FPGA-RICH: a low-latency, high-throughput online particle identification system for the NA62 experiment

Pierpaolo Perticaroli

(INFN Roma, APE Lab)

27th International Conference on Computing in High Energy & Nuclear Physics
CHEP 2024
Krakow – 24th October 2024

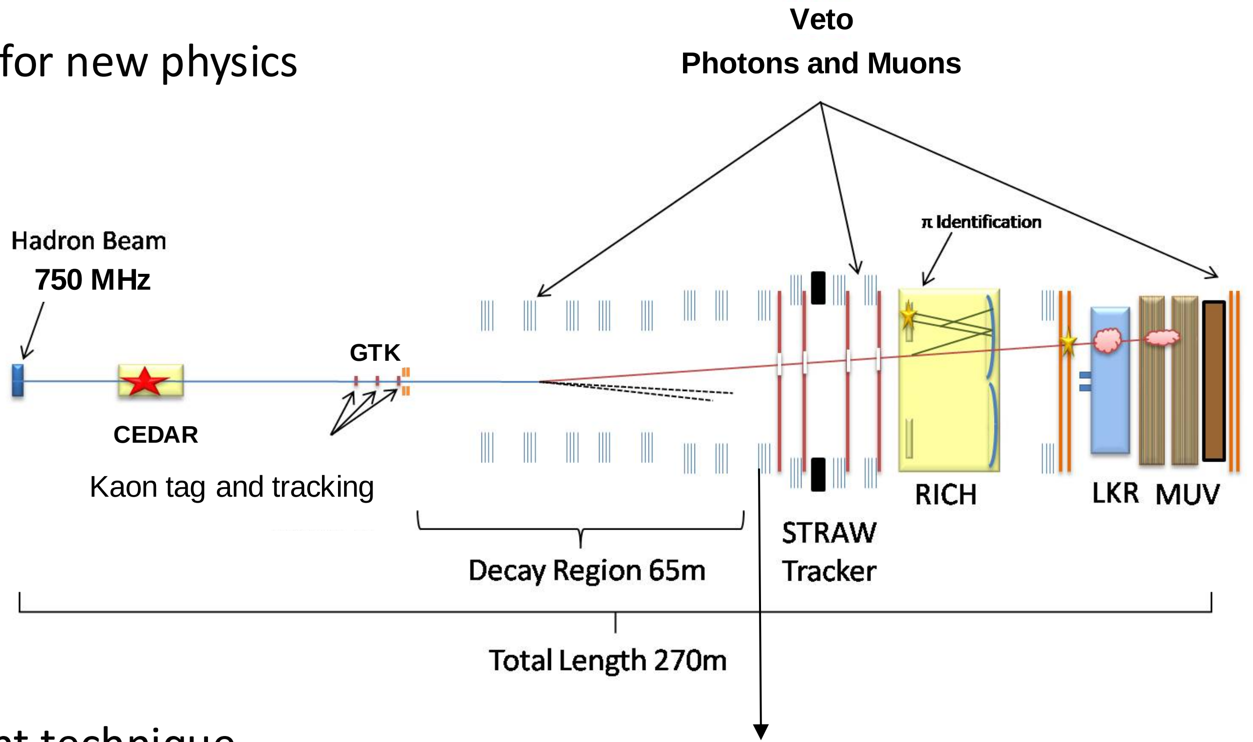
The NA62 Experiment at CERN SPS

- Study of rare Kaon decays and search for new physics
- Latest result:
  ultra-rare $K^+ \rightarrow \pi^+ \nu \nu$ decay observed
  with *bkg-only* hypothesis rejected at significance Z>5
- $\mathcal{B}^{16-22}_{\pi\nu\bar{\nu}} = \left(13.0^{+3.3}_{-2.9}\right) \times 10^{-11}$

  at 68% CL, with 2016-2022 data

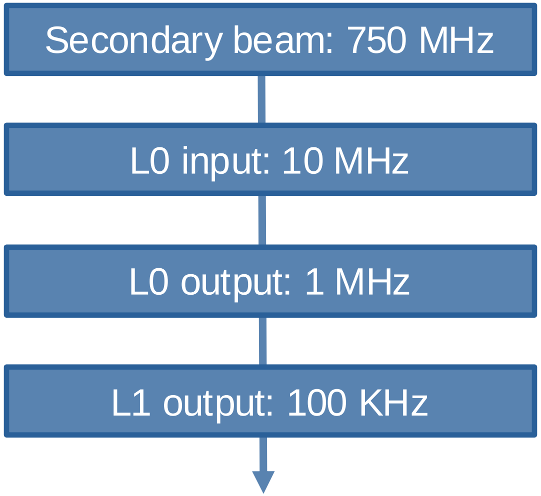
  https://indico.cern.ch/event/1447422/
- Fixed target experiment, decay-in-flight technique.

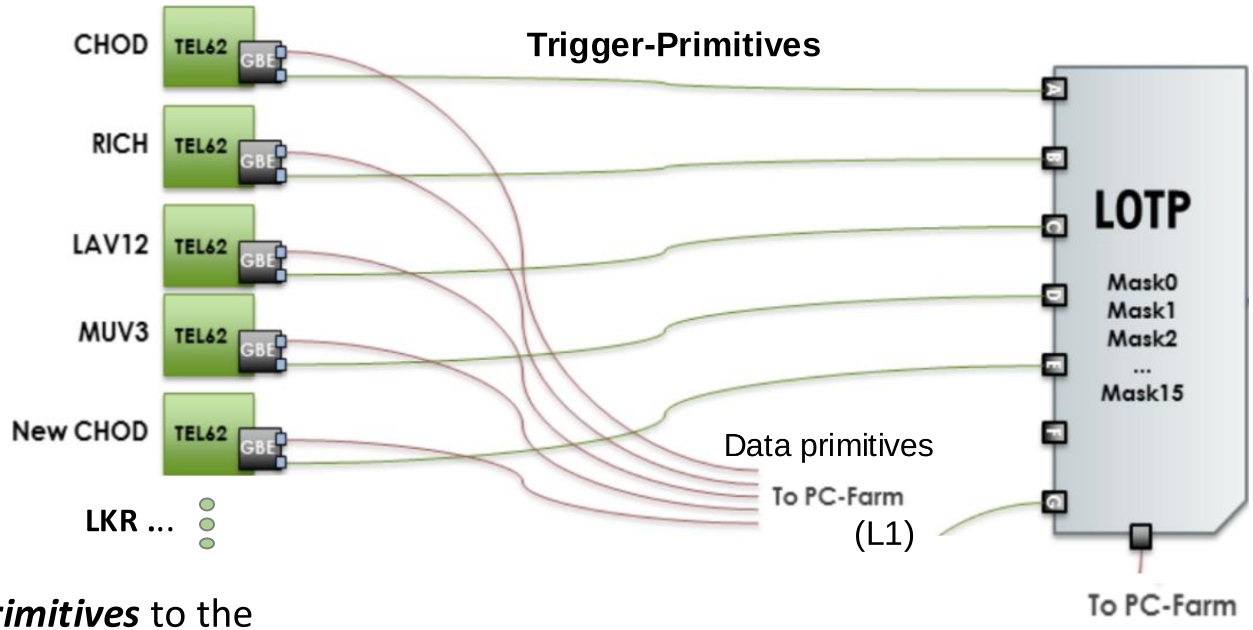
  75 GeV secondary hadron beam
  (6% kaons), nominal rate 750MHz



10 MHz event rate

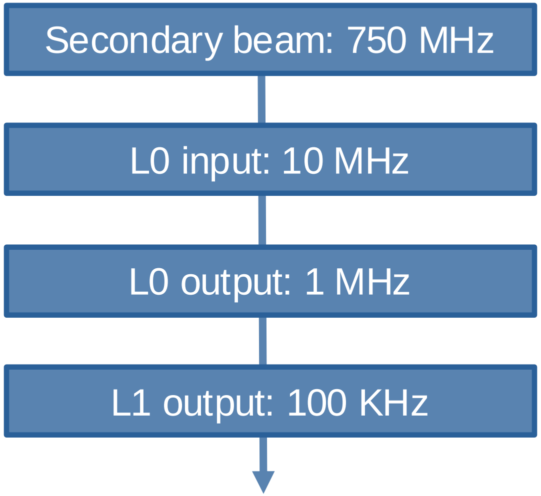
NA62 Data Acquisiton and Low Level Trigger

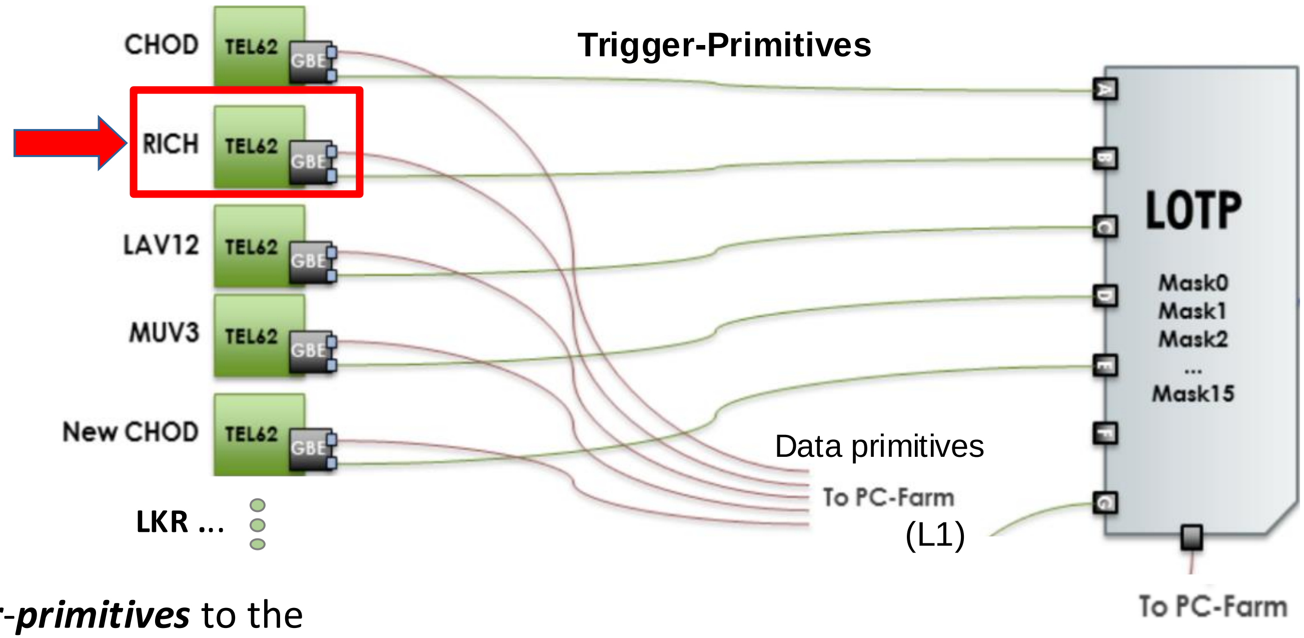




- Some detectors send raw data ***trigger-primitives*** to the **FPGA-based level-0 trigger processor L0TP** over 1GbE UDP links.
- **Read out boards (TEL62)** generate trigger-primitives and buffer detector events while waiting for L0 trigger (max latency 1ms).
- L0TP **checks configurable conditions** (**Masks**) against the physics information inside the primitives (Energy, hit multiplicity, position, ...) to deliver trigger. Data bursts ~ 5s long

