Design Ideas for an Online Data Reduction System for the ePIC dRICH Detector

Luca Pontisso

INFN Roma, APE Lab

for the ePIC Roma1/2 team

ePIC Collaboration Meeting - Lehigh University

July 26th 2024

EPIC dRICH

Compact cost-effective solution for particle identification in the high-energy endcap at EIC

dRICH

INFN

BA, BO, CS, CT, FE , GE, LNS, RM1, RM2, SA, TO, TS

Jefferson Lab

Duke University

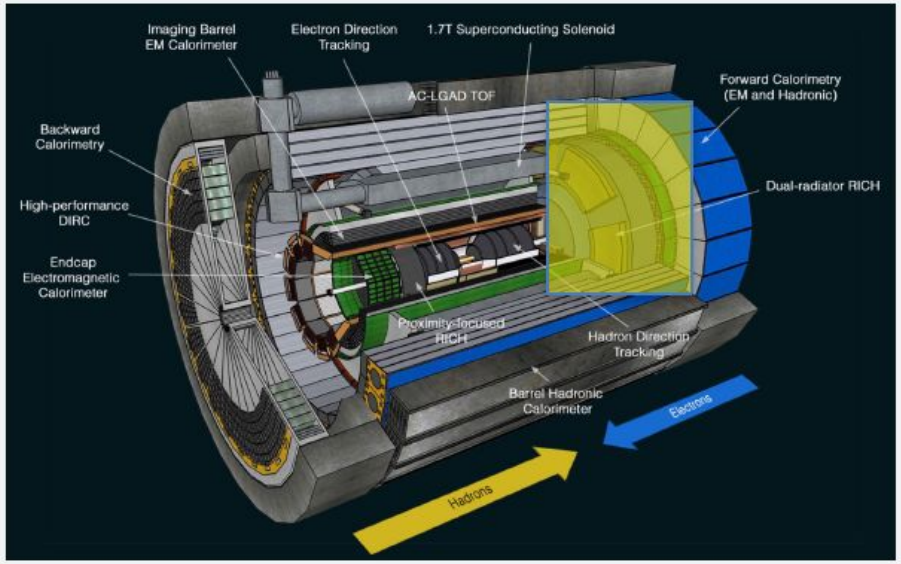
RAMAIAH UNIVERSITY OF APPLIED SCIENCES

NISER

केंद्रीय विश्वविद्यालय हरियाणा CENTRAL UNIVERSITY OF HARYANA NAAC ACCREDITED 'A' GRADE UNIVERSITY

.....

EPIC



EIC RICH Consortium

Brookhaven National Laboratory

Stony Brook University

PURDUE UNIVERSITY

TEMPLE UNIVERSITY

Univerza v Ljubljani

.....

Forward particle detection

Hadron ID in the extended 3-50 GeV/c interval

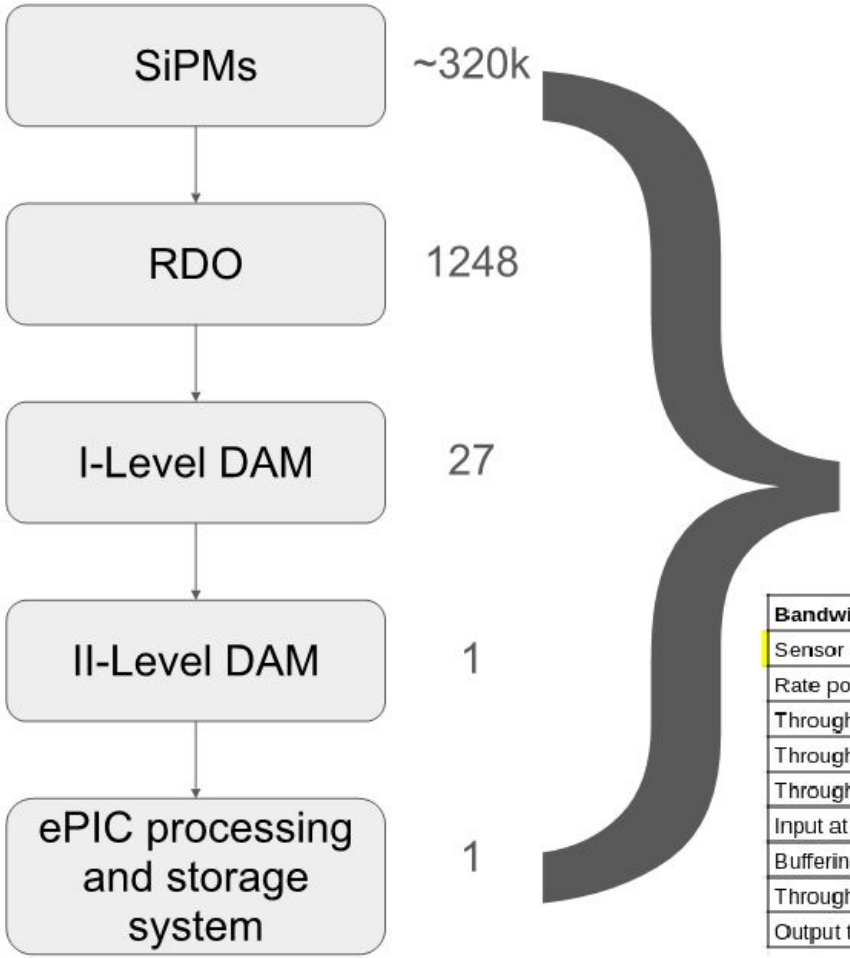
Support electron ID up to 15 GeV/c

Main challenges:

Cover wide momentum range 3 - 50 GeV/c -> dual radiator
Work in high (~ 1T) magnetic field -> SiPM
Fit in a quite limited (for a gas RICH) space -> curved detector

Analysis of dRICH Output Bandwidth

The dRICH DAQ chain in ePIC → the throughput issue



dRICH DAQ parameters	
RDO boards	1248
ALCOR64 x RDO	4
dRICH channels (total)	319488
Number of DAM L1	27
Input link in DAM L1	47
Output links in DAM L1	1
Number of DAM L2	1
Input link to DAM L2	27
Link bandwidth [Gb/s] (assumes VTRX+)	10
Interaction tagger reduction factor	1
Interaction tagger latency [s]	2,00E-03
EIC parameters	
EIC Clock [MHz]	98,522
Orbit efficiency (takes into account gap)	0,92

Bandwidth analysis		Limit
Sensor rate per channel [kHz]	300,00	4.000,00
Rate post-shutter [kHz]	55,20	800,00
Throughput to serializer [Mb/s]	34,50	788,16
Throughput from ALCOR64 [Mb/s]	276,00	
Throughput from RDO [Gb/s]	1,08	10,00
Input at each DAM I [Gbps]	50,67	470,00
Buffering capacity at DAM I [MB]	12,97	
Throughput from DAM I to DAM II [Gbps]	50,67	10,00
Output to each DAM II [Gbps]	1.368,14	270,00

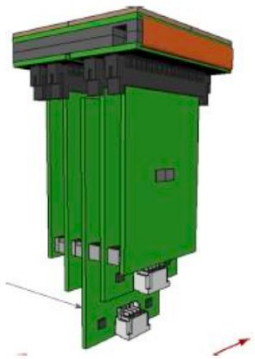
- Sensors DCR:
  3 - 300 kHZ (increasing with radiation damage → with experiment lifetime).
- Full detector throughput (FE):
  14 - 1400 Gbps
- A reduction >1/5 is needed
- EIC beams bunch spacing:
  10 ns → bunch crossing rate of 100 MHz.
- For the low interaction cross-section (DIS) → one interaction every ~ 100 bunches → interaction rate of ~1MHz
- A system tagging the (DIS) interacting bunches can solve the throughput issue (reducing to ~1/100 the data throughput)

Throughput Issue:

1. **Develop a dedicated sub-detector tagging relevant interactions.**
2. **This proposal.**

RDO and ePIC DAQ

P. Antonioli
https://indico.bnl.gov/event/20457/contributions/80658/attachments/49752/85138/20230914-DAQ.pdf



1248

- PDU: 1248
- RDO: 1248
- FEB: 4992

1248

I-level DAM (27)

FELIX

- 47 links to PDU
- 1 link to II-level DAM

27

in exp. hall, rack mounted

PC with 4 FELIX each (??)

27

II-level DAM (1)

FELIX

- 27 links to I-level DAM
- link from central ePIC [clock/trigger]

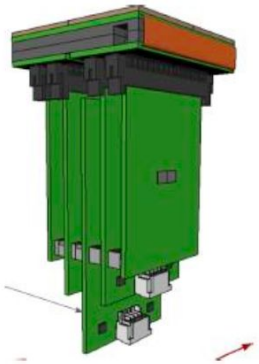
ePIC interaction tagger
able to reach our DAMS in 10 μs!

1

1

RDO and ePIC DAQ

P. Antonioli
https://indico.bnl.gov/event/20457/contributions/80658/attachments/49752/85138/20230914-DAQ.pdf



1248

- PDU: 1248
- RDO: 1248
- FEB: 4992

1248

I-level DAM (27)

FELIX

- 47 links to PDU
- 1 link to II-level DAM

27

in exp. hall, rack mounted

PC with 4 FELIX each (??)

27

II-level DAM (1)

FELIX

- 27 links to I-level DAM
- link from central ePIC [clock/trigger]

ePIC interaction tagger
able to reach our DAMS in 10 μs!

1

1

Data reduction on FPGA

dRICH Data Reduction Stage on FPGA

- Objective: design of a data reduction stage for the dRICH with a ~100 data bandwidth reduction in DAM-I level output to DAM-II level input.
- Make exclusive use of DAQ components (Felix DAMs)
  - Add few DAM units wrt the bare minimum (i.e. 27 Felix) needed to readout the 1248 RDO links to implement a distributed processing scheme.
  - Integration with the dIT (or other detectors) to boost performance and enable other features.
- Online Signal/Noise discrimination using ML
  - Collecting datasets using data available from simulation campaigns
    - Background:
      - e/p with beam pipe gas
      - Synchrotron radiation (MC only, it would be useful to have it reconstructed)
    - Merged (i.e. the Signal): signal + e/p with beam pipe gas background (full)
      - Few events, more statistics would be useful
    - SiPM Noise
      - DCR modelled in the reconstruction stage
      - More statistics of merged reconstructed events with noise would be useful

dRICH Data Reduction Stage on FPGA

- Online Signal/Noise discrimination using ML (continued)
  - Study of Inference Models
    - Restricting our study to inference models that can be deployed on FPGA with reasonable effort (using a High-Level Synthesis workflow)
      - MLP, CNN, GNN NN Models (HLS4ML)
      - BDT (Conifer)
    - Inference throughput (98.5 MHz) is the main concern.
    - HDL optimized implementation is an option.
    - Not necessarily ML-based.
- Deployment on multiple Felix DAMs directly interconnected with the APE communication IP.

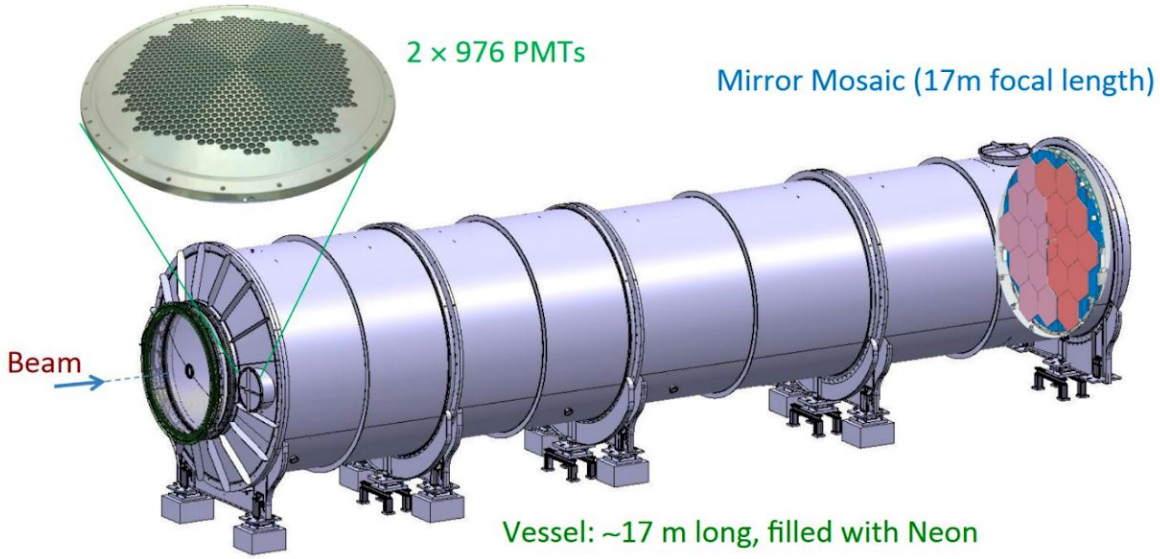
Some Background Activities

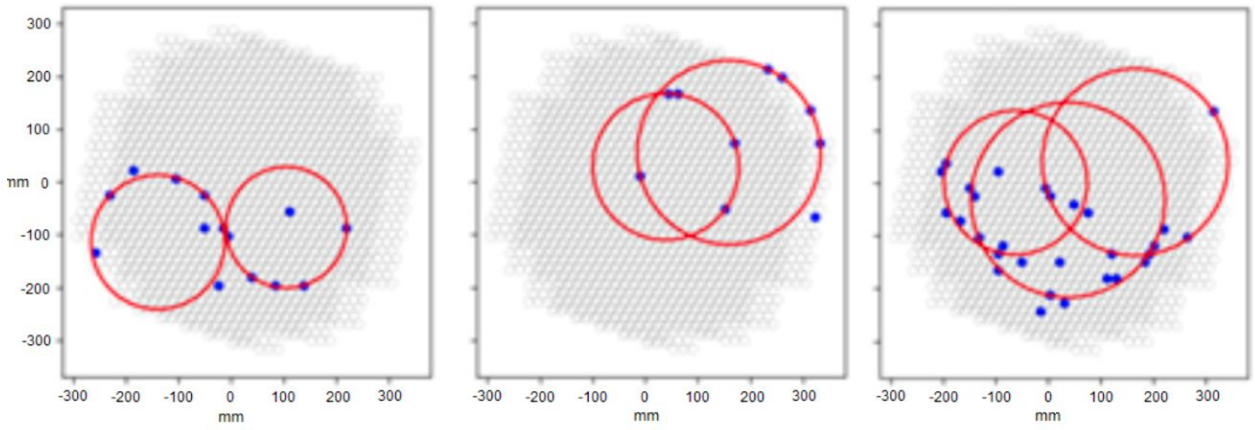
- INFN APE Lab @ Roma1/2: design and development of 4 generations of parallel computing architectures (mainly) dedicated to LQCD (1986-2010) **https://apegate.roma1.infn.it**

- Two recent research activities are relevant for this presentation:
  - FPGA-RICH: online ring counting system based on FPGA for the RICH detector of the NA62 experiment at CERN.
  - APEIRON: a framework offering hardware and software support for the execution of real-time dataflow applications on a system composed by interconnected FPGAs.

