The Level 1 Scouting system of the CMS experiment

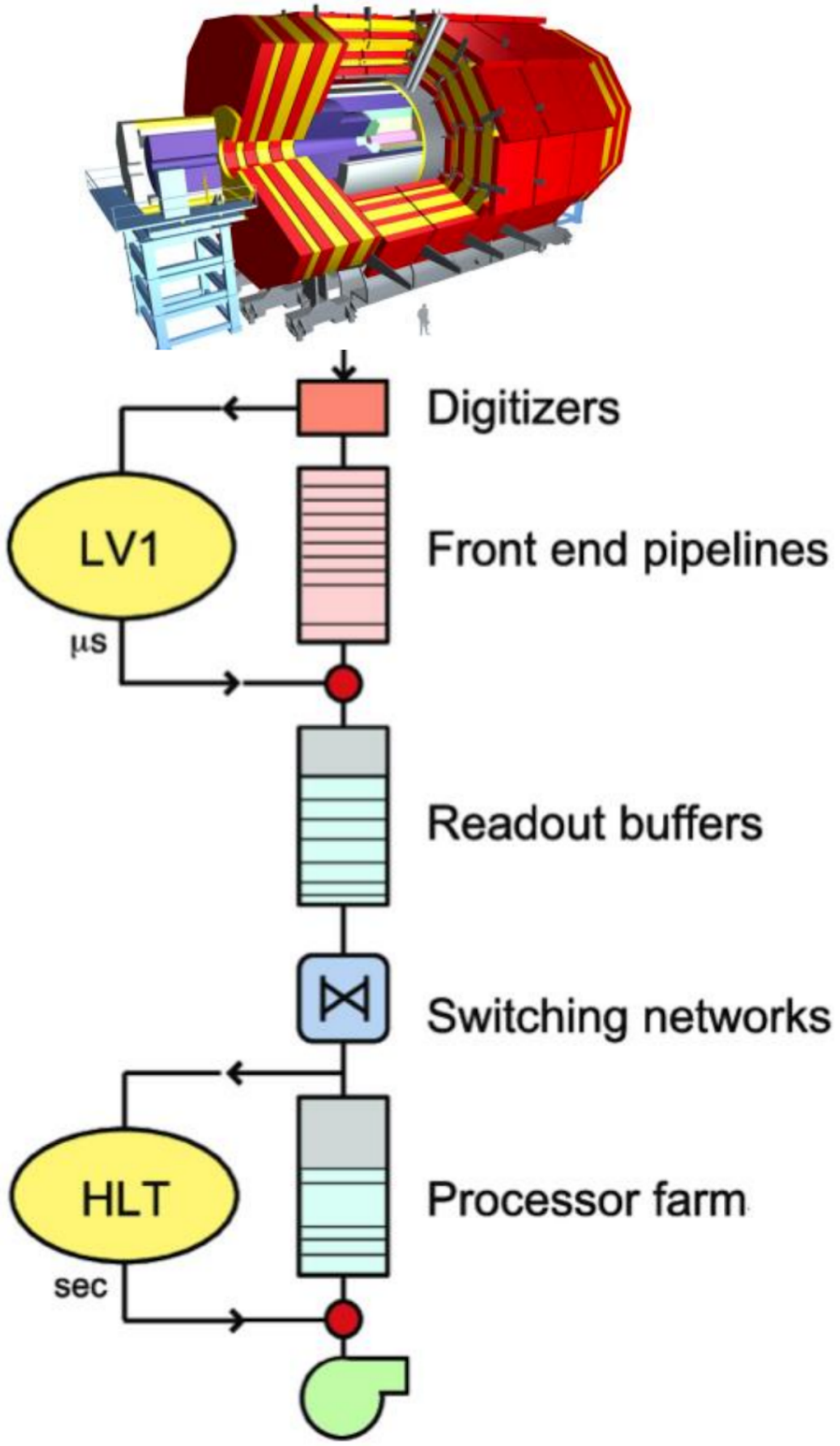
T. James *(CERN)*, on behalf of the CMS L1 Scouting team

21st International Workshop on Advanced Computing and Analysis Techniques in Physics Research

