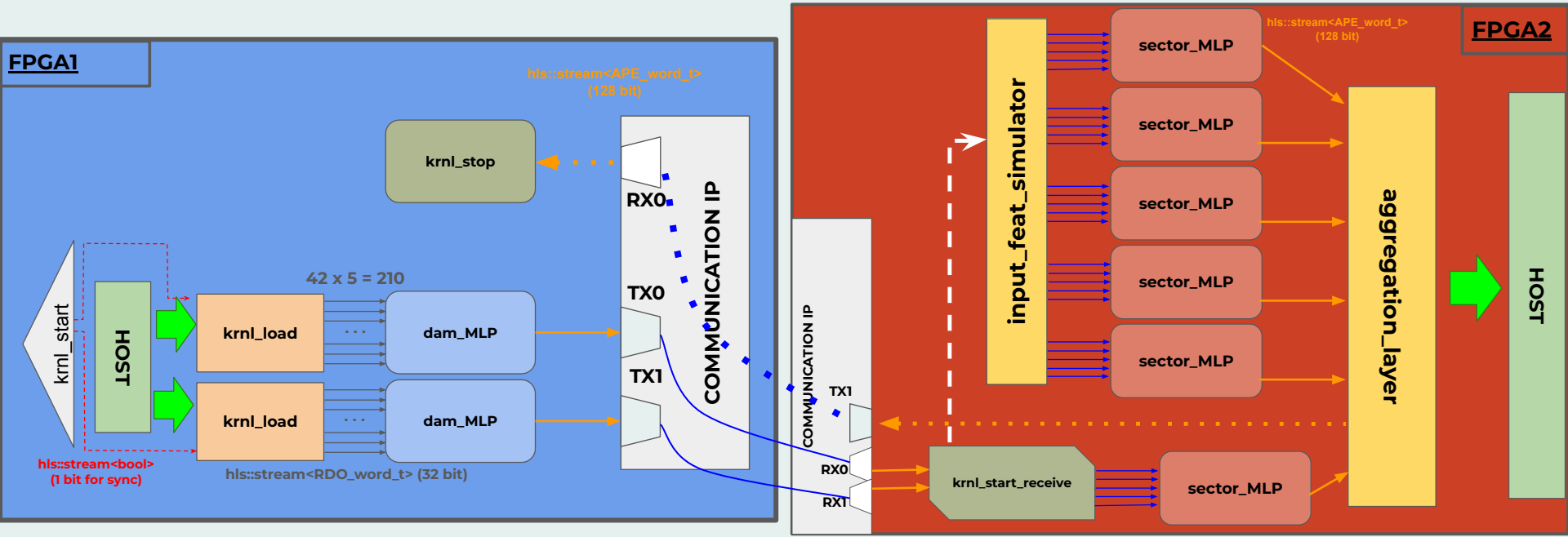
Multi FPGA testbed solution: TP FPGA2

⇒ **krnl_start_receive** implementation: when it receives data from 2 DAM input streams, it mimcs the missing 3 DAM feature streams sending 5 DAM output information to the **sector_MLP** and sends a synchronization signal to the **input_feat_simulator**