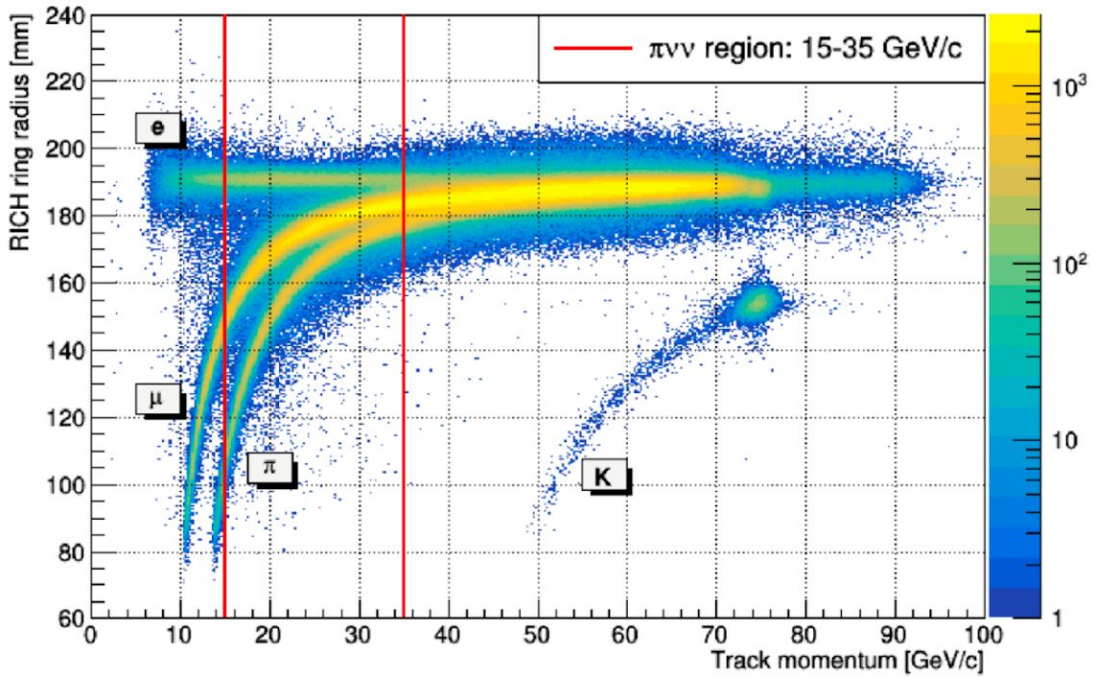
- Other research activities of possible interest for the “GPU approach”:
  - APENet: a high-throughput network interface card based on FPGA used in hybrid, GPU-accelerated clusters with a 3D toroidal mesh topology. [http://doi.org/10.1088/1742-6596/898/8/082035]
  - NaNet: a family of FPGA-based PCIe Network Interface Cards (with GPUDirect/RDMA capability) for High Energy Physics to bridge the front-end electronics and the software trigger computing nodes. [https://doi.org/10.1088/1742-6596/1085/3/032022]

The NA62 Ring Imaging Cherenkov detector (RICH)





- During offline data analysis, it provides PID to distinguish between pions and muons from 15 to 35 GeV
- Uses the Cherenkov rings radius and track momentum
- **L0 primitives contain only number of HIT PMTs**



Workflow for Neural Networks deployment on FPGA

32-bit precision model

QKeras

low-bit precision model

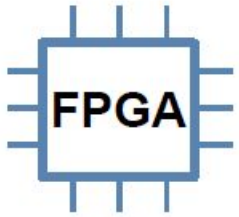
Python Libraries to train model and quantize parameters into **Fixed-point** representation

$2^4\ 2^3\ 2^2\ 2^1\ 2^0\ 2^{-1}\ 2^{-2}\ 2^{-3}$

0 0 0 1 0 . 1 1 0

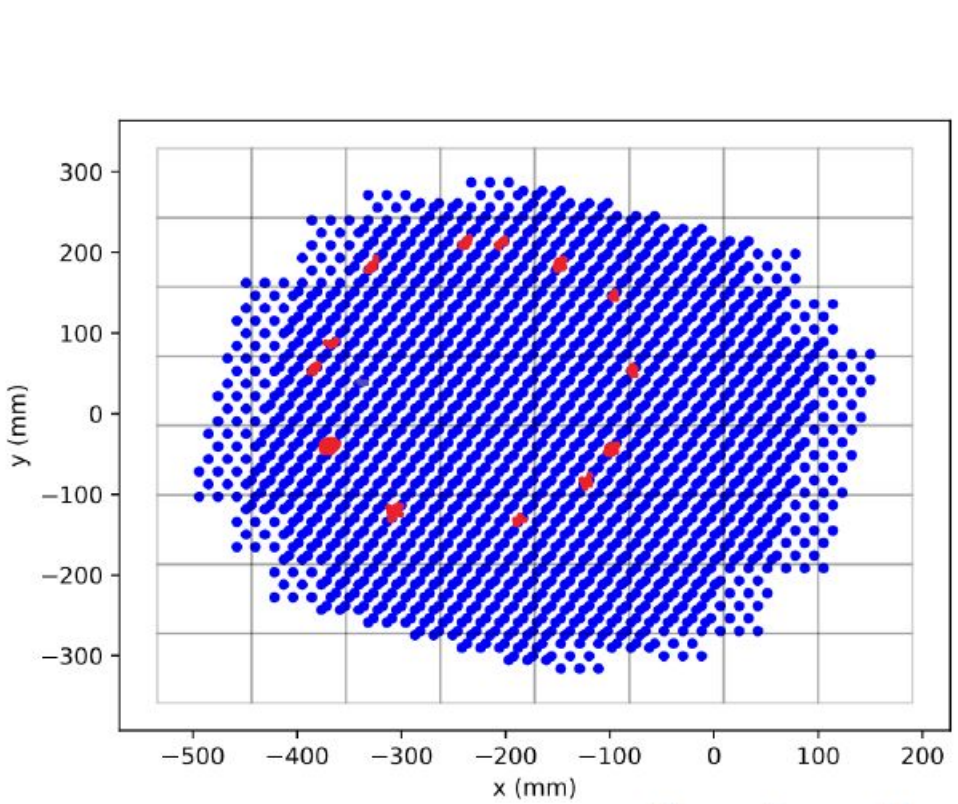
HLS model →

High Level Synthesis: high level C++ code transposed to hardware components, using directives to arrange pipelining, parallelization, memory structures, I/O, etc.



Iterate to find compromise between computational resources, throughput, and NN accuracy

Neural Network Model (actually one of them...)

