CMS REAL-TIME STREAMING

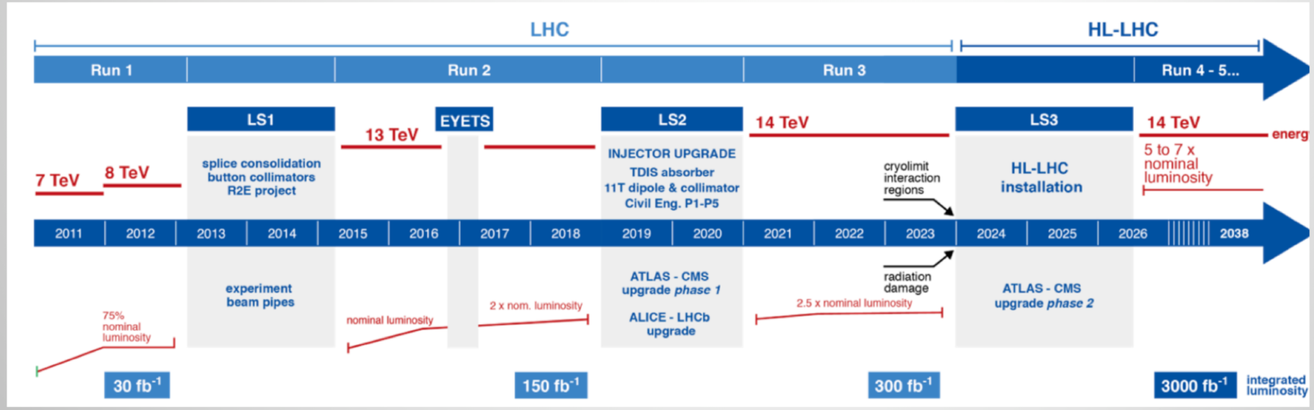
L1 MATCHING INFERENCE ENGINE PROTOTYPE FOR THE PHASE-2 UPGRADE

Micron-Openlab Project

Emilio Meschi – CERN EP/CMD

SUMMARY

- CMS Phase-II Upgrades: Trigger/DAQ
- Phase-2 L1 ML opportunities
- Streaming Inference @ L1
- Preliminary Plan of Work and Opportunities with Real Data

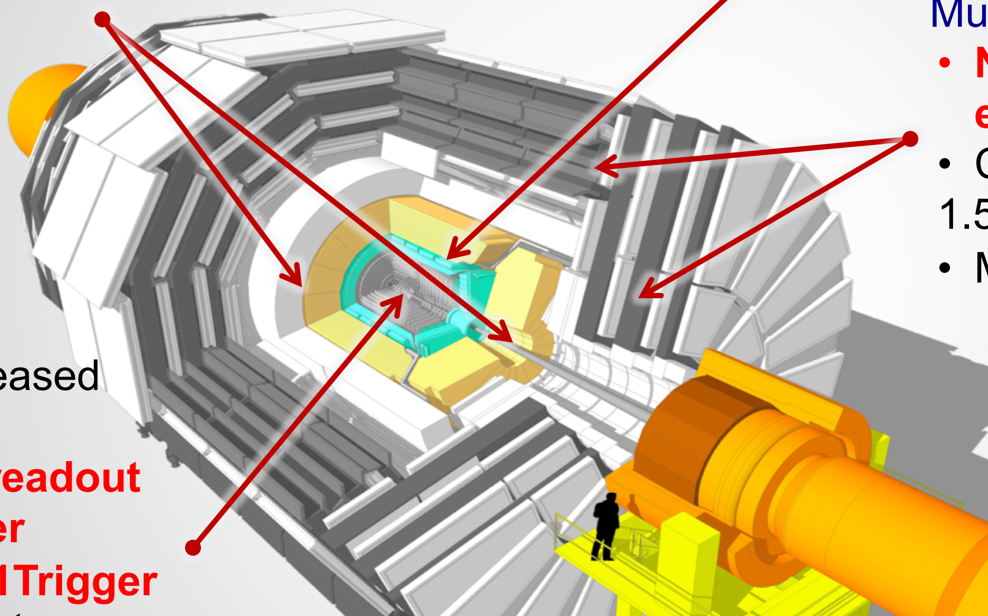


CMS Upgrades for Phase-II

New Endcap Calorimeters
- Rad. tolerant - **extreme transverse and longitudinal segmentation - intrinsic precise timing capability**

Barrel EM&HAD calorimeter
- **New FE/BE electronics**
- **Full granularity to L1 Trigger**

Muon systems
- **New DT & CSC FE/BE electronics**
- Complete RPC coverage 1.5 < η < 2.4
- Muon tagging 2.4 < η < 3

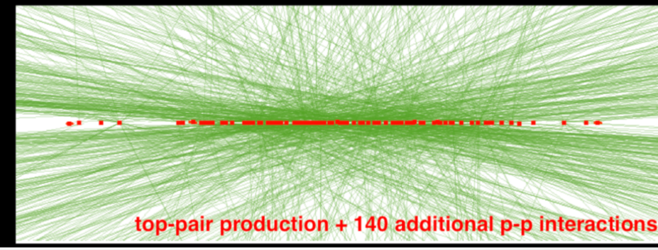


New Tracker
- Rad. tolerant - increased granularity - lighter
- **40 MHz selective readout (Pt≥2 GeV) in Outer Tracker for Level-1Trigger**
- Extended coverage to η ≃ 3.8