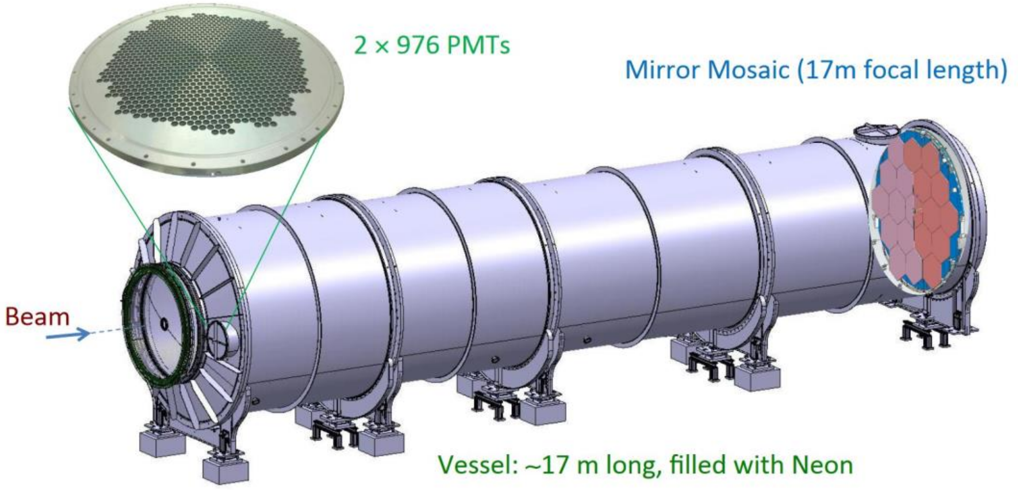
NA62 Data Acquisiton and Low Level Trigger

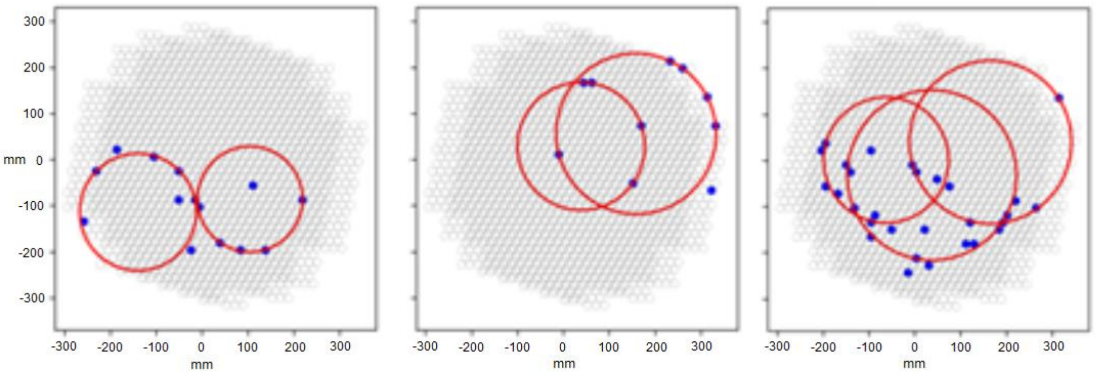




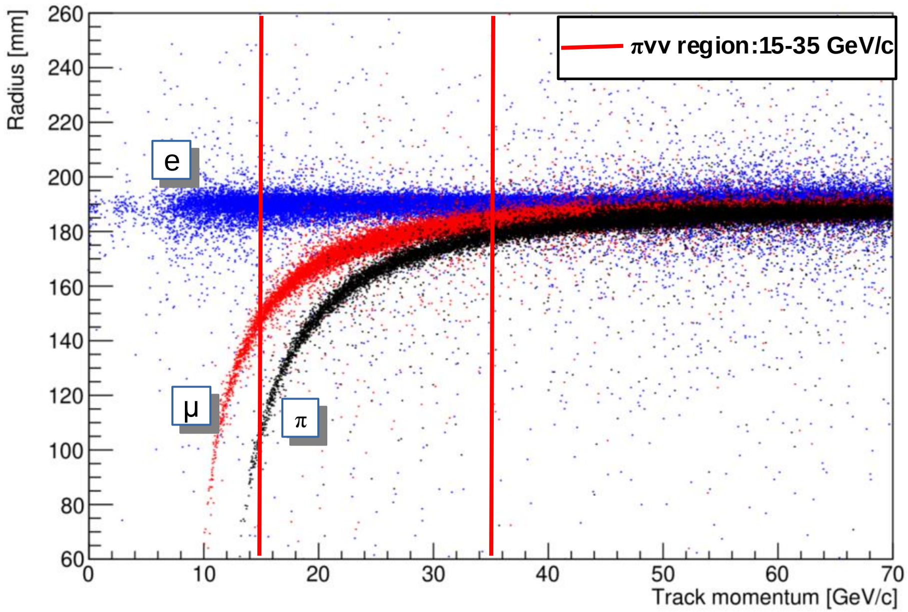
- Some detectors send raw data ***trigger-primitives*** to the **FPGA-based level-0 trigger processor L0TP** over 1GbE UDP links.
- **Read out boards (TEL62)** generate trigger-primitives and buffer detector events while waiting for L0 trigger (max latency 1ms).
- L0TP **checks configurable conditions** (**Masks**) against the physics information inside the primitives (Energy, hit multiplicity, position, ...) to deliver trigger. Data bursts ~ 5s long

The Ring Imaging Cherenkov detector (RICH)

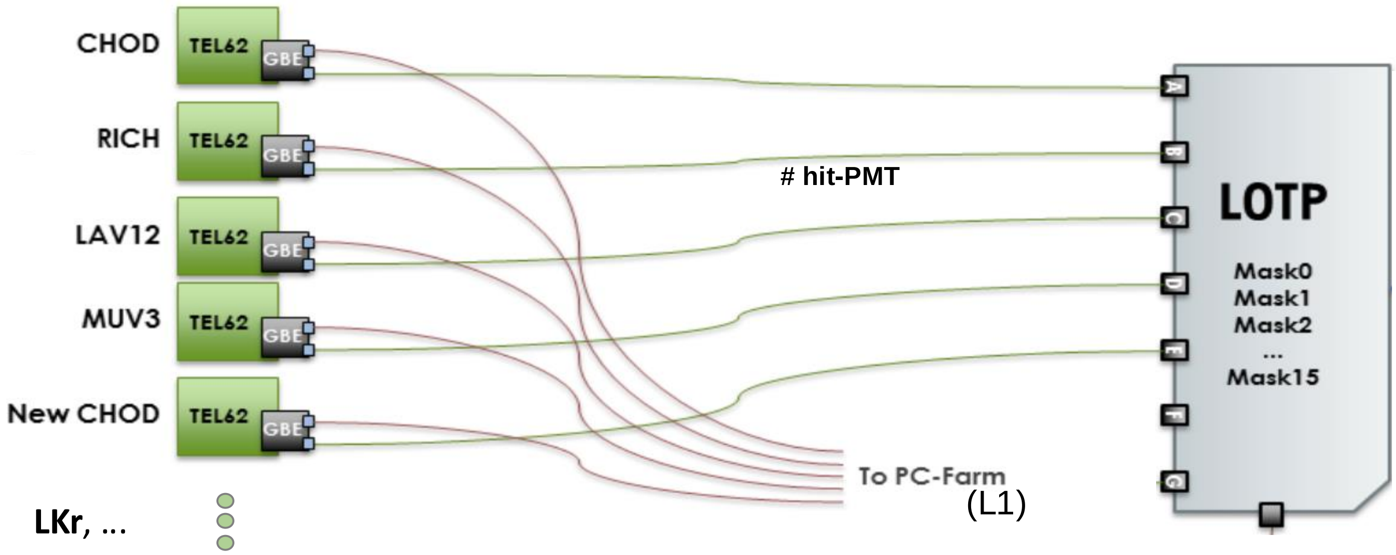


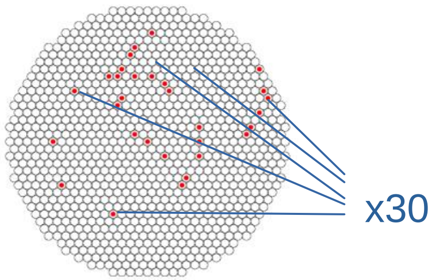


- About 2000 PMT tubes
- During offline data analysis, it provides PID to distinguish between pions and muons from 15 to 35 GeV
- **Current L0 primitives contain only number of HIT PMTs**