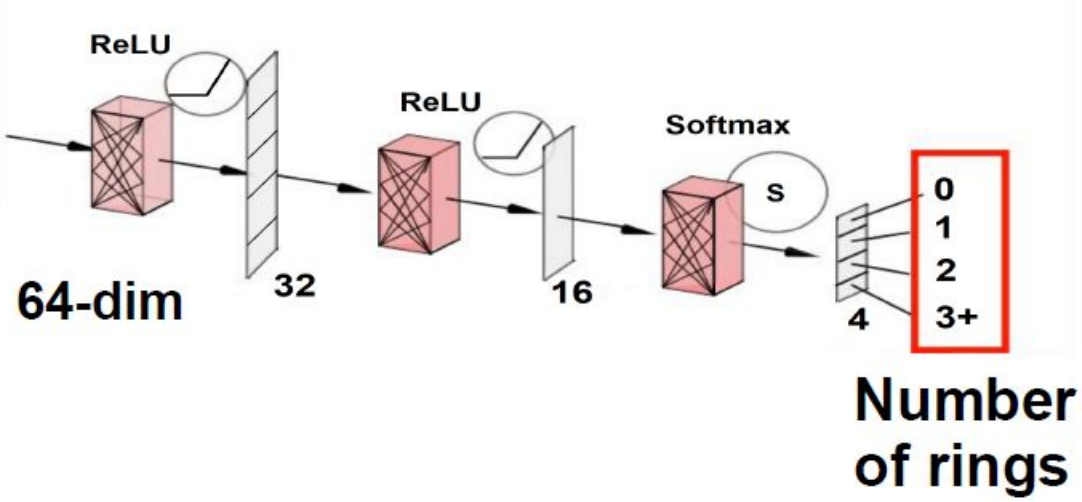


8 x 8 cells

Unrolled

hit-PMTs per cell

0
0
0
⋮
1
0
3
⋮
0
0

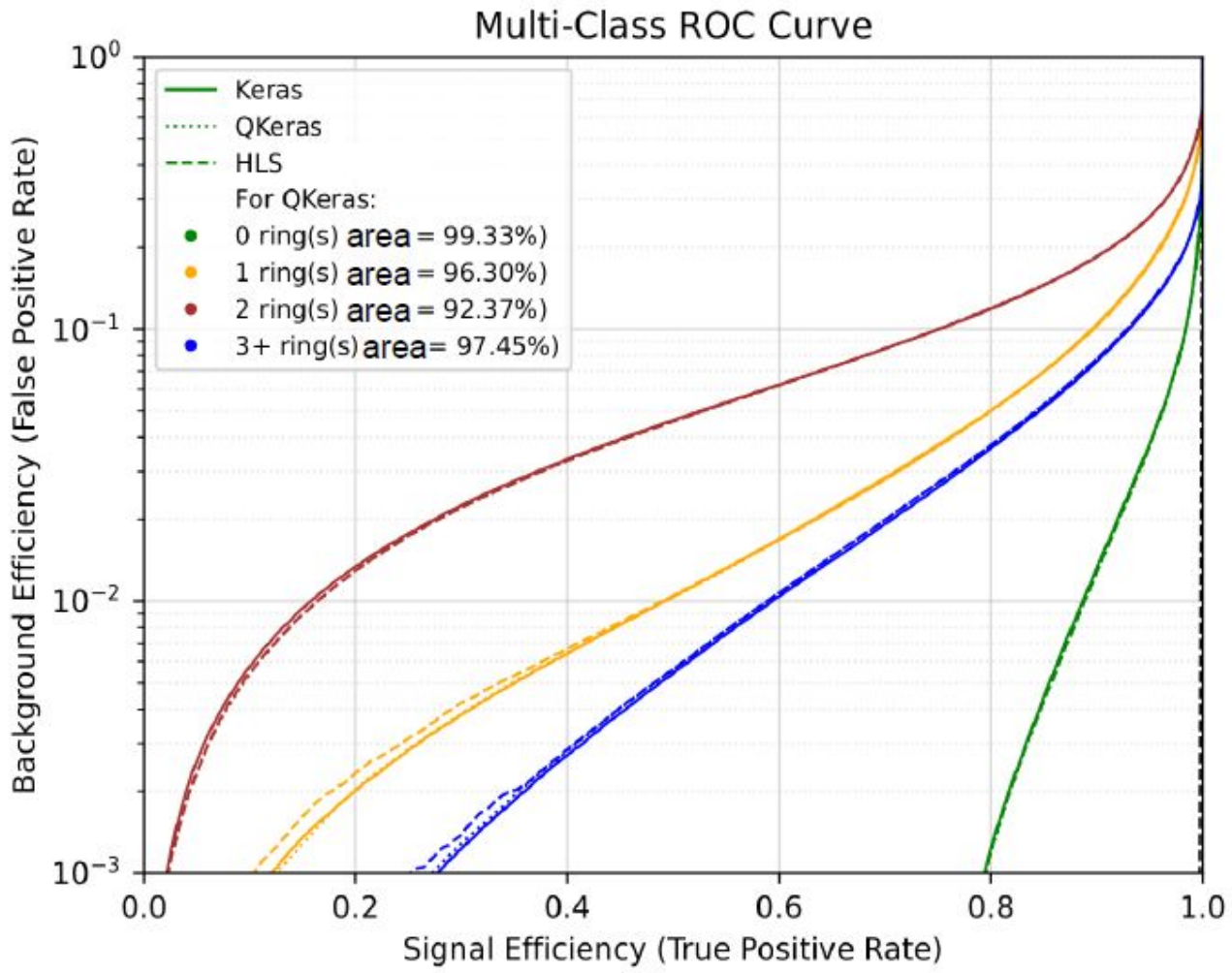


Number of rings

- **Encoding of the 1952 PMTs geometrical positions in the input layer.**

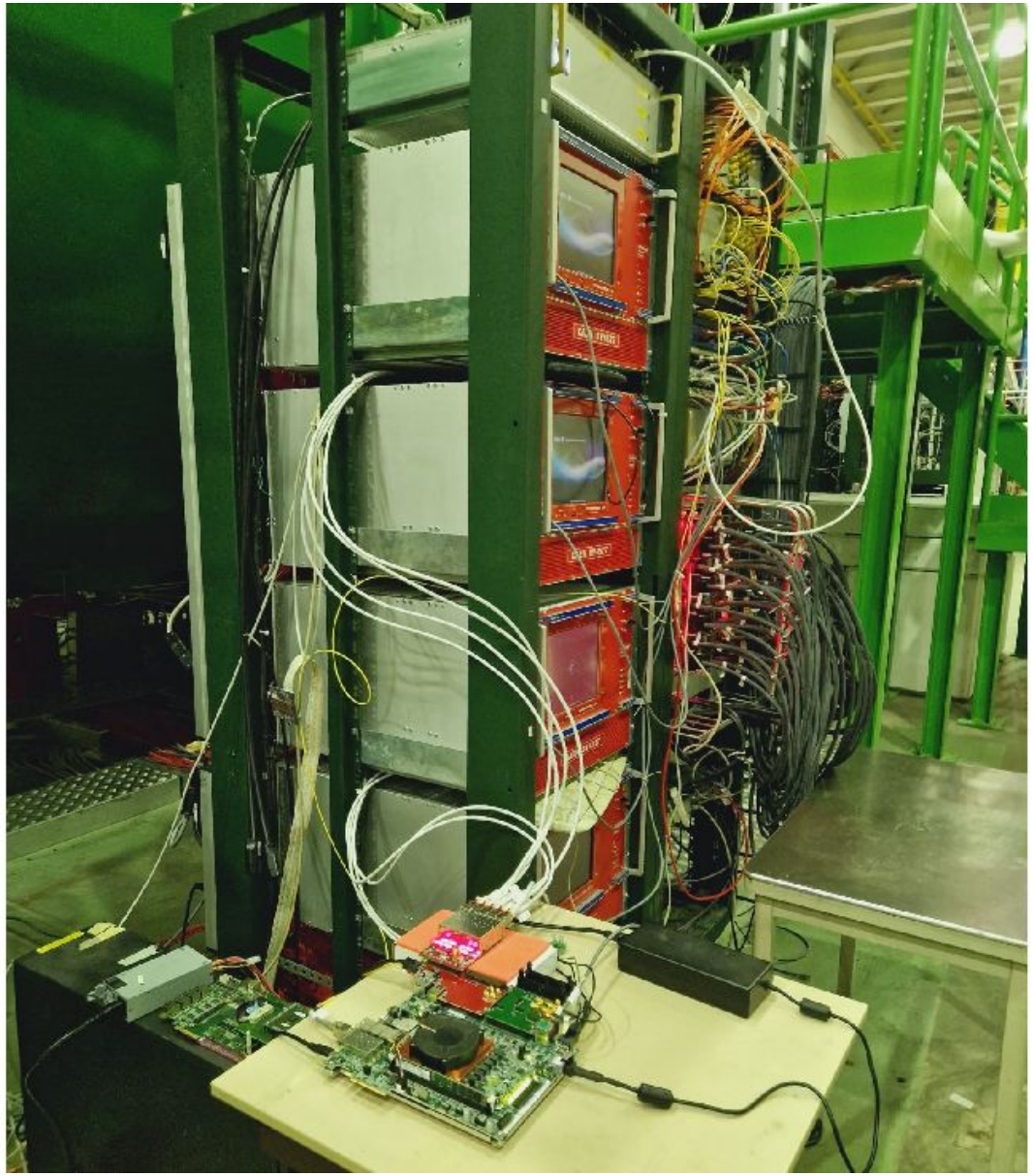
LUT = 14%
Flip−Flop = 6%
DSP = 7%
BRAM = 3%
on Versal VCK190

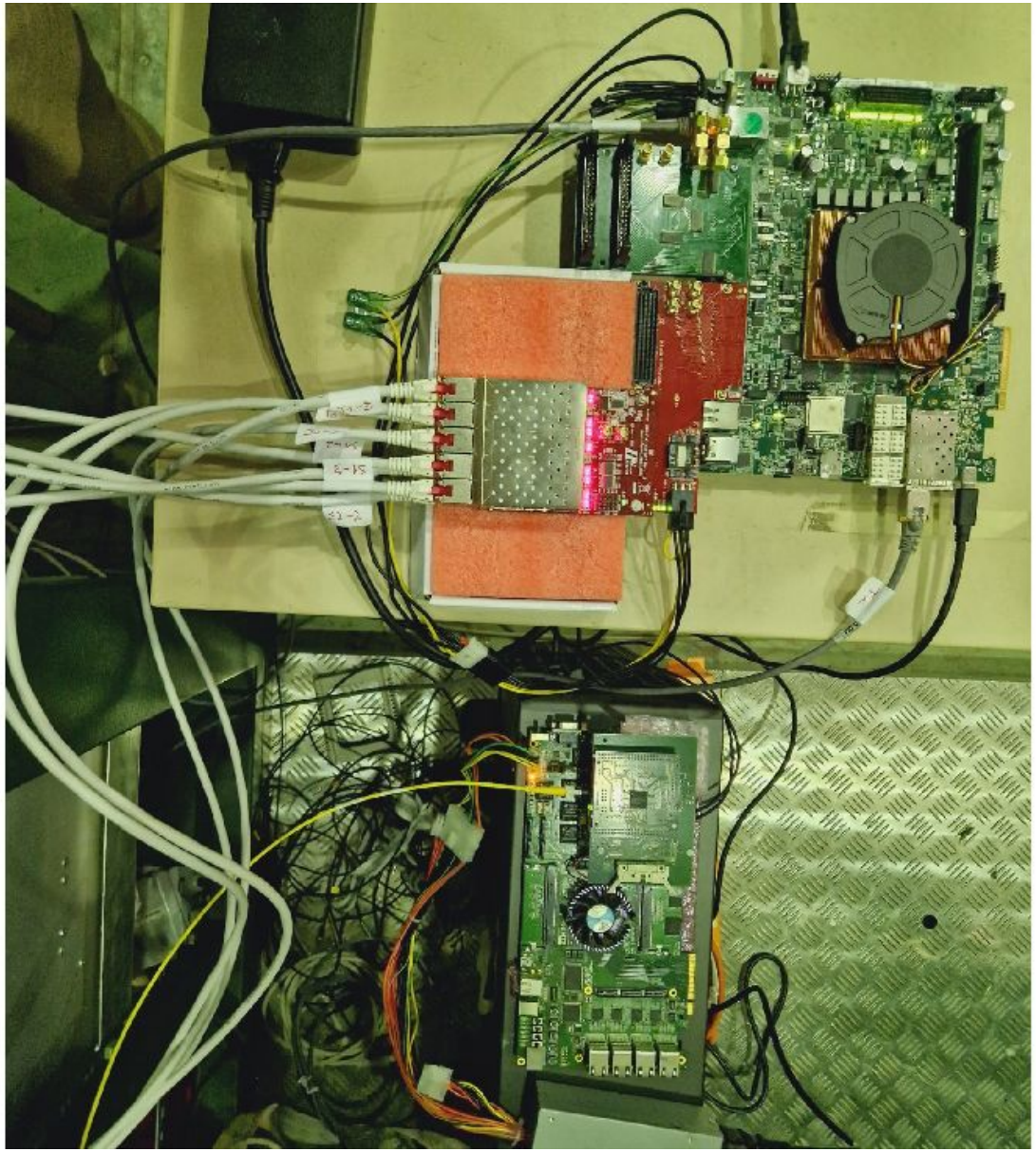
ROC Curve, Throughput & Latency



NN: **avg Throughput ≈ 21 MHz** Latency = 160 ns at 300 MHz clock

Integration of the FPGA-RICH Pipeline

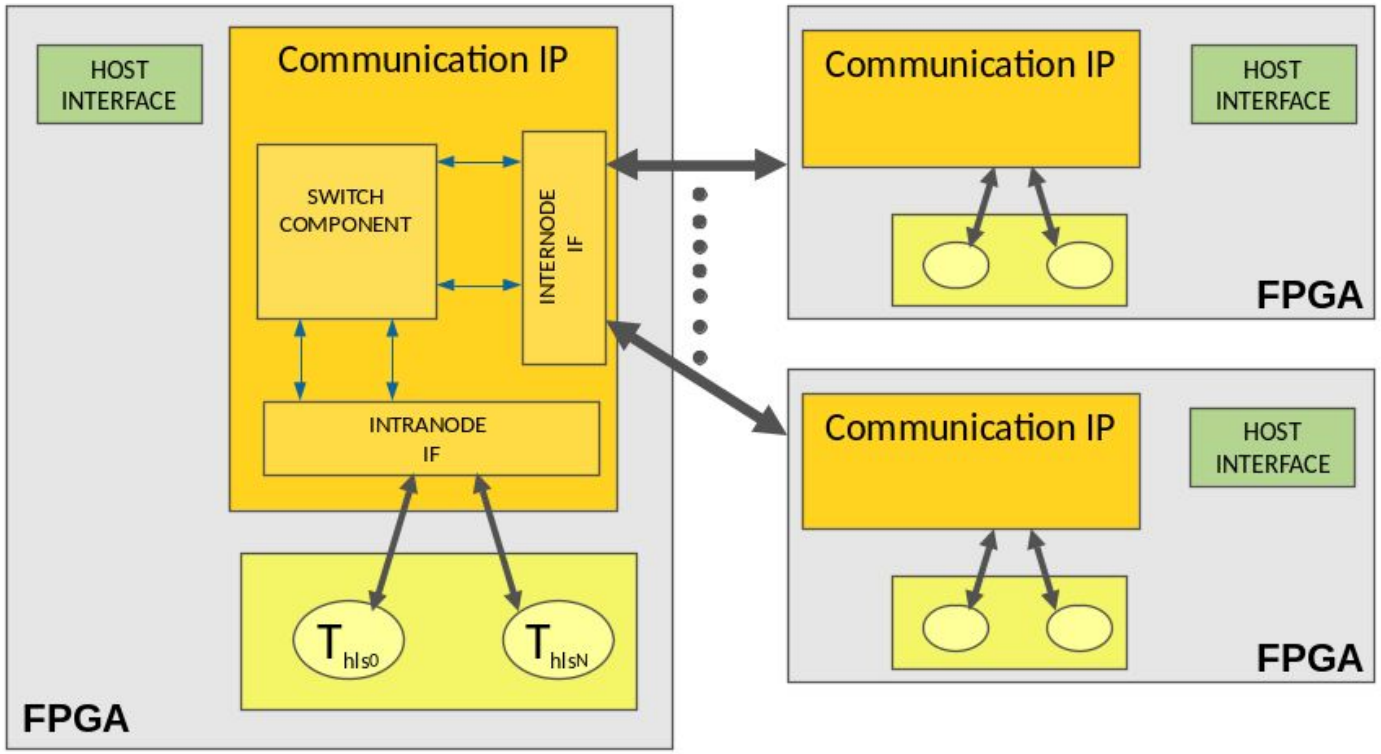


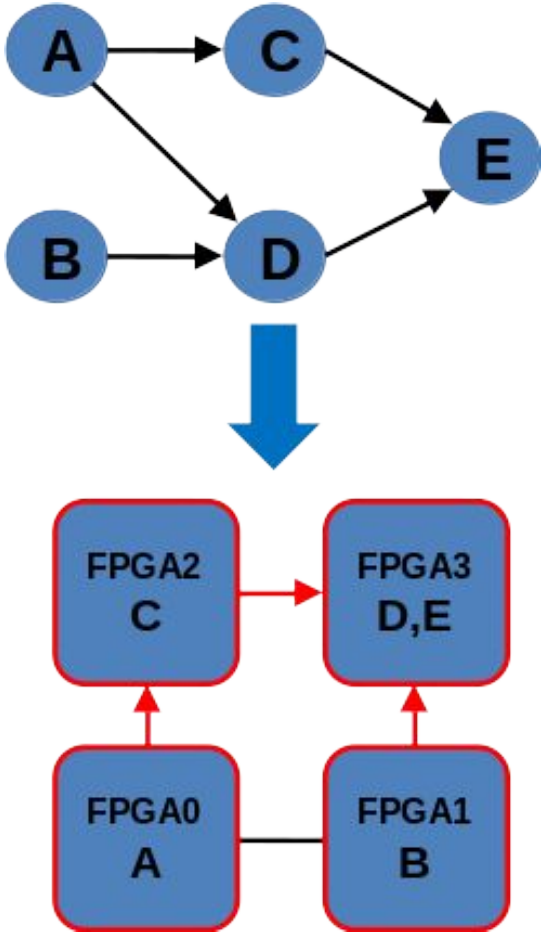


APEIRON: INFN Communication IP

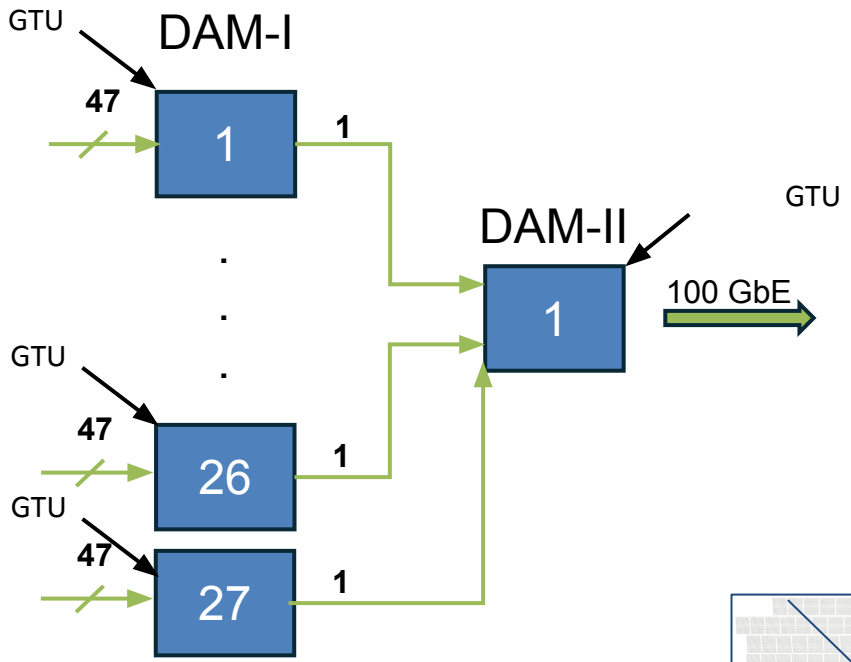
- INFN is developing the IPs implementing a **direct network** that allows **low-latency** data transfer between processing **High Level Syntesis (C++) tasks** deployed on the same FPGA (intra-node communication) and on different FPGAs (inter-node communication).

