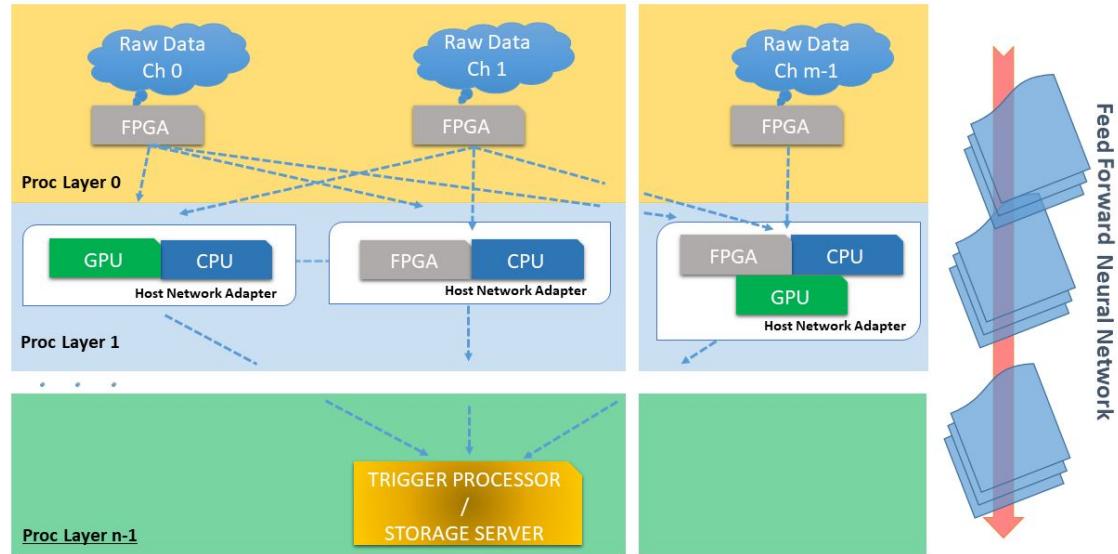
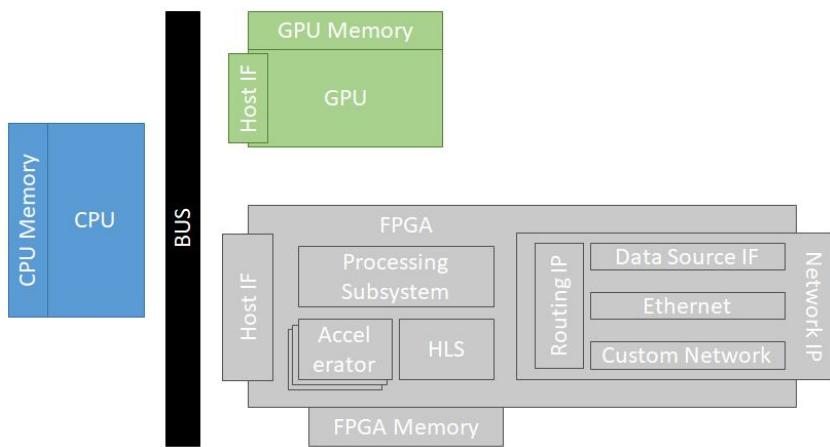


APEIRON

Abstract Processing Environment for Intelligent Read-Out systems based on Neural networks

- m input data channels (from detectors/sub-detectors).
- Distributed processing on heterogeneous computing devices in n subsequent layers.
- Data streams from different channels recombined through the processing layers using a **low-latency, modular and scalable network infrastructure** (configurable in number of channels, topology and size).
- Exploit the specialization of modern computing devices
- Keep the definition of processing and communication the more abstract and device independent as possible.



- Features extraction will occur in the first NN layers on FPGAs: reduced precision and/or DNN compression techniques are to be studied and implemented.
- 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/storage online data reduction stage.

