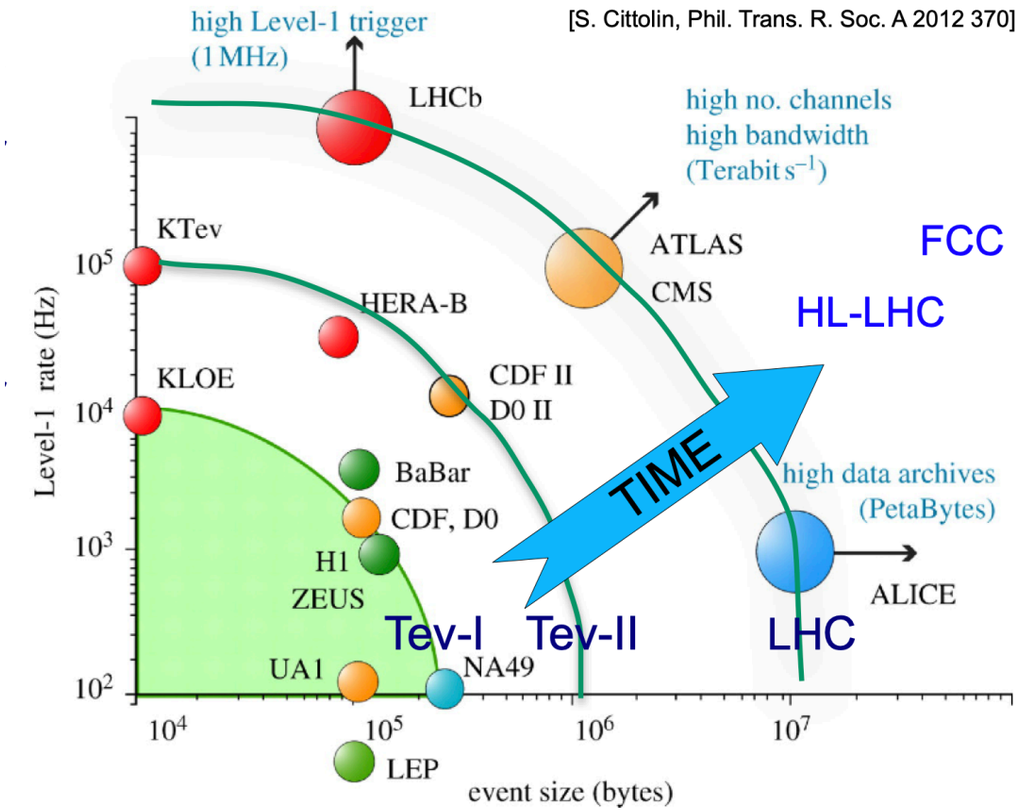
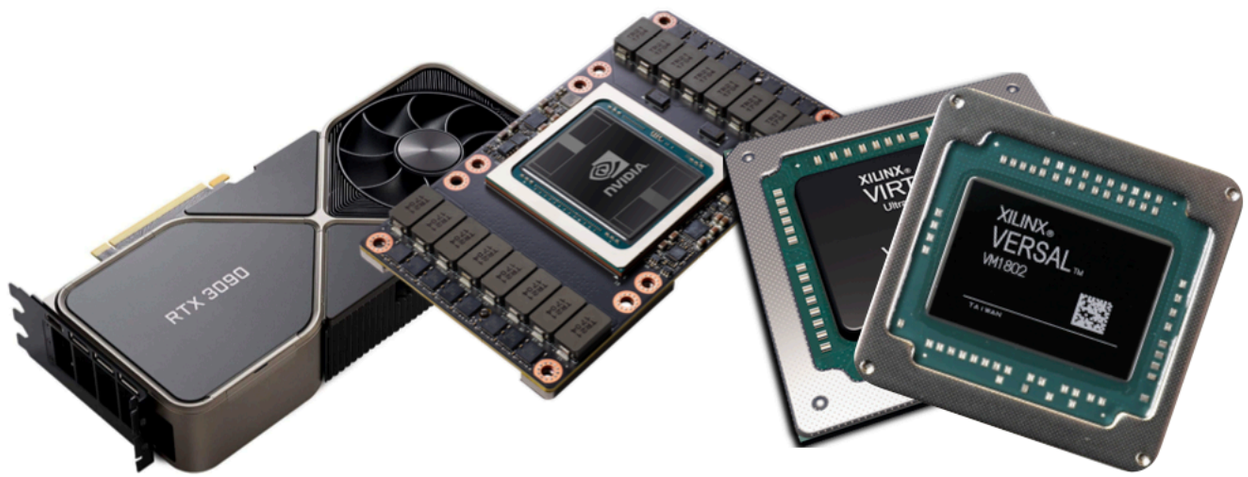
An FPGA arrays test bed for trigger-related applications (but not only)

What's that?

- The idea is to build in Bicocca a "small" array of FPGAs
- (~> 10) interconnected with large bandwidth (400 G Ethernet) for testing applications mostly dedicated to trigger reconstruction and selections for HEP.
- FPGAs are already used at Level-1 trigger for CMS and ATLAS
- we can use this system to test more innovative use cases and build expertise also for usecases outside the accademia
  - One usecase is the use of interconnected fpgas for fast tracking at Level-1 for the LHCb experiment (Retina project)
    - See Punzi's talk at the ICSC Spoke2 kick off meeting





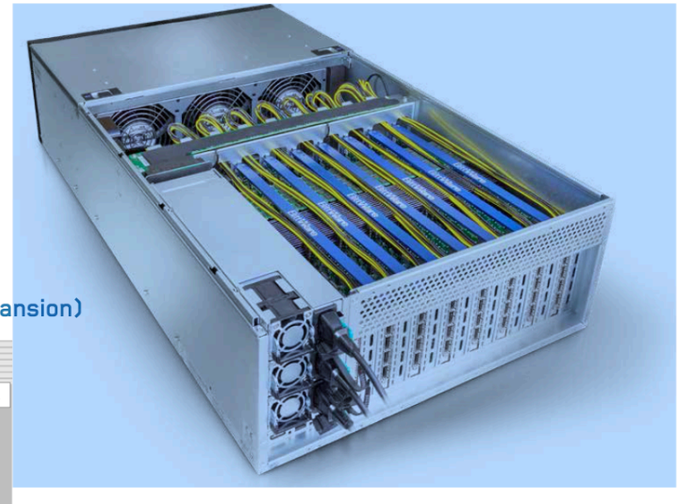
People involved (at the moment)