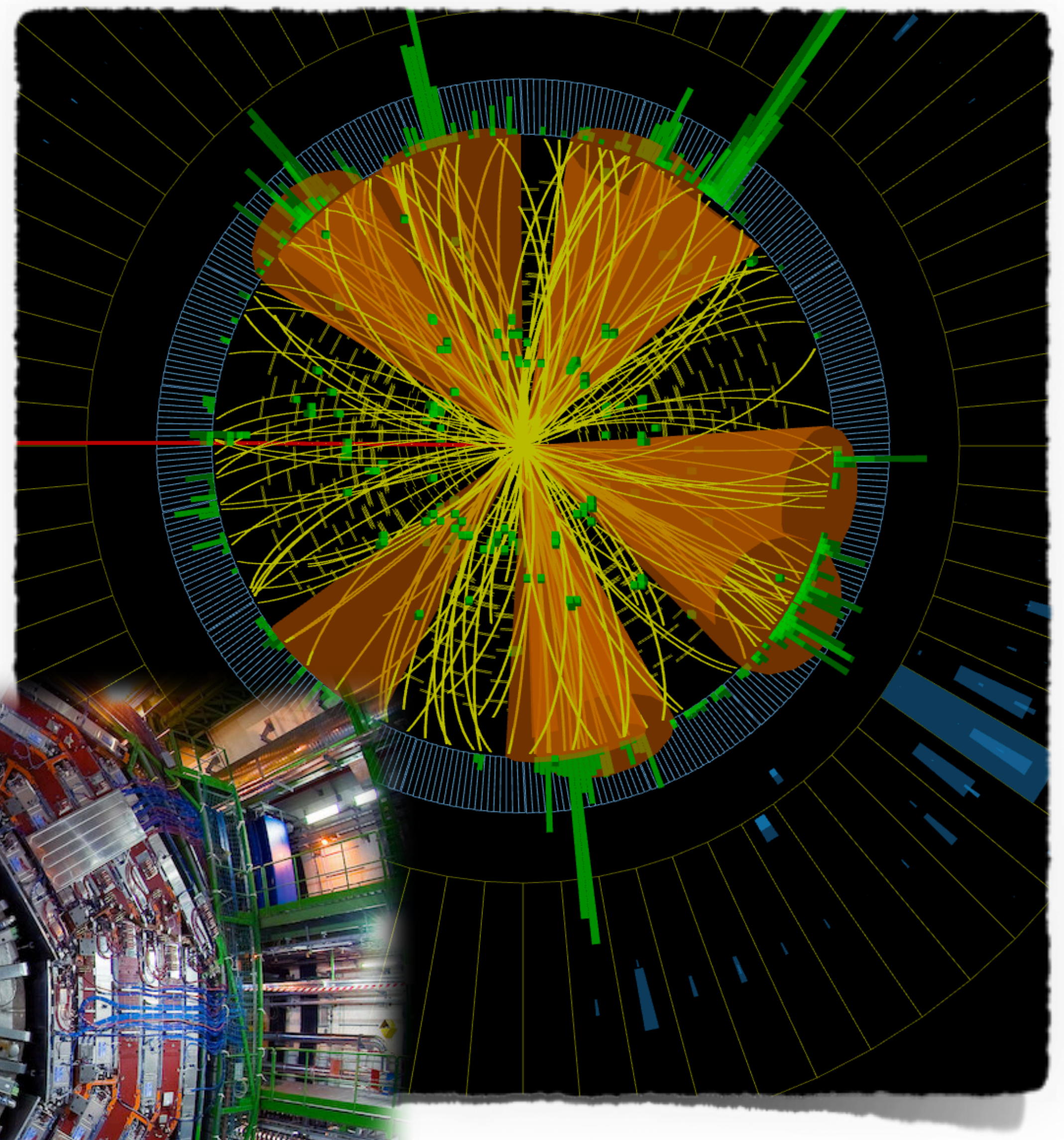
27th Oct 2022

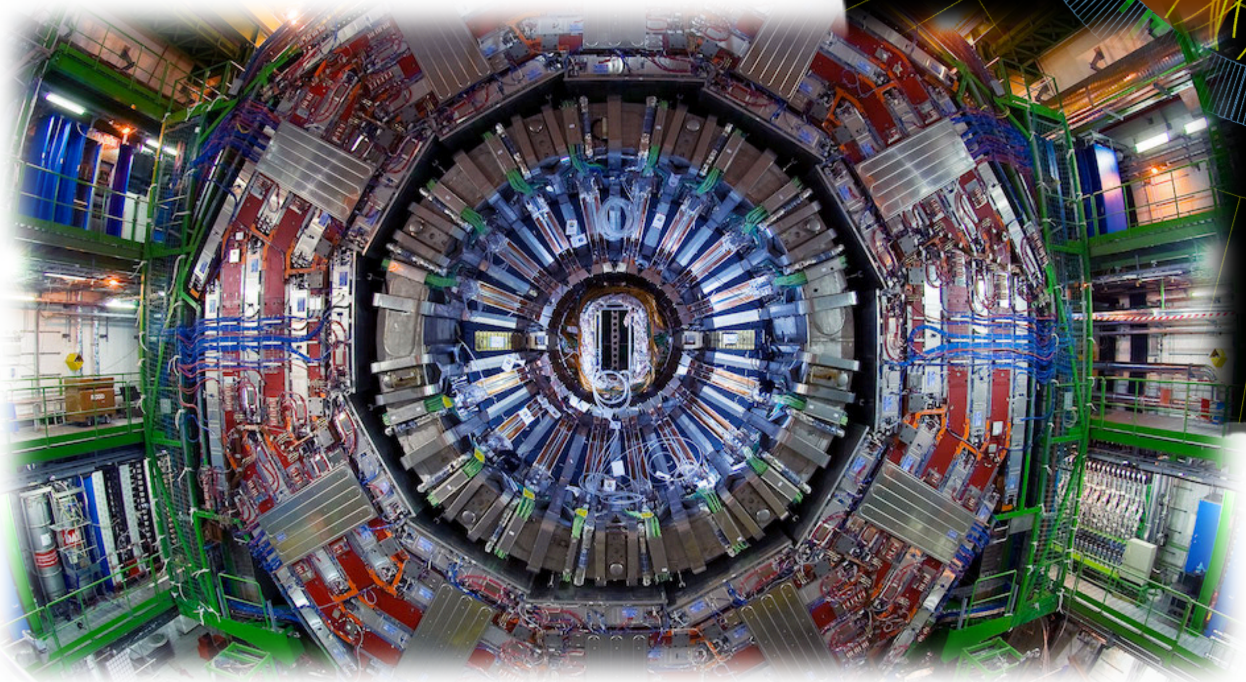
Two stage trigger system

	Phase 1	Phase 2 (High Lumi)
Peak pileup	60	200
BX rate	40 MHz	40 MHz
L1 rate	100 kHz	750 kHz
L1 latency	< 4 μs	< 12 μs
HLT rate	2 kHz	7.5 kHz