Multi FPGA testbed solution

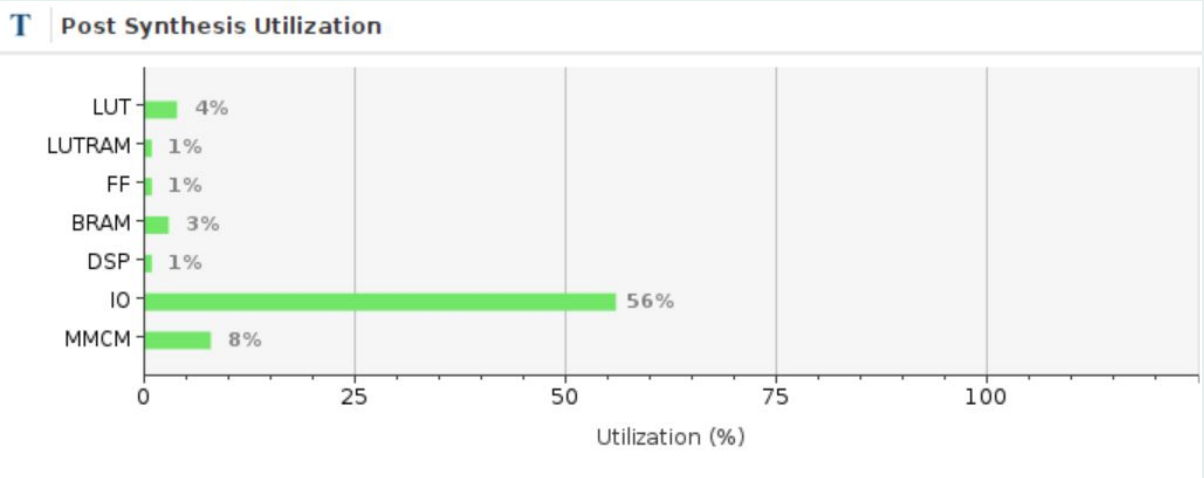
⇒ 2DAM to TP communication



(FPGA) HW Synthesis: TP Resource Usage

- TP NN design (6 Sector NN + Aggregation MLP NN) fits into the available FPGA resources of the AMD Versal™ Premium 1802 board.

➔ **resource usage** to take into account when moving to the target HW (FELIX-155 Xilinx Versal Prime) and **integrating with the standard DAQ firmware.**



(FPGA) HW Synthesis: performance

- **Throughput = 29.999 MHz**

  ⇒ instantiation interval: II~3 cycles (@100 MHz global clock)

Name	Issue Type	Latency (cycles)	Latency (ns)	Interval
top_TP_block_1		10	50.000	1
preprocessing_block		0	0.0	0
hwfunc_Sec0		7	35.000	1
fifo2array		0	0.0	1
relu_ap_fixed_13_10_5_3_0_ap_fixed_11_8_5_3_0_relu_config4_s		0	0.0	1
relu_ap_fixed_12_9_5_3_0_ap_fixed_10_7_5_3_0_relu_config7_s		0	0.0	1
relu_ap_fixed_20_8_5_3_0_ap_ufixed_19_7_5_3_0_relu_config10_s		0	0.0	1
feature_sender		1	5.000	1

- **(Online) Firmware** currently tested with DCR=150kHz dataset**:**

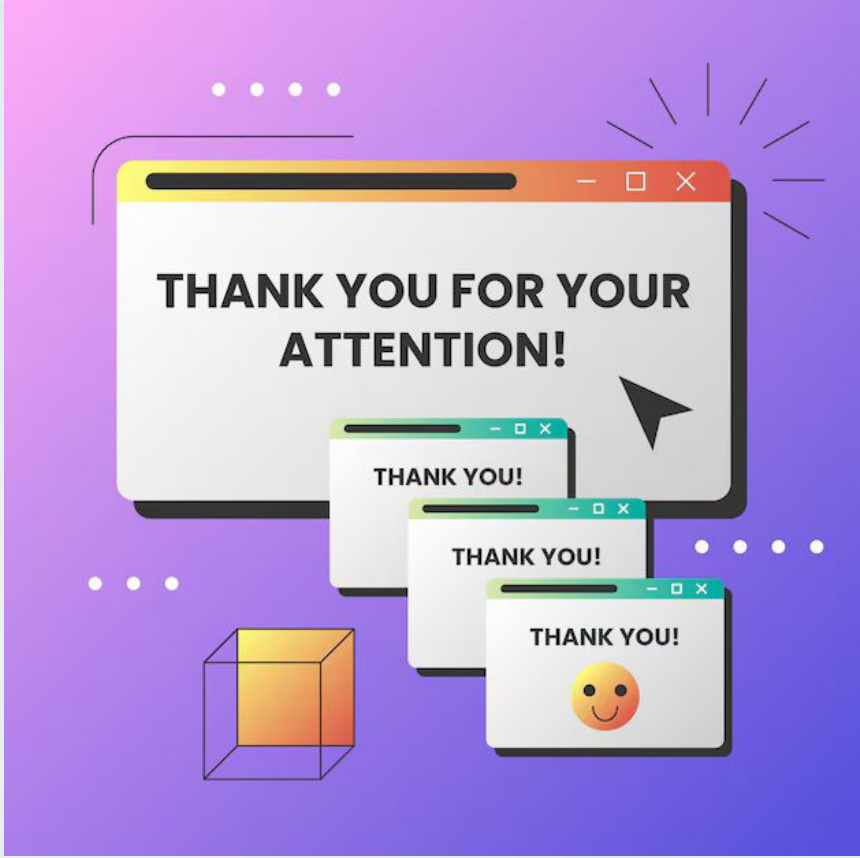
  ⇒ TPR-TNR performance compatible within ~4% with the (offline) quantized model