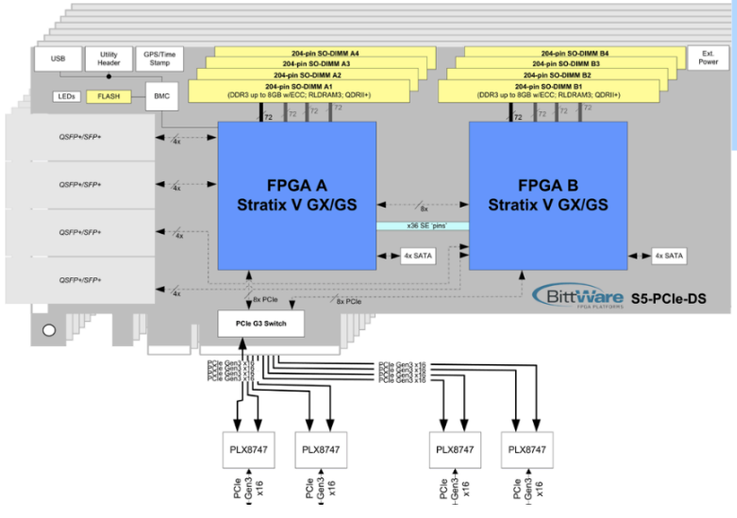
- Mostly trigger driven (so far):
  - ATLAS (Level-1 and HLT), LHCb, and CMS (Level-1 related)
  - Perugia is interested in writing SW for easy transport ML application into the FPGA
  - Bologna more interested in developing a framework to make easier the use of accelarators
- People involved (or that may be interested):
  - MIB : Paolo Dini, Simone Gennai, Maurizio Martinelli, 1 person to hire through PNRR-INFN related funding
  - Pisa: G. Punzi, R. Fantechi, M.Morello, F. Lazzari, G. Bassi , D. Passaro, F. Terzuoli
  - Cagliari: A.Contu
  - Roma : S. Giagu, 1 post doc, 1 phd (?)
  - Napoli: B. Spisso
  - Perugia: M. Mariotti, G. Bianchini, L. Storchi
  - Bologna: R. Travaglini (may be interested at a later time)
  - Padova: J. Pazzini (possibly interested, not yet in the loop)

Where we are

- At the moment we are collecting use cases, interested people and know how.
- We are in contact with E4 company for a gathering information about the price of HW (with different version of FPGAs)
  - These are quite expensive toys to play with …
  - a minimal set up would cost ~200K Euro
- Plan is to test different vendors (Intel and AMD)

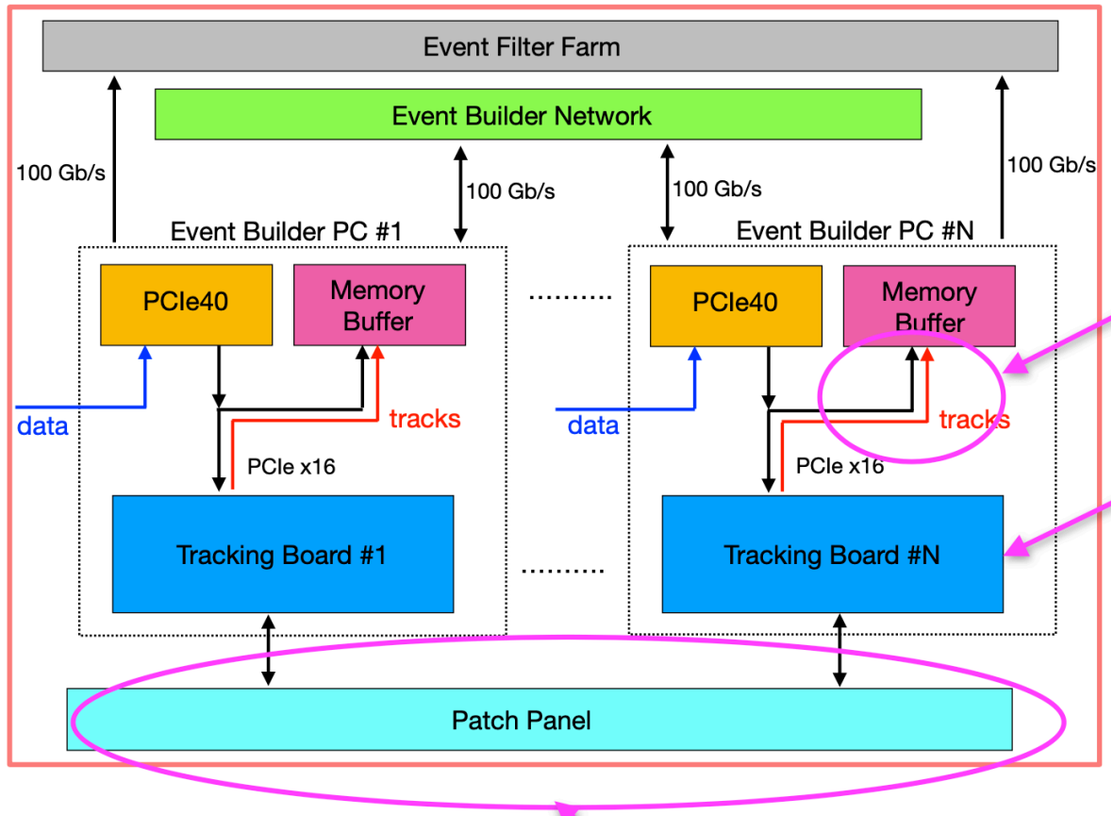
- **24 TeraFLOPS processing: 16x Intel Arria 10 or Stratix V FPGAs**
  - Up to 18 million logic elements (Arria 10 GX)
  - Up to 62,000 multipliers (Stratix V GS)
- **1.28 Terabits/sec I/O**
  - 128x 10GbE, 32x 40GbE, or 32x QDR Infiniband
- **6.5 Terabits/sec memory bandwidth**
  - Up to 64 banks DDR3-1600 (512 GBytes)
  - DDR4, QDRII+, and RLDRAM3 memory options
- **4U or 5U Rackmount PCIe system (server, industrial, or expansion)**