40 MHz Scouting: What does L1 accept miss?



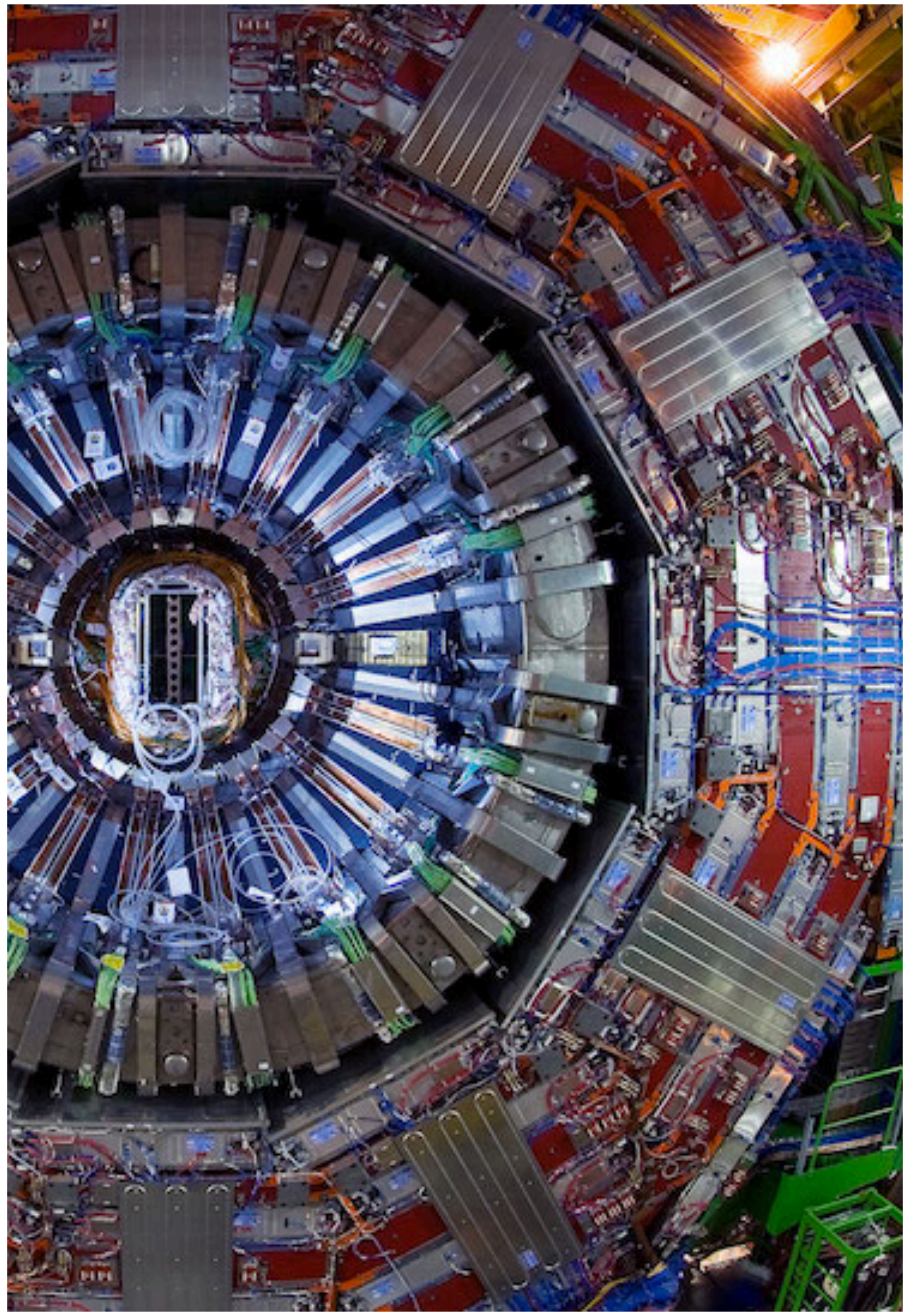


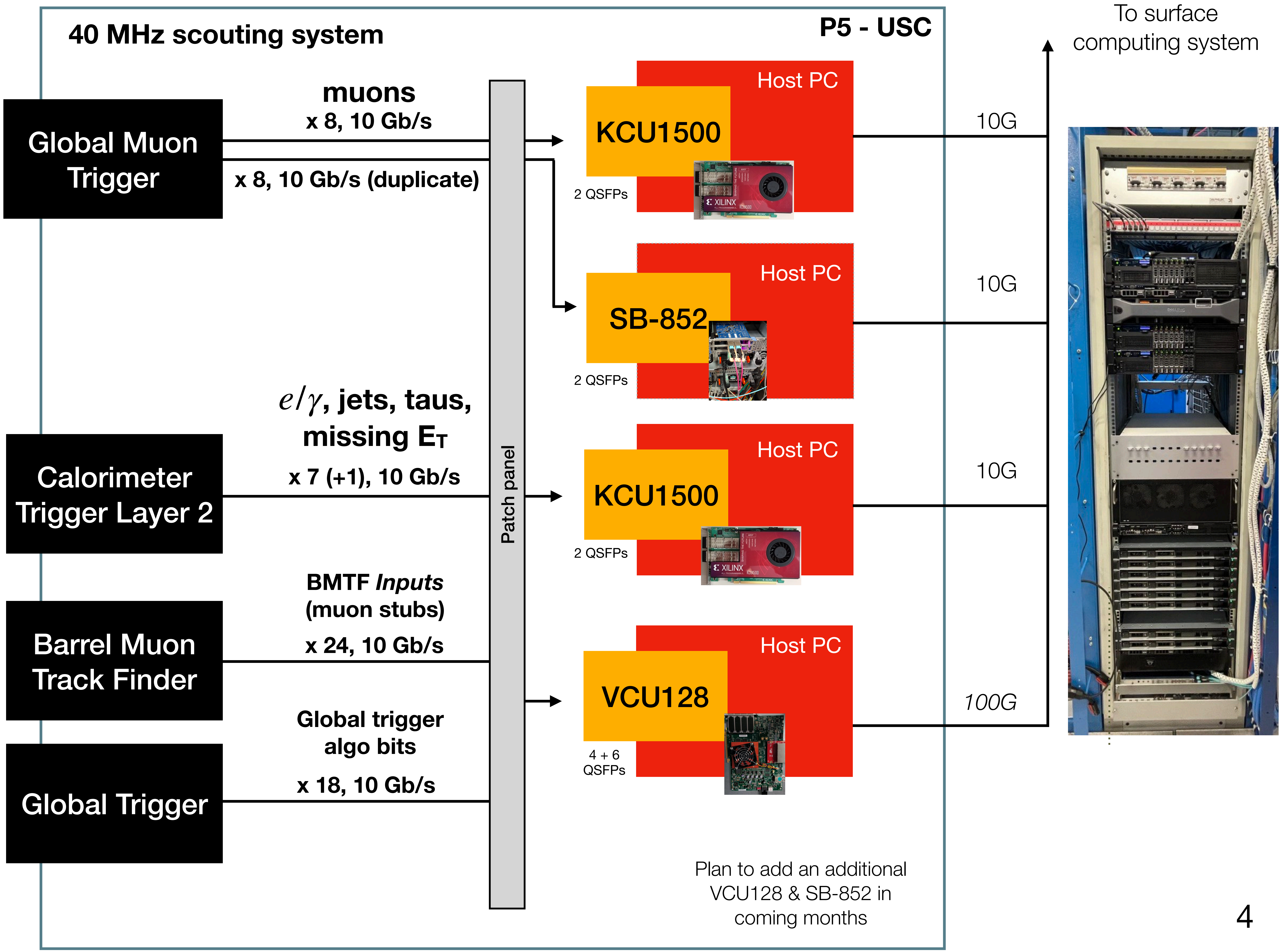
- **Can we acquire L1 trigger data at full bunch crossing rate**
  - subset of detector information, limited resolution
- **Allows for analysis of certain topologies at full rate**
  - semi real-time analysis and/or storing of tiny event record
- Demonstrated for first time at end of 2018

- **Diagnostic and monitoring capabilities**
  - BX-to-BX correlations always available
  - Independent per-bunch lumi measurement

- **Physics cases**
  - Heavy Stable Charged particles over multiple BX
  - Channels where available cuts give low efficiency at attributed rate budget
  - Any long-lived leptonic decays e.g soft displaced muons