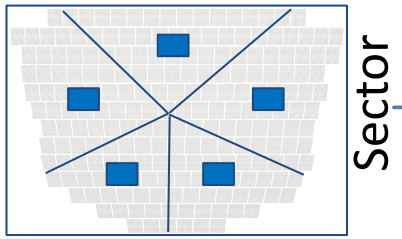
- Inter-node Latency < 1us for packet sizes up to 1kB (source and destination buffers in BRAM)





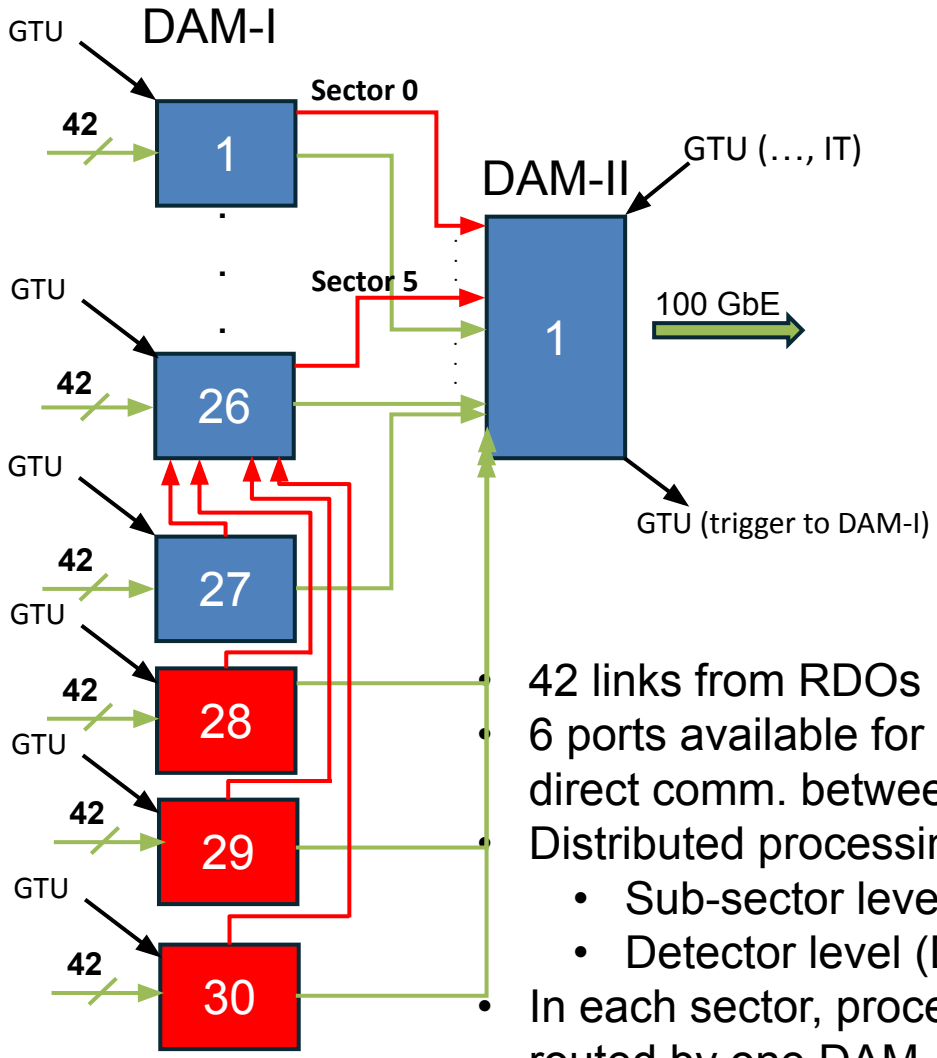
dRICH Data Reduction Stage on FPGA: example deployment





Five DAM-I per sector

Sector

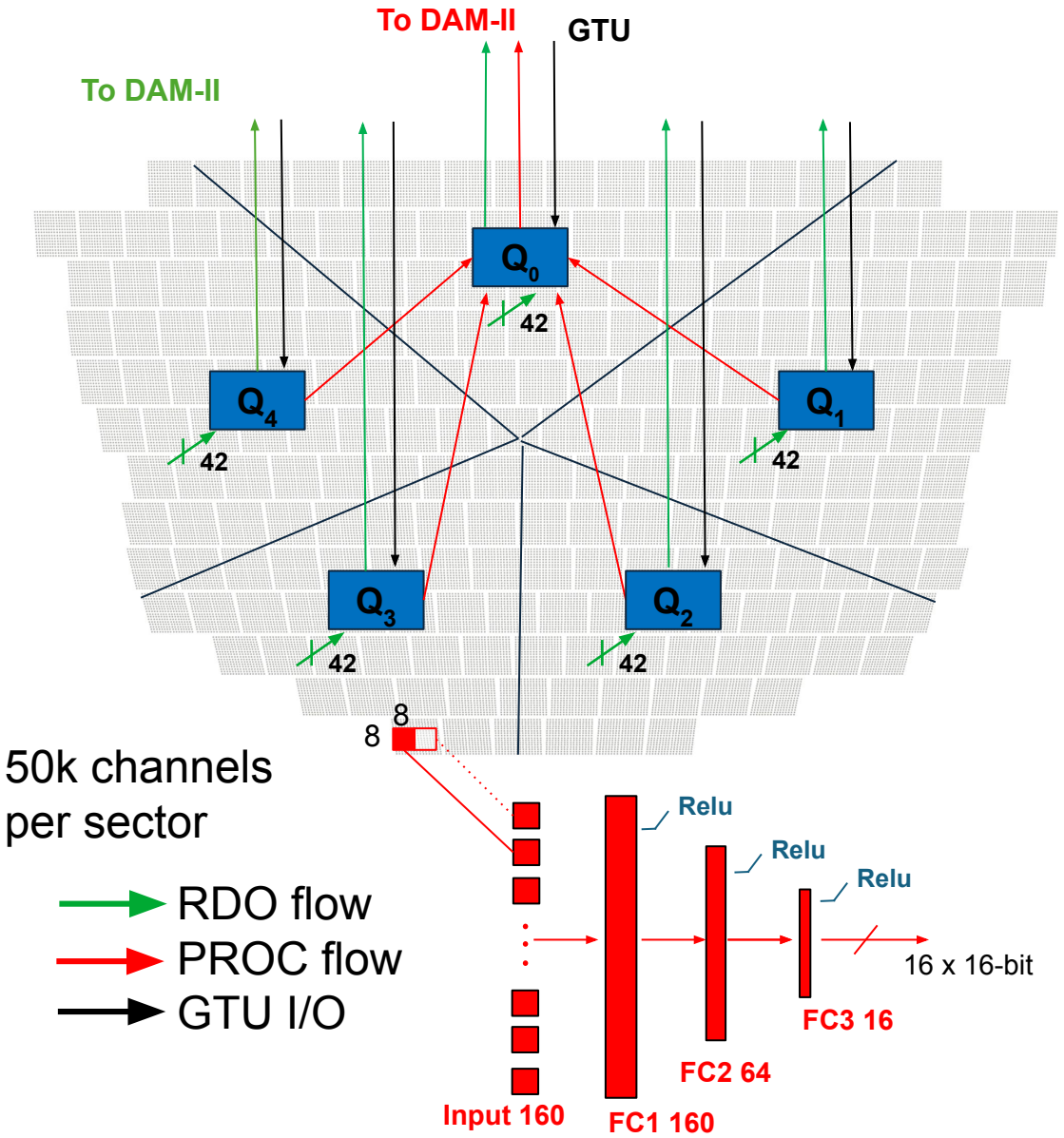


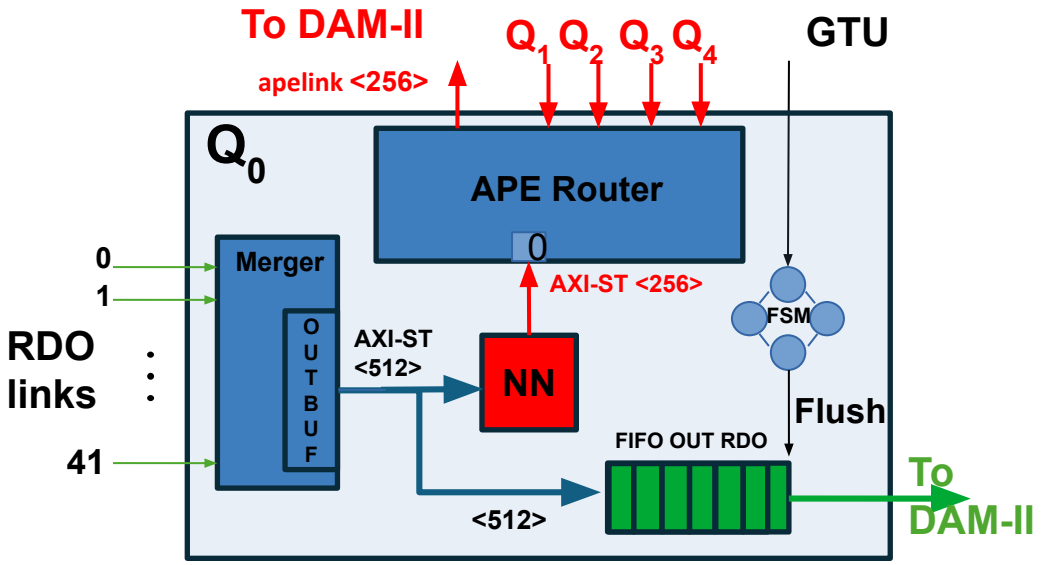
- 42 links from RDOs
- 6 ports available for direct comm. between DAMs
- Distributed processing on FPGAs
  - Sub-sector level (DAM-I)
  - Detector level (DAM-II)
- In each sector, processed data routed by one DAM-I to DAM-II

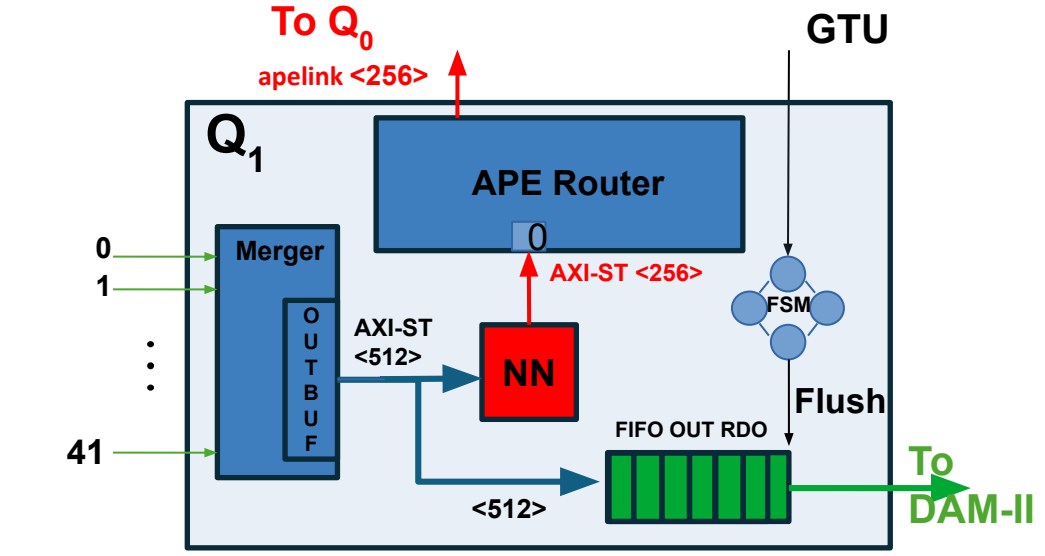
RDO Data

Processed Data

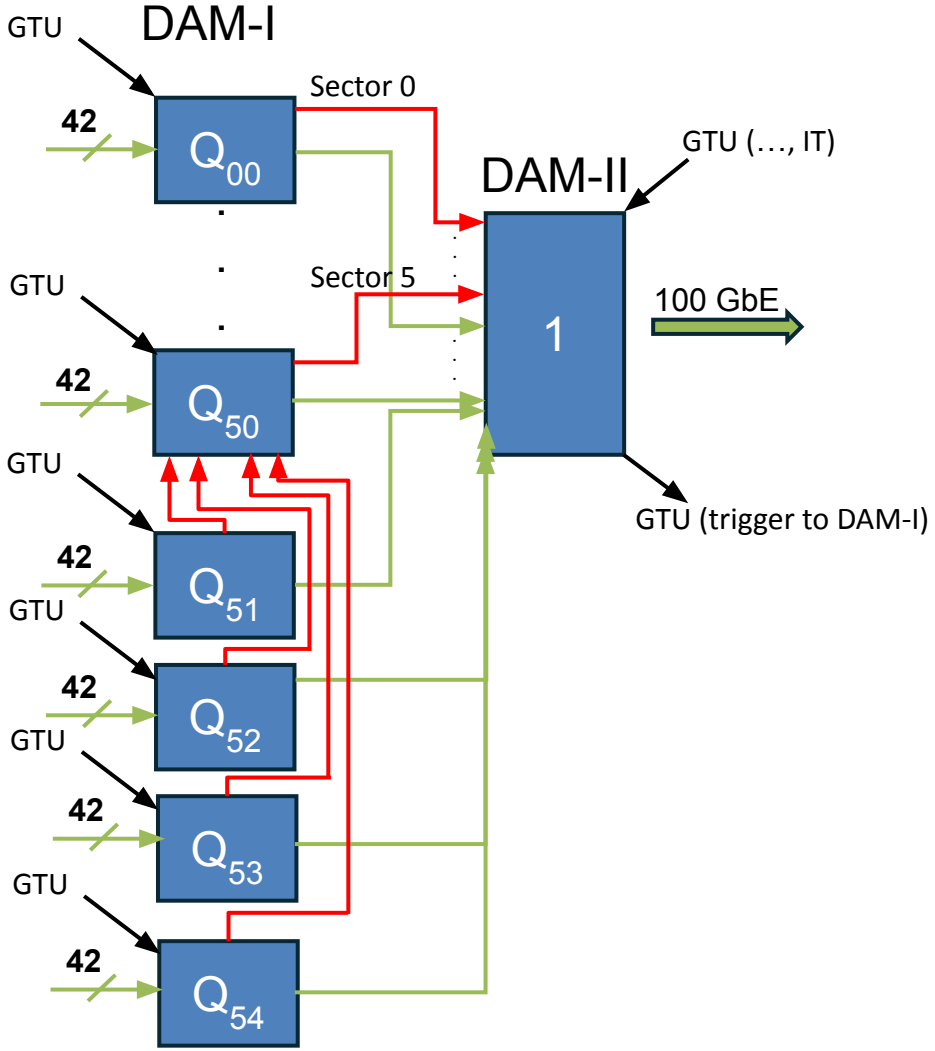
dRICH Data Reduction Stage on FPGA: example deployment