Conclusion

- Implementation of a simplified version of the distributed MLP NN model
- Assessed its performance in terms of **TPR/TNR (ML classification metrics)** and **resources/throughput (HW implementation metrics)**
- Additional **input timing information** to enhance ML performance of the data reduction system
- Partial **validation of 2 DAM to TP communication ⇒ baseline for following developments**

- Ongoing activities on throughput enhancement ⇒ higher clock cycle

- Development of a distributed MLP on two Felix-like FPGA including all the architectural blocks (5 DAM NNs and a full TP) is ongoing ⇒ **validation of the 5DAM to TP communication** and **target board deployment**



Contacts:

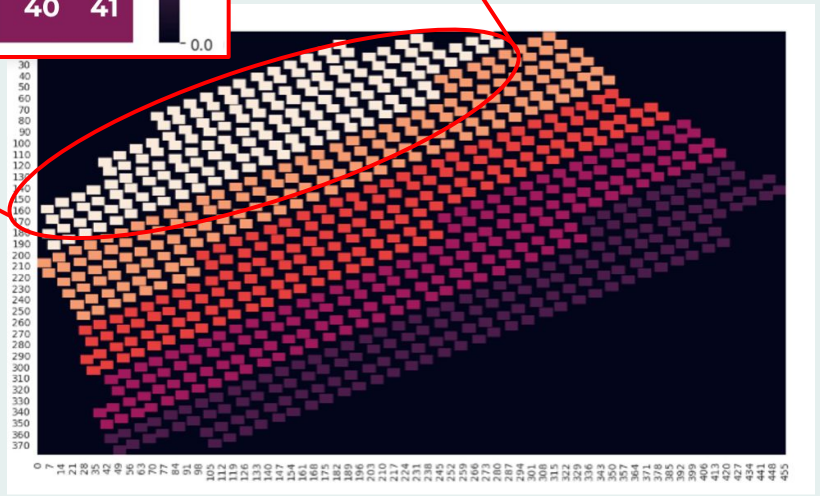
- cristian.rossi@roma1.infn.it
- alessandro.lonardo@roma1.infn.it
- https://apegate.roma1.infn.it

BACKUP

DAM Input ⇒ Subsector PDU Information

- **42 input links** for each DAM, corresponding to the number of expected PDUs per subsector (~210/5).

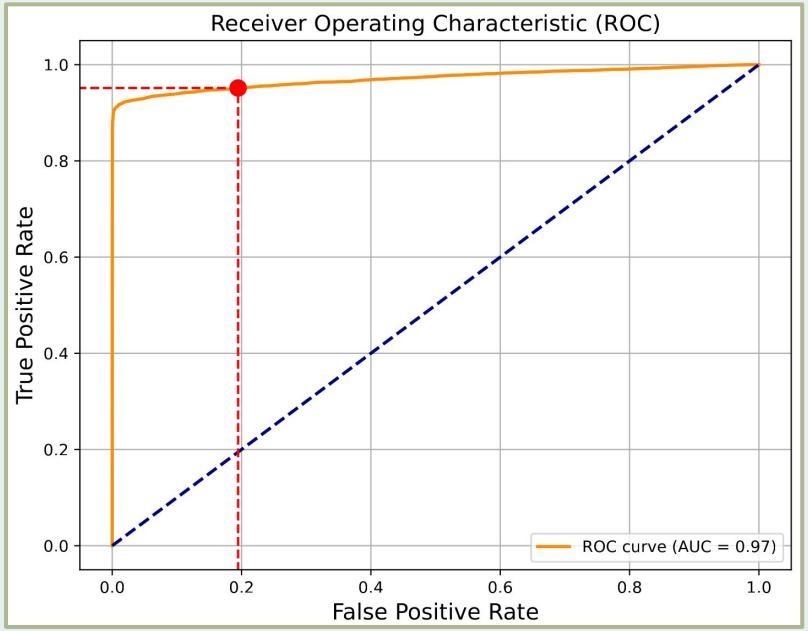
⇒ **Each PDU is input** to a neuron of the input layer of the MLP NN