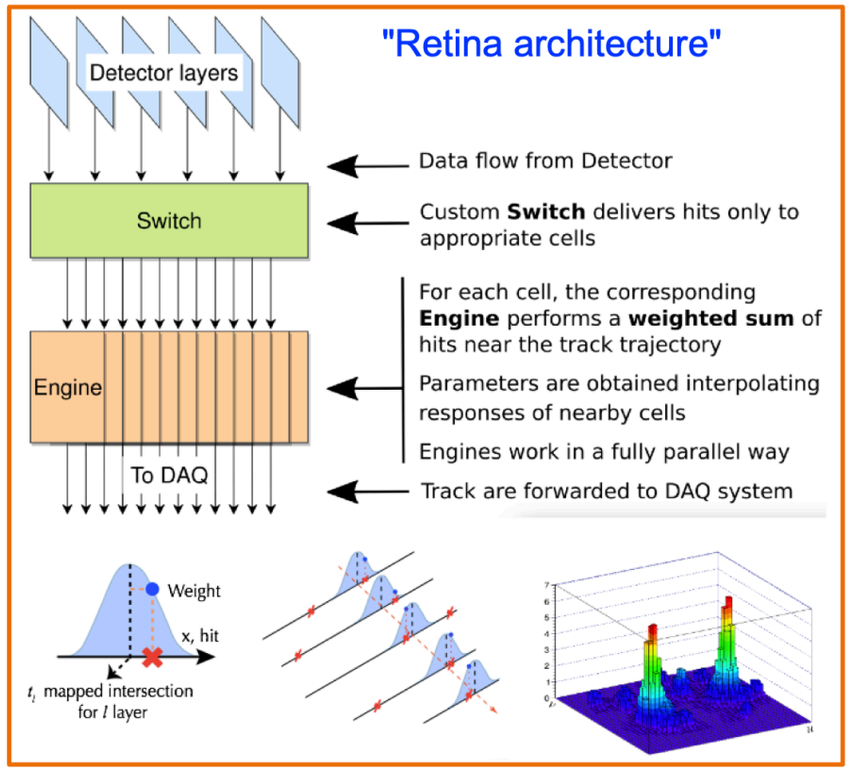
- A Gen-4 version would have 2x bandwidth

Use cases: LHCb



Tracking cards ADD extra "RAW Data" to event

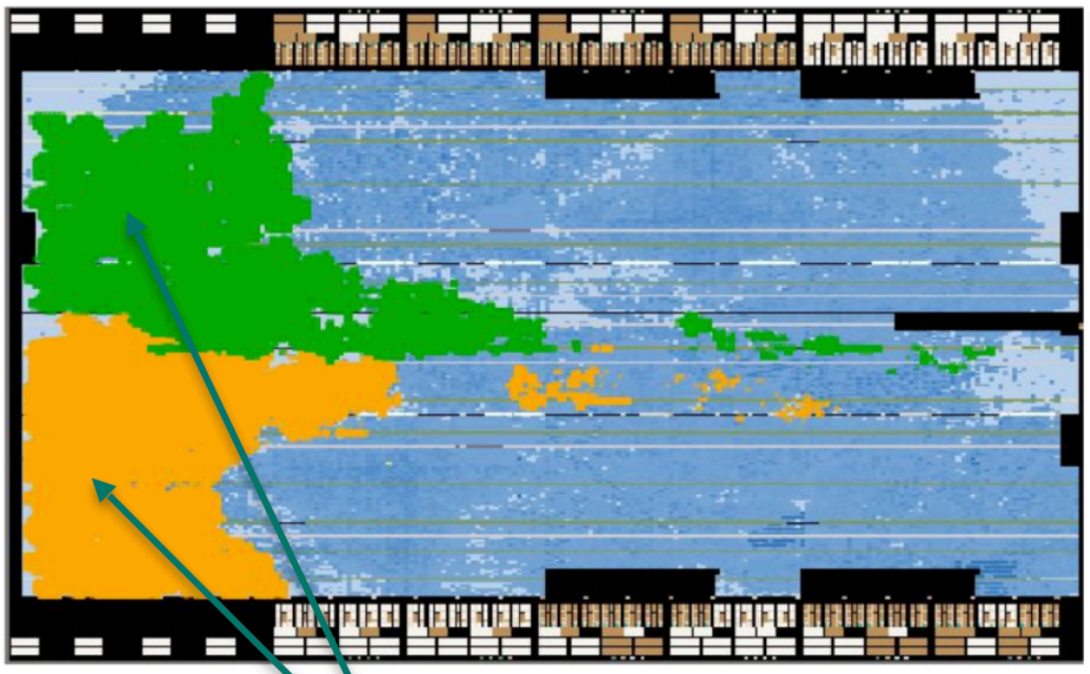
Match every readout card (PCIe40) with a tracking card



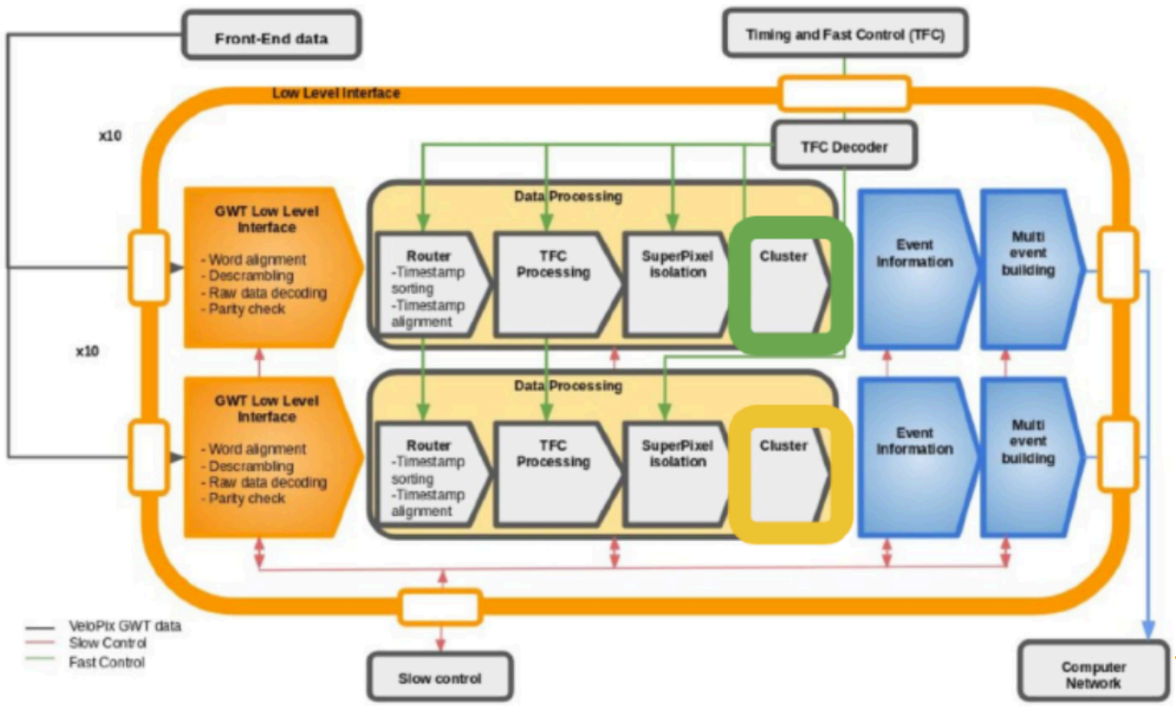
- Architecture for Real-Time (@ 40 MHz) data processing, embedded in array of commercial servers:
  - Tracking cards communicating via fast optical network
  - Very low latency (<1us), makes all boards appear as a single device to the Event Builder
  - Architecture experimentally demonstrated in a testbed setting [F.Lazzari, cdot 2020]

