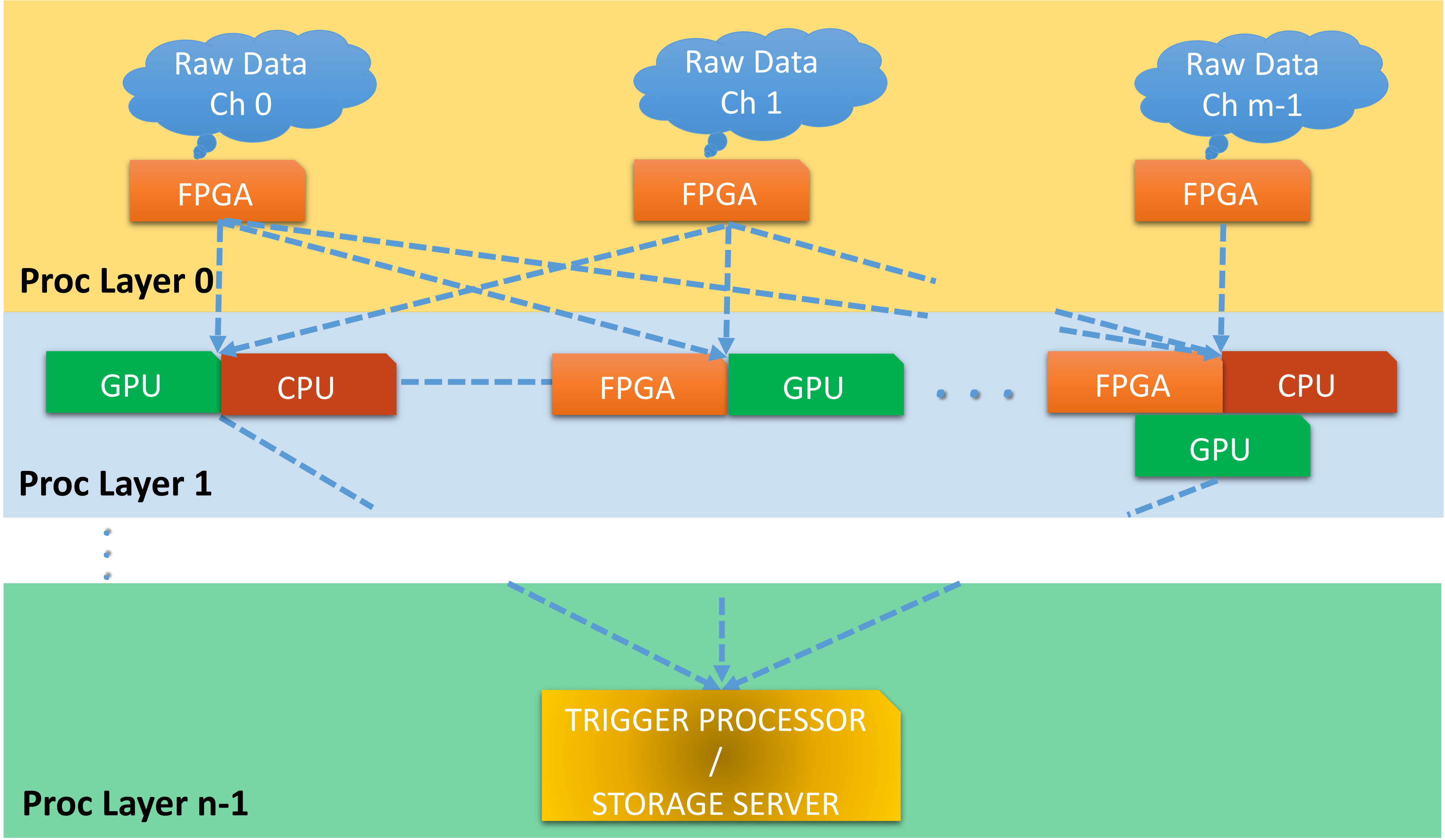
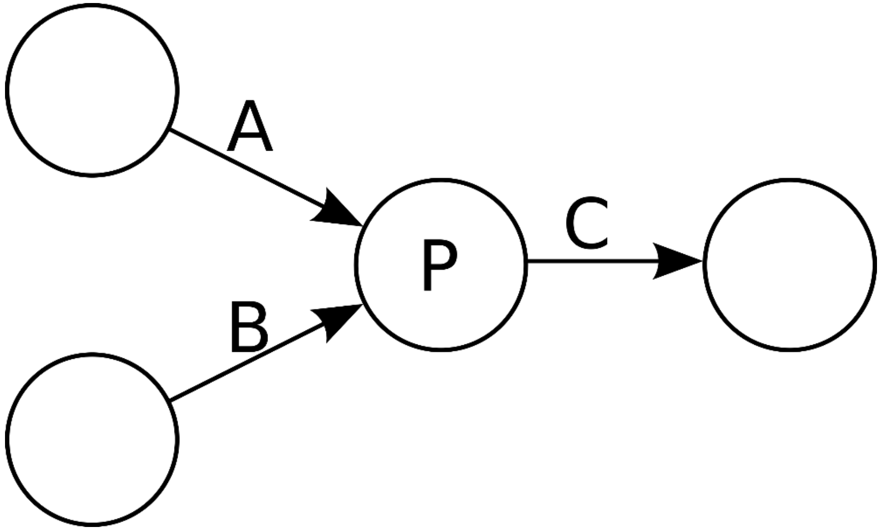
APEIRON Motivation and Concepts