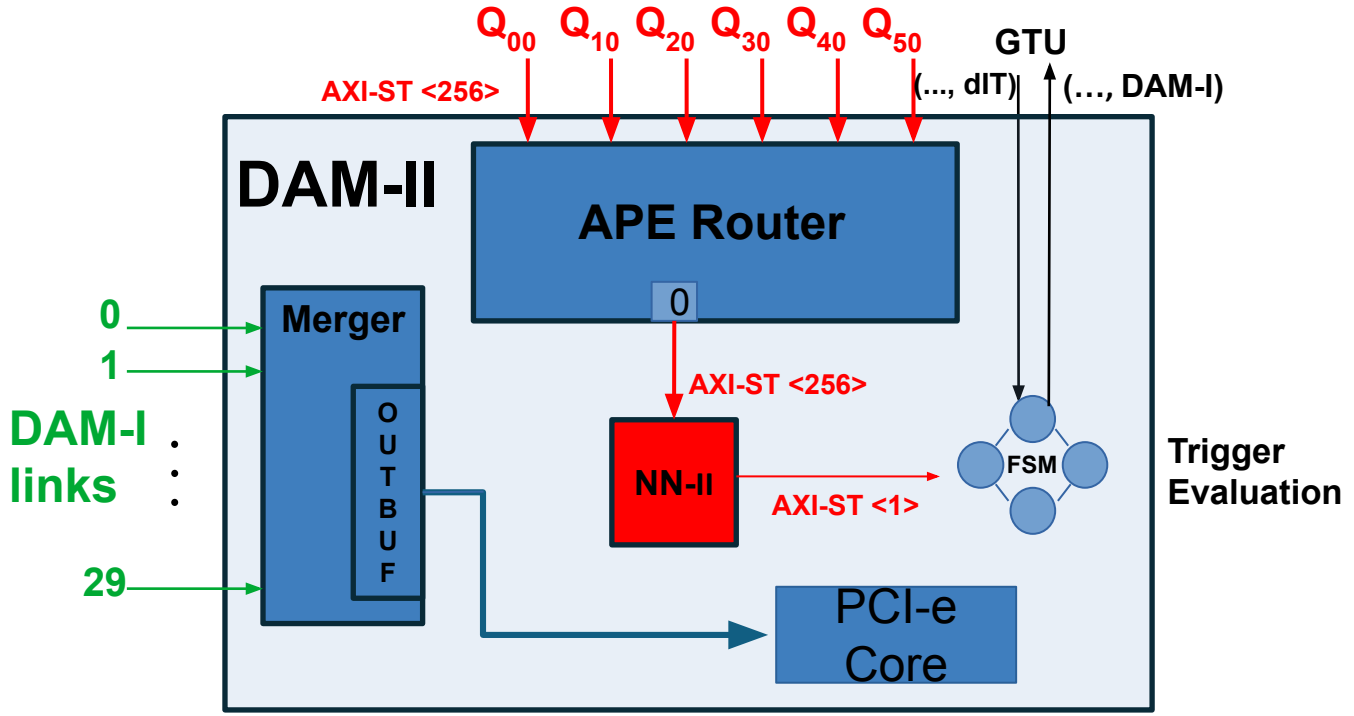


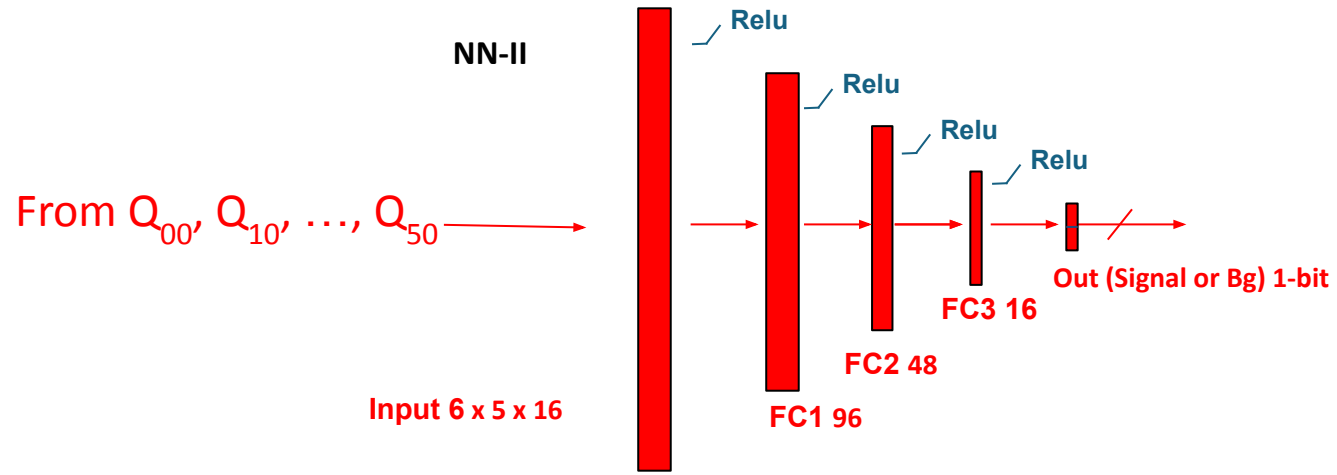




dRICH Data Reduction Stage on FPGA: example deployment





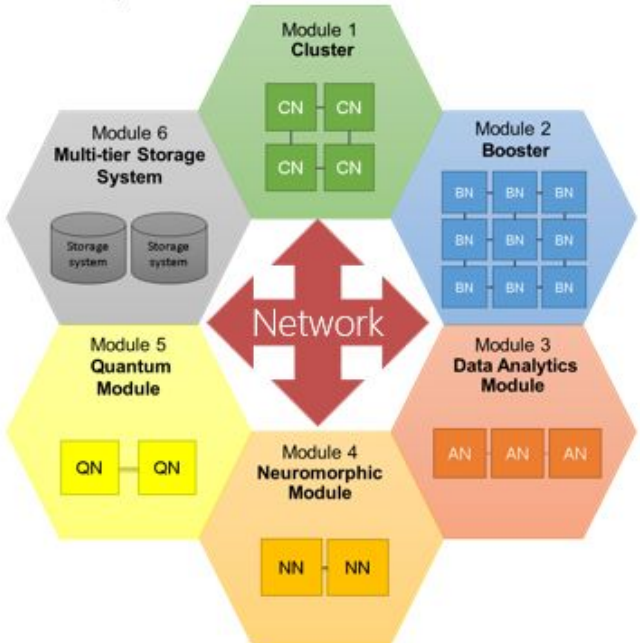


Current status and outlook

- We have started collecting datasets and experimenting with inference models.
- Details of the final deployment will be affected by several factors
  - Final selection on the inference model(s): BDT, MLP, CNN, GNN, …
  - Net amount of FPGA resources available (discounting the “standard” DAQ firmware) in DAMs.
  - Actual additional DAQ resources (DAMs, …) dedicated to the data reduction system.
- Possible additional features
  - Provide services (statistics) for the online monitoring.
  - Having track seeds information from the Interaction Tagger could enable more sophisticated features (e.g. Particle counting, Particle identification)

Background activities useful for the «GPU approach»,
see next slides…

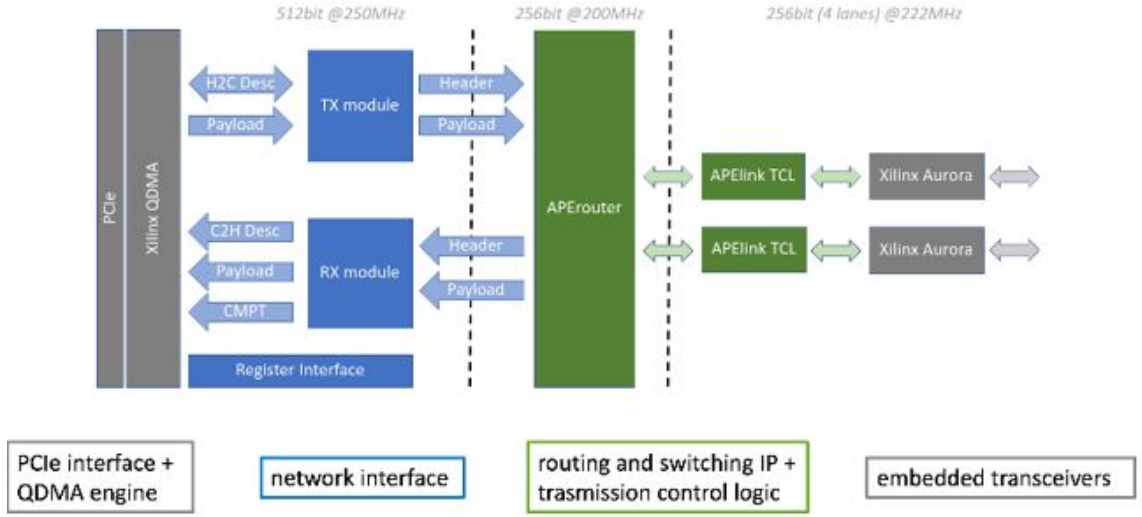
APEnetX Network Interface Card

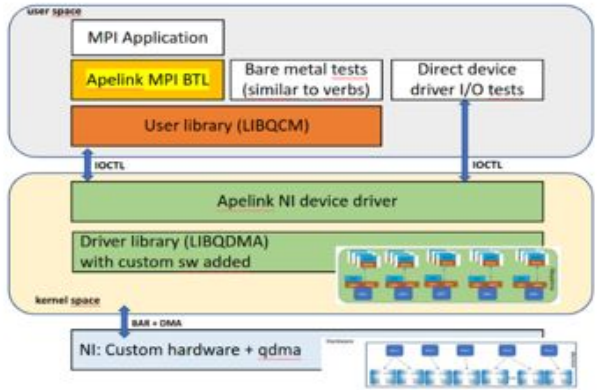


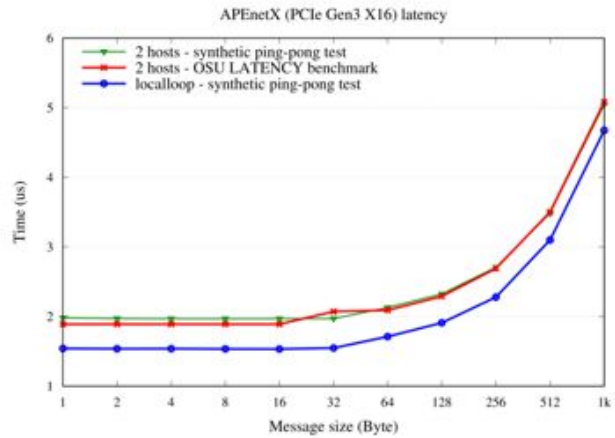
- HPC (High Performance Computing) ; HPDA (High-Performance Data Analytics); AI (Artificial Intelligence)
- Supercomputer: aggregation of resources that are organized to facilitate the mapping of applicative workflows
- HPC is part of the continuum of computing

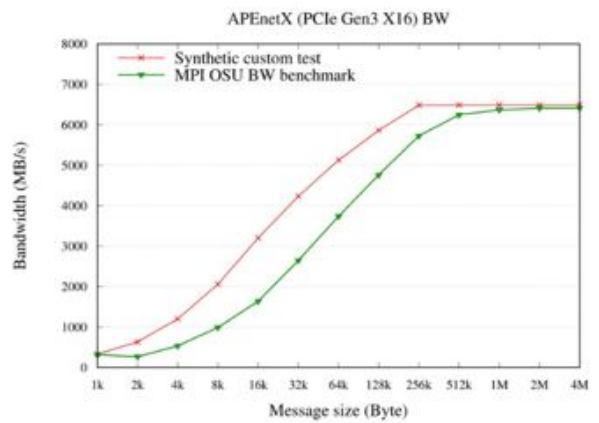
- INFN duties
  - Network Interface Card (APEnetX)
    - PCIe gen4 (GPU+CPU) + BXI link (Xilinx Alveo FPGA)
  - Co-Design through applications (NEST)
- TARGET
  - Developing network IPs to optimize spiking neural network communication

APEnetX in BRAINSTAIN









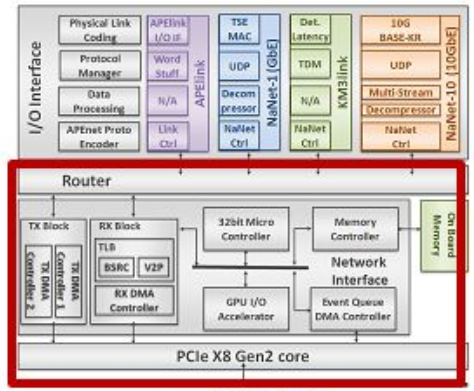
- Xilinx Alveo Board DMA engine
- Providing proprietary software driver and low-level communication library
- NVIDIA GPUDirect RDMA
- Custom OpenMPI BTL
- Bandwidth per channel 57.6 Gbps
- Latency 1.9us
- Validated through HPC-benchmark
- Interoperability with the BXI interconnect
- Proprietary priority management mechanism to improve QoS of the data transmission system

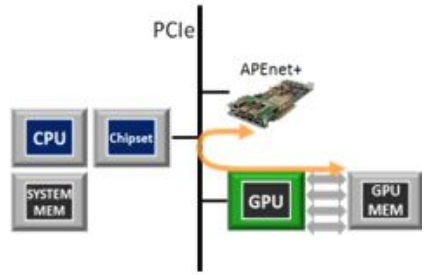
From HPC to HEP: NaNet Project

NaNet: Design and implementation of a family of FPGA-based PCIe Network Interface Cards:

- Bridging the front-end electronics and the software trigger computing nodes.
- Supporting multiple link technologies and network protocols.
- Optimizing data transfers with GPU accelerators.
- Enabling a low and stable communication latency.
- Having a high bandwidth.
- Processing data streams from detectors on the fly (data compression/decompression and re-formatting, coalescing of event fragments, …).