Use cases: LHCb

- In this case the project is in an advanced state, they already have a test bed in Pisa and at CERN, the 1st stage of tracking is running now at LHCb
- they want to test their system with newer generation of FPGAs and have a solid test bed for prototyping their Run-4 tracking system currently under review at LHCb
  - e.g. use part of the FPGAs to generate/simulate events with large throughput and use the others to reconstruct them



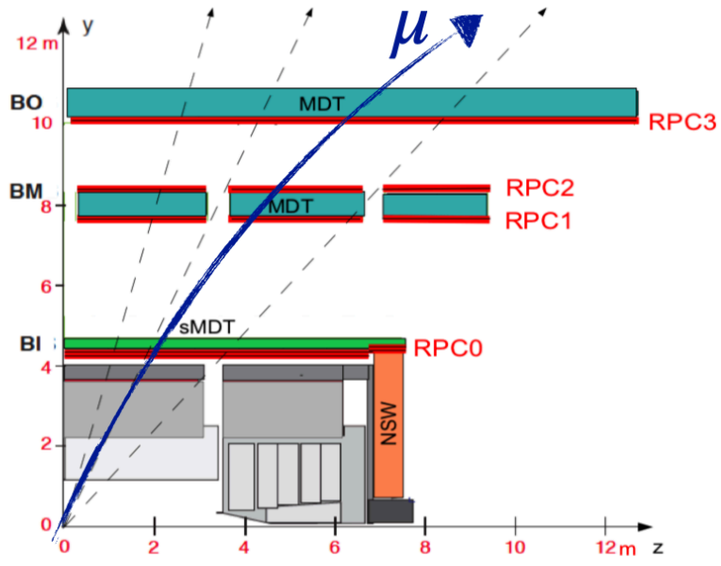
Clustering placement in VELO DAQ FPGA chip



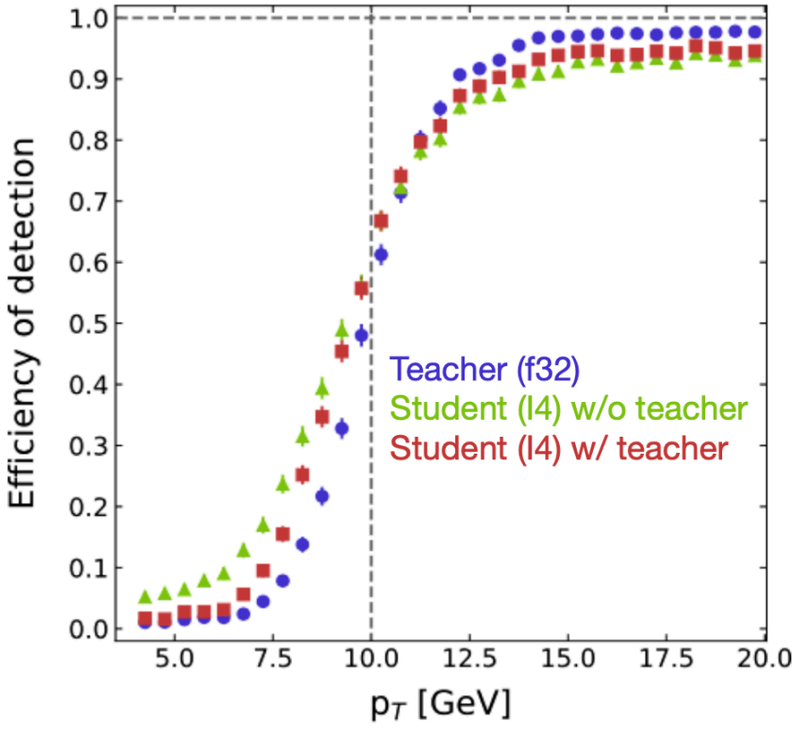
VELO firmware block diagram

Use cases: ATLAS

- Development of ML algorithm for ATLAS trigger system (both Level-1 and HLT).
  - Ultra fast CNN for muon trigger development
    - pT selection ad L1, looking for LLP at HLT



curva efficienza single muon trigger
soglia nominale pt 10 GeV



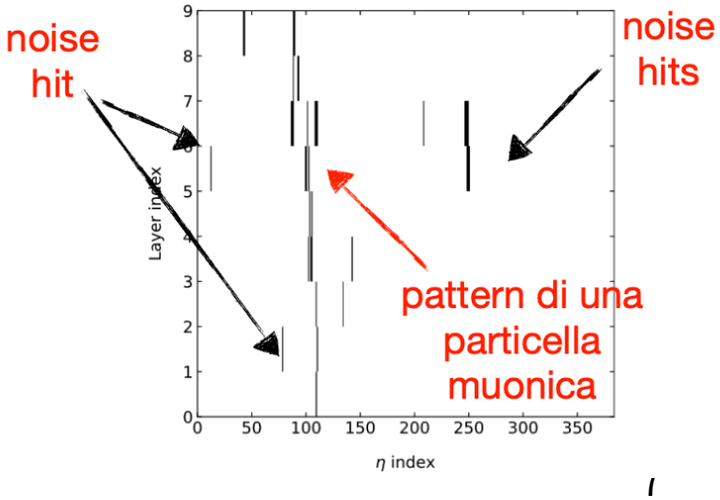
FPGA resource occupation

Table 3 Percentage occupancy relative to the total FPGA available resources (model xcvu13p-fhga2104-2L-e [14])