Trigger/HLT/DAQ
- **Track information in Trigger (hardware)**
- **Trigger latency 12.5 μs L1 output rate 750 kHz**
- HLT output 7.5 kHz

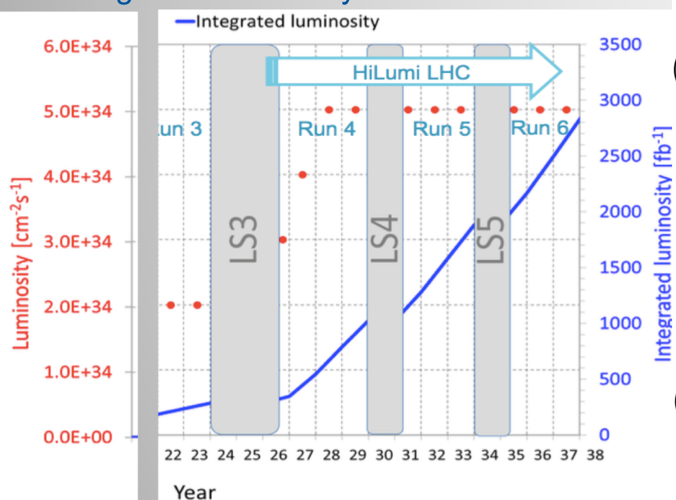
Machine and Experiment

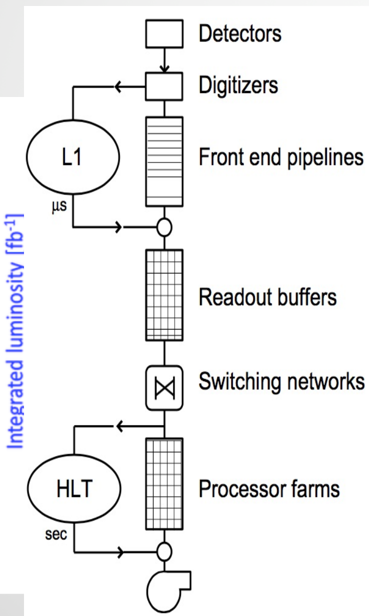
- Physics Program : require challenging performance (at least at the level of Phase I if not improved)
  –> New Tracking, New Calorimetry & New Muons
- Adapt to HL-LHC harsh conditions: **5.5-7e34 PU<200>**
  –> High granularity Tracker and Calorimeters
  very high particles multiplicity/Radiation damage
  –> Increased throughput for DAQ Scales up with PU
  –> Combined Track-Calo for the Trigger Selection



140 (200) baseline (ultimate) PU with ~1.2 PU/mm

250 (>300) fb^{-1}/y and 3 (4) ab^{-1} of total integrated luminosity





Improved triggering with full detector view:
Trigger decision includes calorimeters, muons & tracker (~5us latency)
→ L1Rate = 750 kHz
→ Latency = 12.5 us latency
→ Bandwidth: Phase II ~ 50 Tb/s (1.8 Tb/s in Phase I)

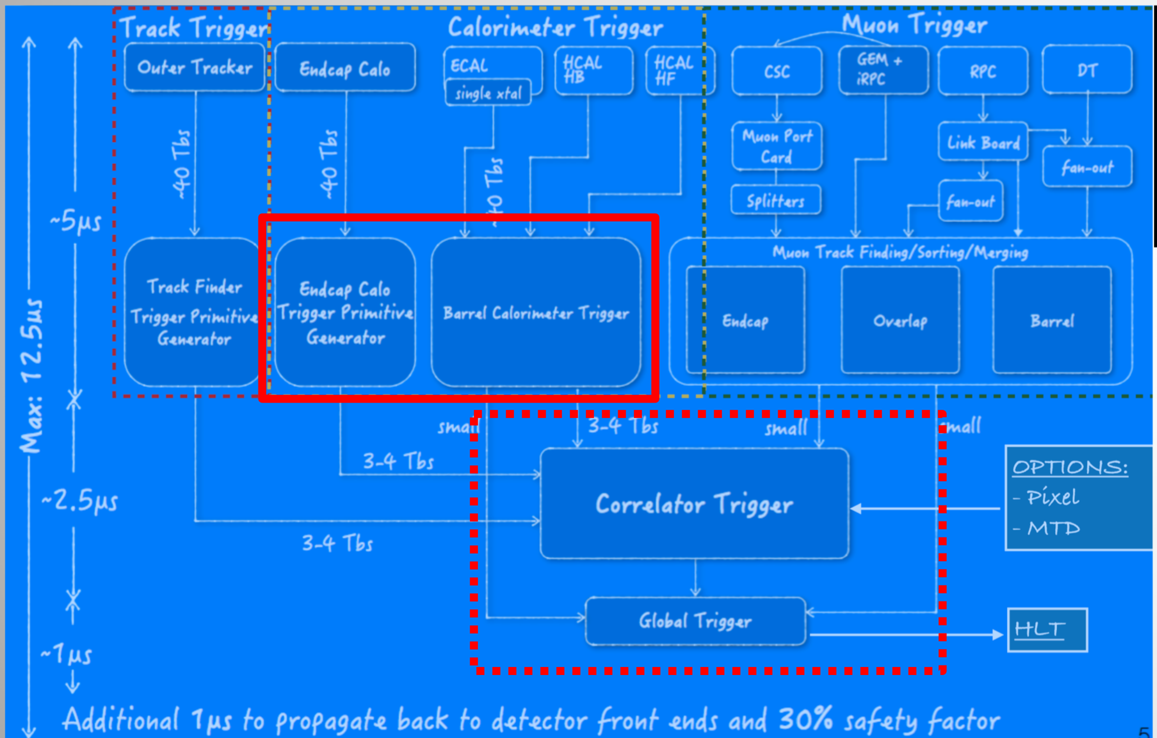
Event Size	7.4 MB
Event Network throughput	44 Tb/s
Event Network buffer (60 seconds)	333 TB
HLT accept rate	7.5 kHz
HLT computing power [c]	9.2 MHS06
Storage throughput	61 GB/s
Storage capacity needed (1 day)	5.3 PB