- The study of rare decays Physics needs collecting a large statistics of interesting events with hard-to-find signatures out of an overwhelming background.
- Trigger-less approach involves the handling of high volume of data and high costs.
- Need to investigate new techniques to improve online particle identification and further suppress background events in trigger systems, or to perform an efficient online data reduction for trigger-less ones.
- Distribute processing over the whole chain in subsequent layers, from data readout to low level trigger or storage servers, following a streaming approach.
- Combine data streams from different channels along the processing layers.
- Adopt a modular and scalable network infrastructure.
- Exploit the specialization of modern computing devices (CPU, FPGA, GPU), but…
- keep processing and communication definition the more abstract and device independent as possible to ease development, validation and maintenance.
- Deep (Convolutional) and Spiking Neural Networks as reference approach for trigger.
- Apply all of this to relevant Physics use cases.

General Architecture



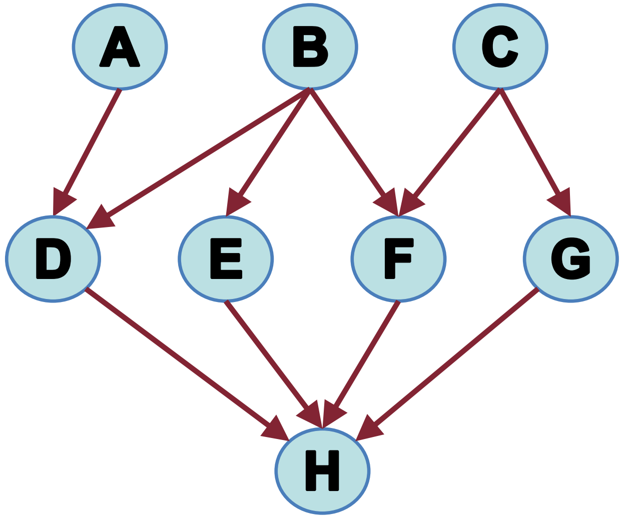
Dataflow Programming Model

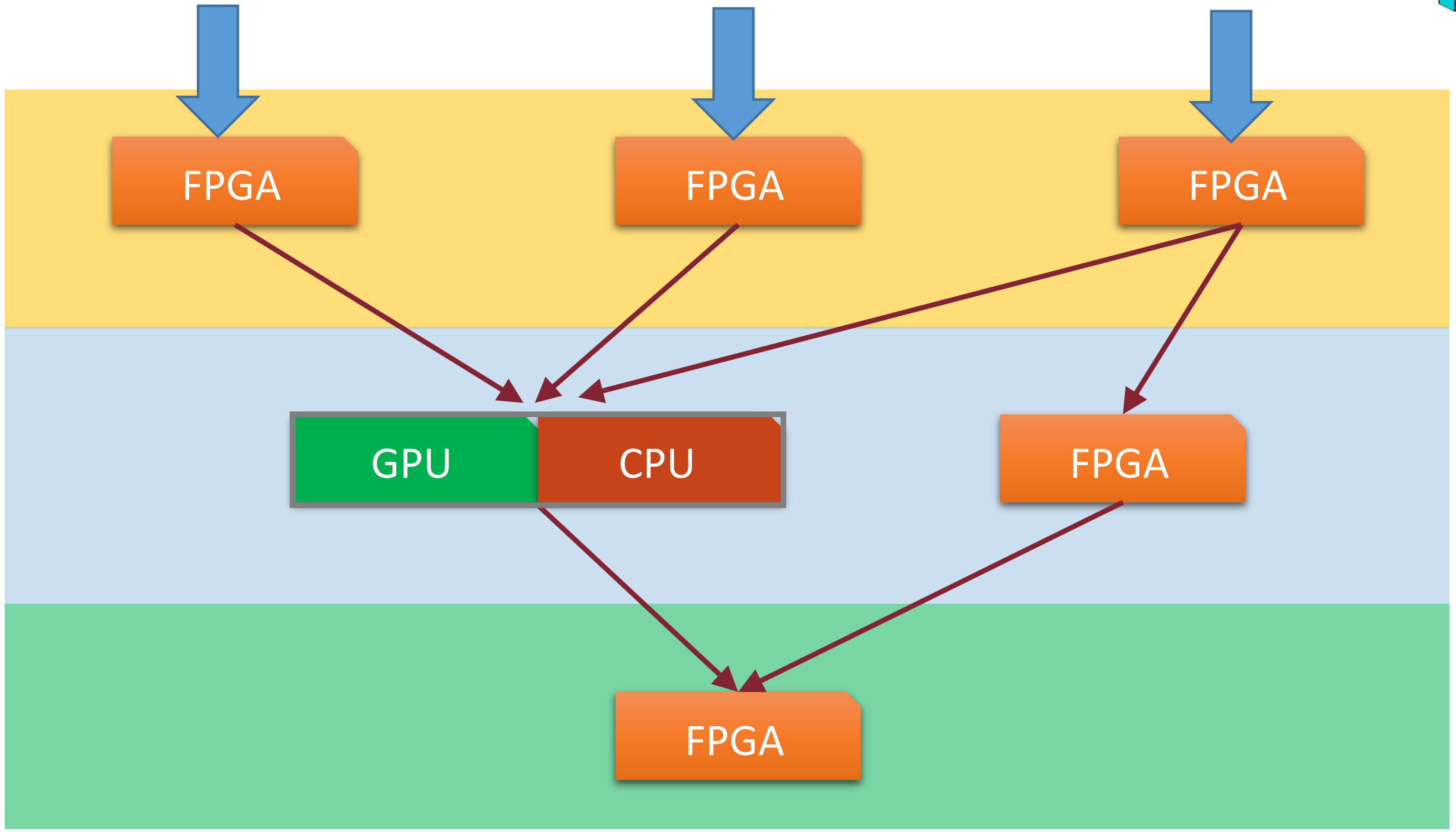


A Kahn process network of three processes without feedback communication. Edges A, B and C are communication channels. One of the processes is named process P. (from Wikipedia)