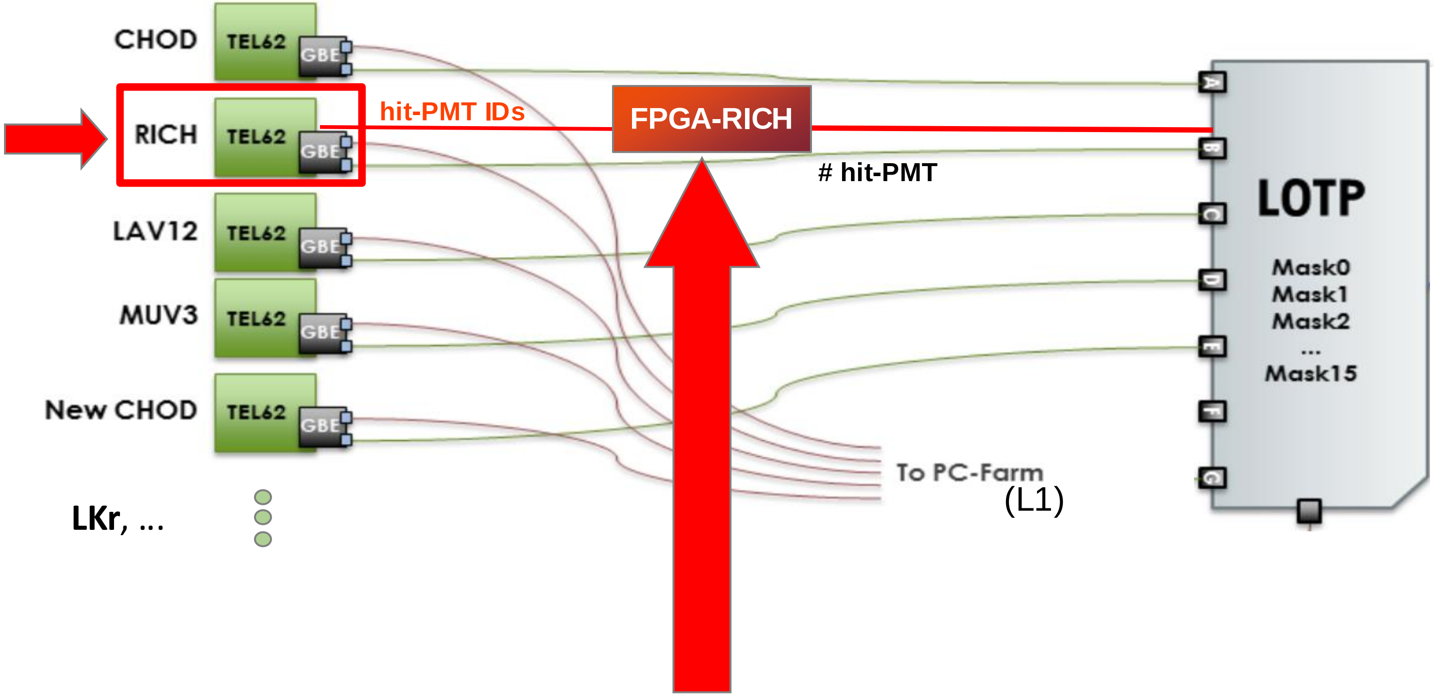
Smart primitives: FPGA-RICH

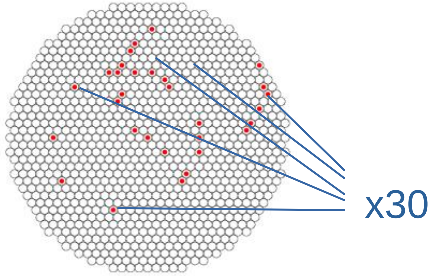




RICH primitives: Number of hit-PMTs

Smart primitives: FPGA-RICH

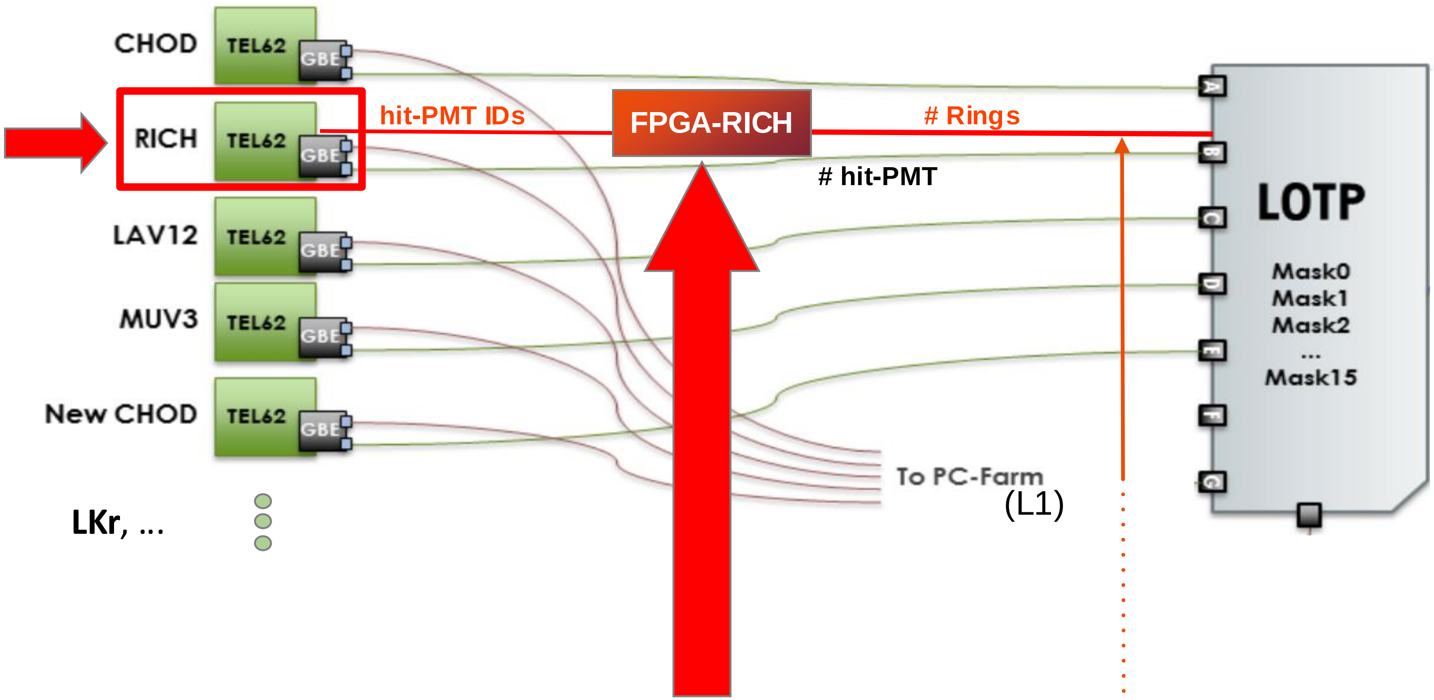


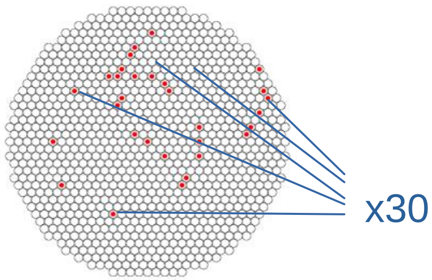


RICH primitives: Number of hit-PMTs

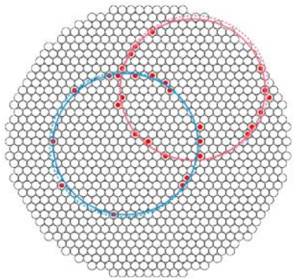
FPGA-RICH: **(partially) reconstruct the rings geometry online** using an AI algorithm on FPGA, to generate a refined primitive stream for L0TP.

Smart primitives: FPGA-RICH





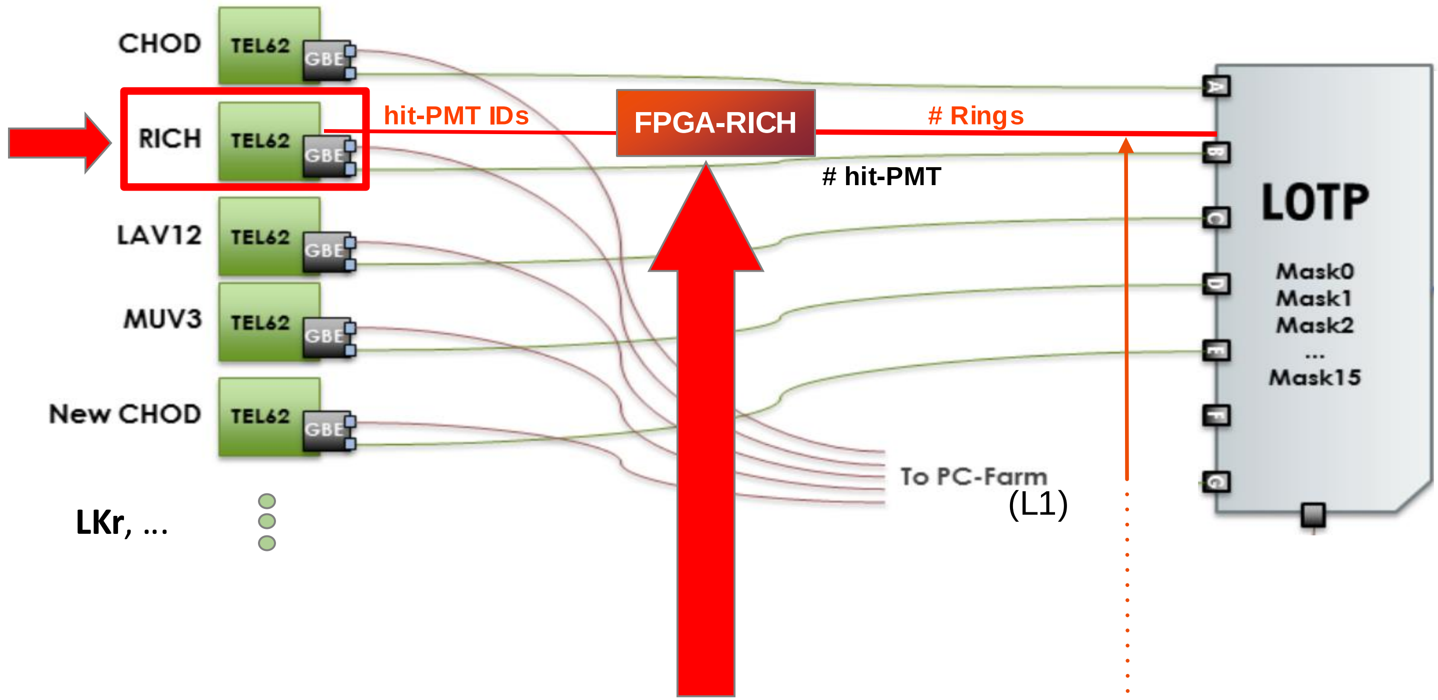
RICH primitives: Number of hit-PMTs

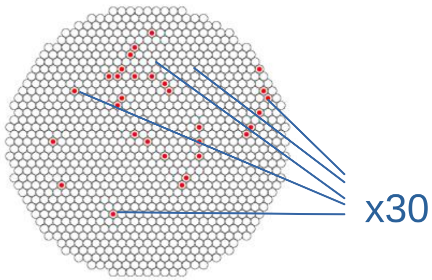


FPGA-RICH primitives:
Number of rings (0, 1, 2, 3+)
(more in the future, e.g. # of e^-)

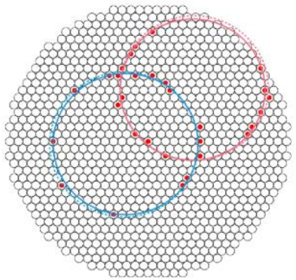
FPGA-RICH: **(partially) reconstruct the rings geometry online** using an AI algorithm on FPGA, to generate a refined primitive stream for L0TP.