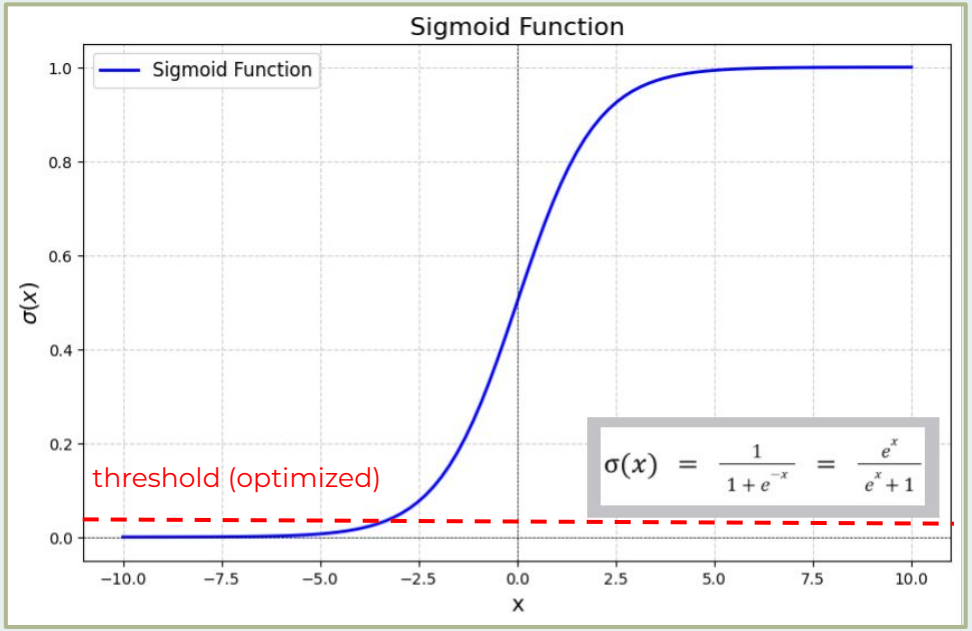


⇒ **42 input neuron** for the input layer of the MLP NN

("Answer to the Ultimate Question of Life, the Universe, and Everything", cit.)

dRICH Data reduction: ROC studies for TPR maximizing





$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad >= 80\% \; [\Rightarrow \text{reduction factor} >= 5]$$

$$\mathrm{FNR} = 1 - \mathrm{TNR}$$

Time-Information Input for Multi MLP model

- We start to evaluate how to implement the **timing information** to enhance the performance of the data reduction system.
- dRICH hits **timing distribution**

  ⇒ **binwidth connected** to the bit resolution available from Coarse Counter and Fine Counter

- **Time normalized to first hit in each event**
- Distribution shows:

  ⇒ **~2 ns wide peak** (⇒ primary interactions)

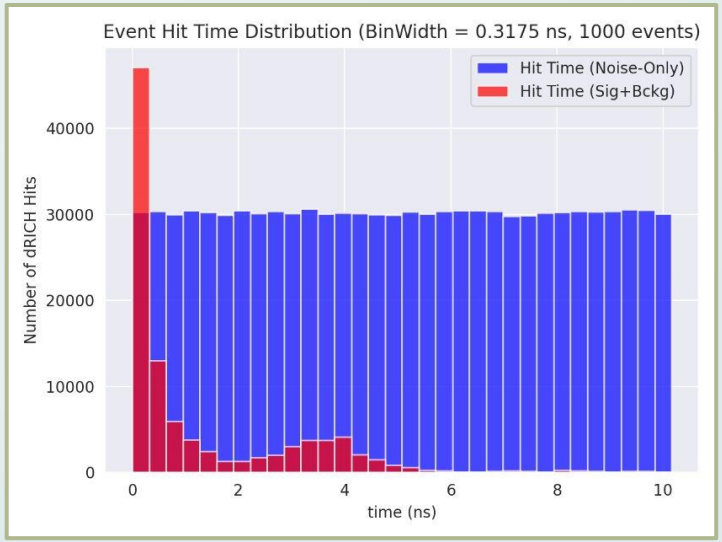
  ⇒ **tail + second peak (~4 ns)** (⇒ secondary interactions)