- Programming Model based on **Kahn Process Networks (KPNs):**
  1. Determinism: for the same input history the network produces exactly the same output
  2. Monotonicity: partial information of the input stream to produce partial information of the output stream
  3. Processes can run concurrently and synchronize through blocking read on input channels
- Task expressed in high level language (C/C++)
- Validation of processing definition can be done on any execution platform

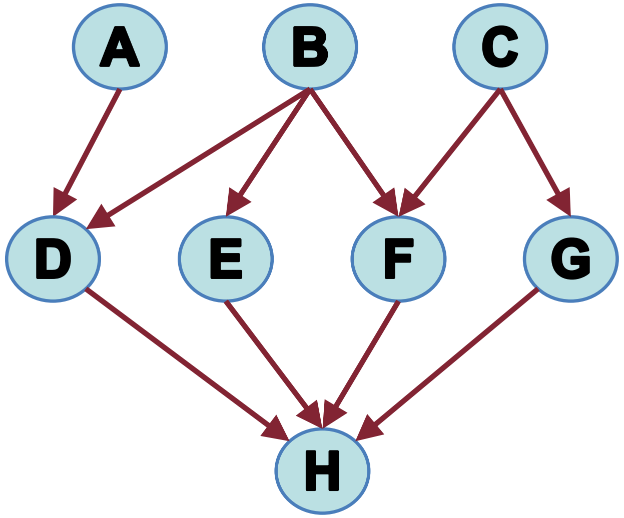
Process Network vs Execution Platform

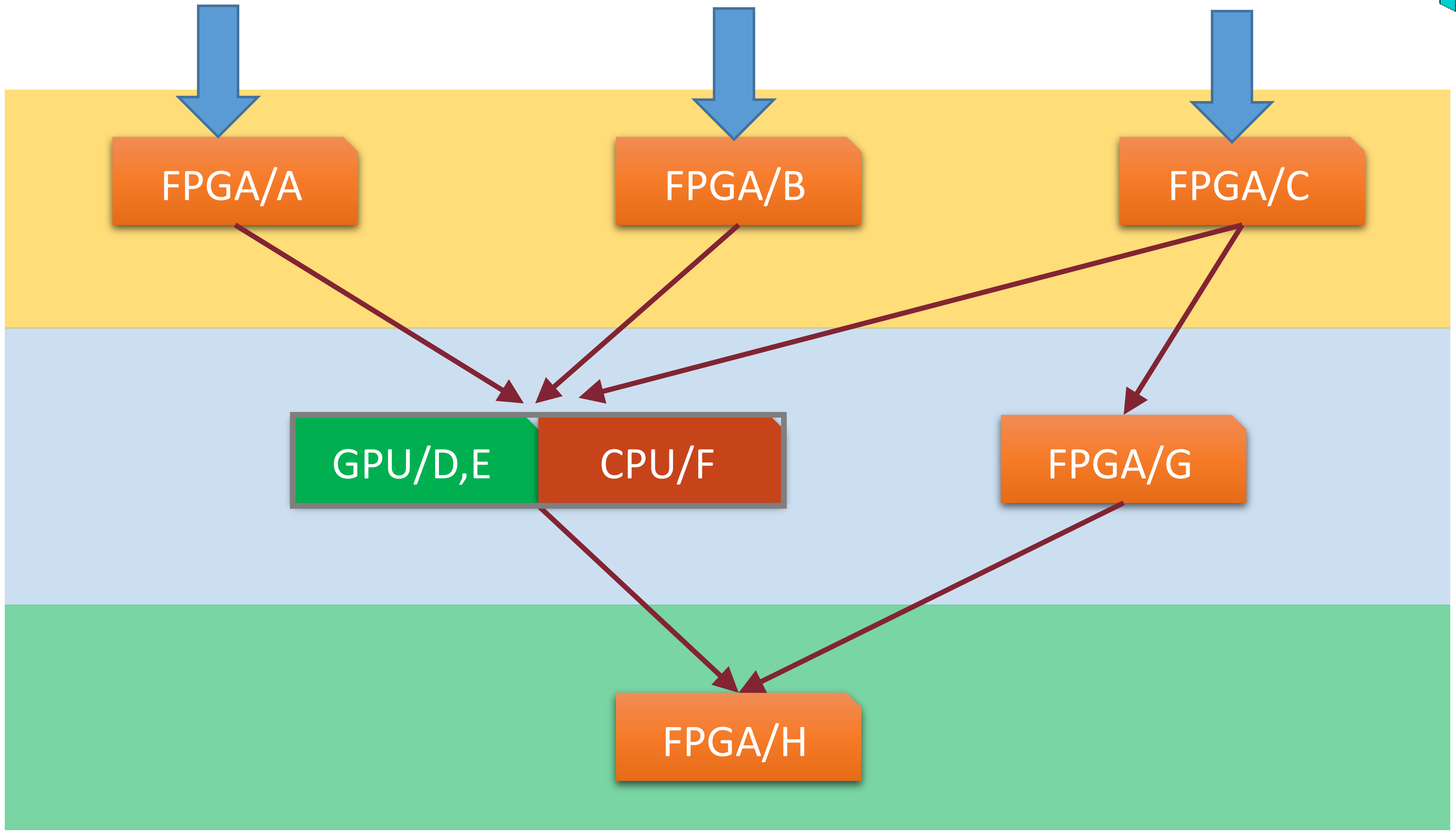




3 Processing layers, 3 data channels

Mapping of Process Network to Execution Platform