Smart primitives: FPGA-RICH





RICH primitives: Number of hit-PMTs



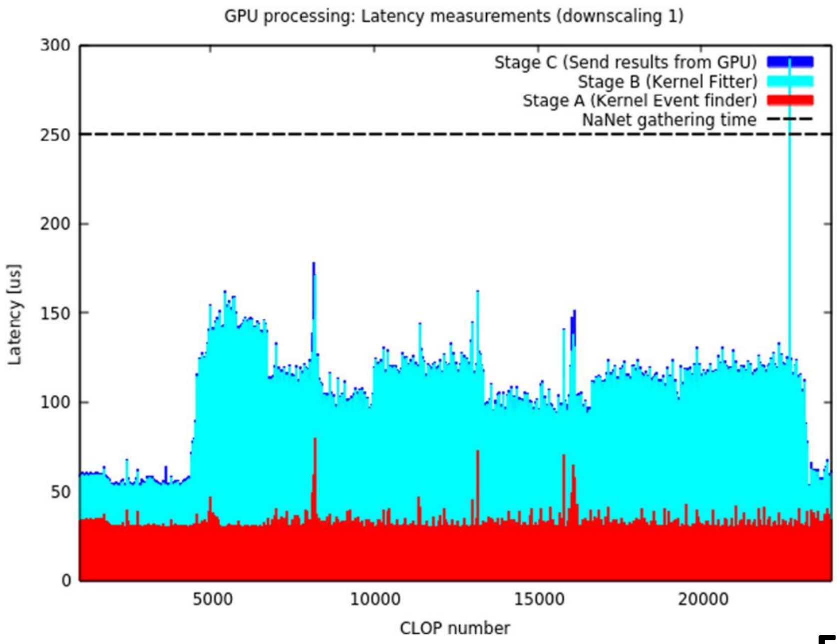
FPGA-RICH primitives:
Number of rings (0, 1, 2, 3+)
(more in the future, e.g. # of e^-)

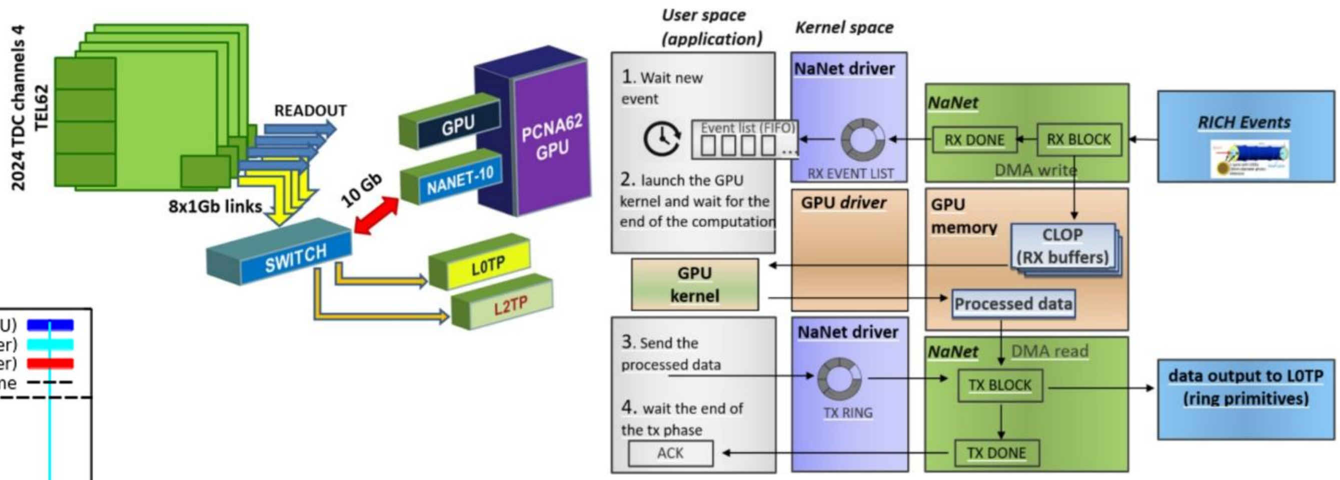
FPGA-RICH: **(partially) reconstruct the rings geometry online** using an AI algorithm on FPGA, to generate a refined primitive stream for L0TP.

The main challenge is the processing throughput (10 MHz).

- Past approach using GPU and a custom DMA capable NIC (NaNet)

http://dx.doi.org/10.1088/1742-6596/1085/3/032022



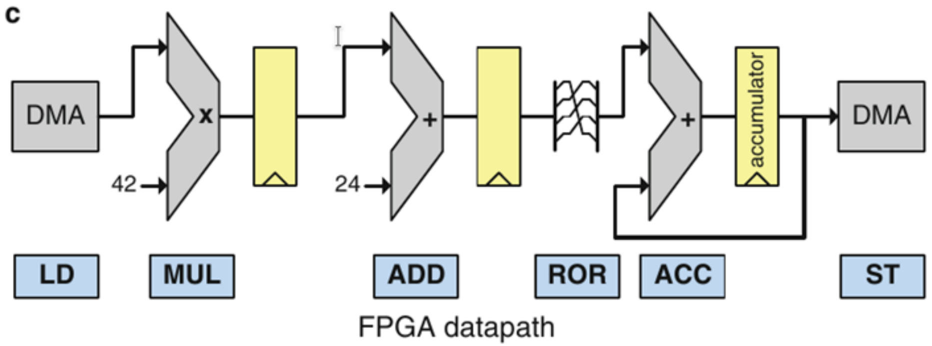


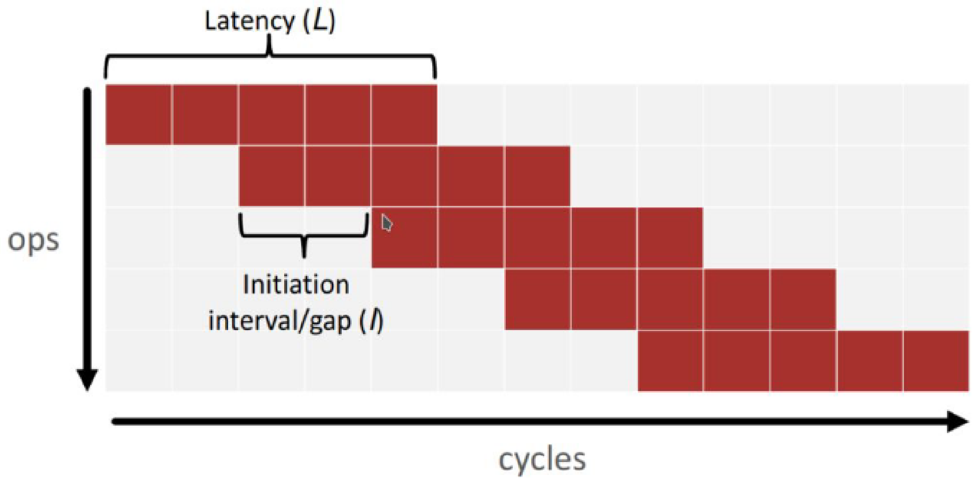
- To sustain high throughput, GPU's parallel architecture has to be exploited on multiple data → need to 'halt' event data stream through a buffering phase, accumulate, then transfer to GPU memory
- **High latency ~100 µs relatively to other primitive generating sub-detectors (~ 1µs)** → complicates L0TP checks and buffering for time alignment

FPGA provides low-latency, full streaming solution working as any other sub-detector

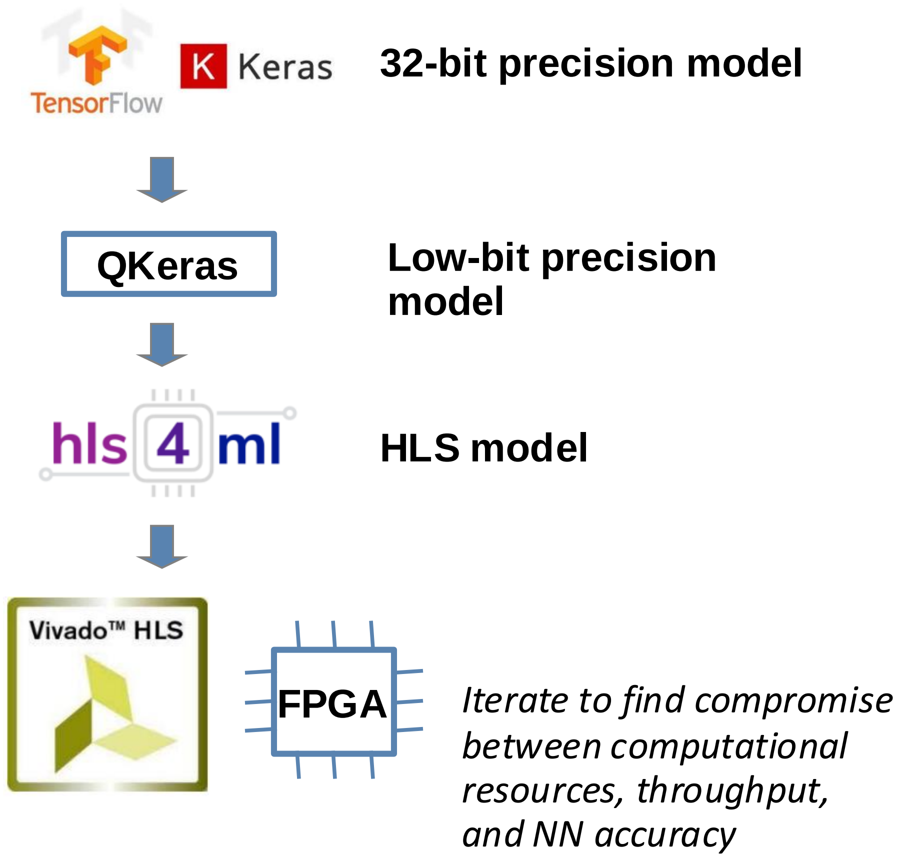
Why FPGA are good for real-time inference?

- Customizable I/O and deterministic latency make them well suited for TDAQ systems.
- Improvements to silicon manufacturing process made them very interesting for heavy computation as well.
- In our case, the challenge is **the processing throughput** → a pipelined design can potentially produce a new output at each clock cycle.
- ***Initiation interval (II)****:* Number of clock cycles before the function can accept new input data. **The lower the II, the higher the throughput**
- The greater the number of pipeline stages, the greater the latency.
- High level synthesis tools allows to describe datapaths in FPGA using high level software languages (C/C++, OpenCL, SYCL,...).