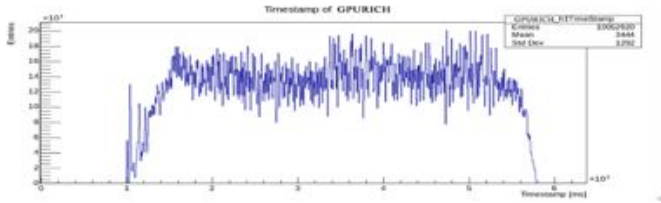
NaNet architecture



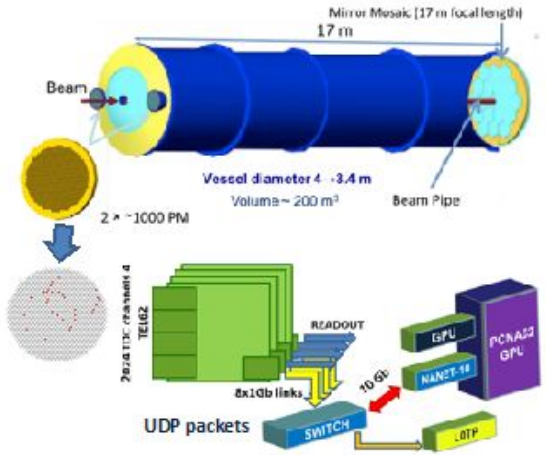


CPU+GPU

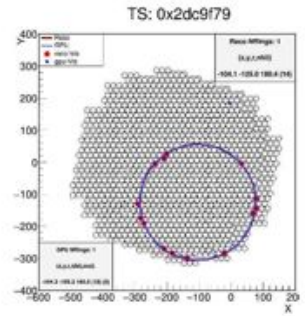
GPU-RICH generated primitives



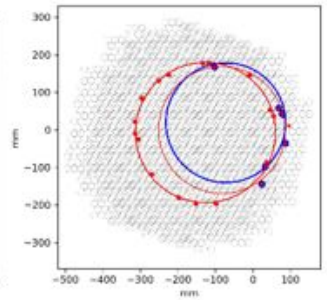
GPU-RICH overview



GPU 1 ring == Reco 1 ring



GPU 1 rings
Reco 2 rings



NA62 KM3NeT NA62

	NaNet-1	NaNet³	NaNet-10	NaNet-40
Year	Q3 - 2013	Q1 - 2015	Q2 - 2016	Q3 - 2019
Device Family	Altera Stratix IV	Altera Stratix V	Altera Stratix V	Altera Stratix V
Channel Technology	1 GbE	KM3link	10 GbE	40 GbE
Transmission Protocol	UDP	TDM	UDP	UDP
Number of Channel	1	4	4*	2
PCIe	Gen2 x8	Gen2 x8	Gen3 x8**	Gen3 x8
SoC	NO	NO	NO	NO
High Level Synthesis	NO	NO	NO	YES
nVIDIA GPUDirect RDMA	YES	YES	YES	YES
Real-time Processing	Decomp.	Decomp.	Decomp. Merger	?

Thanks

Contacts
Presenter: luca.pontisso@roma1.infn.it
Roma1: alessandro.lonardo@roma1.infn.it
Roma2: roberto.ammendola@roma2.infn.it

BACKUP SLIDES

The NA62 Experiment at CERN SPS

- Measurement of the K^+ decay:

  $BR(K^+ \to \pi^+ \nu \bar{\nu})$

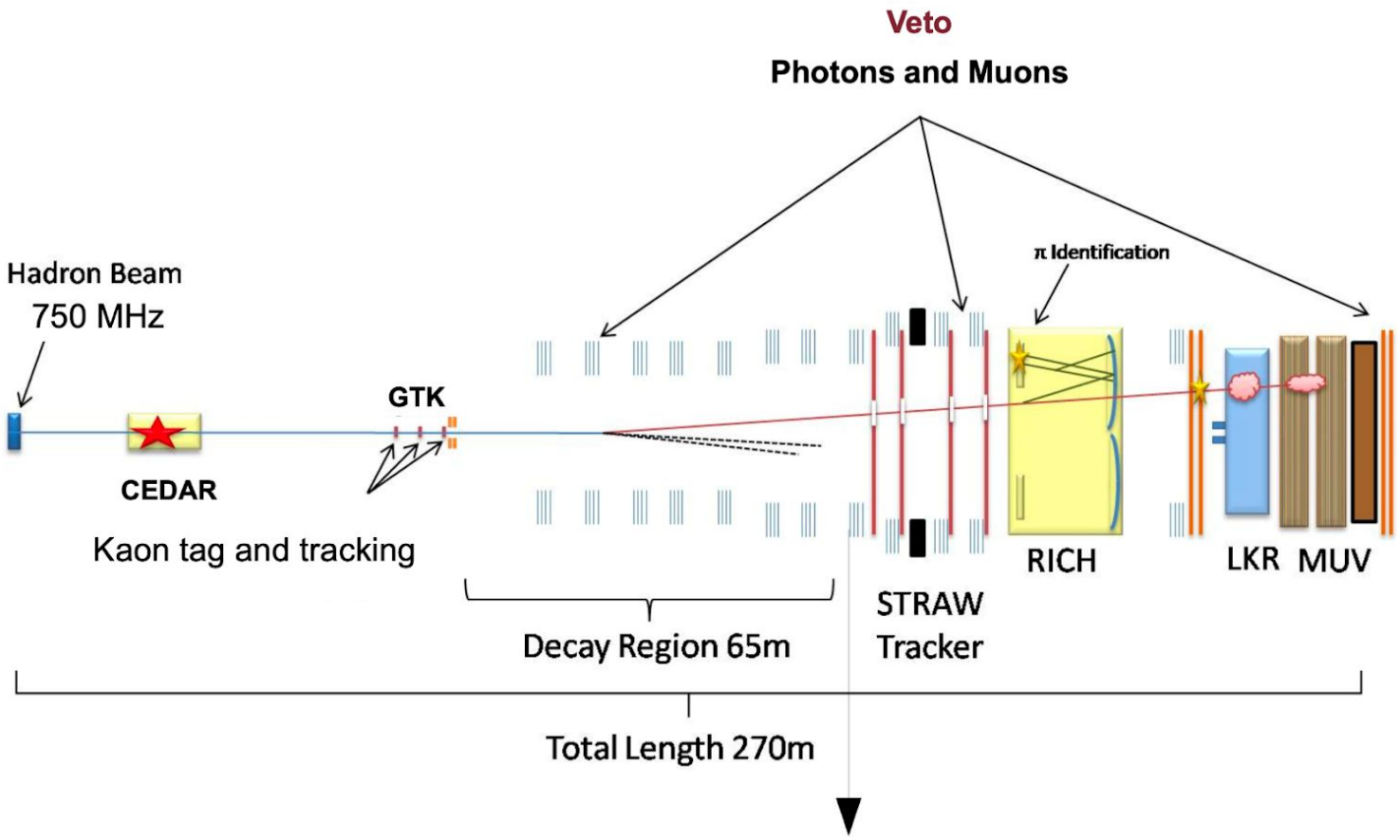
- **Ultra-rare** channel, **SM prediction**:

  $BR_{SM} = (8.60 \pm 0.42) \times 10^{-11}$

  Run I NA62 measurement:

  $BR_{NA62} = (10.6^{+4.0}_{-3.4}|_{stat} \pm 0.9_{syst}) \times 10^{-11}$

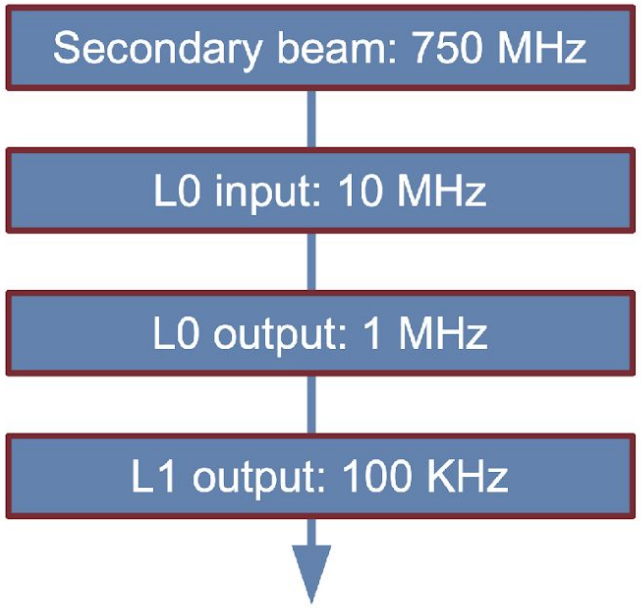
- **Fixed Target experiment**:
  75 GeV secondary hadron beam
  (6% kaons).

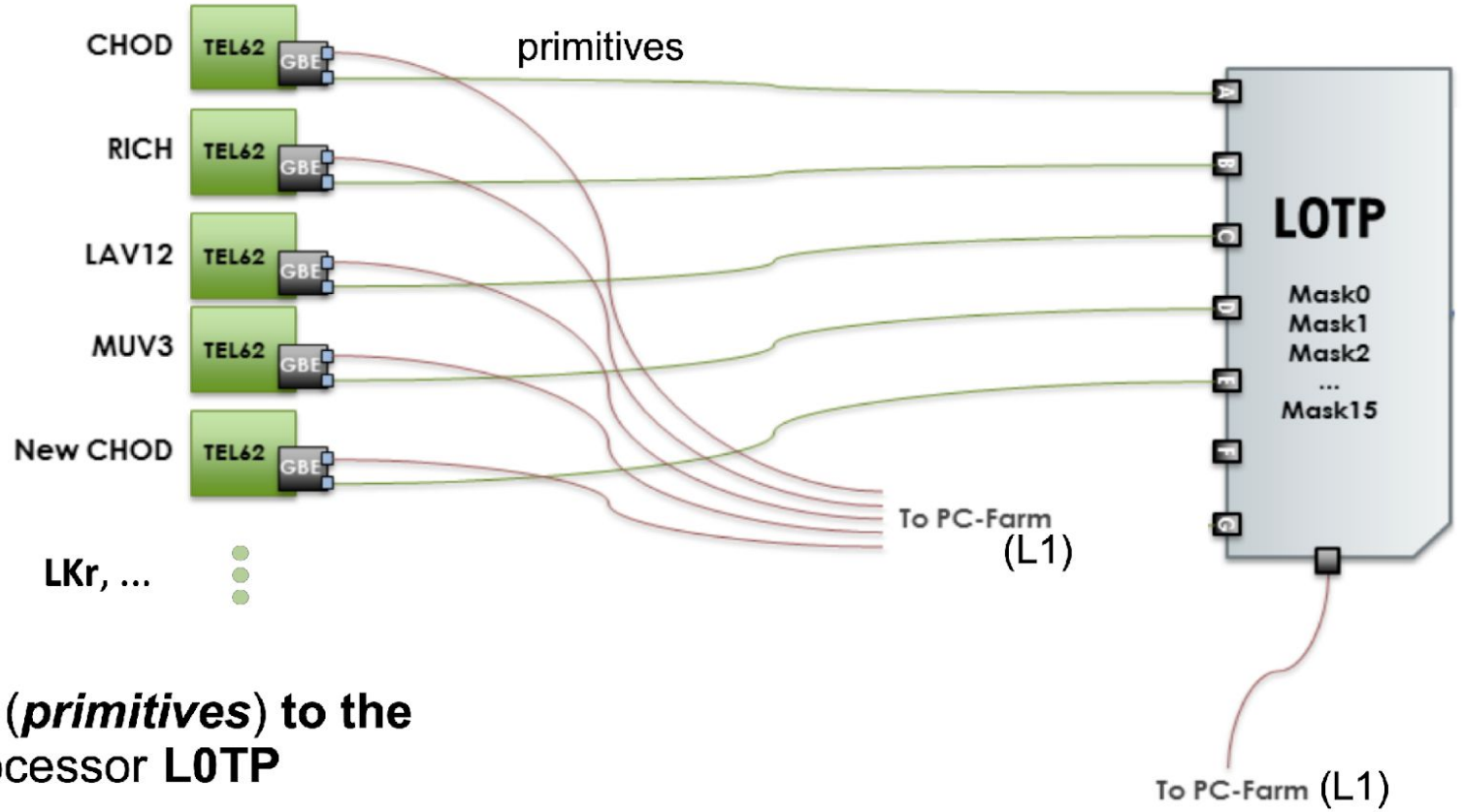


10 MHz event rate

Need highly selective filtering sytem

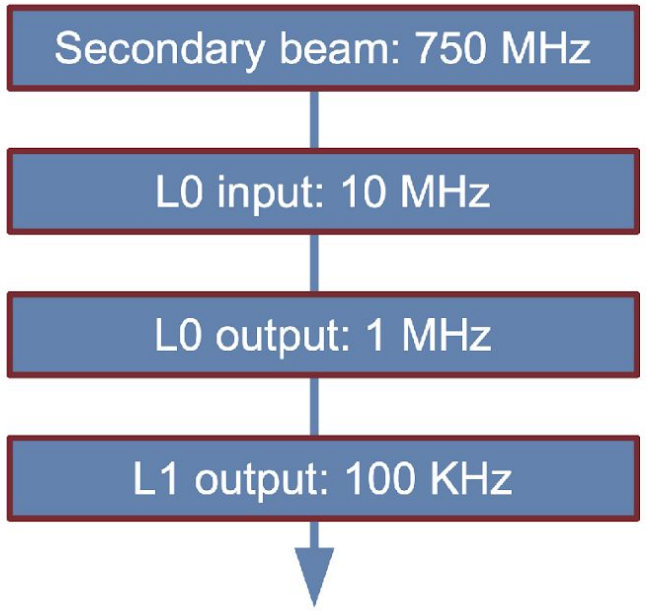
The NA62 Data Acquisition and Low Level Trigger