About 50k point-to-point bidirectional 10 Gbps optical links (lpGBT) on-to-off-detector with varying fractions devoted to trigger data

Overview

→ Achieving the best possible physics selectivity from the upgraded CMS detector → Organisation of the dataflow and overall infrastructure of the trigger revisited to benefit from the sub-detector upgrades

–> Building from Phase I experience: flexible architecture to adapt to LHC conditions & physics program, large computing power (FPGA) for highest bkg rejection and global detector view (high-speed links) to mitigate pile-up, calculate global quantities etc.



Possible Phase II Trigger architecture

→Sophisticated clustering algorithms deployed in the detector back-end electronics.

→ Building trigger objects @ Correlator level. Bringing HLT @ Level-1 (including higher-level objects: PF)

Offline capabilities = Increased selectivity achieved

Processing step	Time (μs)
Input data received by CT	5
Trigger objects received by GT	7.5
L1A received by TCDS	8.5
L1A received by front-ends	9.5

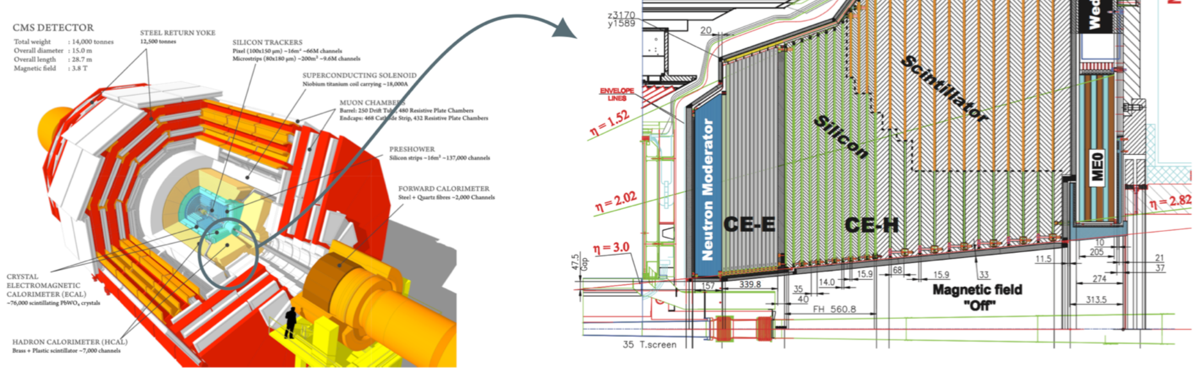
High Granularity Endcap Calorimeter

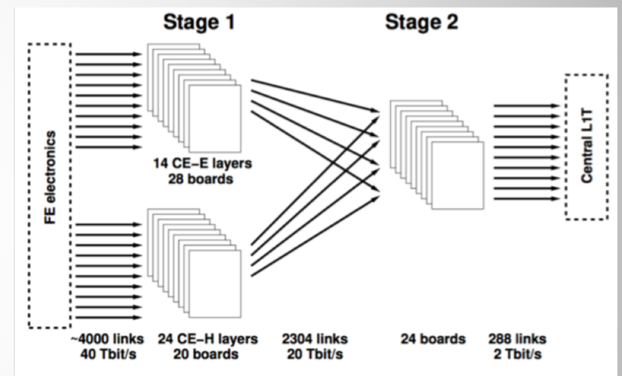
⇒ Electromagnetic calorimeter (**CE-E**): Si, Cu & CuW & Pb absorbers, 28 layers, 25 χ & $\sim 1.3\lambda$

⇒ Hadronic calorimeter (**CE-H**): Si & Scintillator, steel absorbers, 24 layers, $\sim 8.5\lambda$

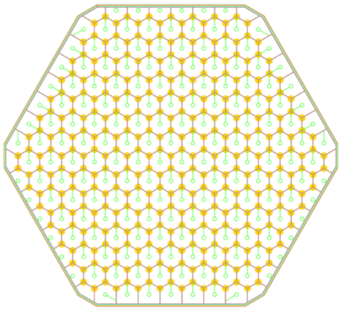
Key Parameters:

- HGCAL covers $1.5 < \eta < 3.0$
- $\sim 600\ \mathrm{m}^2$ of silicon sensors
- $\sim 500\ \mathrm{m}^2$ of scintillators
- 6M Si channels, 0.5 or 1 cm^2 cell size
- Intrinsic timing capabilities ($\sim$ 25ps resolution)

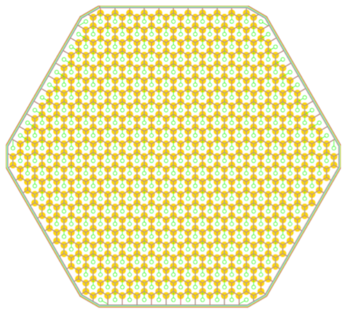




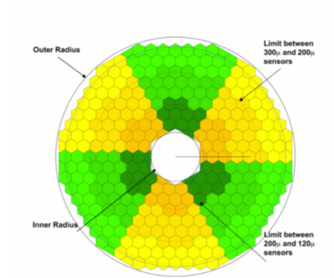
HGCal Sensors and Baseline Trigger Primitives



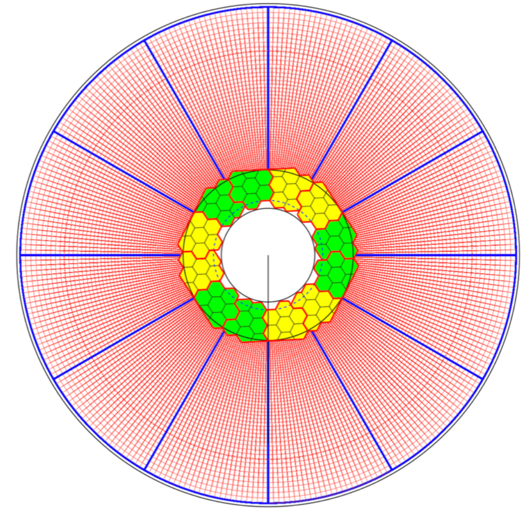
(a) Big Silicon sensors, 1.18 cm^2



(b) Small Silicon sensors, 0.52 cm^2



(c) Layout of a layer where only silicon sensors are present



(d) Layout of wafers and Scintillator tiles in a layer where both are present: the 22^{nd} layer of CE-H

Table 8.5: Baseline Endcap Calorimeter cluster definition.

Quantity	N bits	Comment
E_T	2×16	with and without PU subtraction
Endcap	1	
f_{EE}	13	E_T fraction in EE
f_{BH}	12	E_T fraction in BH
L_{max}	6	Max energy layer
η	11	Shower start
ϕ	11	Shower start
z	10	Shower start
N_{cells}	8	
Quality	12	
Extra flags	12	
Minimum total	128	

HGCAL Clusters = 128 bits

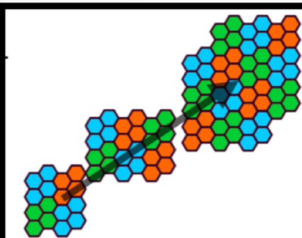


Table 8.3: Concept for the header data sent to the central L1T correlator per BX.

Quantities	Bits	Total bits
Total energy, BX number, number of clusters	16, 8, 8	32
Energy map 15 (η) ×72 (ϕ)	16	17 280
Total		17 312

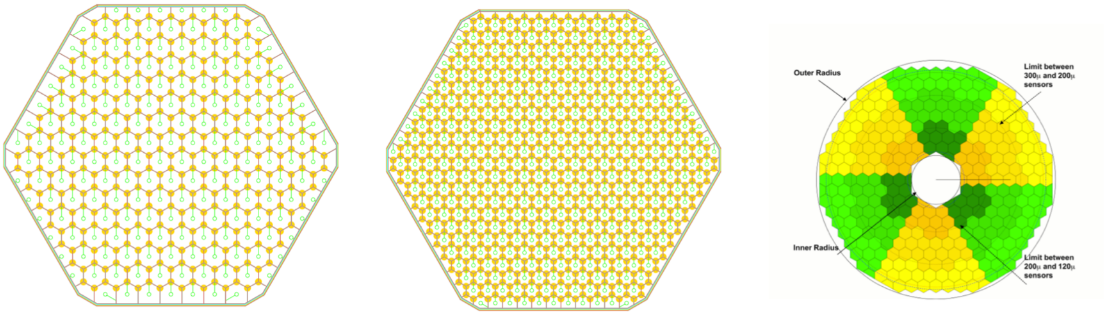
HGCAL Towers = 32 bits

WHAT KIND OF TRIGGER PRIMITIVES

Clusters: formed from 4 adjacent Trigger Cells (2x2cm2) built at the Front-End Level. Current approach in HGCAL TDR: 2D clustering/Layer & 3D clustering (combining 2D clusters)