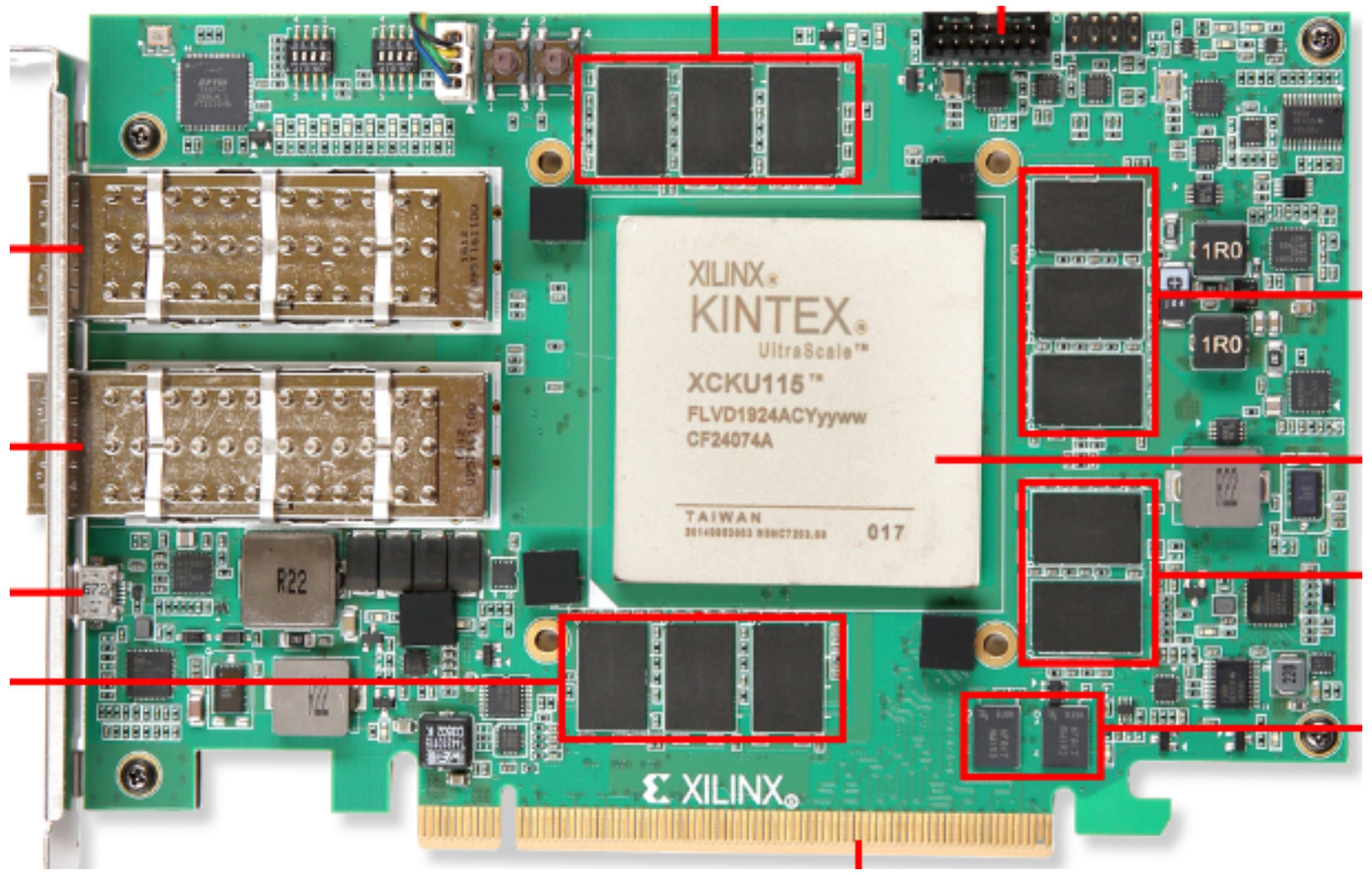
L1 Scouting Demonstrator Run 3





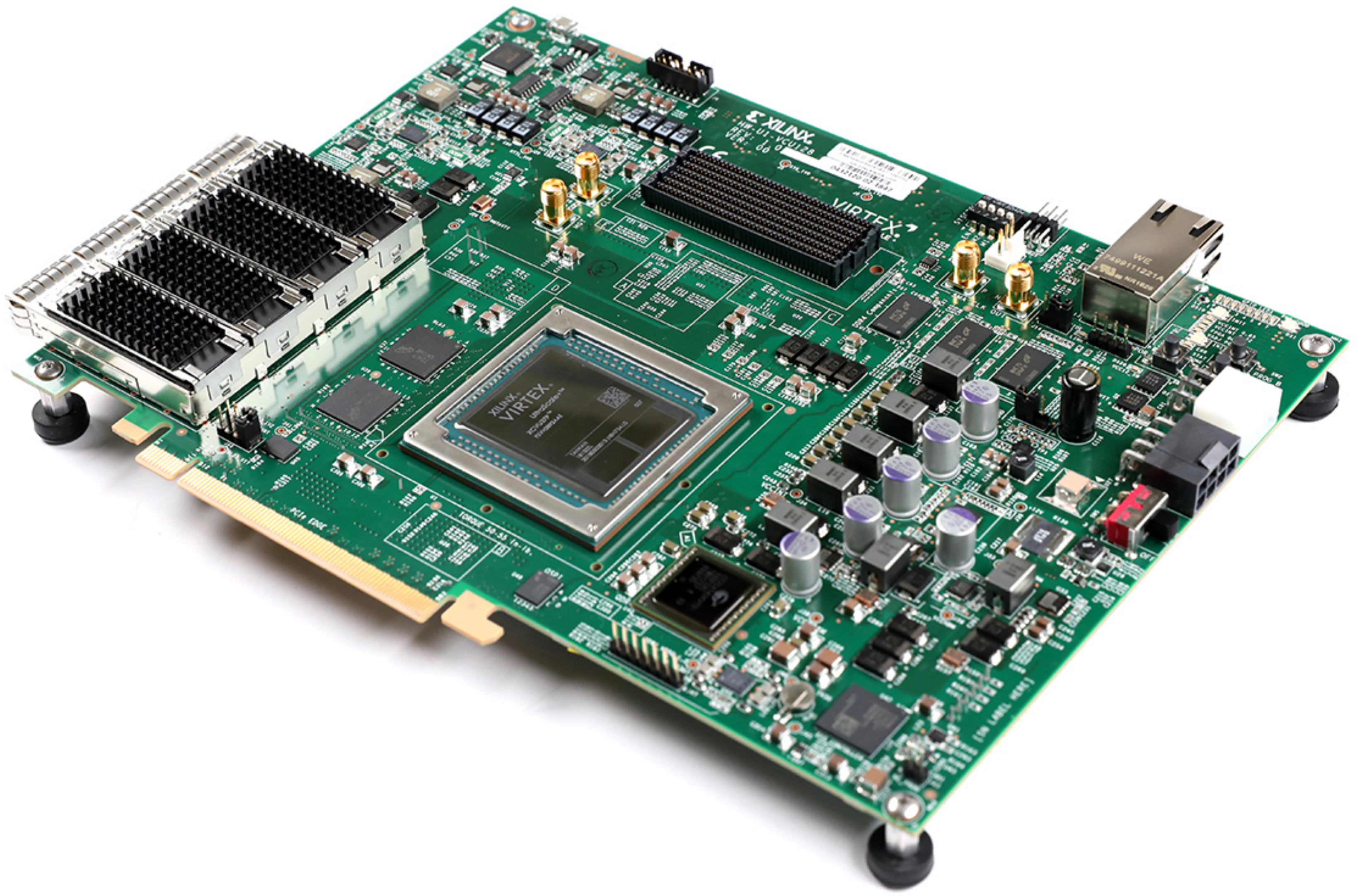
Hardware: rule of three

Xilinx KCU1500



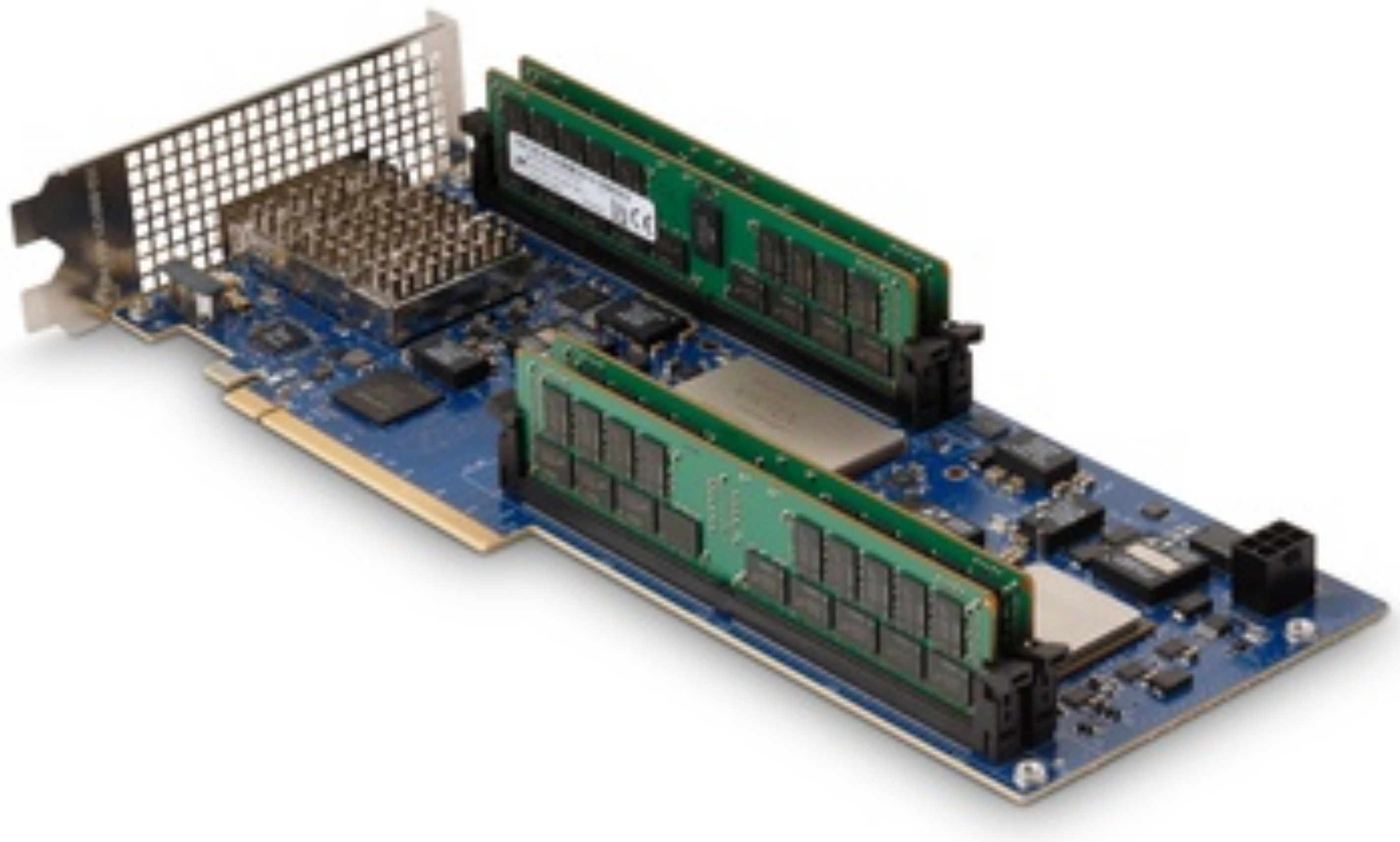