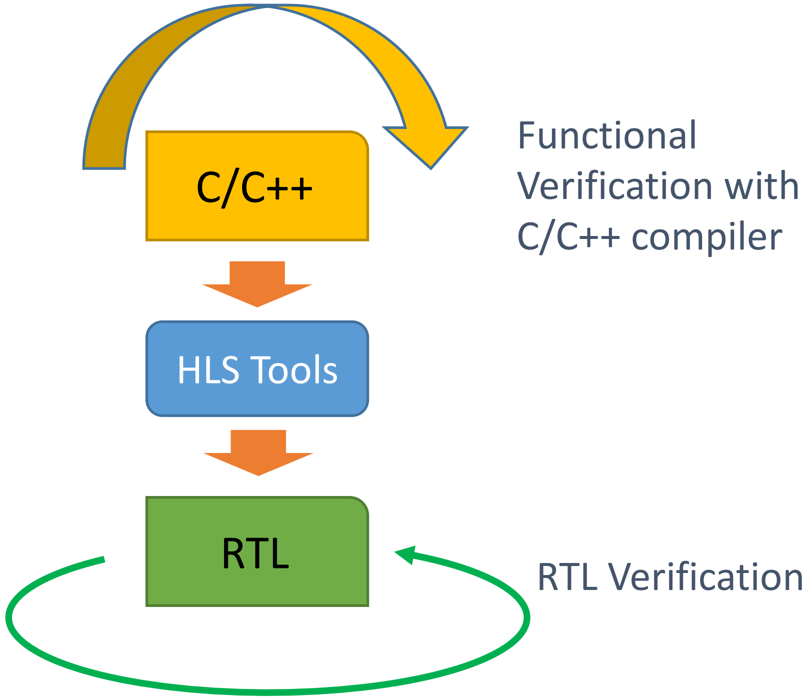


3 processing layers, 3 data channels

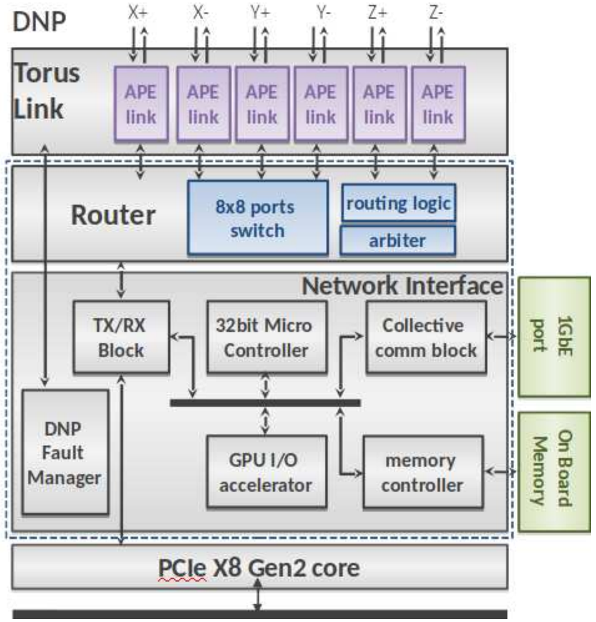
Strict loop between definition of processing, heterogeneous hardware platform, mapping among them and performance evaluation.

High Level Synthesis Tools

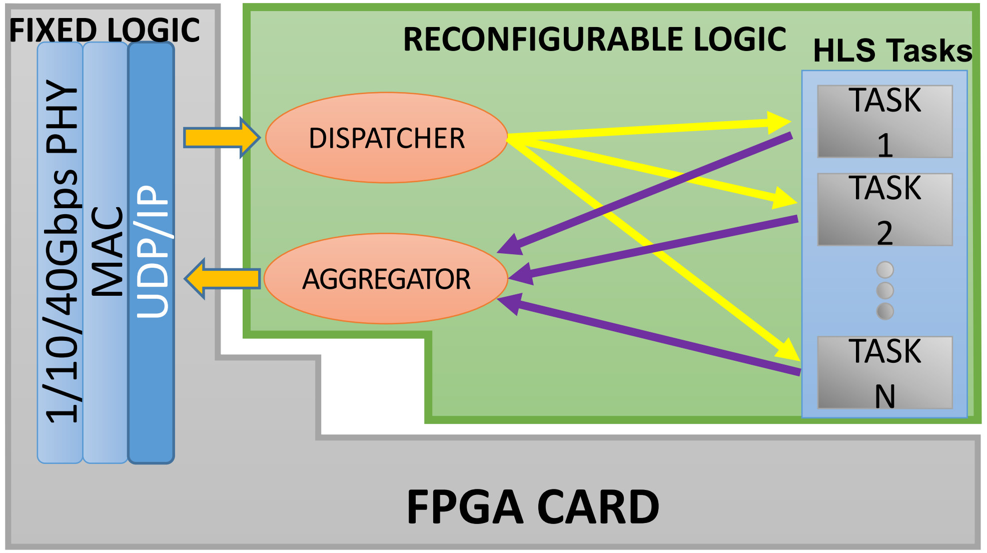
- Taking an abstract behavioural or algorithmic description of a digital system and creating a corresponding RTL structure
- Enabling C/C++ code to be directly targeted into programmable devices (FPGAs) without the need to writing VHDL/Verilog code
- Providing users with a faster path to IP creation and reuse
- Availability of libraries for math functions, arbitrary precision data types, linear algebra, DSP …