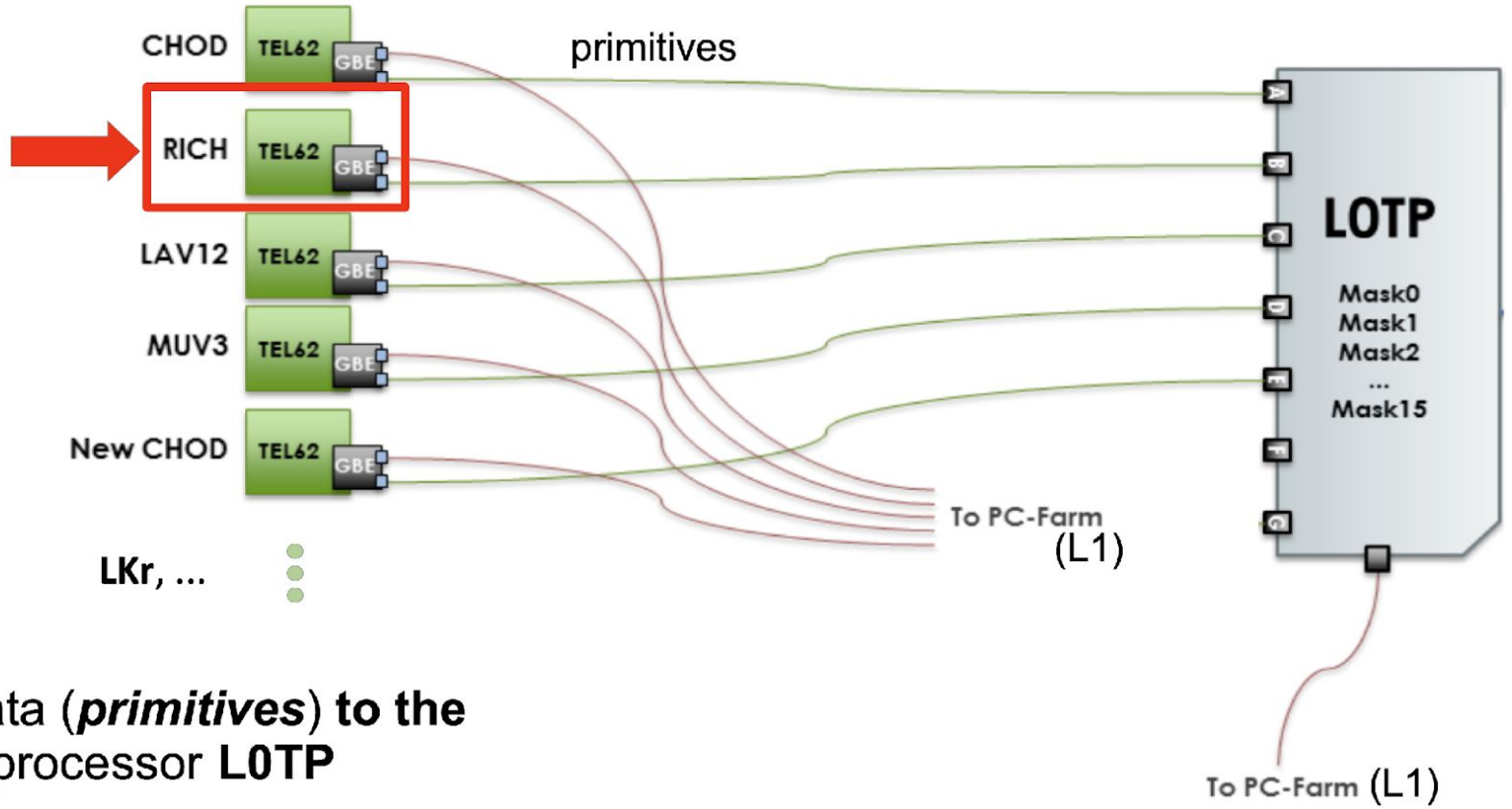




- Some detectors send raw data (***primitives***) **to the** FPGA-based level-0 trigger processor **L0TP**
- Primitives are generated from **TEL62 read out boards**
- L0TP **checks conditions** (**Masks**) to determine if an event should be selected and sent to L1
- Masks rely on the **physics information inside the primitives** (Energy, hit multiplicity, position, ...)

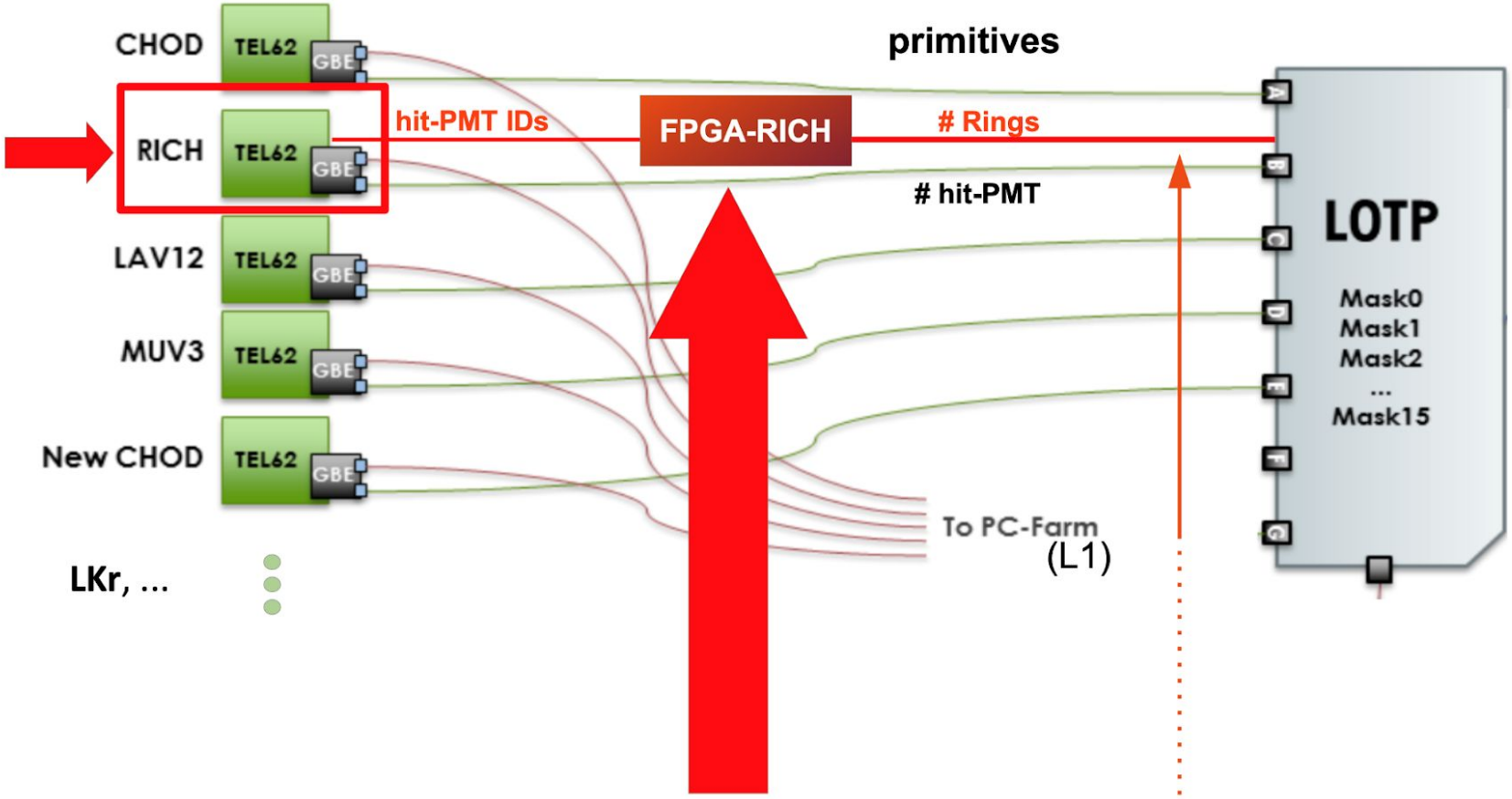
The NA62 Data Acquisition and Low Level Trigger

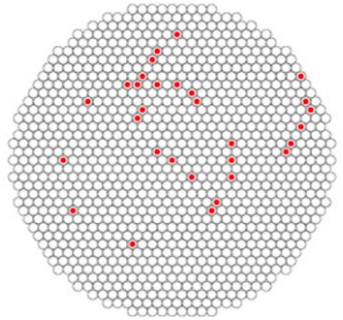




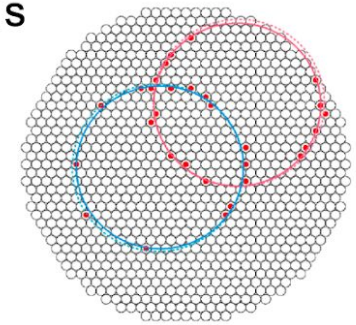
- Some detectors send raw data (***primitives***) **to the** FPGA-based level-0 trigger processor **L0TP**
- Primitives are generated from **TEL62 read out boards**
- L0TP **checks conditions** (**Masks**) to determine if an event should be selected and sent to L1
- Masks rely on the **physics information inside the primitives** (Energy, hit multiplicity, position, ...)

Smart Primitives: FPGA-RICH





RICH primtives: # of hit-PMTs

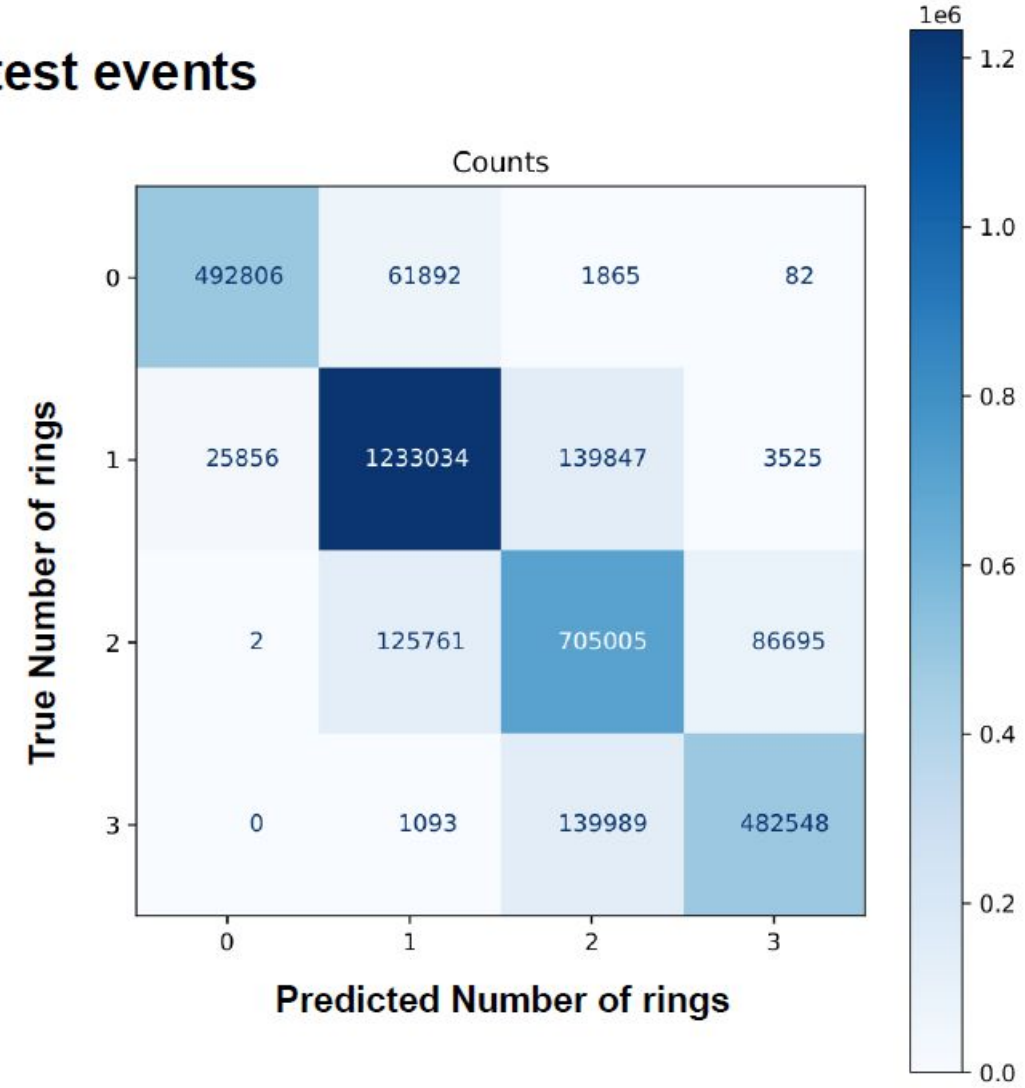


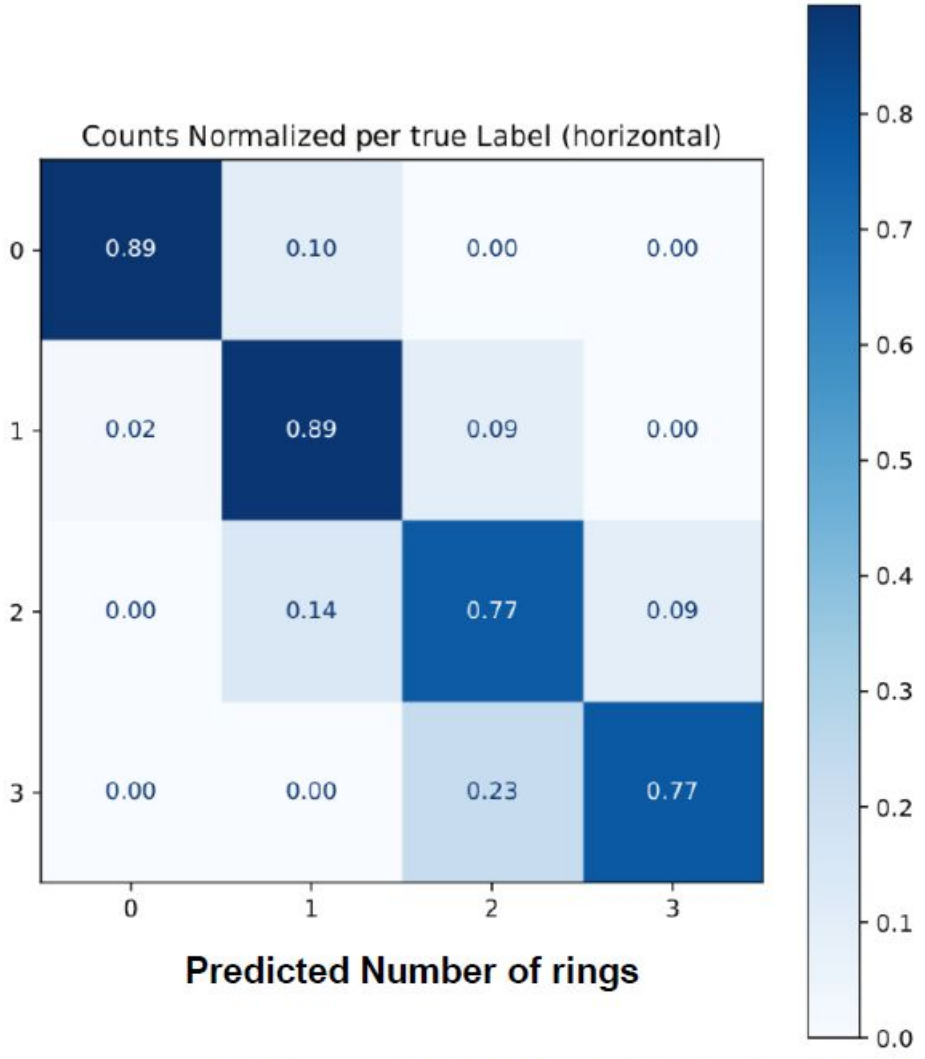
FPGA-RICH primtives:
# of rings (0, 1, 2, 3+)
(more in the future, e.g. # of e^-)