- PCIe Gen3x8 x2
- KU115
- 2x QSFP

Xilinx VCU128



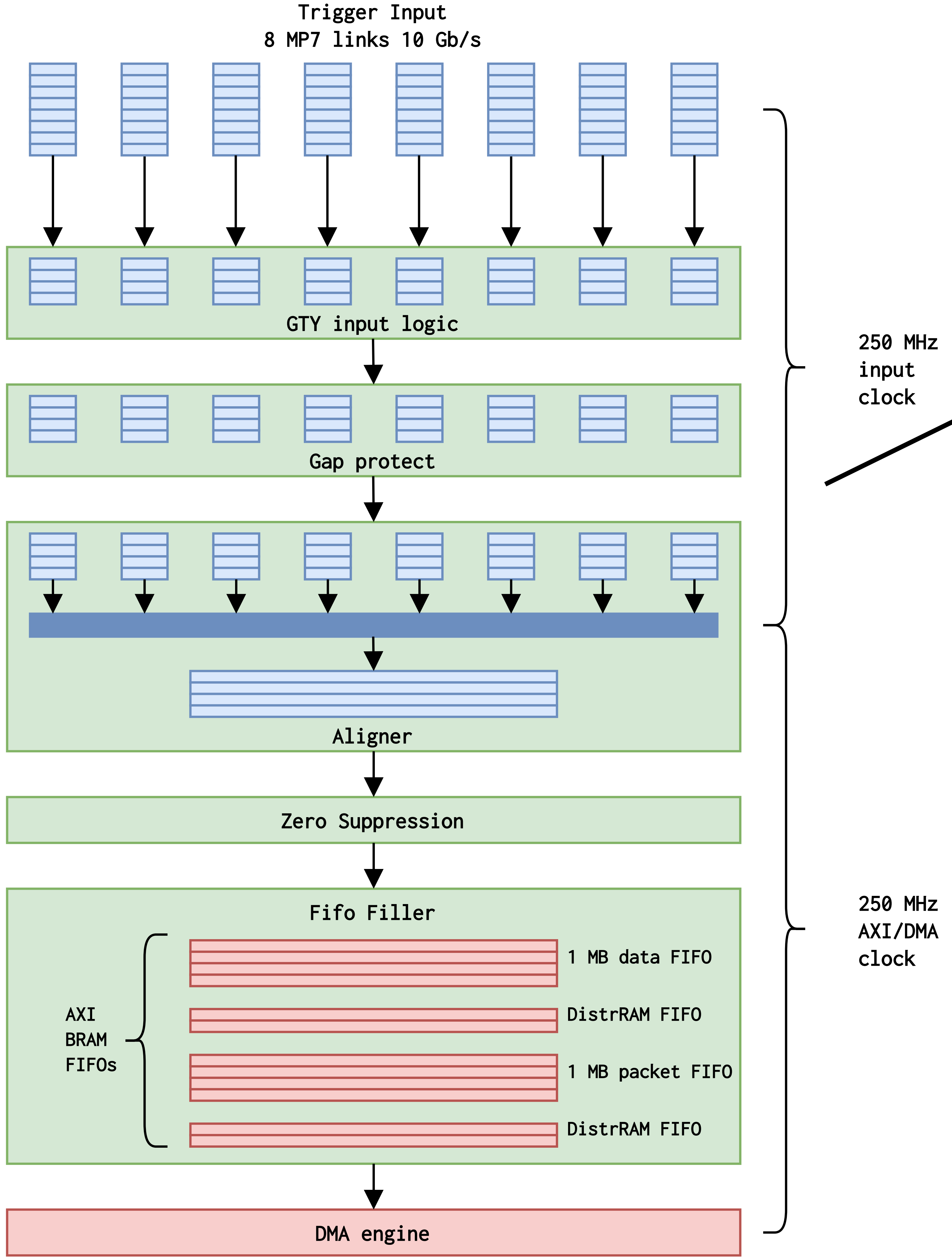
- PCIe Gen3x16 or PCIe Gen4x8
- VU37P (w/ 8GB HBM)
- (4 + 6 w/mezzanine) QSFP