Model (9 × 16)	BRAM	DSPs	FF	LUT
Teacher (%)	20.9 %	258.0 %	69.4 %	15.3 %
Student 32 bit (%)	3.2	31.0	8.4	2.7
QStudent 4 bit (%)	0.2	0.05	0.4	1.7

Inference time per event on FPGA
Xilinx Ultrascale+ XCV13P

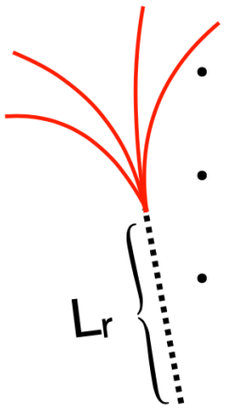
- Teacher fp32: 5 ms (Tesla V100 GPU)
- Student 4 bit: 438 ns (hls4ml implementation)
- Student 4 bit: 84 ns (our VHDL implementation)



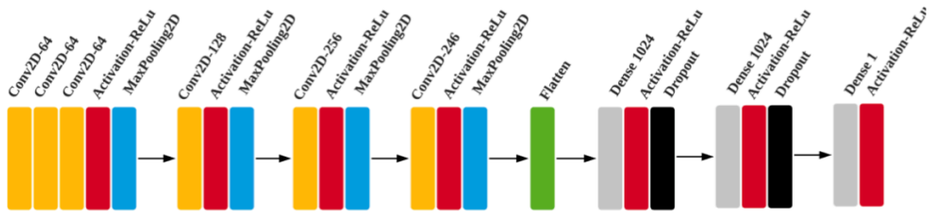
Use cases: ATLAS

- Development of ML algorithm for ATLAS trigger system (both Level-1 and HLT).
  - Ultra fast CNN for muon trigger development
    - pT selection ad L1, looking for LLP at HLT

High Level Muon Trigger: modello DNN addestrato ad identificare particelle a lunga vita media a partire dei pattern di hit rilasciati in un rivelatore dello spettrometro muonico

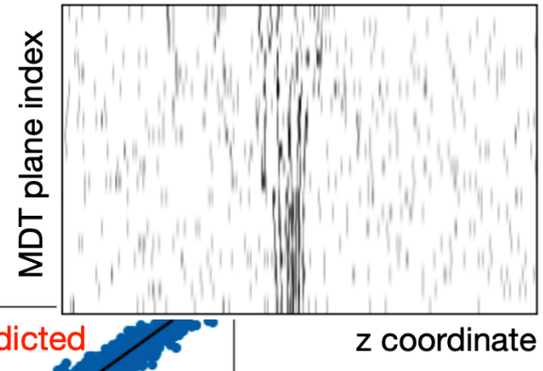


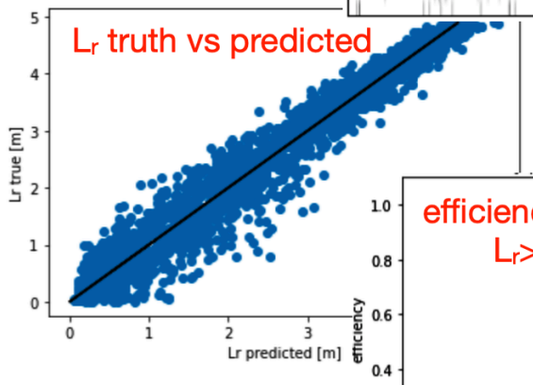
- benchmark: particelle neutre che decadono in multi-muoni a differenti distanze dal vertice primario di interazione (LLP)
- hits pattern nello spettrometro rappresentati come immagini binarie
- immagini usate per addestrare una CNN a predire la distanza di decadimento radiale (L_r) della LLP