APEIRON 2020

RM2 Direct Contribution

Milestone Concordate		data	% compl.
M1.1: Specifica architetturale della piattaforma generale di esecuzione di APEIRON e specifica implementativa del modello di programmazione dataflow ADE		30/06/2020 (+ 3 mesi)	50%
M2.1: Specifiche Host Interface e Network Interface. Specifiche interfaccia blocco acceleratore		31/12/2020	100%
M3.1: Versione prototipale del software stack di APEIRON. Rilascio del prototipo del driver per Host Interface basato sulle specifiche; definizione della Libreria User Space per canali di comunicazione; specifica del formato ADE; linee guida per CNN e SNN plastiche su APEIRON		31/12/2020 (+ 3 mesi)	75%
M4.1: Definizione dell'architettura minimizzata per FPGA della CNN che implementa la Partial Particle Identification in NA62 e sua implementazione prototipale		31/12/2020	100%
M5.1: Definizione e procurement dei componenti per il testbed dedicato allo use case Partial Particle Identification in NA62		31/12/2020	100%

- Impatto COVID-19: no accesso al LAB, altre attività hanno impegnato i partecipanti in % di tempo maggiori del previsto.
- Tutti i fondi assegnati sono stati impegnati grazie alla operatività ininterrotta dell'amministrazione.
- Sottomesso proposta di progetto EU (H2020-JTI-EuroHPC-2019-1 “TEXTAROSSA”) per 270k€ di finanziamento, attualmente in fase di valutazione.

Use case of APEIRON in NA62: Partial Particle Id in the LOTP+ (I)

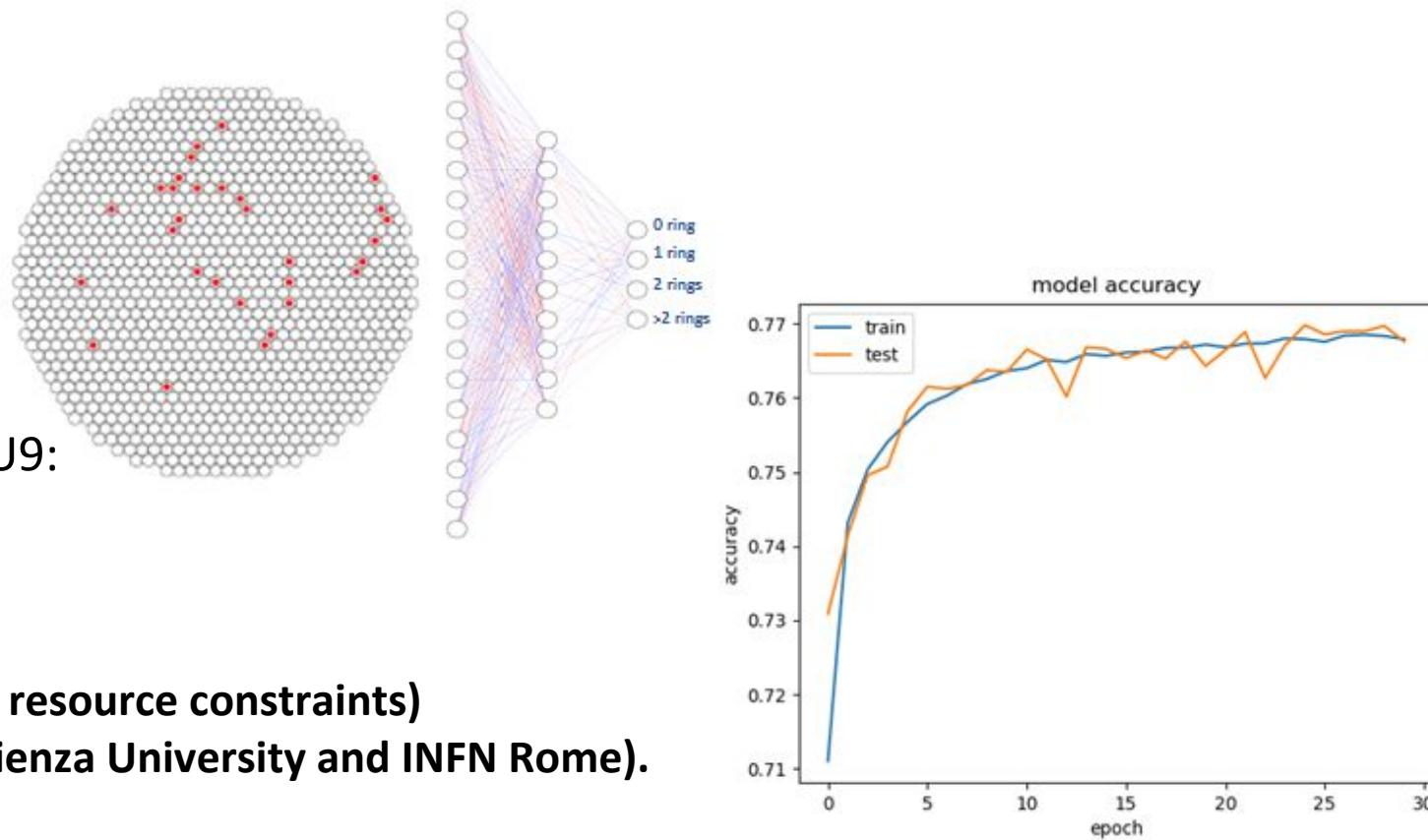
- Counting the number of rings per event in NA62 RICH using a Neural Network implemented on the FPGA.
- Preliminary step towards Partial Particle Identification in LOTP+ : 0, 1, 2, >2 rings, how many are electron rings.