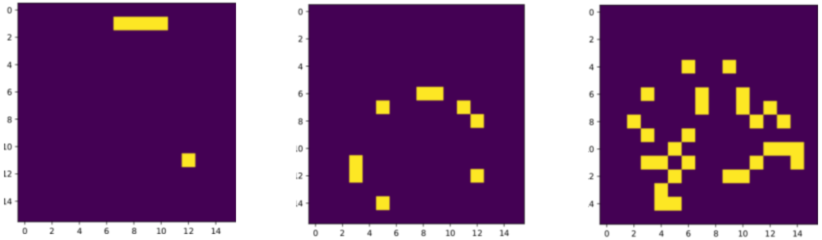
NN Implementation Workflow and Dataset



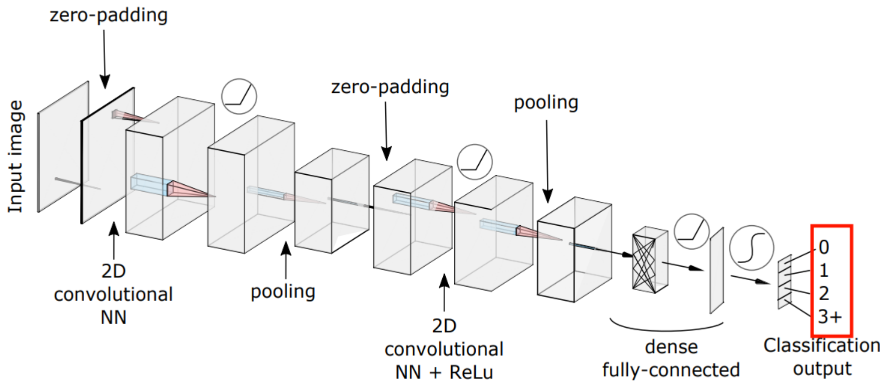
- **DATASET:**
  - **Training (3M events) and Test dataset (2M)** obtained from real data from CERN EOS, using the NA62 analysis framework and a custom analyser. Dumping rings number, radius, number of e^- with checks on radius or EoP, from different offline analysis algorithms
  - **Ground-truth**: Number of rings from offline trackless reconstruction algorithm that uses only PMT hits
  - **Train to be as good online as the best offline algorithm**

NN Architectures: Convolutional Model

- **Input representation: 16x16 images**



- **Output: 4 classes (0, 1, 2, 3+ rings)**
- **Quantization (fixed point):**
  – Weights and biases: 8 bits <8, 1>
  – Activations:16 bits <16, 6>
- **FPGA resource usage (Alveo U200)**
  - LUT 5.2%, FF 1.5%, DSP 4.8%,
  - BRAM 0.05%
- **Latency: 388 cycles @ 220MHz**
- **Initiation Interval (II): 369 cycles**
- **Throughput: 0.6 MHz**



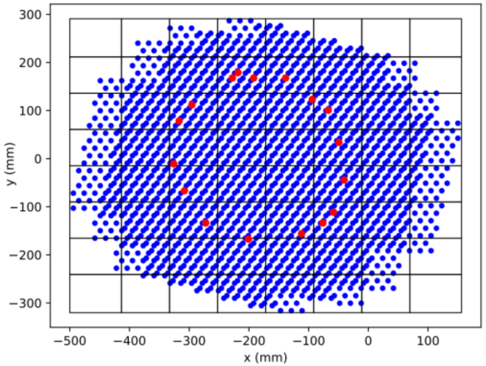
Very small NN: 2 x Conv (8 size filter, K=3x3) + Mpool (Stride=2)
2 x Dense(128→16, 16->4)

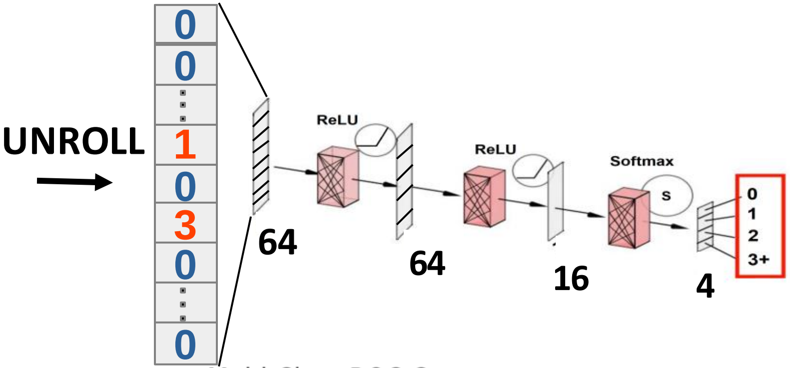
Class	0 (0 rings)	Efficiency 88.4	Purity 95.4
Class	1 (1 rings)	Efficiency 88.5	Purity 87.3
Class	2 (2 rings)	Efficiency 78.3	Purity 70.3
Class	3 (3+ rings)	Efficiency 74.3	Purity 85.1

Efficiency = TP / (TP + FN)
Purity = TP / (TP + FP)

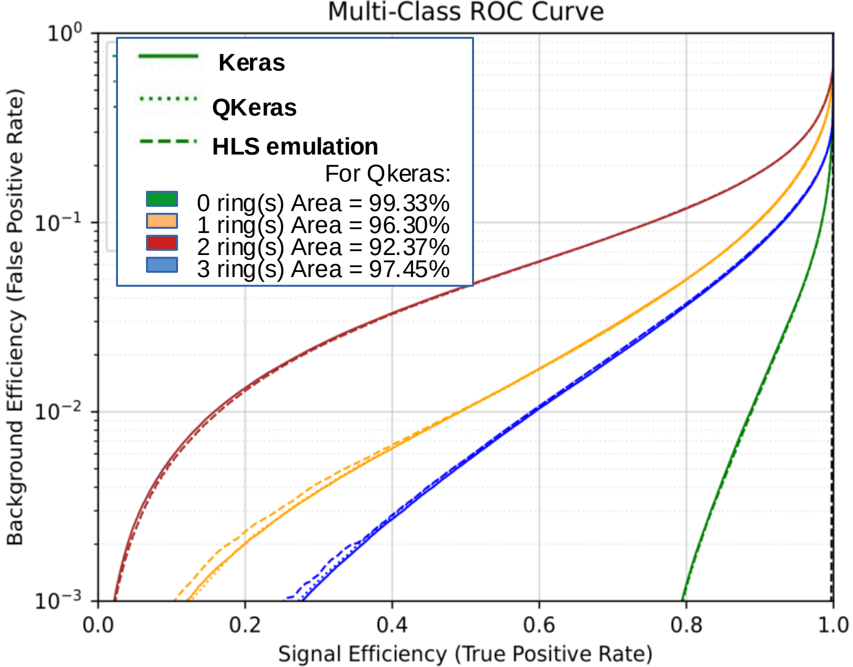
NN Architectures: Dense Model

- **Quantization <fixed point>:**
  - Weights and biases: 8 bits <8, 1>
  - Input: 6 bits (unsigned int),
  - Activations:16 bits <16, 5>
- **FPGA resource usage (Versal VCK190)** - LUT 7.2%, FF 2.2%, DSP 7.4%, BRAM 0.0%
- **Latency: 28 cycles @ 300MHz**
- **Initiation Interval (II): 9 cycles**
- **Throughput: 33 MHz**



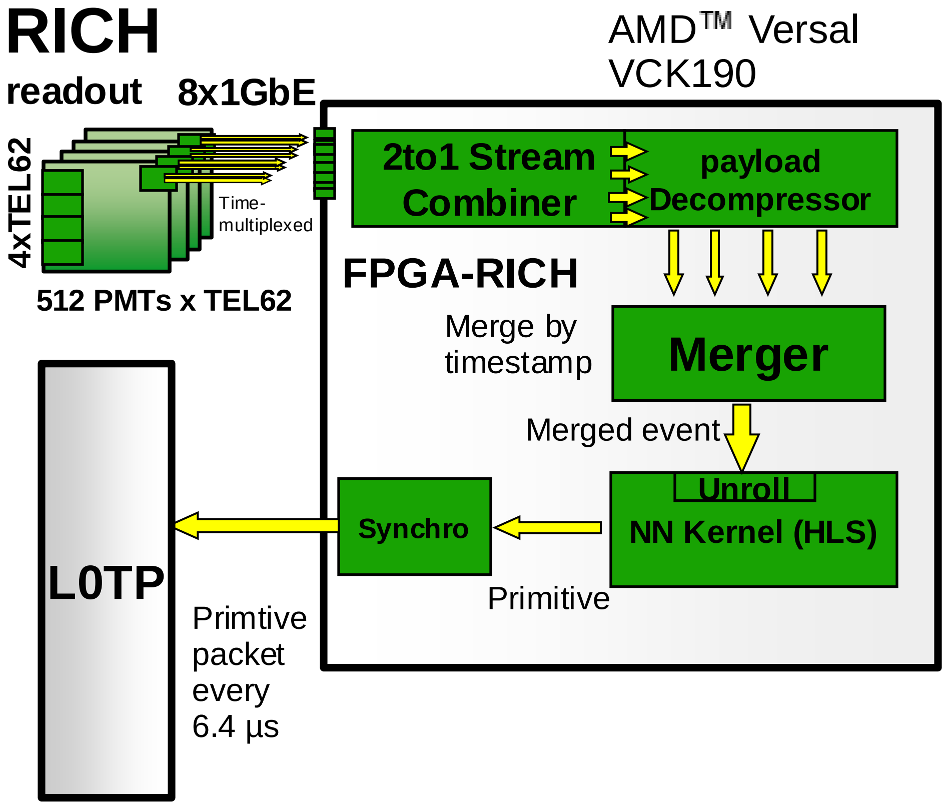


Class 0 (0 rings)	**Efficiency 88.9**	**Purity 95.0**
Class 1 (1 rings)	**Efficiency 88.9**	**Purity 86.5**
Class 2 (2 rings)	**Efficiency 76.3**	**Purity 72.2**
Class 3 (3+ rings)	**Efficiency 77.1**	**Purity 84.6**



Final throughput including input construction from RICH data stream depends on event hits-number:
≈ 23 MHz for avg event, latency 160 ns, at 300 MHz