394 MHz → LSB = 2.54 ns

Calibration needed to use it at DAM level

31 29	28 26	25 24	23 9	8 0
Col ID	Pix ID	TDC ID	Coarse Counter	Fine Counter

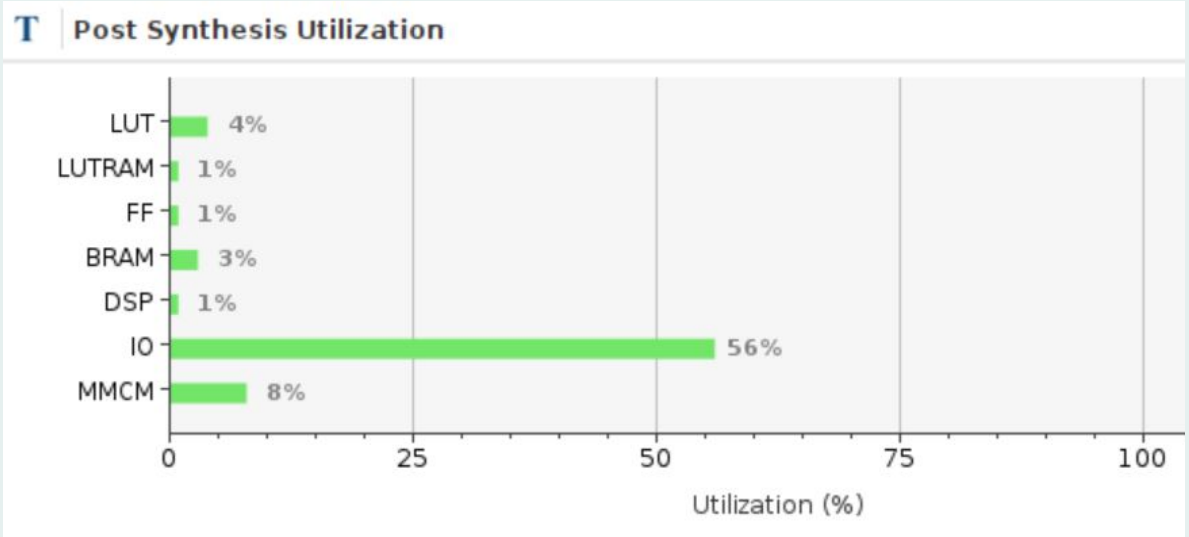
TIME



(FPGA) HW Synthesis: 5DAM Resource Usage

- 5 DAM NN design fits into the available FPGA resources of the AMD Versal™ Premium 1802 board.