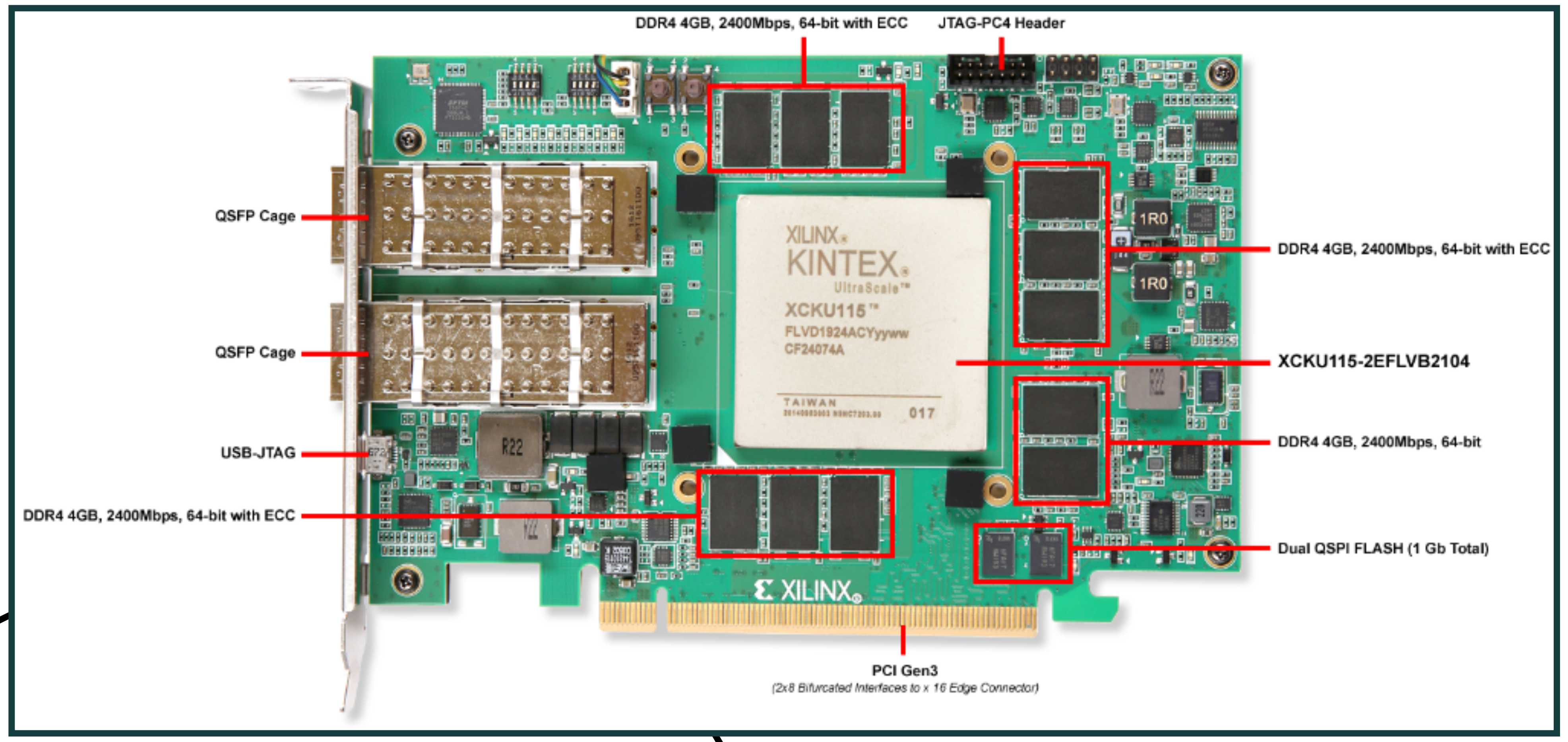
Micron SB-852

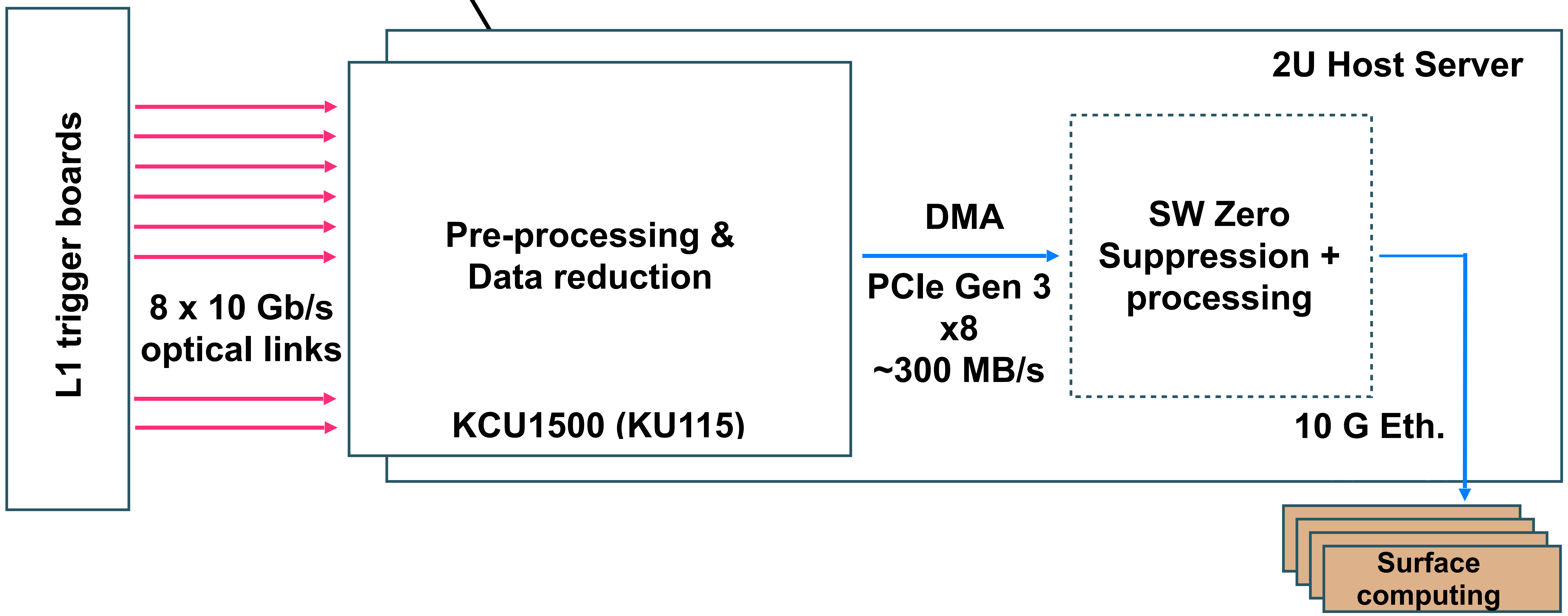


- PCIe Gen3x16
- VU9P
- 2x QSFP
- 64 GB DDR4

CMS 40 MHz Scouting with Xilinx KCU1500



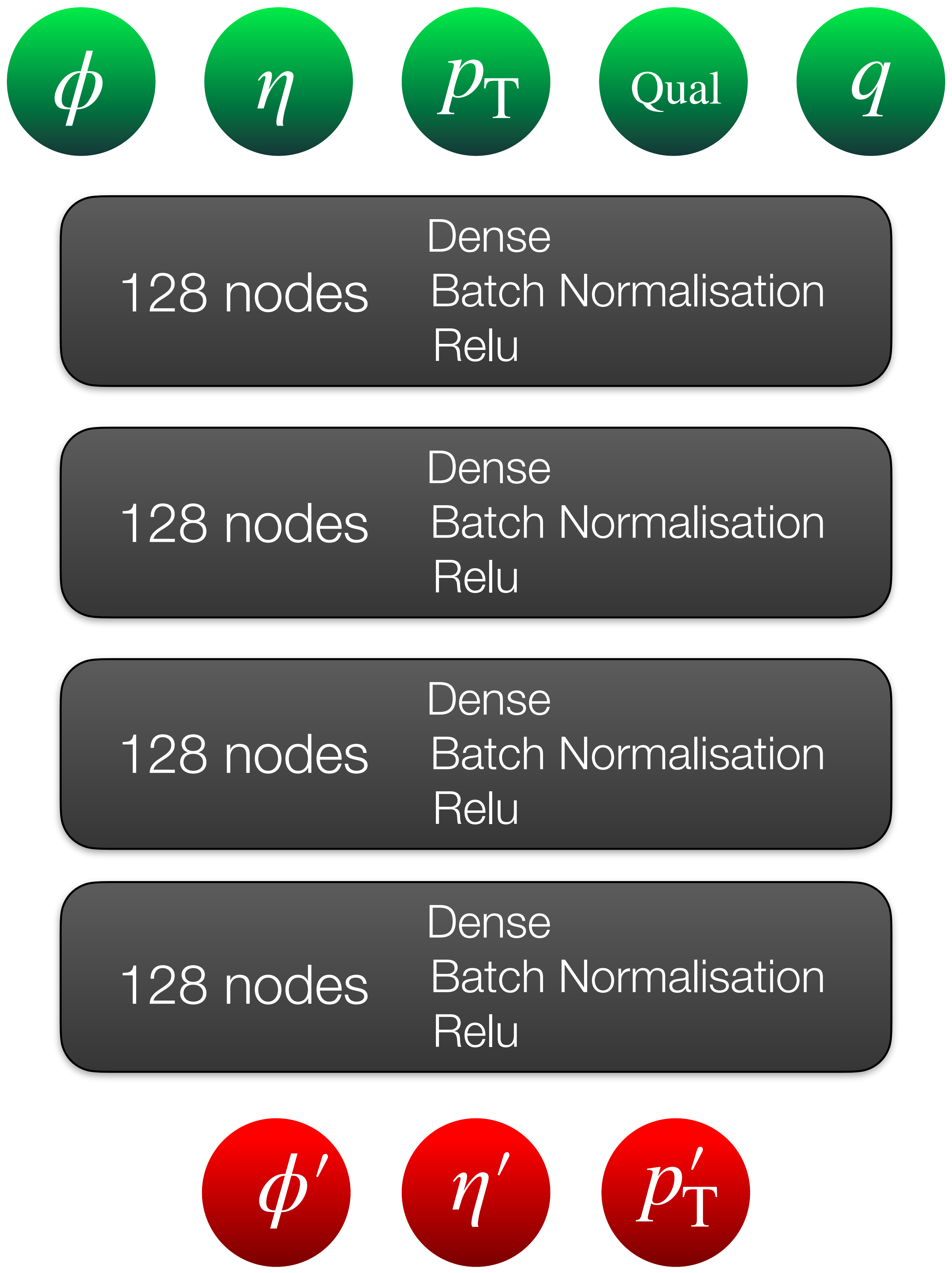




Why ML for scouting?