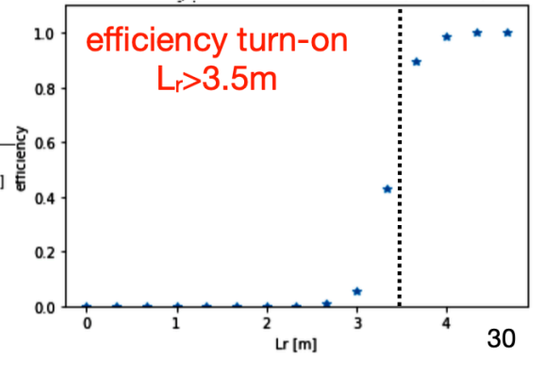


VGG Convolutional Neural Network architecture

example decay with 10 tracks in the MS

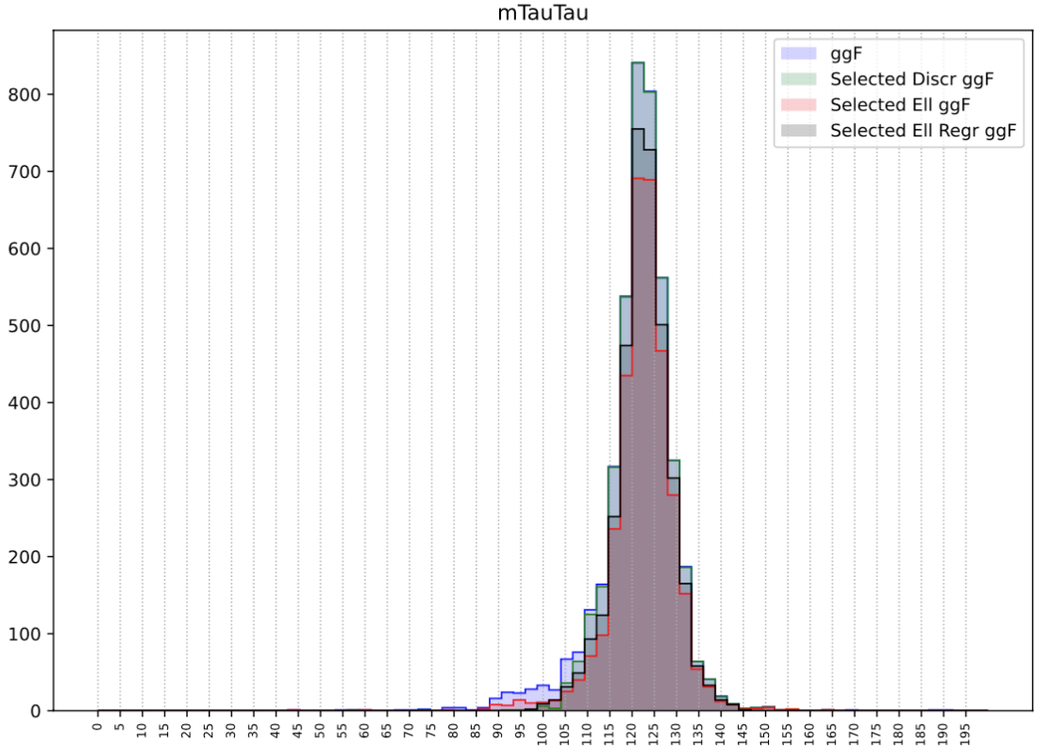


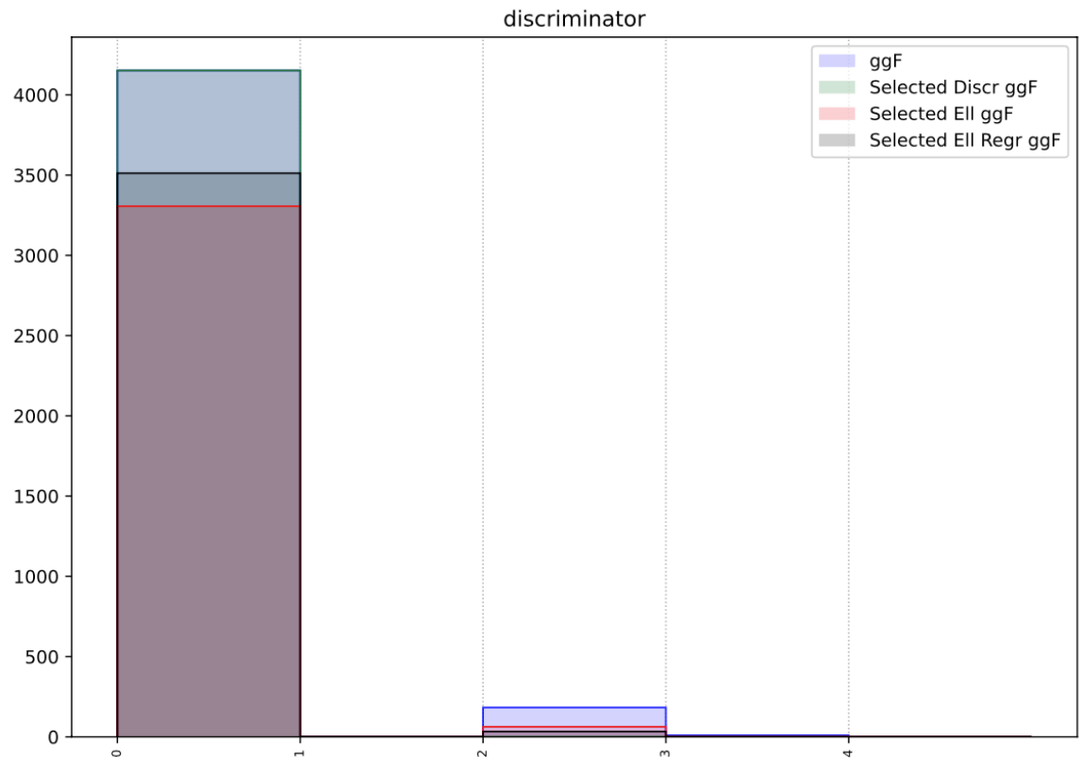




Use cases: CMS

- MIB: mostly interested in developping ML algorithms for HL-LHC @ Level-1
  - Possible contact point with the usecase from Perugia (see later in this talk)
- Interest in developping a ML based trigger for HH->bbtautau final state
  - Either a di-tau+mass cut or a topological one
  - at the moment working on the offline version of the model …





Use cases: CMS

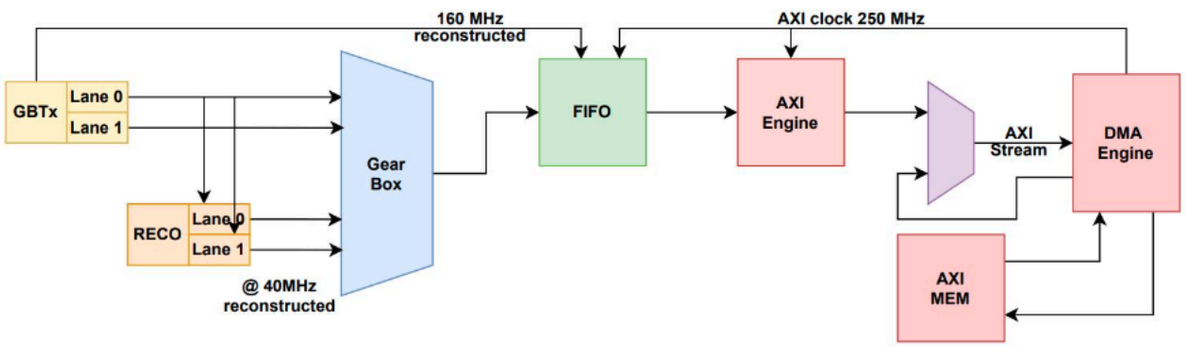
- CERN group may be interested as well
  - Use of the cluster for fast generation of events at 40 MHz Scouting analysis for HL-LHC
  - Test of complex ML models distributed on more than one FPGA
    - For complex L1 reconstruction/classification problems
  - Test of different SW for access the FPGA used as a service to accelarate part a ML model evaluation that would run on the associated CPU
  - Use of the cluster for emulation of L1 trigger decision for those algorithms that use ML@L1