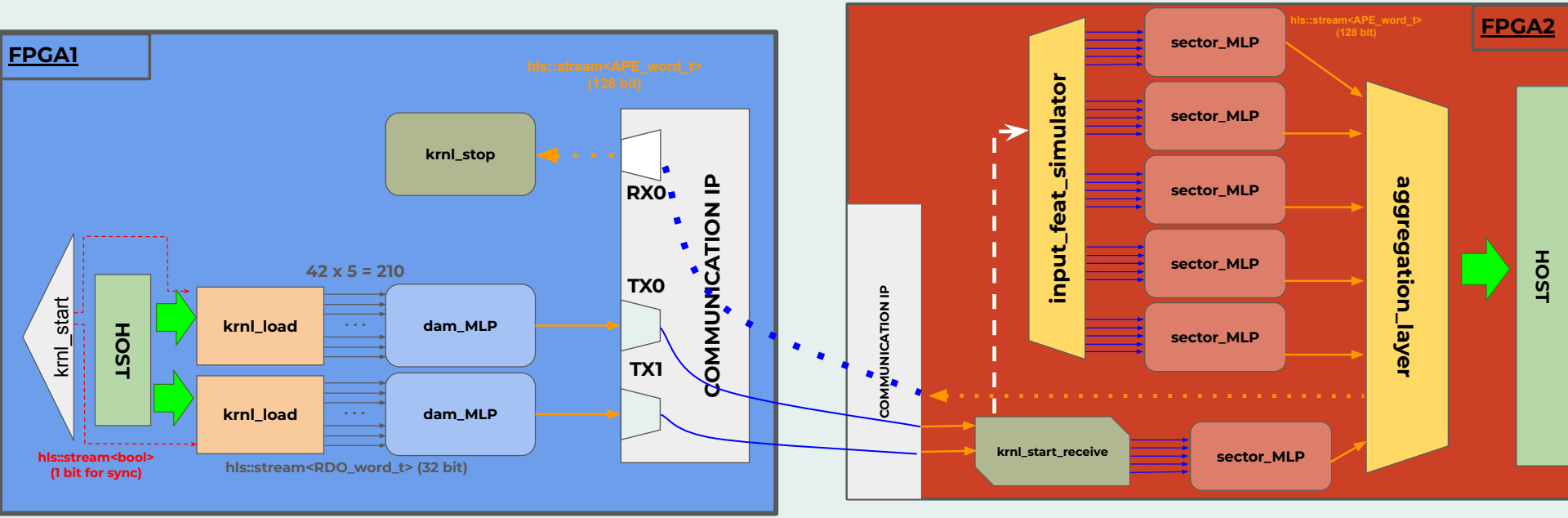
⇒ **resource usage** to take into account when moving to the target HW (FELIX-155 Xilinx Versal Prime) and **integrating with the standard DAQ firmware.**



dRICH: Analysis of Output Bandwidth

Multi FPGA testbed solution

⇒ 2DAM to TP communication

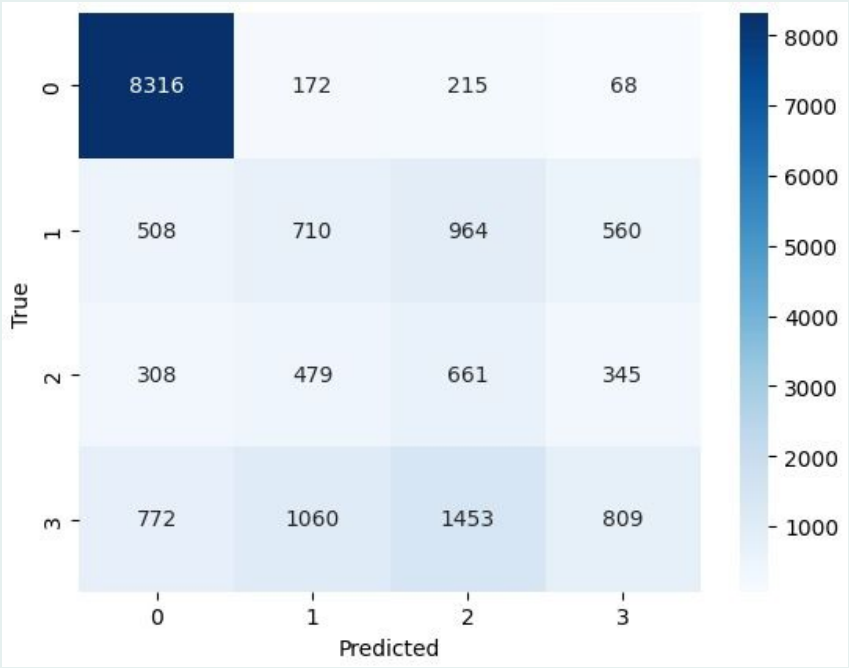


(?) Particle Counting with the Same Architecture?

- Maintaining the same NN design (due to HW constraints related to the dRICH readout architecture), we tried to evaluate other ML tasks:

  ⇒ **particle counting:**

  - **classes: 0, 1, 2, 3+ particles**



- **Unfortunately,** while presenting a good capability in tagging 0-particle events (e.g. Noise-Only Events), it seems incapable of distinguish between events presenting different numbers of MonteCarlo (MC) reconstructed particles
- **:(**
- [maybe offline? working directly on the dRICH global information with no subsectors constraints?]