FPGA-RICH: **reconstruct the rings geometry online** using an AI algorithm on FPGA, to generate a refined primitive stream for L0TP selection masks

Neural Network Sensitivity

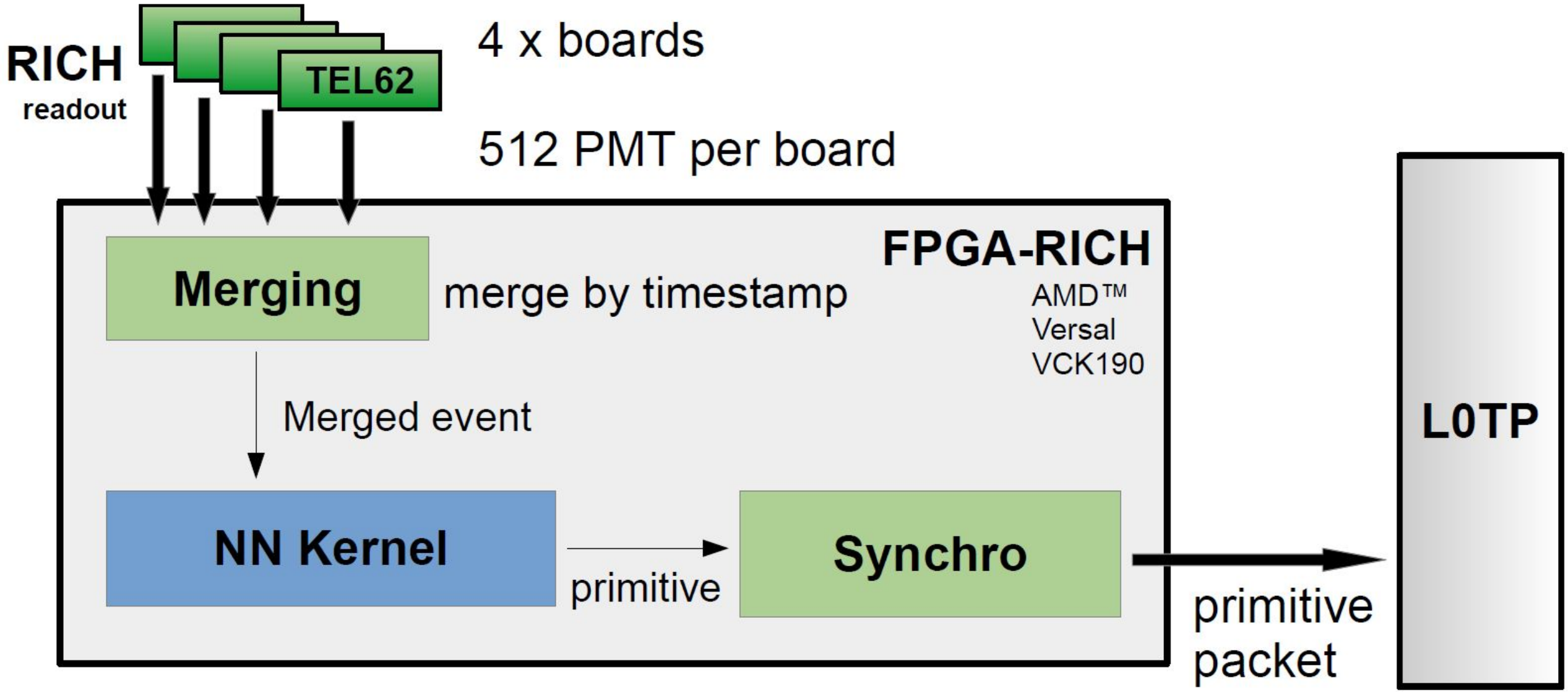
3.5 M test events





NN git: https://baltig.infn.it/ape-lab/fpgarich

Integration of the FPGA-RICH Pipeline

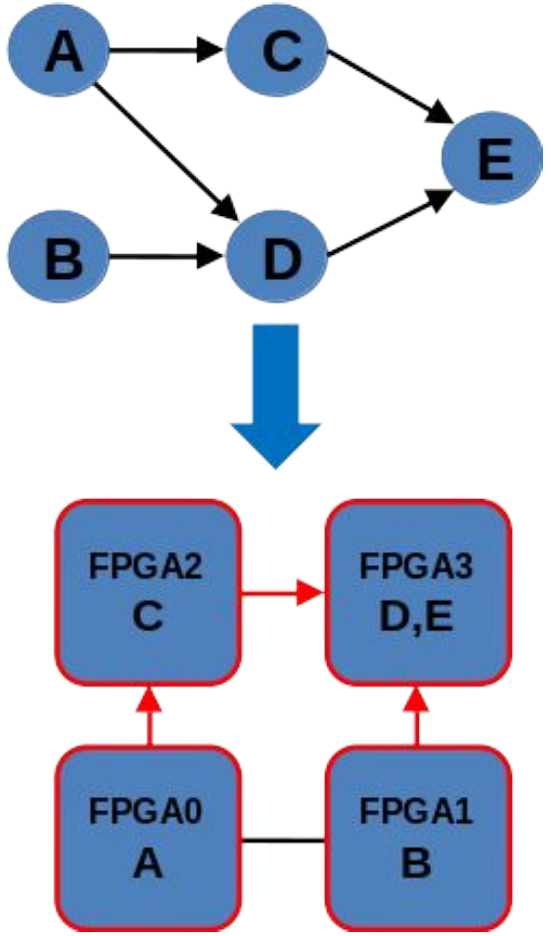


APEIRON: an overview

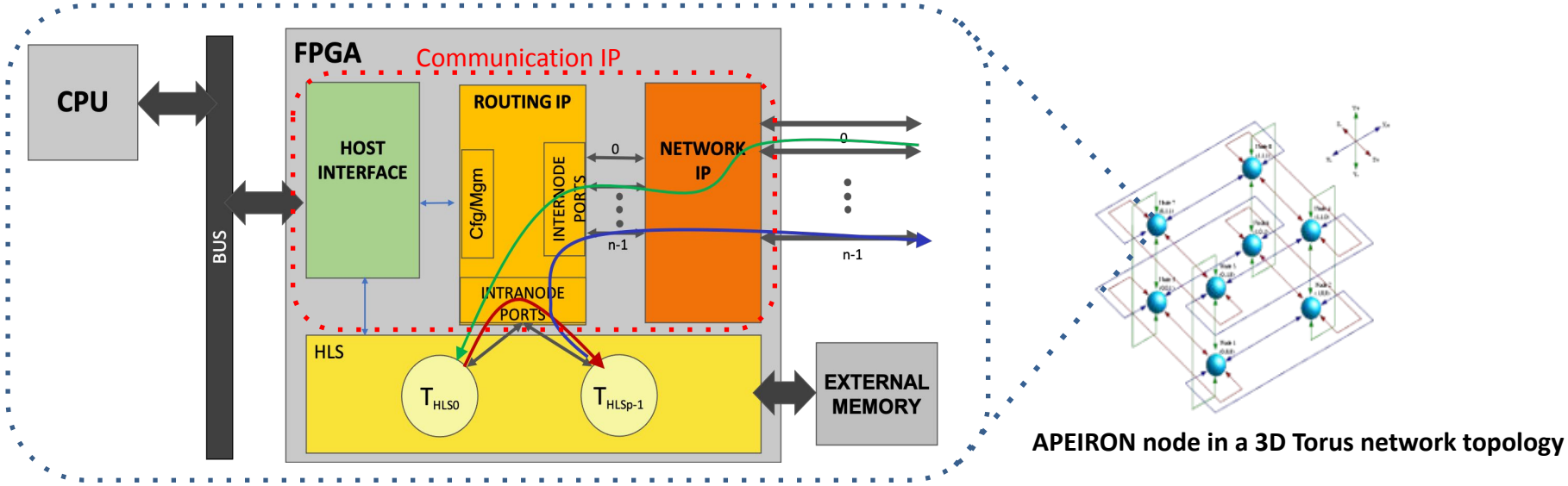
- **Goal:** develop a framework offering hardware and software support for the execution of real-time dataflow applications on a system composed by interconnected FPGAs .

  - Map the dataflow graph of the application on the distributed FPGA system and offers runtime support for the execution.
  - Allow users with no (or little) experience in hardware design tools, to develop their applications on such distributed FPGA-based platforms
    - Tasks are implemented in C++ using High Level Synthesis tools (Xilinx Vitis).
    - Lightweight C++ communication API
      - Non-blocking *send()*
      - Blocking *receive()*

  - **APEIRON is based on Xilinx Vitis High Level Synthesis framework and on INFN Communication IP (APE Router)**



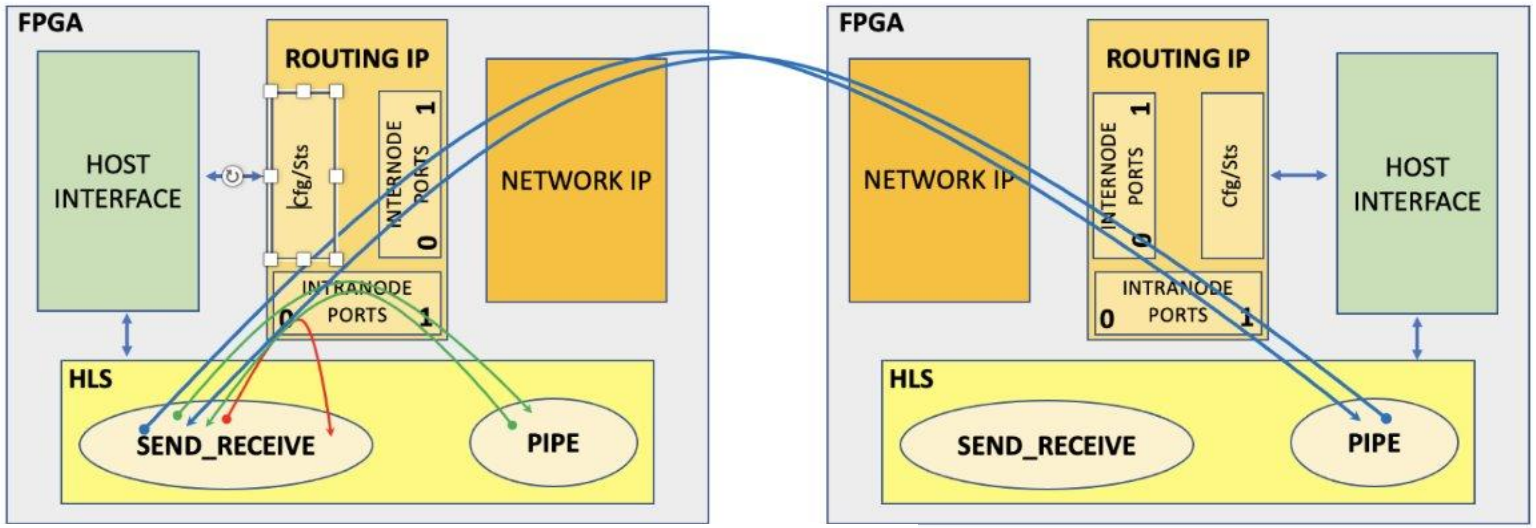
APEIRON: the Node

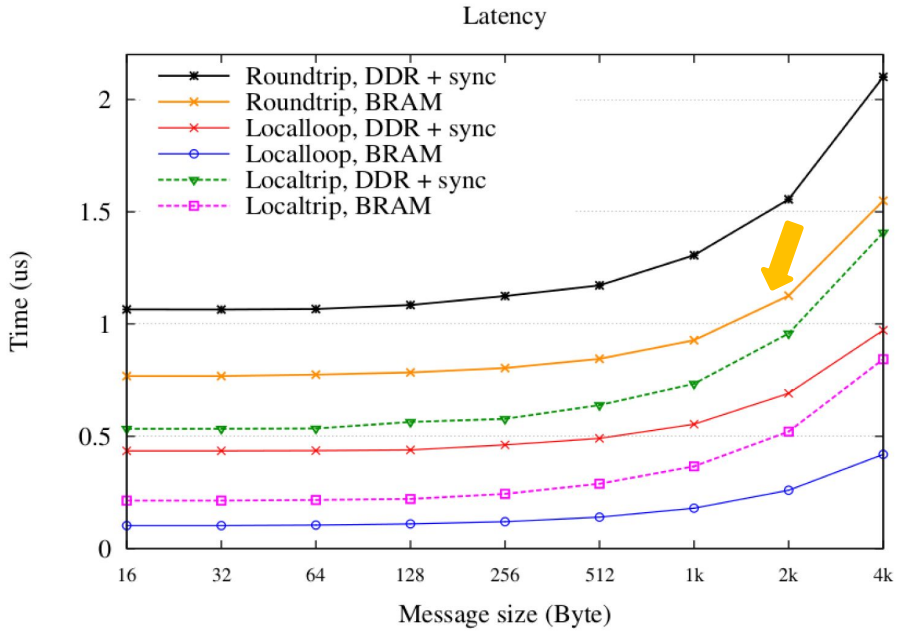


APEIRON node in a 3D Torus network topology

- **Host Interface IP**: Interface the FPGA logic with the host through the system bus.
  - Xilinx XDMA PCIe Gen3
- **Routing IP**: Routing of intra-node and inter-node messages between processing tasks on FPGA.
- **Network IP**: Network channels and Application-dependent I/O
  - APElink 40 Gbps
  - UDP/IP over 10 GbE
- **Processing Tasks**: user defined processing tasks (Xilinx Vitis HLS Kernels)

APEIRON: Communication Latency





Test modes

- Local-loop (red arrow)
- Local-trip (green arrows)
- Round-trip (blue arrows)

Test Configuration

- IP logic clock @ 200 MHz
- 4 intranode ports
- 2 internode ports
- 256-bit datapath width
- 4 lanes inter-node channels

Inter-node LATENCY (orange line) < 1us for packet sizes up to 1kB (source and destination buffers in BRAM)