First test model with 3 dense layers:

- Input: 64 hits per event
- 18-bit fixed precision
- Trained on 80000 sample events
- Accuracy: ~ 75%
- Latency: 40 cycles (@100MHz)
- Initiation Interval: 8 cycles
- Low ratio of used resources on the Xilinx VU9:
LUTs 3.3%, DSP blocks 9.7%, 0.9% BRAM.

Integration with LOTP+ design is feasible.

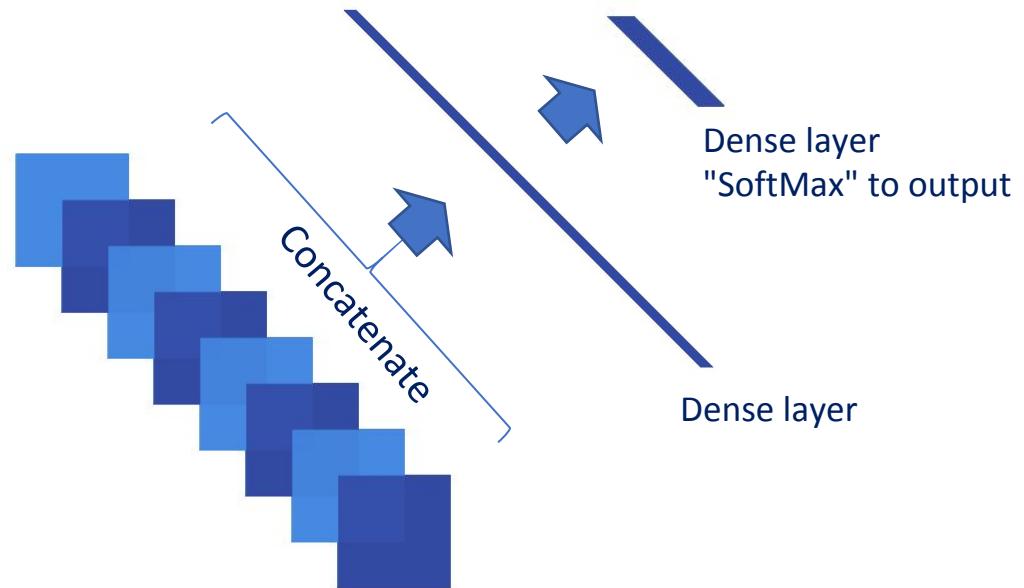
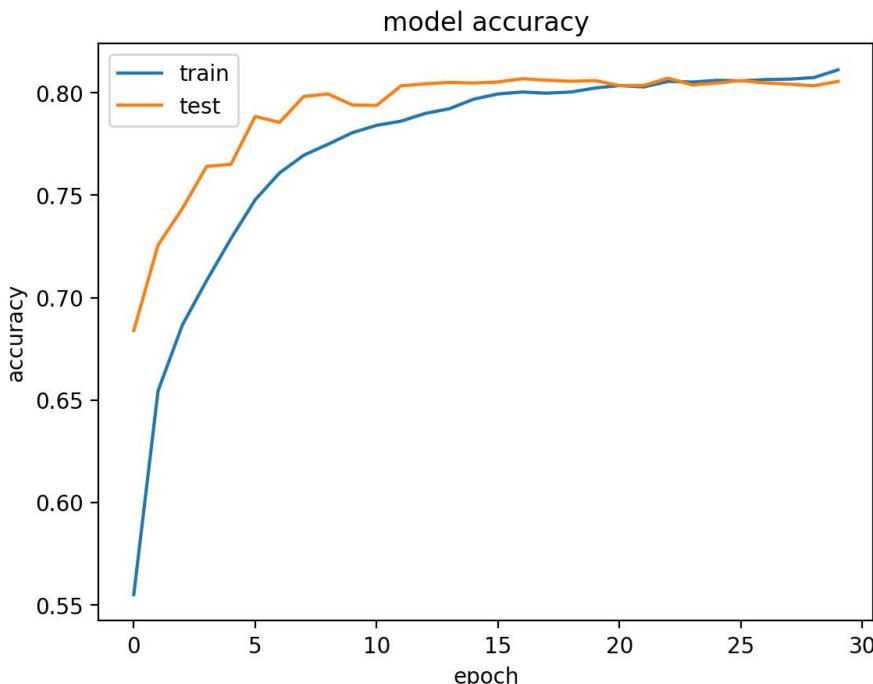


- Working to improve accuracy (within the FPGA resource constraints)
- Activity in collaboration with A. Ciardiello (Sapienza University and INFN Rome).

Use case of APEIRON in NA62: Partial Particle Id in the LOTP+ (II)

Test model with CNN+dense layers (trying to improve accuracy):

- Accuracy: ~ 80%
- Input: image 49x49 (fixed-point precision)
- CNN layer filters processed separately then merged
- It needs a preprocessing stage to convert PMTs channels to image (also blur, resize and normalization)



Implementation estimates on VU9 FPGA:

- Clock period 11ns
- Resources: 2% BRAM, 5% DSP and 2% LUT
- Data and weights represented with 18-bit fixed-point
- Disabled all HLS optimizations to be able to synthesize the design
 - Processing latency per event 17835 cycles!!!
 - Investigating the critical issues preventing the optimized C++ to firmware flow (HLS4ML and HLS translation stages)

APEIRON 2021

Attività RM2

1. Mainly development of the APEIRON FPGA IPs (first release)
 - Host Interface
2. Collaboration in development of:
 - Network IP
 - Application-specific accelerators

Anagrafica RM2 (0.3 FTE)

- Roberto Ammendola 30%

Richieste Finanziarie

- Missioni: 1k€

Anagrafica RM1 (1.2 FTE)

- Alessandro Lonardo (Resp.) 30%
- Andrea Biagioni 20%
- Ottorino Frezza 20%
- Pier Stanislao Paolucci 20%
- Paolo Valente 20%
- Piero Vicini 10%

Richieste Finanziarie RM1

- Missioni: 4k€
- Consumo: 2k€
- Inventario: 10 k€
 - Xilinx Versal dev kit
 - Rack