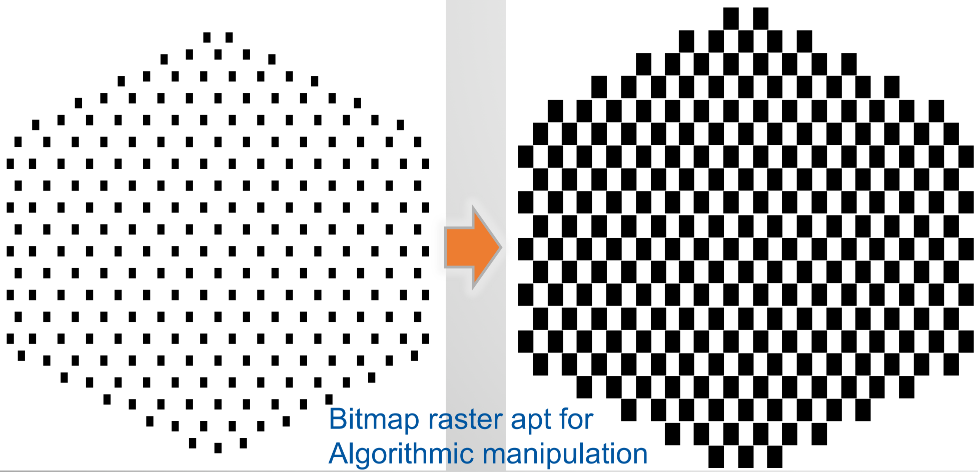
Towers: Energy maps (15 eta x 72 phi) *Not physical towers!*

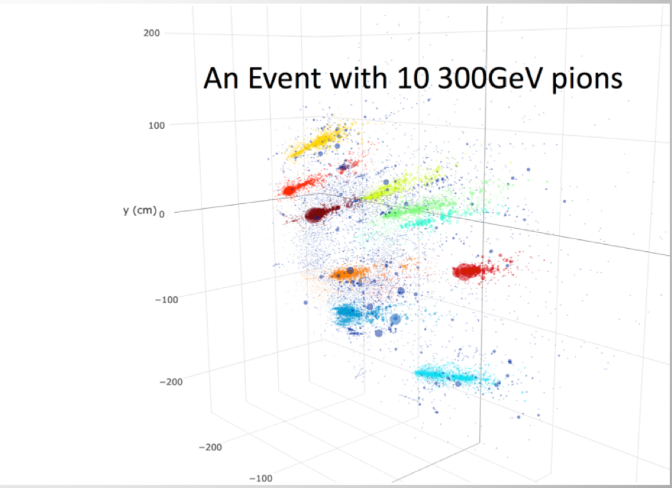
HGCal Trigger Primitives with a CNN



(a) Big Silicon sensors, 1.18 cm^2 (b) Small Silicon sensors, 0.52 cm^2 (c) Layout of a layer where only silicon sensors are present

- Particles going through HGCAL will produce showers of secondary particles that will be detected as 4D images (5D including timing)
- Convolutional Neural Networks (CNNs) can be applied to HGCAL Trigger Primitives to
  - Classify showers
  - Give best estimate of energy
  - Reconstruct shower axis
- First Challenge: hexagonal symmetry (in general CNN implementations assume a square lattice)





CNN for HGCal

First attempts: EM shower classification

Fixed size x-y grid: for each layer

- Adjust grid pace such that at most one trigger cell falls in each grid cell
- Center grid around best-matching 2d-cluster

This is obviously unrealistic

- 20x20 matrix of tc_energy for each of the first 29 layers (EE)

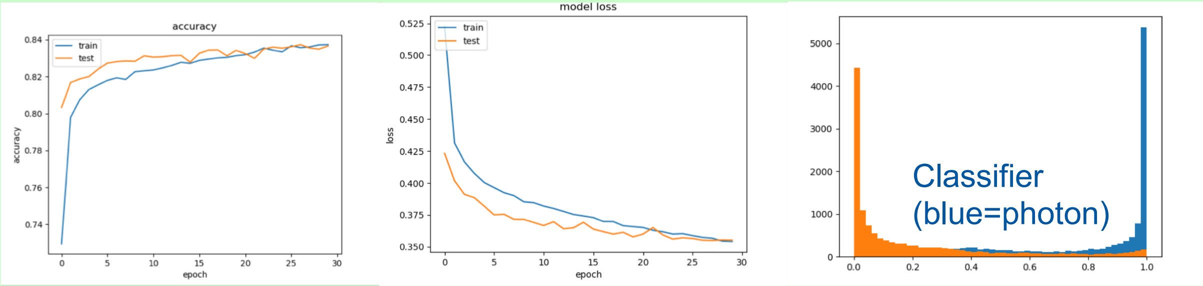
Simple 3d- CNN fed with 20x20x29x1deep

Same technique for classification and energy estimation

Can be used for position offset ?

Vertex estimator ?