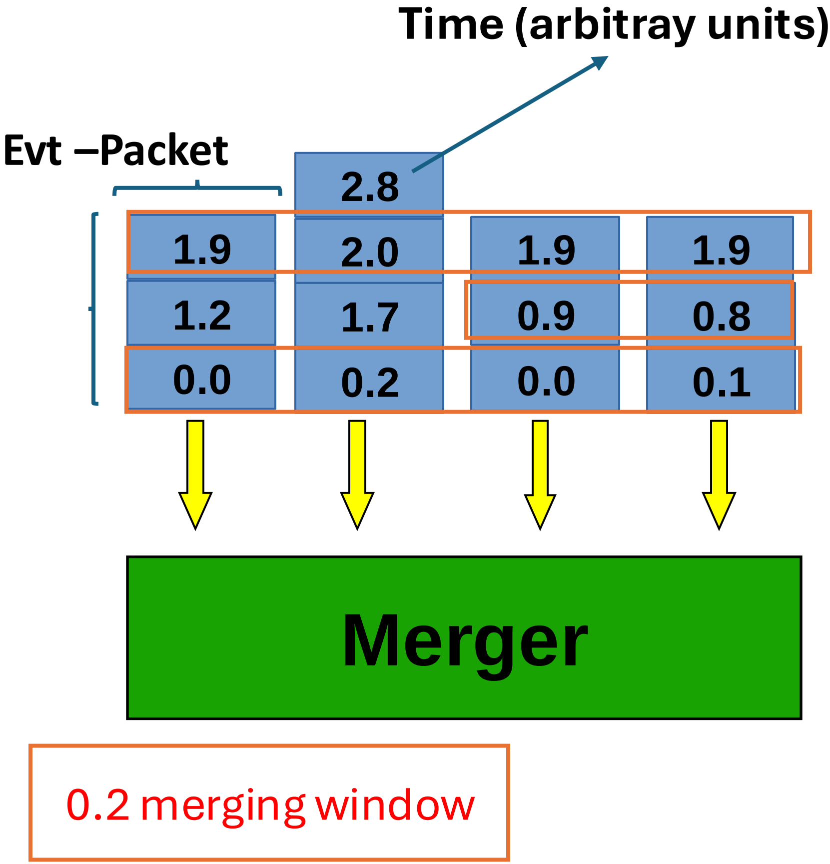
Integration of the FPGA-RICH Pipeline



- **Retro-fit of the RICH readout**
- **Custom firmware on the TEL62** boards for FPGA-RICH dataflow. Send compressed UDP packets of PMT-hits event data through 8x1GbE links (2xTEL62, time-multiplexed) every 12.8 µs
- Each TEL62 handles 512 PMT channels of ≈2000 total, so each stream's events are fragments of a full physics event
- **Merging stage**
  merge four boards event-fragments by timestamp into full RICH physics events for NN Kernel
- **Synchronization stage**
  accumulate primitives in packets and send them every 6.4us as required by L0TP.

Merging Stage

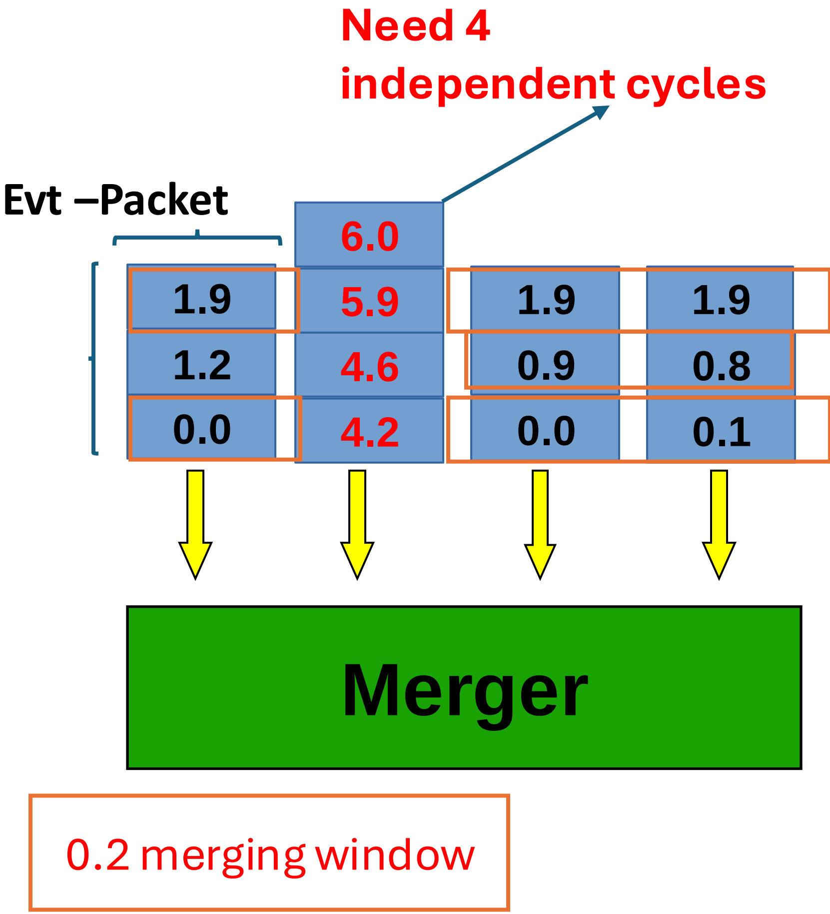
- **Merge by timestamp**:
  take smallest timestamp among incoming streams as *base* and merge events from other streams in a fixed time window
- **Limited clock cycle budget for a set of packets**:
  Event-packets every ≈ 12.8 us. Have to consume a set of packets every ≈ 3800 clock cycles on average, at 300 MHz
- **Sensible to time misalignment:**
  Any corrupted and time misaligned stream has to be ”merged” only with itself in a slow non-parallel operation, wasting clock cycles



Merging Stage

- **Merge by timestamp**:
  take smallest timestamp among incoming streams as *base* and merge events from other streams in a fixed time window
- **Limited clock cycle budget for a set of packets**:
  Event-packets every ≈ 12.8 us. Have to consume a set of packets every ≈ 3800 clock cycles on average, at 300 MHz
- **Sensible to time misalignment:**
  Any corrupted and time misaligned stream has to be ”merged” only with itself in a slow non-parallel operation, wasting clock cycles

Can constrict throughput

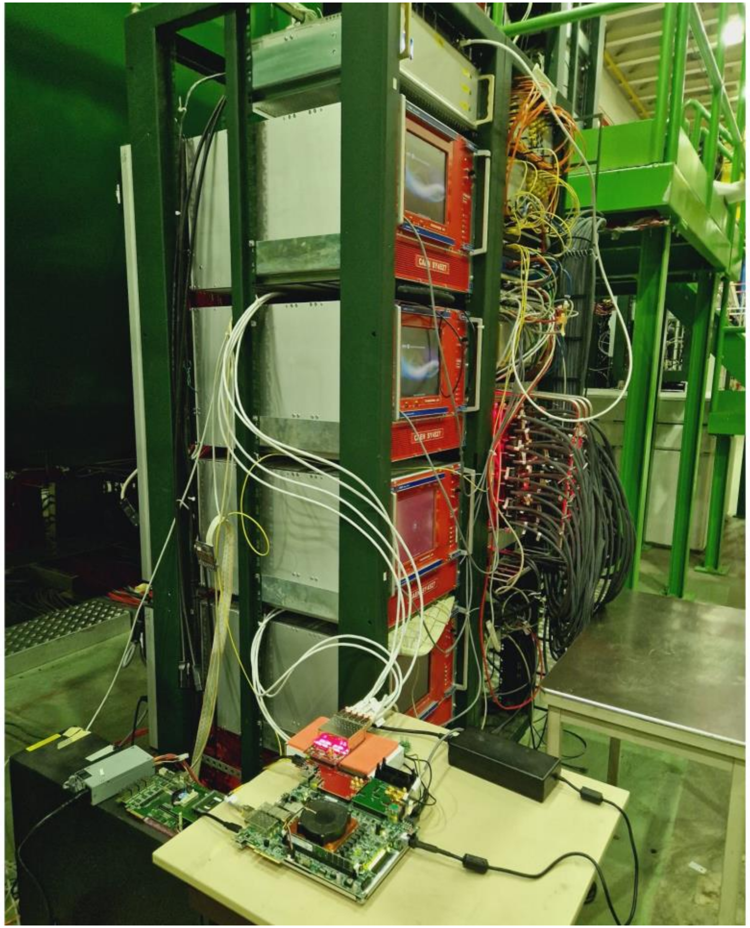


Pipeline Validation

- Tested in the lab with fpga-based emulator of TEL62 streams, using artificially generated events and dumps of real TEL62 data
  - **With artificial events measured:**
    Latency ~ 1 μs Throughput > 9.38 MHz
    at 150 MHz clock

- **Completed synthesis at 300 MHz**

- **Deployed at the experiment** and tests ongoing with beam in parasitic mode (independent from standard experiment dataflow)



Current status and outlook

- Issues affecting custom TEL62 firmware for FPGA-RICH dataflow currently compromise merger time alignment. Pipeline works for part of the burst, then stalls when event rate rises

- Difficult to change firmware during run, but we are working around issues at our end by flushing corrupted data

See Ottorino Frezza's talk from past Monday

FUTURE WORK

- **Integrate FPGA-RICH with L0TP+ FPGA**
(FPGA-RICH Utilization: LUT = 14%, BRAM = 3%, DSP = 7% FF=6% (VCK190))

- **Expand PID capabilities**: e.g. predict number of electrons, combining data stream from calorimeter

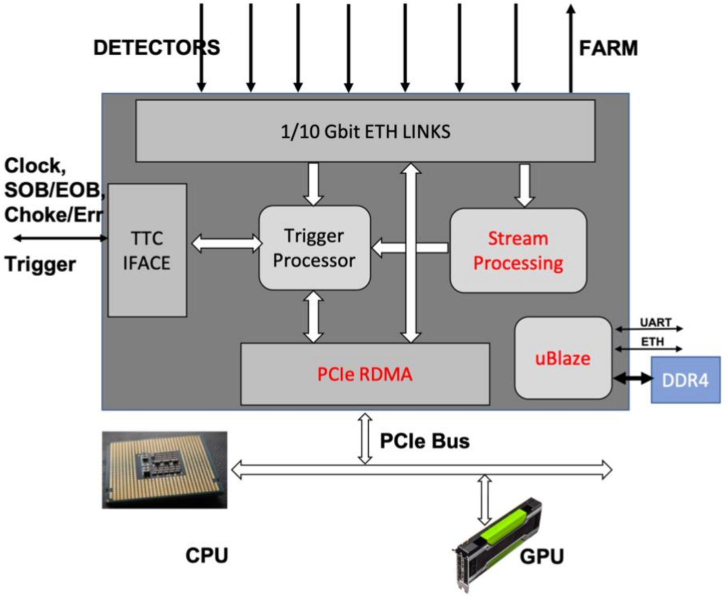
Thank you!