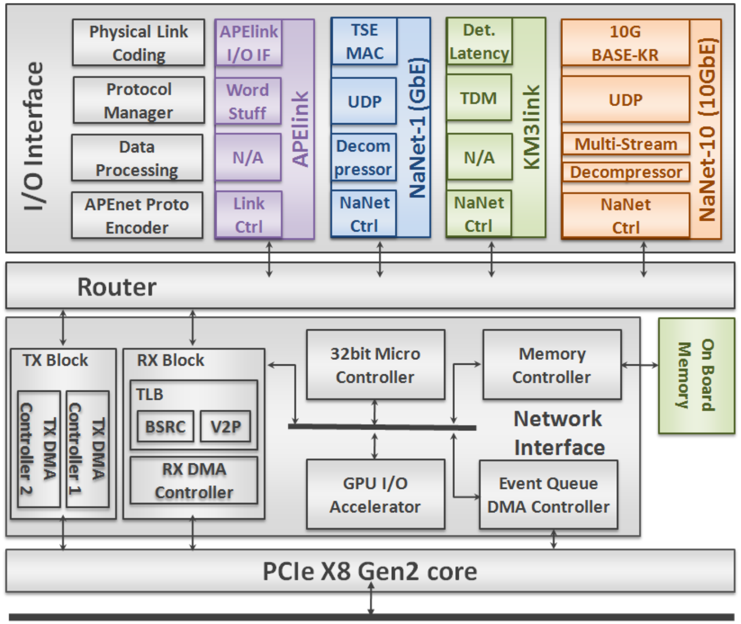
APE Group Network IPs



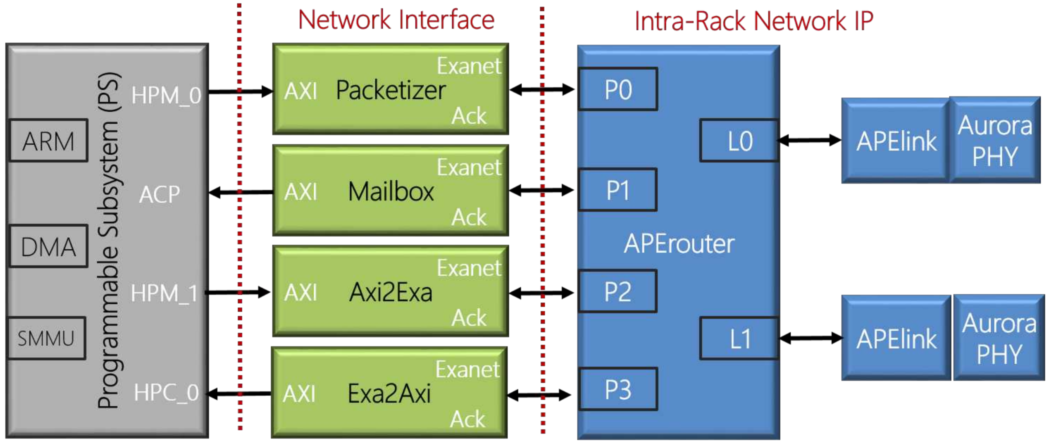
APENet



APEIRON

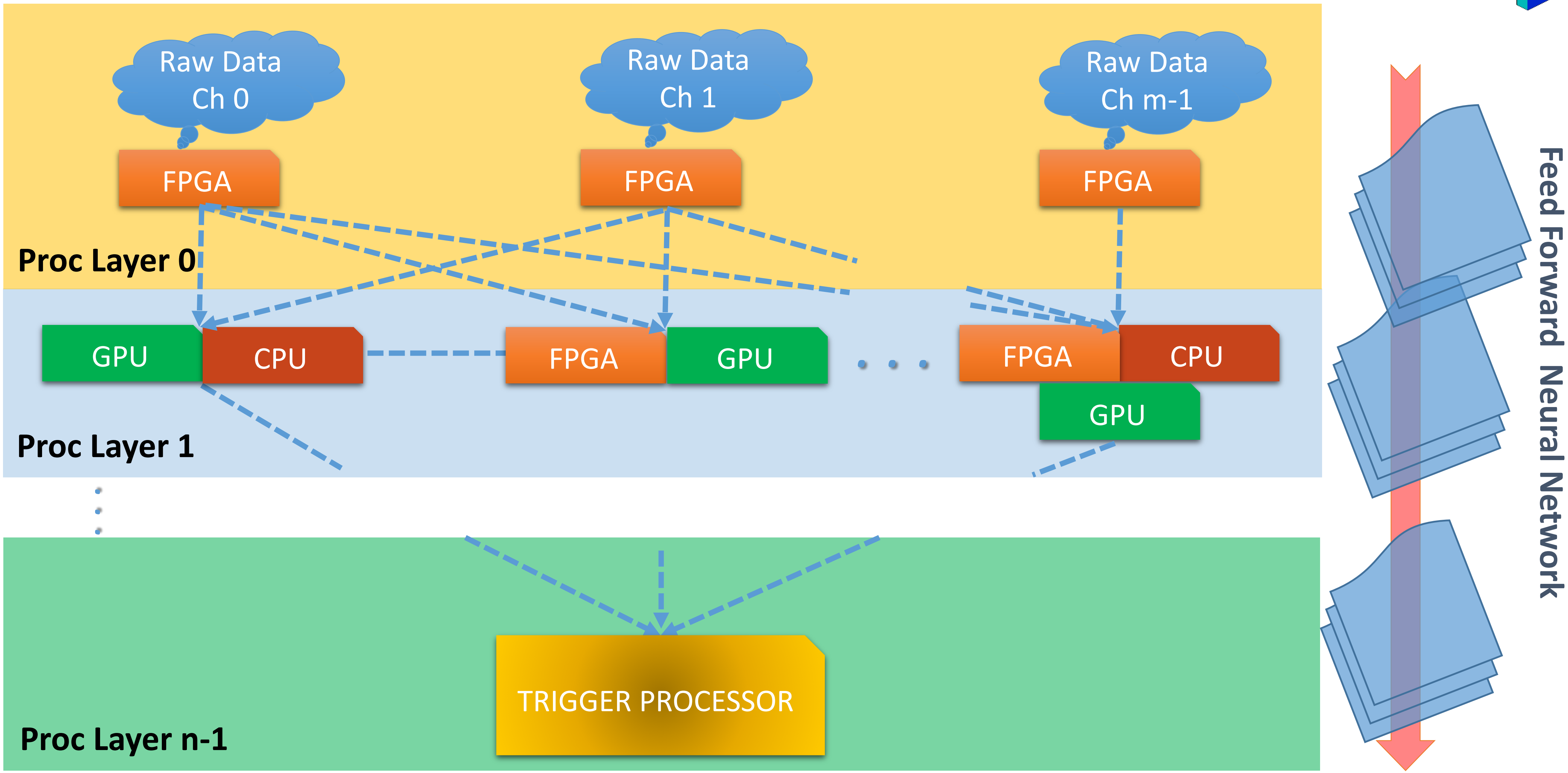


NaNet



ExaNet

Enabling the Use of NN in Low Level Trigger

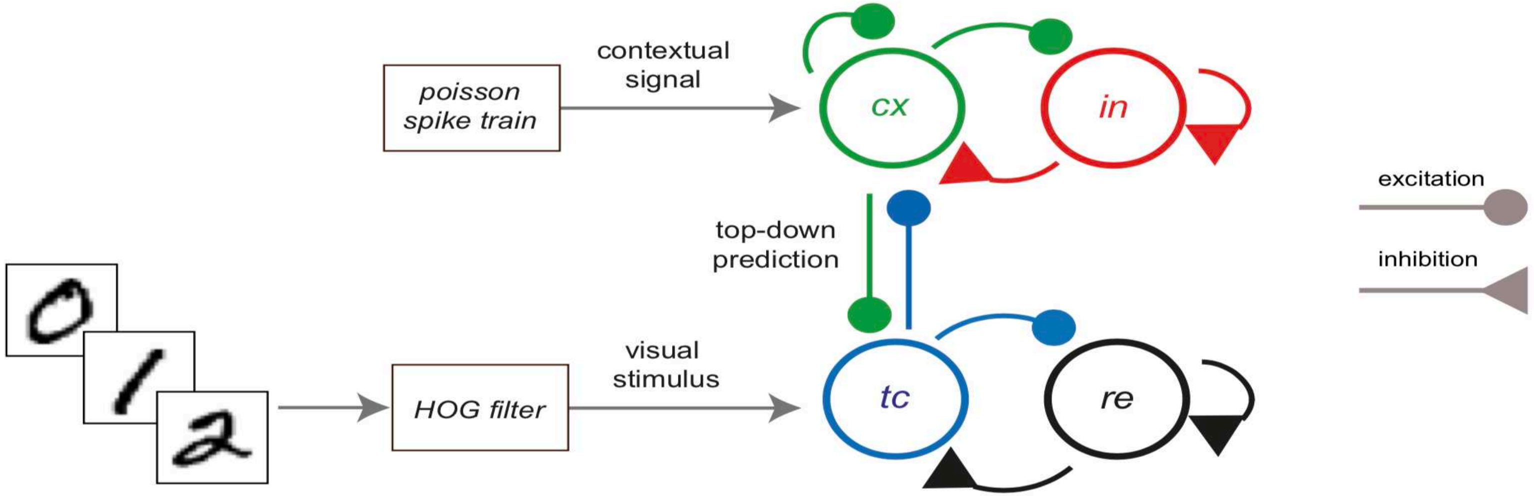


Enabling the Use of NN in Low Level Trigger

- Convolutional Neural Network (CNN) represented as a KPN: a process implements a layer, communication between layers occurs via channels.
- Distribution of processes possible at any scale:
    - Device (shared memory communication channels)
    - System (host bus communication channels)
    - Multi-System (network communication channels)
- Features extraction will occur in first NN layers (e.g. conv+ReLU+Pool), and will be implemented on FPGA devices in first processing layer, kind of «automatic primitive definition» through machine learning.
- This implementation must be lightweight to face the limited memory and floating point resources of the (possibly many) FPGA devices directly attached after the digitazation stage: study reduced precision and/or DNN compression techniques.
- More resource-demanding CNN layers implemented in subsequent processing layers.
- Classification produced by the CNN in last processing layer (e.g. pid) will be input for the trigger processor.