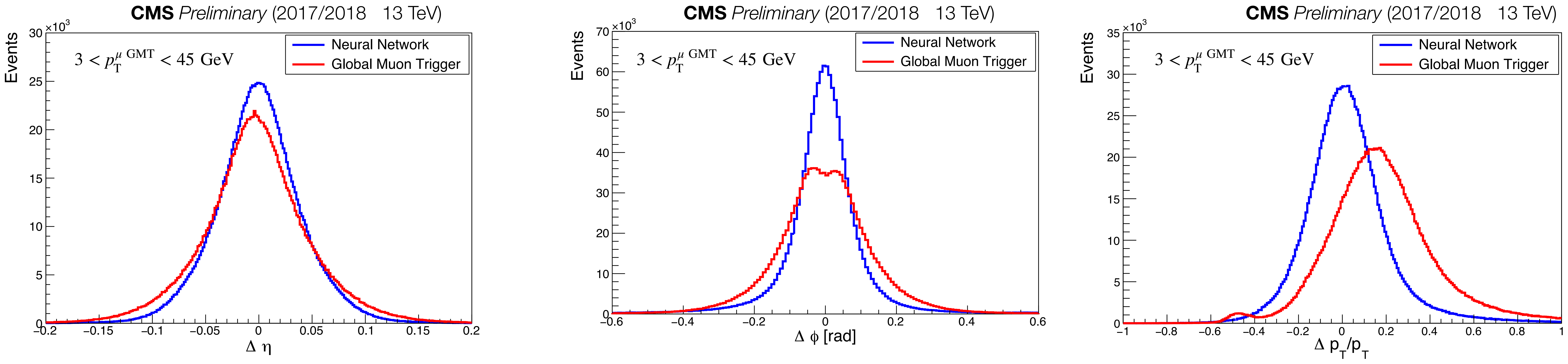
- Trigger objects calibrated for a given efficiency at a threshold
  - For triggering, not physics analysis
- **Use the offline objects as target** to re-calibrate the parameters of the trigger level objects
- We have full offline reco & trigger objects for Zero Bias and Triggered events

- **Inputs** - L1 objects e.g μGMT muons:
- **Target** - Offline fully reconstructed objects
- Use of classical ***fully connected*** neural networks to 'recalibrate' L1 information to improve their utility for an online analysis



μGMT re-calibration with Neural Network

- NN shown to universally improve precision of ϕ, η and p_T, able to achieve ~2x improvement in track parameter precision for some interesting areas of phase-space



- Trained with *Zero-bias* dataset 2017, 2018, re-run with Run 3 trigger emulation for up-to-date muon trigger algorithms

- $\Delta\eta$, $\Delta\varphi$, Δp_T is the difference between the prediction (or μGMT *extrapolated*) values, and the offline muon tracks for matched muons ($\Delta R<0.1$ at 2nd muon station)

Micron Deep Learning Accelerator (MDLA)

- Proprietary **Inference Engine firmware**, scalable and programmable solution to deep learning inference
- Offers ~Tera MAC (multiply-accumulate operations) /s
- Board configured with MDLA *Compiler*