BACKUP SLIDES

New functionalities

L0TP+ reproduces all L0TP functions but considering the huge amount of FGPA resources (only 30 % BRAM, 17 % LUT used in L0TP+) there is room to add several capabilities to the original design.

- **DATA LINKS**:
  the system is able to support ten 25GbE links through the FMC+ daughercard, and additional QSFP28, and FireFly ports can be used to connect additional data links from the detectors via 100 Gbps low latency links.
- **MICROCONTROLLER:**
  a 32-bit MicroBlaze Soft-Core Micro Controller was integrated for debug and configuration purposes. Applications can be deployed onto it either bare metal or by Xilinx Petalinux.
- **PCIe HOST INTERFACE**
- **STREAM PROCESSING MODULE:**
  with the outlook of processing primitive streams and thus improving the efficiency of the trigger (e.g. online PID in RICH via HLS4ML Neural Networks)



Conclusions

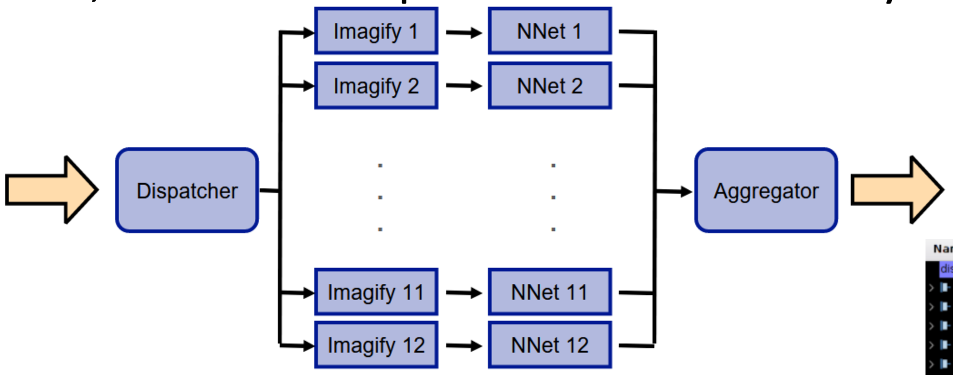
- New study of $K^+ \to \pi^+ \nu\bar{\nu}$ decay using NA62 2021–22 dataset:
  - Improved signal yield per SPS spill by 50%.
  - $N_{bg} = 11.0^{+2.1}_{-1.9}$, $N_{obs} = 31$
  - $\mathcal{B}_{21-22}(K^+ \to \pi^+ \nu\bar{\nu}) = \left(16.0^{+5.0}_{-4.5}\right) \times 10^{-11} = \left(16.0\ \left({}^{+4.8}_{-4.2}\right)_{stat}\ \left({}^{+1.4}_{-1.3}\right)_{syst}\right) \times 10^{-11}$

- Combining with 2016–18 data for full 2016–22 results:
  - $N_{bg} = 18^{+3}_{-2}$, $N_{obs} = 51$ (using 9+6 categories for BR extraction)
  - $\mathcal{B}_{16-22}(K^+ \to \pi^+ \nu\bar{\nu}) = (13.0^{+3.3}_{-2.9}) \times 10^{-11} = (13.0\ \left({}^{+3.0}_{-2.7}\right)_{stat}\ \left({}^{+1.3}_{-1.2}\right)_{syst}) \times 10^{-11}$
  - Background-only hypothesis rejected with significance Z>5.
- **First observation of $K^+ \to \pi^+ \nu\bar{\nu}$ decay: BR consistent with SM prediction within 1.7σ**
  - **Need full NA62 data-set to clarify SM agreement or tension.**

2023–LS3 data-set collection & analysis in progress…

Convolutional model issue → Kernel replication

Throughput is not enough to sustain L0 rate, but we can replicate the network multiple times, also on multiple devices if necessary.



Resources usage for 12 replicas:

- LUT 74%
- FF 17%
- DSP 61%
- BRAM 1.4%

Processing time @220MHz: 137 ns per event

Processing throughput: 7.2 MHz

APEIRON applications:

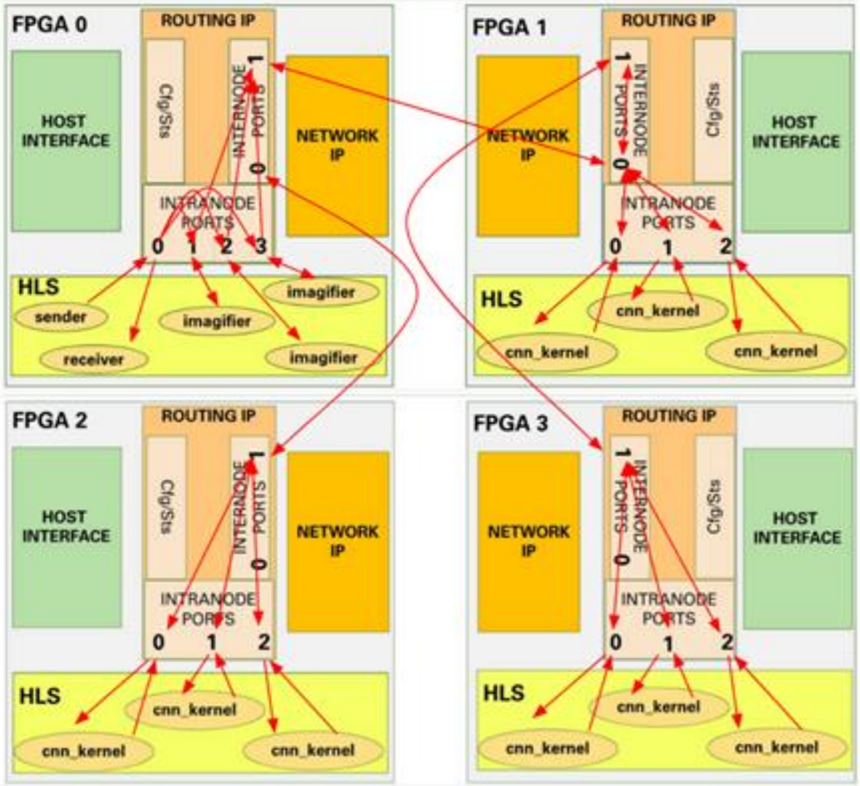
- RAIDER (TEXTAROSSA)

KPI	CNN CPU tensorflow	CNN CPU+GPU tensorflow
purity/efficiency (per class)	efficiency: - 0: 93% - 1: 83% - 2: 75% - 3+: 83% purity: - 0: 88% - 1: 90% - 2: 71% - 3+: 78%	efficiency: - 0: 93% - 1: 83% - 2: 75% - 3+: 83% purity: - 0: 88% - 1: 90% - 2: 71% - 3+: 78%
time to solution [s]	158.521	125.963
throughput [events/s]	**189250**	**238165**
energy to solution [J]	11091.919	17497.783 (8724.648 GPU)
energy efficiency [events/J]	**270.467**	**154.305**

KPI	RAIDER @200 MHZ [4 FPGA, 9CNNs]
time to solution [s]	0.554
throughput [events/s]	**4873646.209**
energy to solution [J]	165.277 (101.055 FPGA)
energy efficiency [events/J]	**16336.183 (26718.126 FPGA)**

x20

x100



Generation strategy of training and validation data sets

- Dataset for training and validation obtained using the NA62 analysis framework
- Analyser called RingDumperAPE
- Single run or in batch (run list) from CTRL trigger sample
- Output: Histograms + Events dumped on plain text files

RICH Hit list (TDCEvent)
RICH trackless reconstruction (TRecoRICHEvent)
Downstreamtrack reconstruction (Downstreamtrack
)
L0TP (TNA62L0Data)
Event Labels

- Different labels are dumped to be used as ground truth
  1. Number of rings from RichReco
  2. Number of rings from Downstreamtrack
  3. Number of electrons from RichReco (based on ring radius only)
  4. Number of electrons from Downstreamtrack (based on MostLikelyHypothesis)
  5. Number of electrons as 4 + check on the radius + check on Energy over momentum ratio (EOP)

Electron radius = [185,195] mm
Eop = [0.90,1.10]

Mostlikely hypothesys = multiple are rejected

- Event rejection criteria can be optionally activated
  - Formal check on the reconstructed tracks and rings (e.g. chi2)
  - Event characteristics e.g. NHit, Momentum, etc

Preliminary Number of electrons classification

- Batch processing on 2017-2018 data
- Label used is number 5 on slide 24
- Momentum < 35 GeV/c
- Additional requirement is: number of rings from RICH Reco == number of tracks from Downstreamtrack

```
Total Events   163905
Total events of class 0 is    84628   (51.63 %)
Total events of class 1 is    76822   (46.87 %)
Total events of class 2 is     2432   (1.48 %)
Total events of class 3 is       23   (0.01 %)
Total events classified as 0 is    75533   (46.08 %)
Total events classified as 1 is    75209   (45.89 %)
Total events classified as 2 is    11920   (7.27 %)
Total events classified as 3 is     1243   (0.76 %)
Class  0  Efficiency 82.6  Purity 92.5  OverContamination  7.5  UnderContamination   0.0
Class  1  Efficiency 80.6  Purity 82.3  OverContamination  0.2  UnderContamination  17.5
Class  2  Efficiency 74.6  Purity 15.2  OverContamination  0.0  UnderContamination  84.8
Class  3  Efficiency 91.3  Purity  1.7  OverContamination  0.0  UnderContamination  98.3
```

L0TP Masks

Mask label	Definition	Downscaling
Not μ	RICH *Q1*!MUV3	200
$\pi\nu\bar{\nu}$	RICH*Q1*!QXUTMC*!MUV3*!LKr(E > 31 |> 1cl)	1
μ – exotics	RICH*2*MO2*!LKr > 10GeV	3
$\pi\mu$	RICH*QX*MO1*LKr > 10GeV	5
Dielectron	RICH*QX*LKr > 20GeV	8
Multi-Tracks	RICH*QX	100
$\mu\mu$	RICH*QX*MO2	2
μ exotic (!KTAG at L1)	RICH*Q2*MO1*LKr > 10GeV	5
ν_μ	RICH*Q1*!Q2*MOQX	15