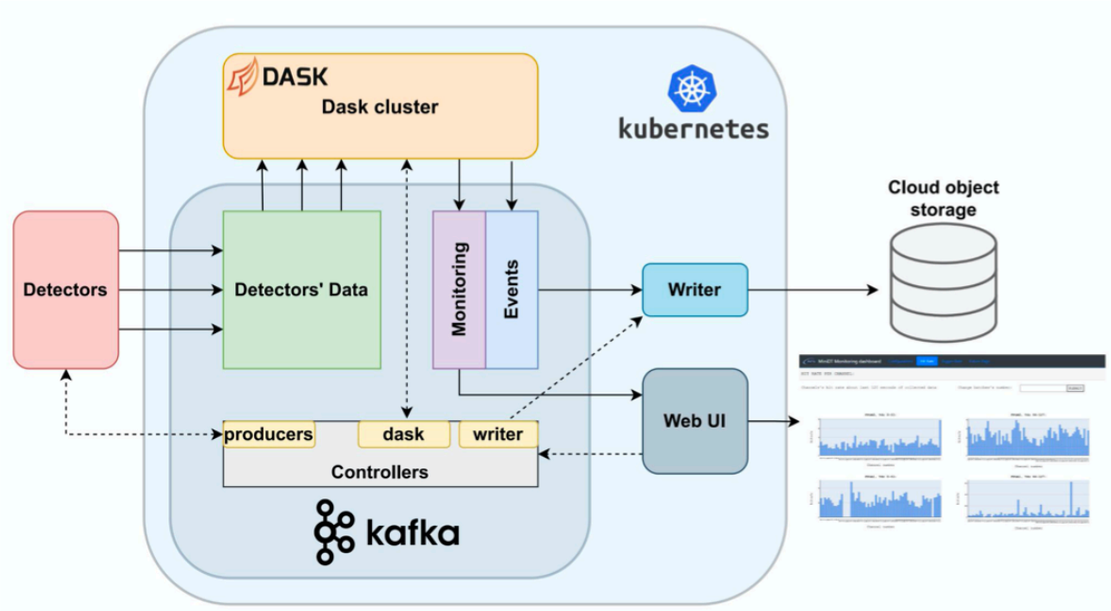
- With hls4ml we can make bitfiles for Alveo & Zynq cards quite easily
  - Alveo U200 / U250 are quite close in resources to VU9P / VU13P
  - ➡ if it fits in the L1T it should fit an Alveo
- We could just run the NN on an FPGA
  - Guaranteed agreement - same exact model firmware
  - Faster execution, profit from batching
- Probably this would only be an ‘extra’
  - Cannot assume that every site has Alveos to use

From Sioni Summers talk at ML@L1Workshop

Use cases: CMS

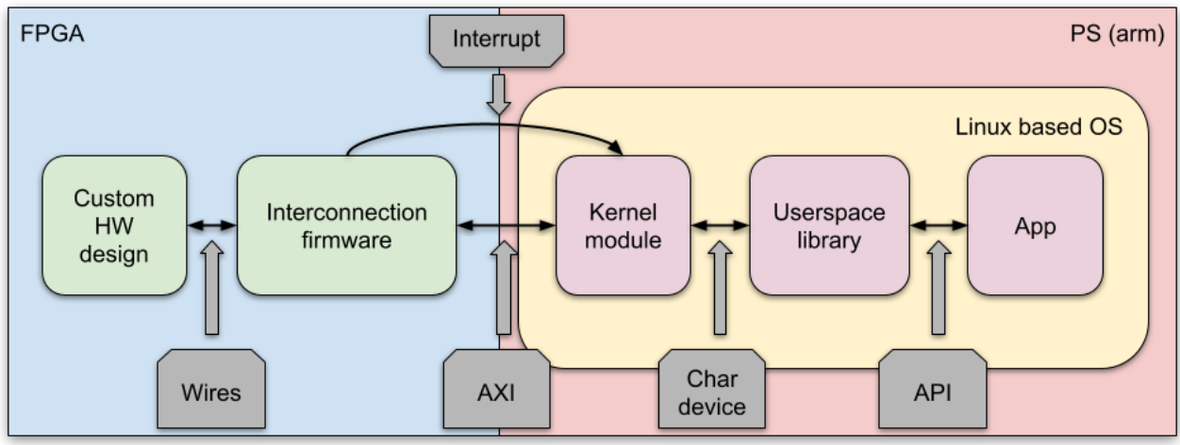
- Padova project on Muon scouting at 40 MHz, may be interested in FW development for FPGA
  - Not clear if really interested, did not receive feedback …





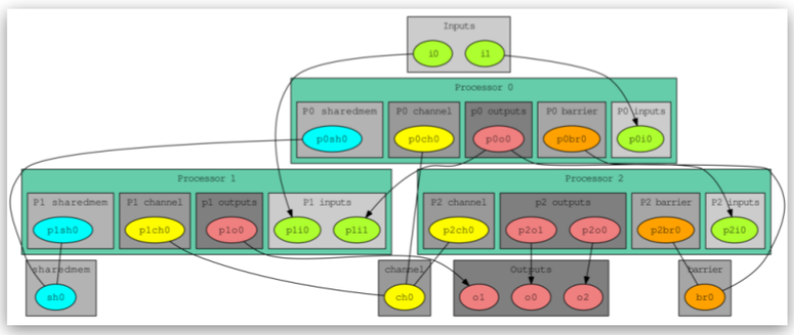
Fast machine learning inference with FPGA

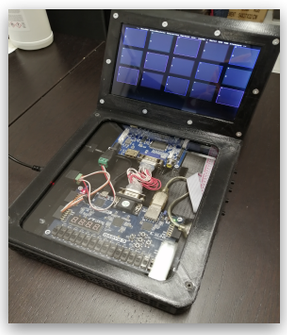
- From a machine learning model trained with standard frameworks, synthesized in FPGA as a graph of heterogeneous and interconnected processors
  - Features:
    - Optimized resource usage (LUTs, DSP)
    - Highly customizable
    - Available at a high level (Jupyter Notebooks, PYNQ)
    - Vendor independent (Xilinx-Amd, Altera-Intel)
- Development of accelated systems on hybrid processors (ARM & FPGA)
  - Exploiting FPGA clusters with these approach for benchmarking and supporting real scenarios