Non-converting photon vs electron



Plan of Work

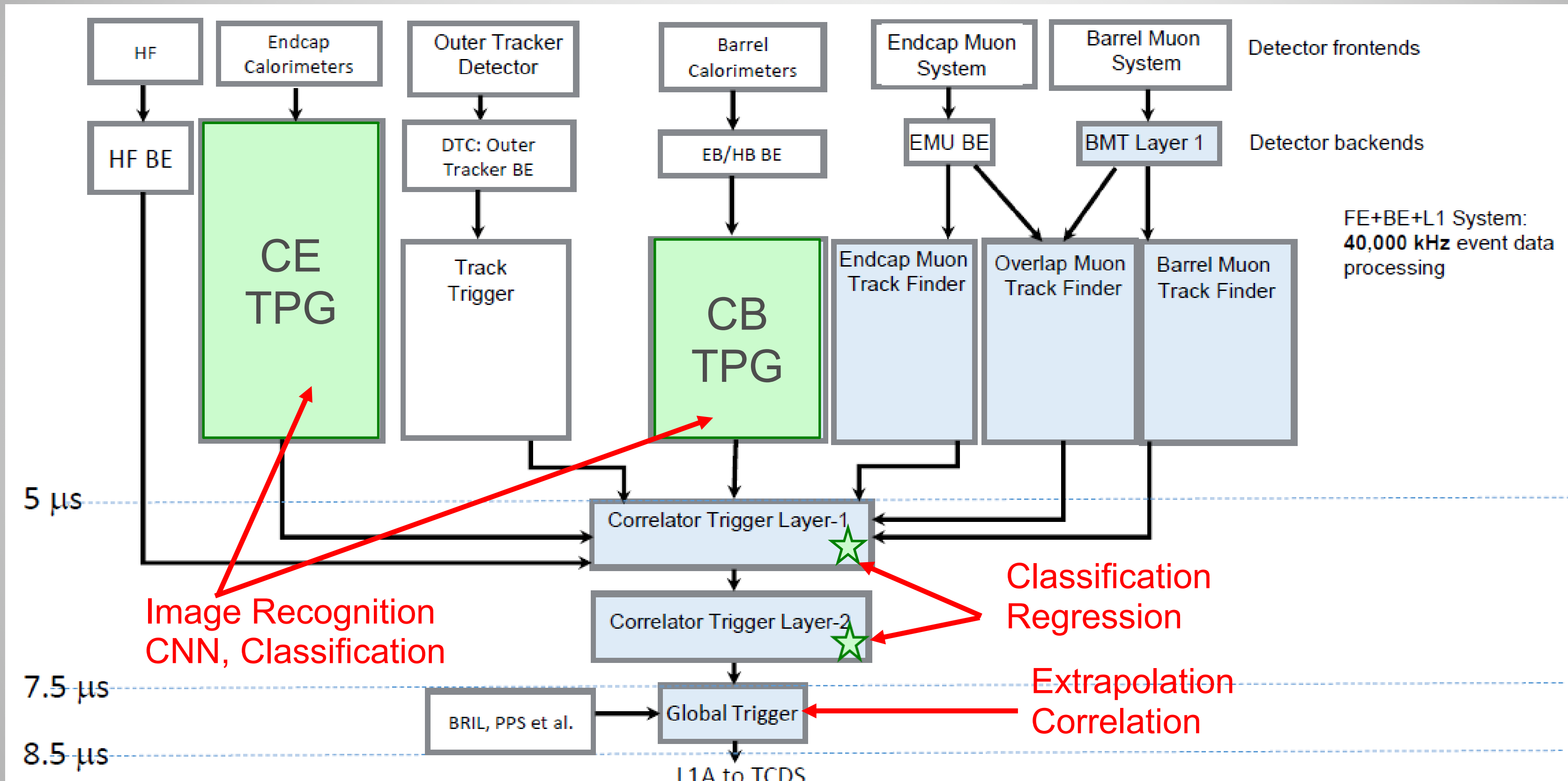
1. Use unsupervised clustering algorithms, optimized to arrange the recorded particle hits into candidate particle showers
2. Use supervised algorithms to accomplish tasks such as particle identification and energy measurement on the clustered showers

Denoising techniques could also be investigated to mitigate the effect of pileup

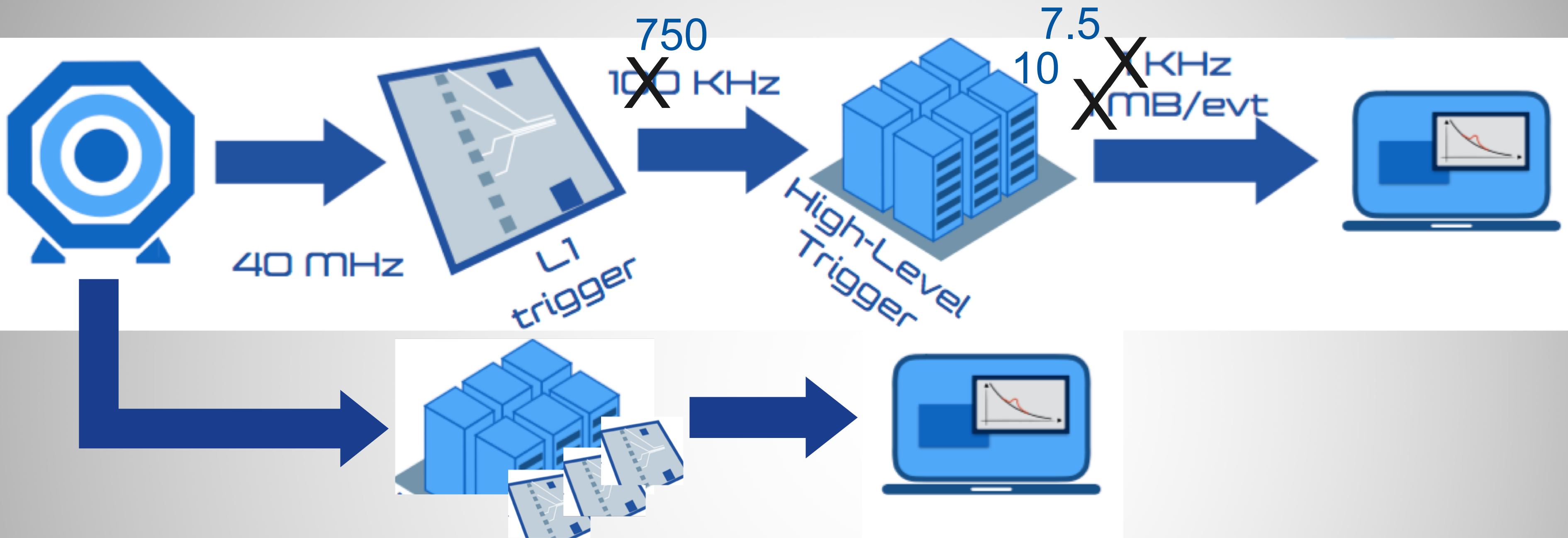
3. Optimize the network architecture not only with respect to accuracy, but also to **execution time**