Table 2.1. Trigger masks from 2018 run

where ! stands for negation and the specific meaning of each condition is:
RICH: at least 2 in-time hits in RICH
Q1: at least 2 in-time CHOD quadrants hits
Q2: at least 2 in-time opposite CHOD quadrants hits
QX: at least 2 in-time opposite quadrants hit CHOD
UTMC: ("upper tight multiplicity cut") less than 5 hits in CHOD
MO1: 1 outer muon in MUV3 (at least 1 single -or double- PM outer tiles)
MO2: 2 outer muons in MUV3 (coincidence of 2 single -or double- PM outer tiles)
MOQX: cross di-muons in MUV3 (coincidence of outer tiles in opposite quadrants)
MUV3: any MUV3 primitive

SOURCE ID		COUNTER | FORMAT		TOTAL NUMBER OF HITS			
SOURCE SUB-ID		NUM OF EVENTS		TOTAL MGP LENGTH			
Event data							
Event data							
Event data							
...							
32	24	23	16	15	8	7	0

EVENT TIMESTAMP							
Reserved		EVENT FINE TIME		EVENT NUMBER OF HITS			
Padding 5 bits	HIT #0 PM ID (9bits)		HIT #1 PM ID (9 bits)		HIT #2 PM ID (9 bits)		
Padding 5 bits	HIT #3 PM ID (9bits)		HIT #4 PM ID (9 bits)		HIT #5 PM ID (9 bits)		
Padding 5 bits	HIT #6 PM ID (9bits)		HIT #7 PM ID (9 bits)		HIT #8 PM ID (9 bits)		
Padding 5 bits	...		...		...		
32	24	23	16	15	8	7	0

MTP header	Source ID		MTP assembly timestamp high					
	Source sub-ID		Number of primitives in MTP		Total MTP length			
Timestamp word	0x0000		Primitive timestamp high					
Primitive data	Primitive ID				Timestamp low		Fine time	
Primitive data	Primitive ID				Timestamp low		Fine time	
	...							
Timestamp word	0x0000		Primitive timestamp high					
Primitive data	Primitive ID				Timestamp low		Fine time	
	...							
Bits	31	24	23	16	15	8	7	0

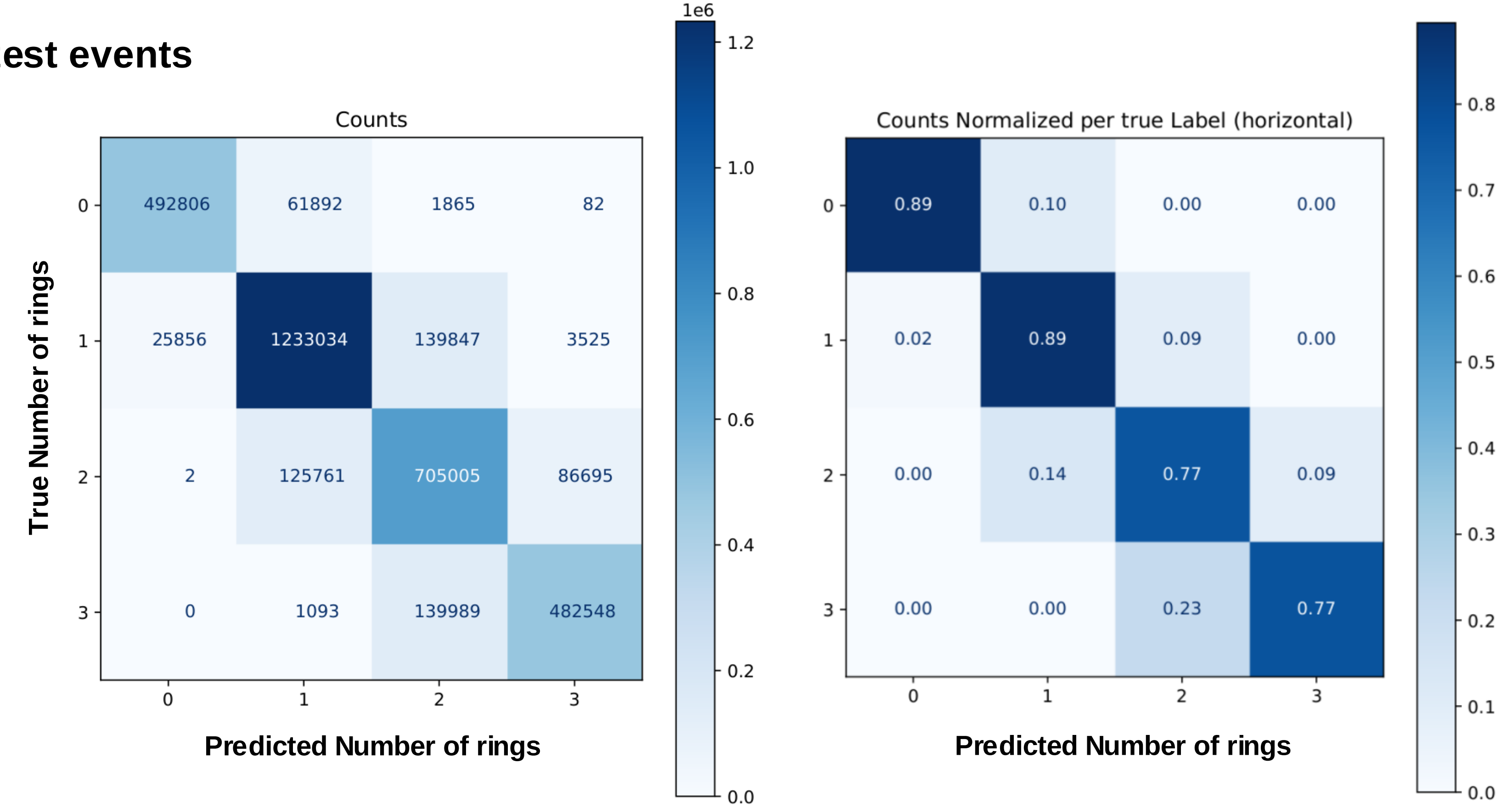
PATTERN

128b word

	STR 3 MGP	STR 2 MGP	STR 1 MGP	STR 0 MGP	STR 3 HITS	STR 2 HITS	STR 1 HITS	STR 0 HITS	RESERVED	WINDOW	TOT HITS	TIMESTAMP				FT
	STR 1; HIT 1		STR 1; HIT 0		STR 0; HIT 5		STR 0; HIT 4		STR 0; HIT 3		STR 0; HIT 2		STR 0; HIT 1		STR 0; HIT 0	
	STR 2; HIT 0		STR 1; HIT 8		STR 1; HIT 7		STR 1; HIT 6		STR 1; HIT 5		STR 1; HIT 4		STR 1; HIT 3		STR 1; HIT 2	
	STR 3; HIT 3		STR 3; HIT 2		STR 3; HIT 1		STR 3; HIT 0		STR 2; HIT 4		STR 2; HIT 3		STR 2; HIT 2		STR 2; HIT 1	
	PADDED BITS										STR 3; HIT 6		STR 3; HIT 5		STR 3; HIT 4	
bit range	127...120	119...112	111...104	103...96	95...88	87...80	79...72	71...64	63...56	55...48	47...40	39...32	31...24	23...16	15...8	7...0

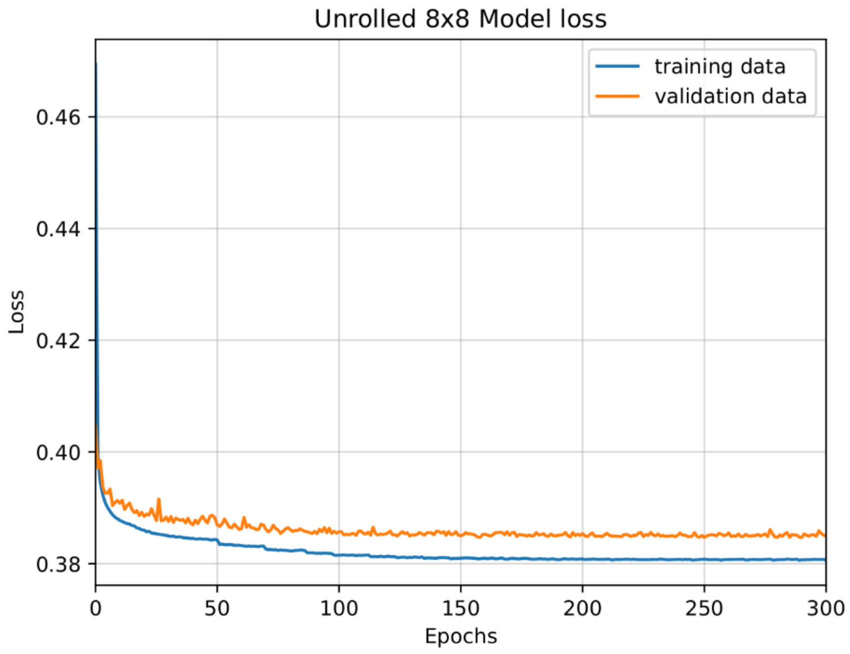
Neural Network Sensitivity (or Efficiency)

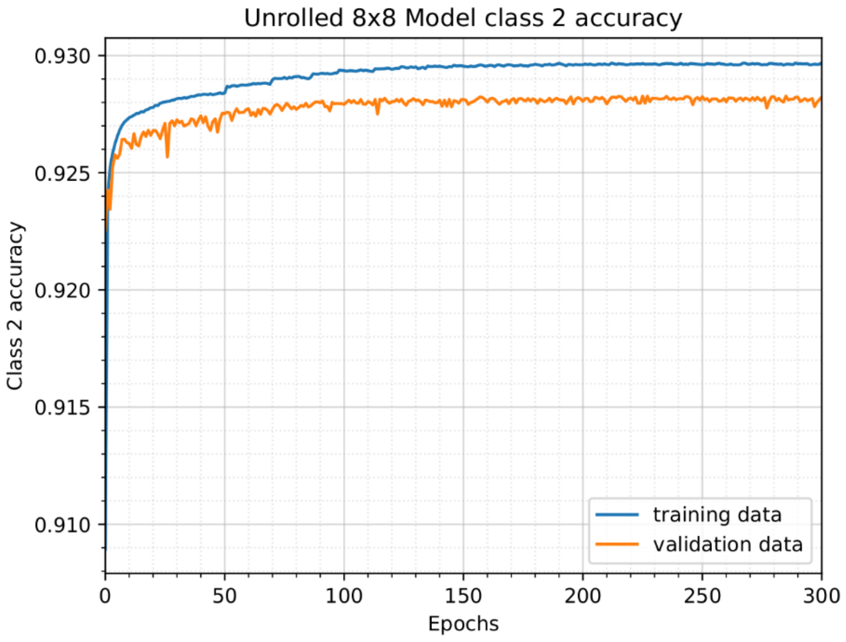
3.5 M test events



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

Training Loss and Accuracy





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