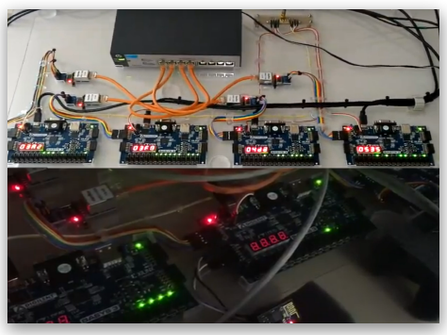


Fast machine learning inference with FPGA

- The BondMachine is an open source (https://github.com/BondMachineHQ) software ecosystem for the dynamical generation of computer architectures that can be synthesized on FPGA.
  - High level programming language (Golang) for both the hardware and software
  - Functional style programming
  - Computational graph and Neural Networks
  - Architecture generating compiler





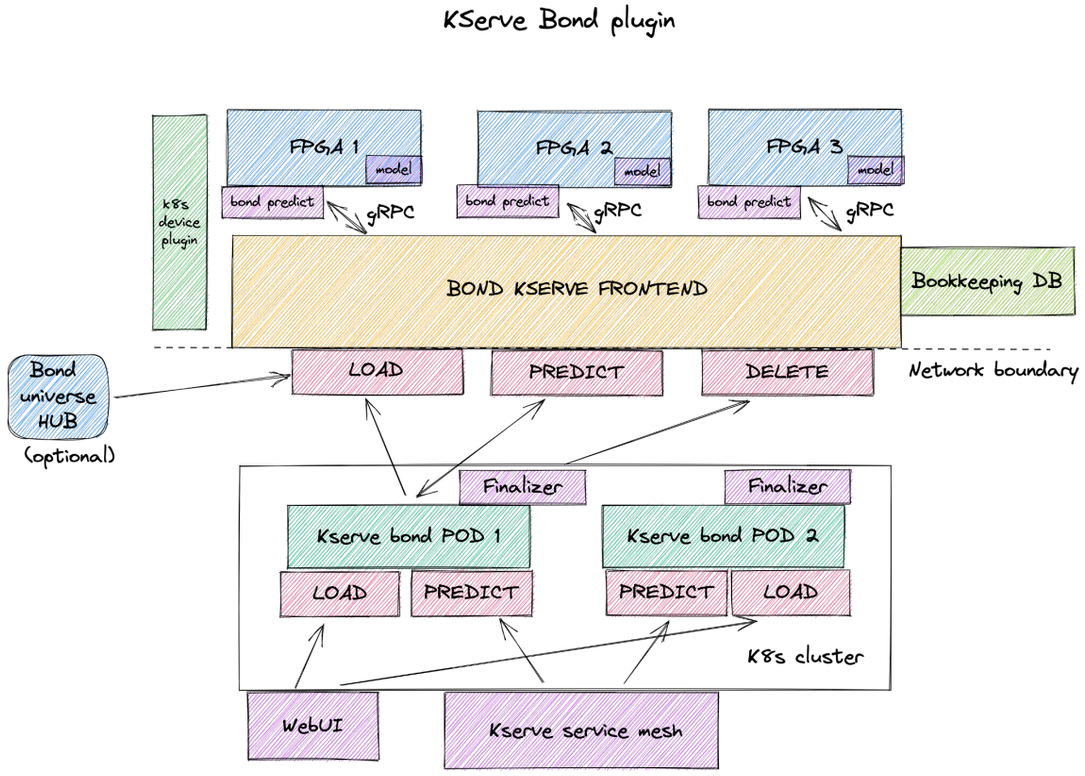


History and Major Highlight

- CCR
  - 2015 First ideas
  - 2016 Poster
  - 2017 Talk
  - 2022 Talk
- **InnovateFPGA 2018 Iron Award, Grand Final at Intel Campus (CA) USA**
- Invited lectures at FPGA workshops ICTP 2019 and 2022
- Golab 2018 talk and ISGC 2019 PoS
- Article published on Parallel Computing, Elsevier 2022 DOI:10.22323/1.351.0020

System of inference as a service

- Spin off of the previous project
  - Mirko Mariotti, Giulio Bianchini, Diego Ciangottini
- Chep2023 - KServe inference extension for a FPGA vendor-free ecosystem https://indico.jlab.org/event/459/contributions/11826/



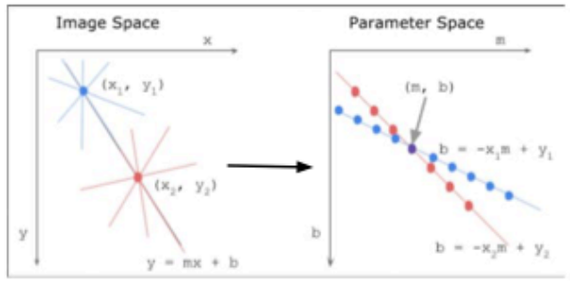
Back up

Hough Transform for ATLAS track reconstruction for HL-LHC

- See talk of Fabrizio Alfonsi at ACAT 2022
- https://indico.cern.ch/event/1106990/contributions/4991260/attachments/2533440/4359527/poster_acat2022_fabrizio_alfonsi_pre_20221015.pdf

Event Filter Tracking

ATLAS is studying the performance of the Hough Transform (HT) tracking algorithm to use it for the future Inner Tracker detector. To exploit it, the ATLAS environment requires that the position of the particle be defined with the polar coordinates radius "r" and azimuth angle "ϕ".



Generic example of HT algorithm: two points in the left are connected in the right thanks to a change in coordinate system

Two versions of HT implemented on FPGA are under investigation as candidates for **(raw) particle tracking for filtering**, the Flexible version and the Low-Resources version. The FPGA boards used for implementation tests are Xilinx commercial demonstrators VC709, used by both versions, and VCU1525.

Z-slices	μ 1-2 Gev	μ 2-4 GeV	μ > 4 GeV	π 1-2 GeV	π 2-4 GeV	π > 4 GeV
19	95.9 %	100 %	98.6 %	88.8 %	92.7 %	95.2 %
6	96.6 %	100 %	98.6	89.3 %	93 %	95.9 %

Efficiency of candidate tracks extracted with respect to truth. Tests done for muons and pions tracks. Low p_t shows low efficiency, issue in investigation. Candidate tracks and corresponding hits range between 550-950 and 10-13 respectively, leading to a processing time ranging from 3 to 6 μs. These results are related to the 8 outer-most layers of the barrel region only.

ALGORITHM PERFORMANCE AND TIME