The final target is to implement the CNN-based particle reconstruction both in the online and offline phases of the data reconstruction, on standard **CPU** as well as on dedicated hardware architectures, such as the **FPGA**s for the **L1 trigger** or **GPU**s in the **HLT**

Summary: areas of interest



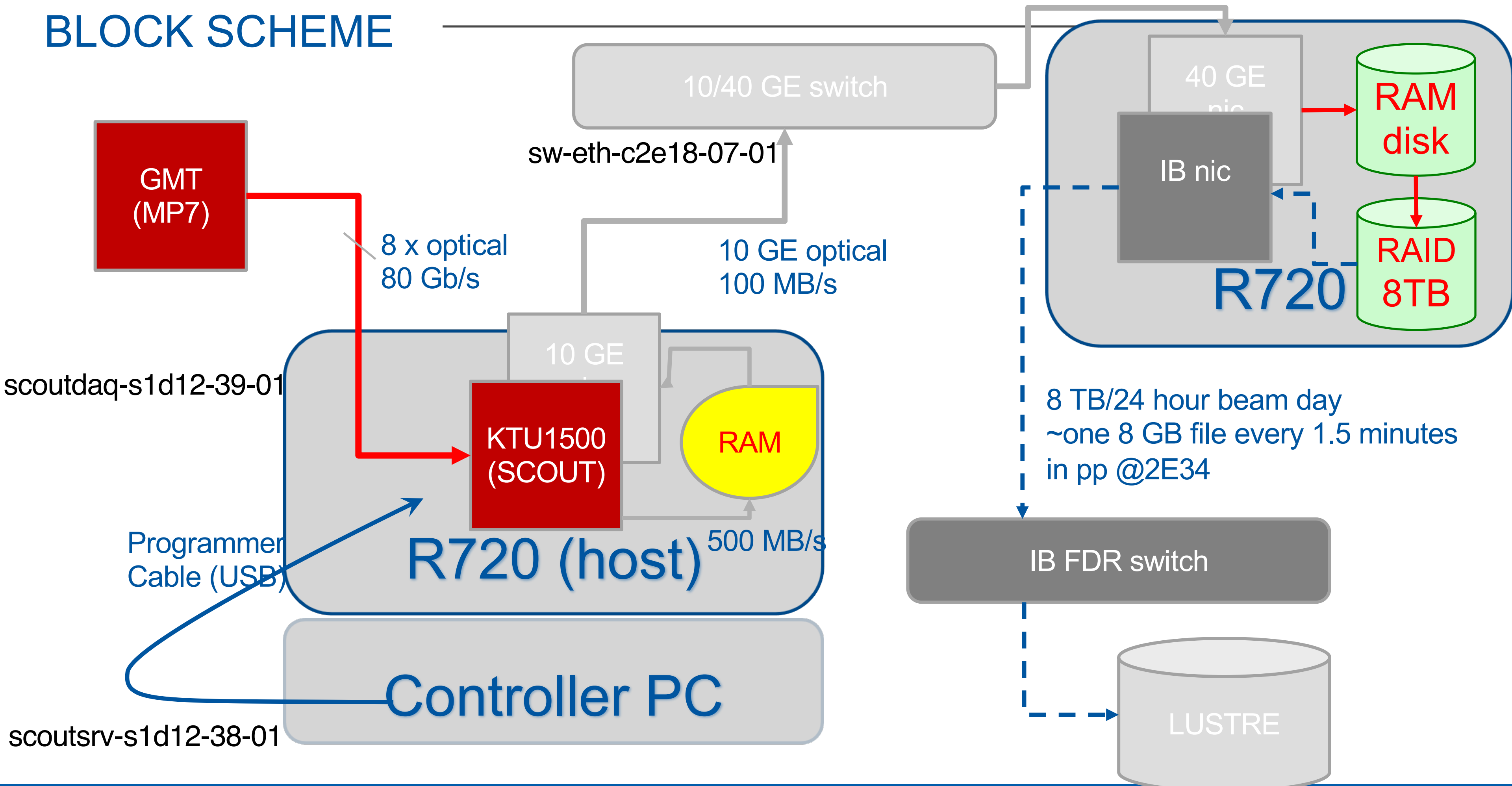
40 MHz Scouting for Phase-2



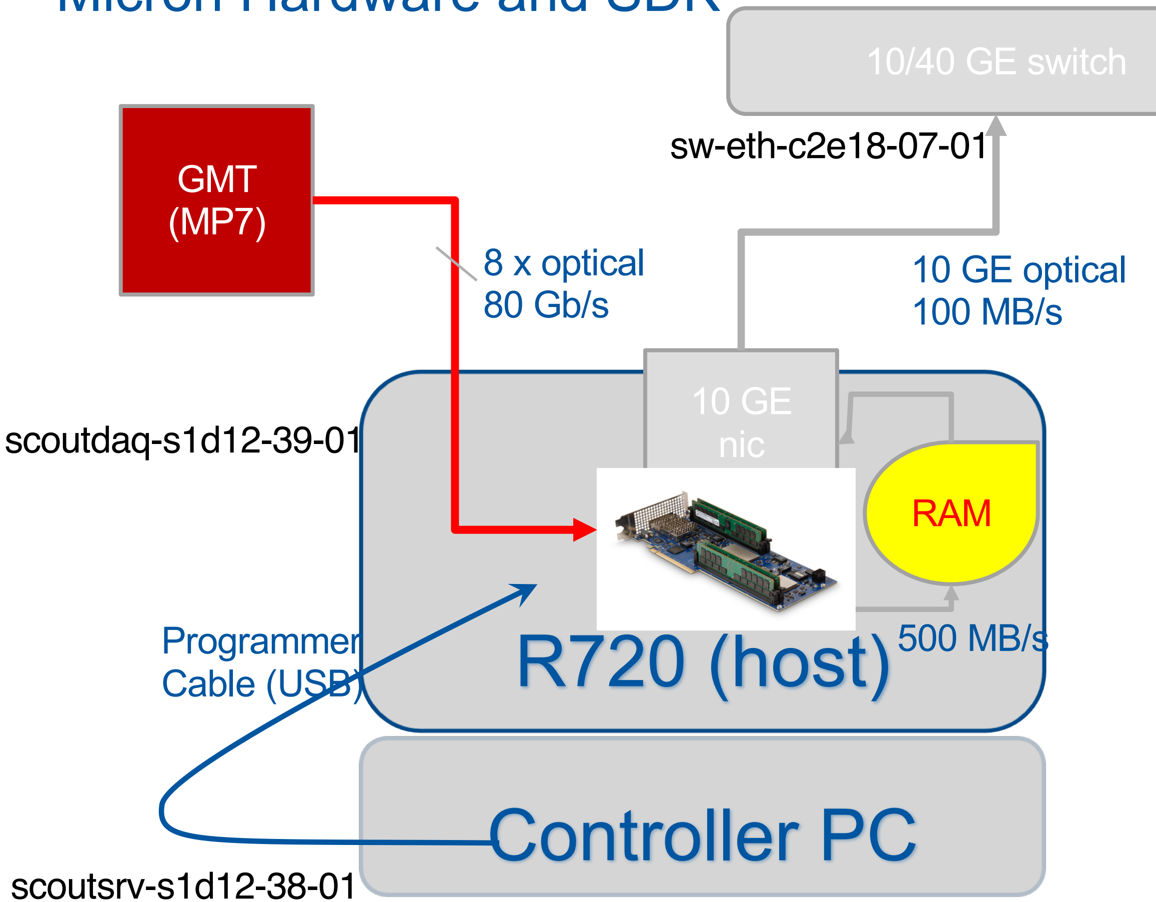
40 MHz Scouting
FPGAs
GPU
HPC low-latency
Interconnects

Real Time indexing of features
Analysis by query

uGMT SCOUTING BLOCK SCHEME



POSSIBLE APPLICATION OF Micron Hardware and SDK



Run-3 L1 SCOUTING

- Implement extrapolation inference using pico SDK
- Extend implementation to other estimators
- Horizon: run on real data in Run 3 (2021)

Preliminary plan of work (phases 1 and 2)

Phase 1

- Identify and study in detail suitable cases that lend themselves to a ML approach
  - Use the configuration for the Phase-2 upgrade of the CMS experiment
  - Team with groups working on ML approach for HGCAL, Barrel Calorimeter, Correlator (PF)
  - Goal: identify at least one case study for functionality tests
- Generate and use simulated data to test different ML approaches for the identification of physical features in data from a particular subsystem of CMS
  - Collaborate to the design and training of different types of NN compatible with the Micron co-processor
  - Goal: prepare test data for benchmark
- At the same time:
  - Familiarize with pico SDK
  - Simple tests with hardware
- Benchmark: measure performance of the trained NN, both in terms of physical accuracy and execution time, using Micron hardware
  - Goal: attain desired physics performance while minimizing execution time
  - Not yet with specific L1 input format/interface

Phase 2

- Setup test with point-to-point links (specific FW)
  - Possibly using “scouting” test case
- Optimization of performance for the main use case from Phase 1
  - Identify and reject ”redundant” data, if any
  - Identify optimisations of the network in view of meeting the latency requirements
- Goal: demonstrate performance in realistic conditions both for physics and latency