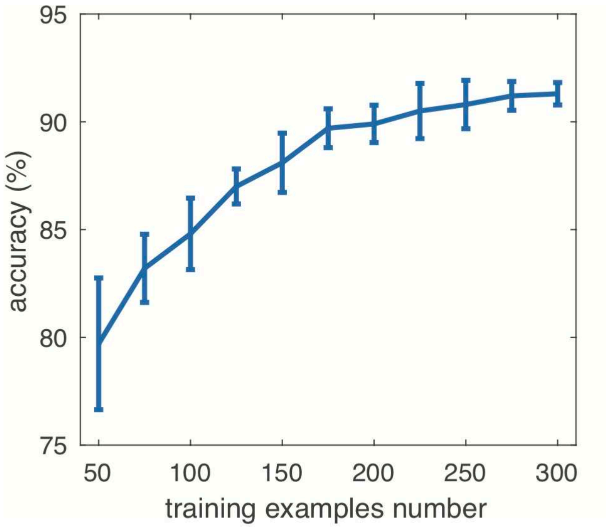
Fast learning from few examples, in a brain inspired thalamo-cortical spiking model

Scientific Reports (2019). *C. Capone, E. Pastorelli, B. Golosio, P.S. Paolucci*.



Current status:

- Neural Network trained to classify handwritten characters (MNIST dataset). The learning is incremental.
- after 10 examples per digit, 85% classification accuracy
- **Small memory footprint compared to CNN → promising for FPGA implementation**



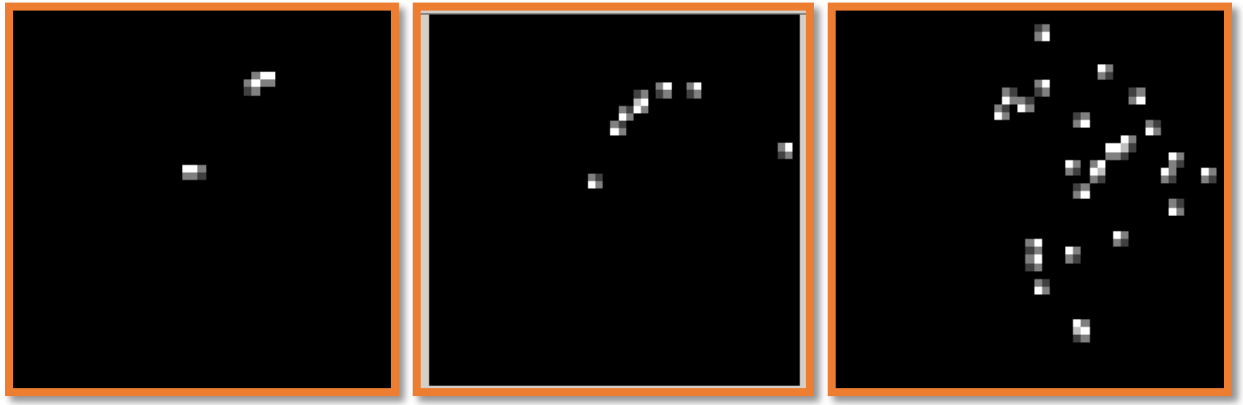
L0TP+: synergies and opportunities

Upgrade of the FPGA-based Level-0 Trigger Processor of the NA62 experiment at CERN for the post-LS2 data taking (2021-2024), and more:

- Avoid obsolescence of current platform (Altera Stratix-IV → Xilinx Ultrascale+).
- Exploit higher performances (clock frequency, memory, high speed serial links) and **new design flow (High Level Synthesis) introduced with recent FPGAs.**
- Be ready to support (at least) x4 beam intensity foreseen in future experiment developments (NA62x4 and KLEVER) through many 10GbE/GBT channels.
- Add new functionalities, e.g.:
  - **Support tightly coupled CPUs and/or GPUs through PCI Express to implement software triggers, leveraging the NaNet design.**
  - **Use the considerable computing power of the Xilinx Ultrascale+ to improve trigger performances (next slide)**
- INFN Roma1/2, Pisa, Torino. Roma1 is coordinating the activities.

Use Case 1: Partial Particle Identification Using RICH Data in Na62

- Partial reconfiguration of trigger firmware starting from a high level language description (C/C++) enabled by modern High Level Synthesis (HLS) tools, but to what extent this methodology can be applied in the L0 trigger must be verified.
- Case study: partial particle identification in the RICH detector with a CNN in the FPGA



Rings:

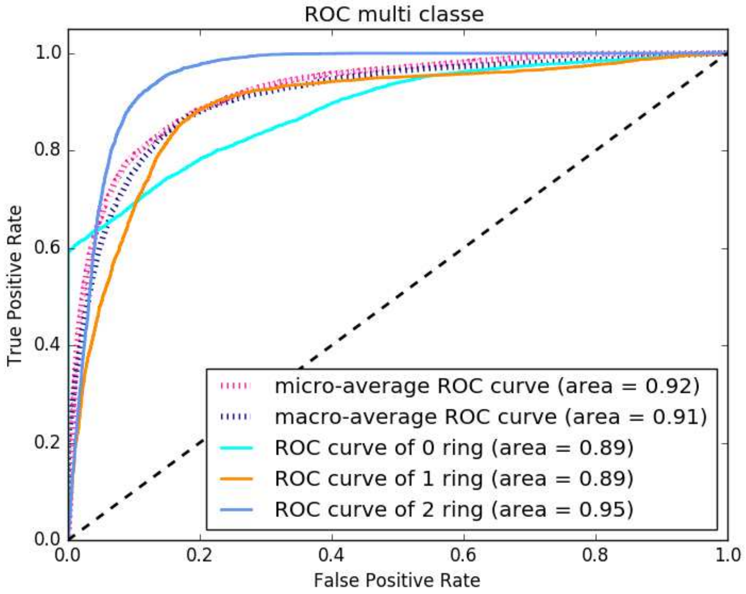
0 1 2

TASK

- Count the number of rings.
- RICH hit maps transformed into 46x46 images.
- In collaboration with A. Ciardiello

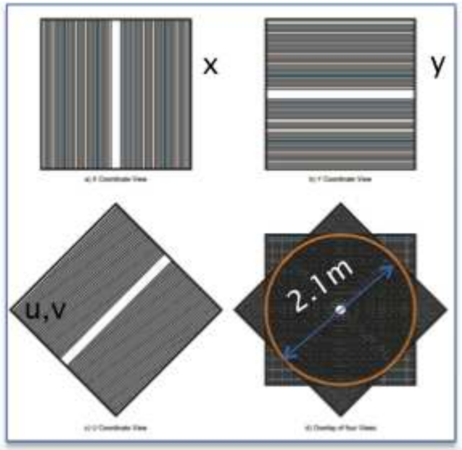
NEXT STEPS

- Classification for type of particles (electrons, K or Pi from the beam, others)
- Implement the “minimal” CNN on the FPGA using HLS



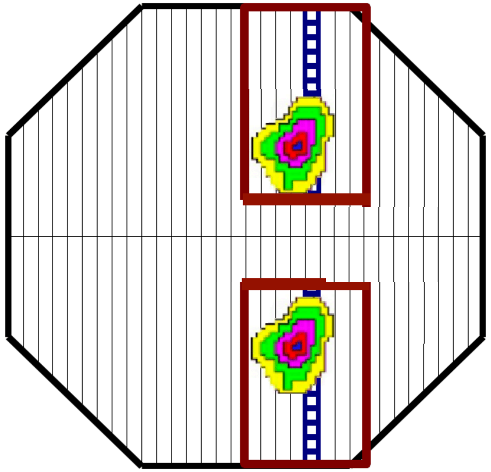
Use Case 2: Full Particle identification in NA62 L0 Trigger

Straw tube hits image



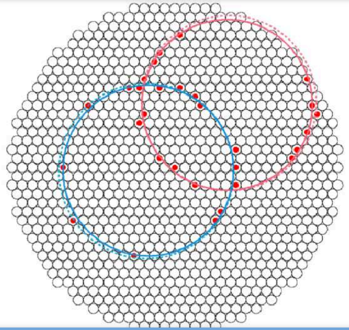
Use ML on straw tube image to identify one or more LKr regions of interest

LKr hits image



Readout a region of LKr, use ML on the image to classify clusters (e, μ, π)

Send a PID code



Partial PID on RICH data using CNN

Send a PID code

Decision

Use the combined info to get PID decision

Workpackages, Project Schedule & Financial Requests

- WP1 - Framework Definition
- WP2 - HW Developments
- WP3 - SW Developments
- WP4 - Physics Use Cases
- WP5 - System Integration and Benchmarking

- Year 1
  1. Architecture specification.
  2. Dataflow programming framework (DPF) prototype.
  3. List of relevant candidate Physics Use Cases (PUC) along with the preliminary definition of the associated processing.
  4. Procurement of a small scale testbed.

- Year 2
  1. Delivery of HW IPs.
  2. Delivery of a first release of DPF.
  3. List of selected PUC.
  4. Implementation of selected PUC processing using DPF validated on a test platform.
  5. Definition of the execution platform for the selected PUC.
- Year 3
  1. Integration of execution platforms for PUC.
  2. Benchmarking of PUC using the framework.
  3. Dissemination of results.

Financial request: 100k€ in 3 years.

People

Nome	**Profilo**	**Sezione**	**Percentuale**
Alessandro Lonardo	Tecnologo	Roma1	30%
Roberto Ammendola	Tecnologo	Roma2	30%
Paolo Valente	Dirigente di Ricerca	Roma1	20%
Piero Vicini	Primo Ricercatore	Roma1	20%
Andrea Biagioni	Tecnologo	Roma1	20%
Ottorino Frezza	Tecnologo	Roma1	20%
Francesca Lo Cicero	Tecnologo	Roma1	20%
Gianluca Lamanna	RTDb	Pisa	20% in sovrapposizione con NA62
Pier Stanislao Paolucci	Ricercatore	Roma1	20% in sovrapposizione con HBP
Mauro Raggi	Prof. Associato	Roma Sapienza	20% in sovrapposizione con NA62
Francesco Simula	Ricercatore	Roma1	20% in sovrapposizione con NA62
Paolo Cretaro	Assegnista	Roma1	10% in sovrapposizione con NA62
Giulia De Bonis	Ricercatore	Roma1	10% in sovrapposizione con NA62

APEIRON @ Tor Vergata

- Anagrafica: R. Ammendola 30%
- Richieste: missioni 1 kEuro
- Attivita`:
  - responsabilita` del WP2 (HW Developments)
  - sviluppo interfaccia PCIE GEN3/4 e interfaccia di rete