Keras

TensorFlow

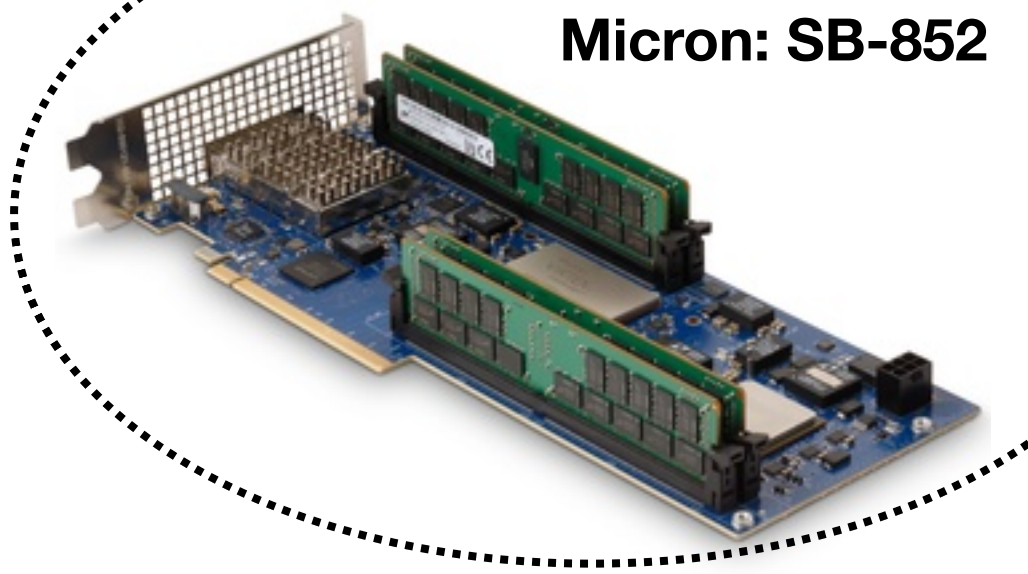
Caffe2

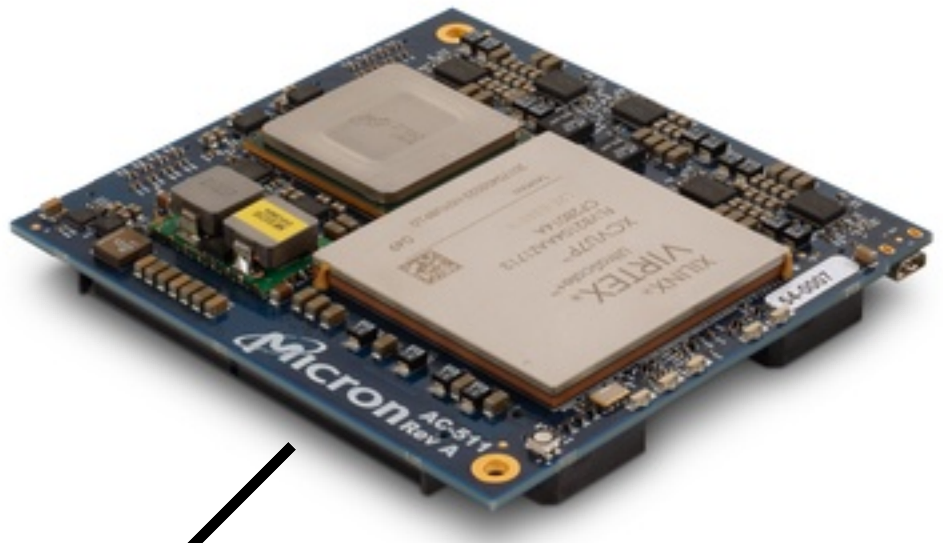
PyTorch

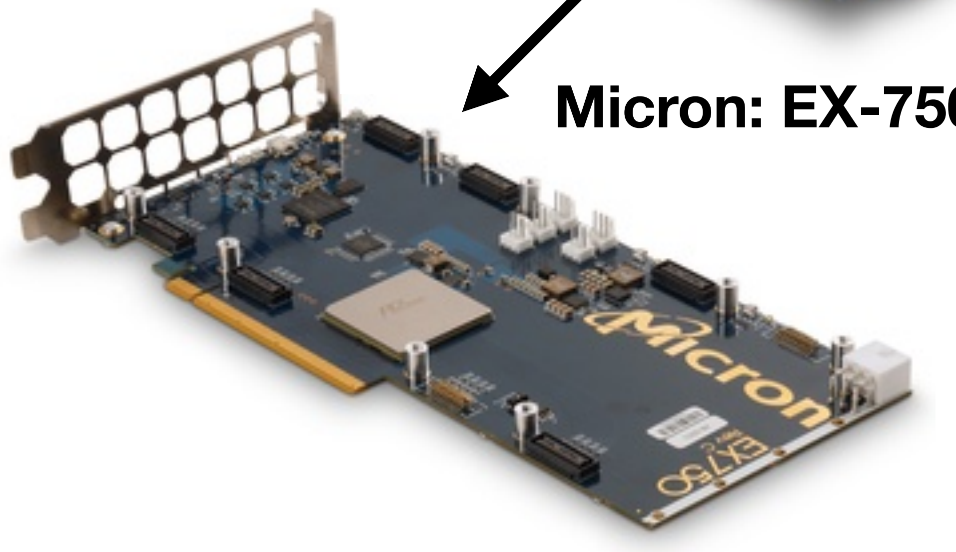
ONNX

Micron Deep Learning Complier

Micron: SB-852



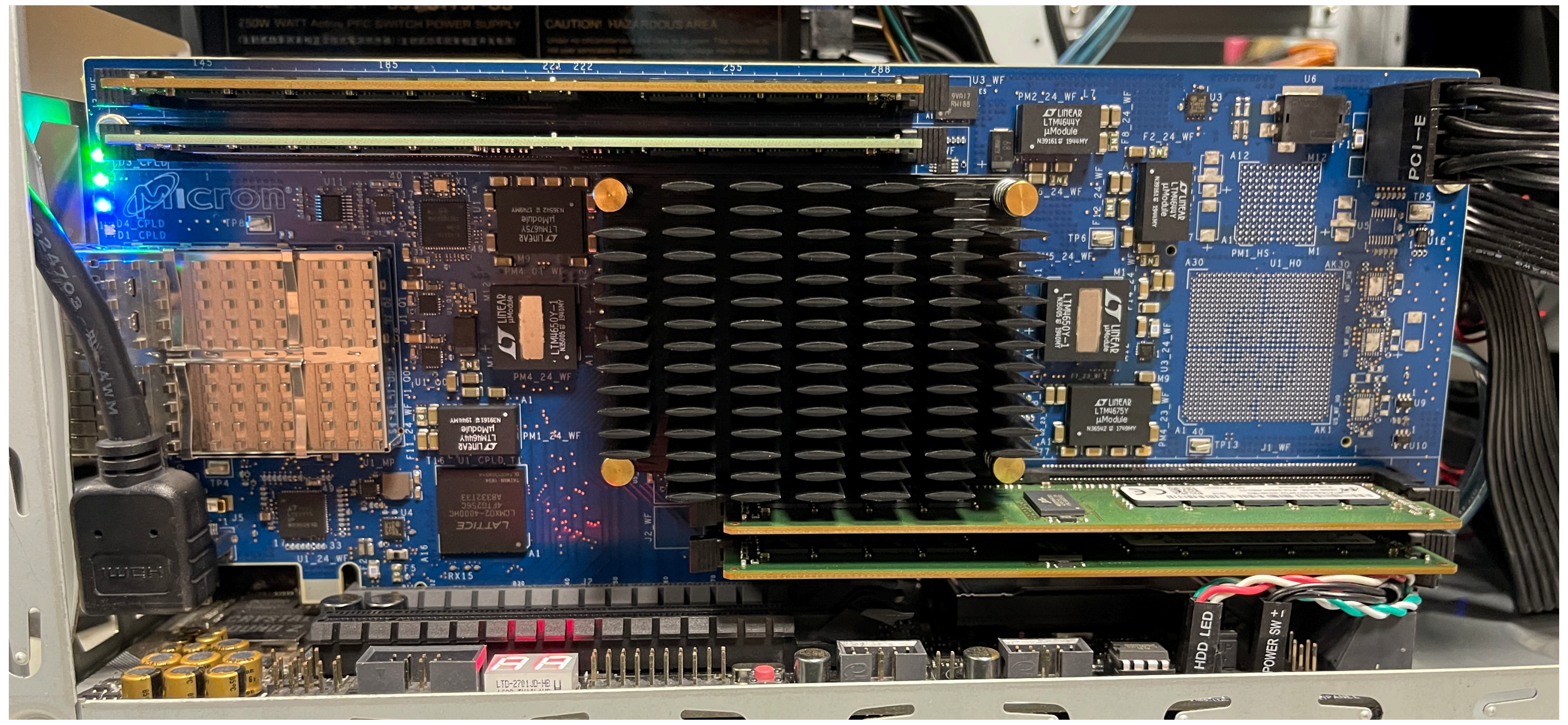




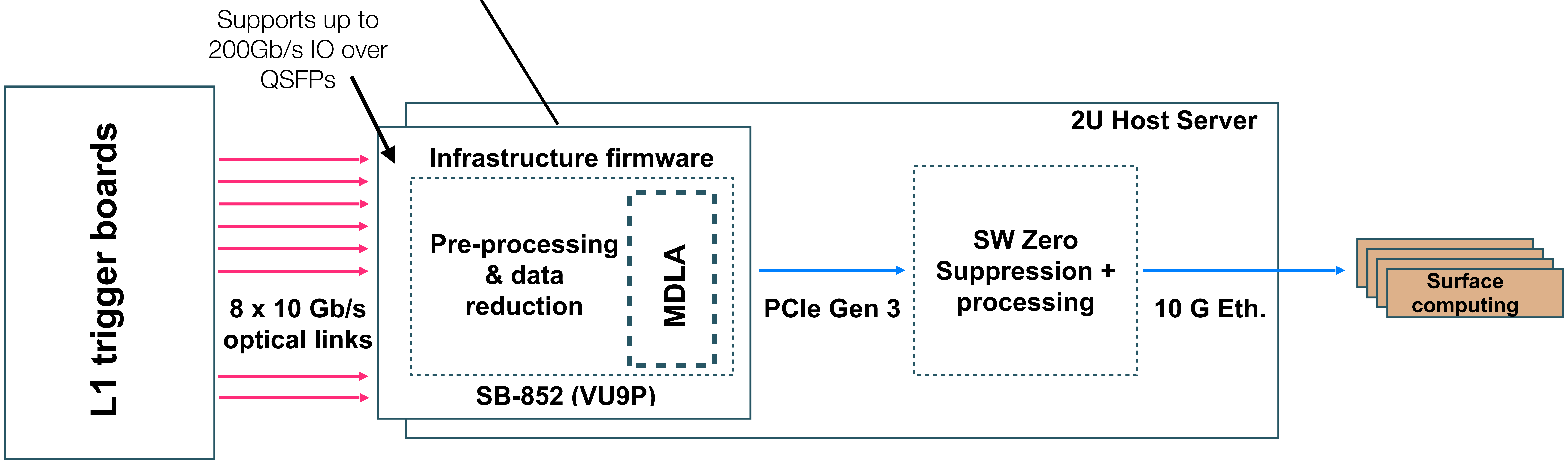
Micron: EX-750 + AC-511(s)

User friendly API reports diagnostics of interest: latency, precision, bandwidth

CMS 40 MHz Scouting with Micron SB-852



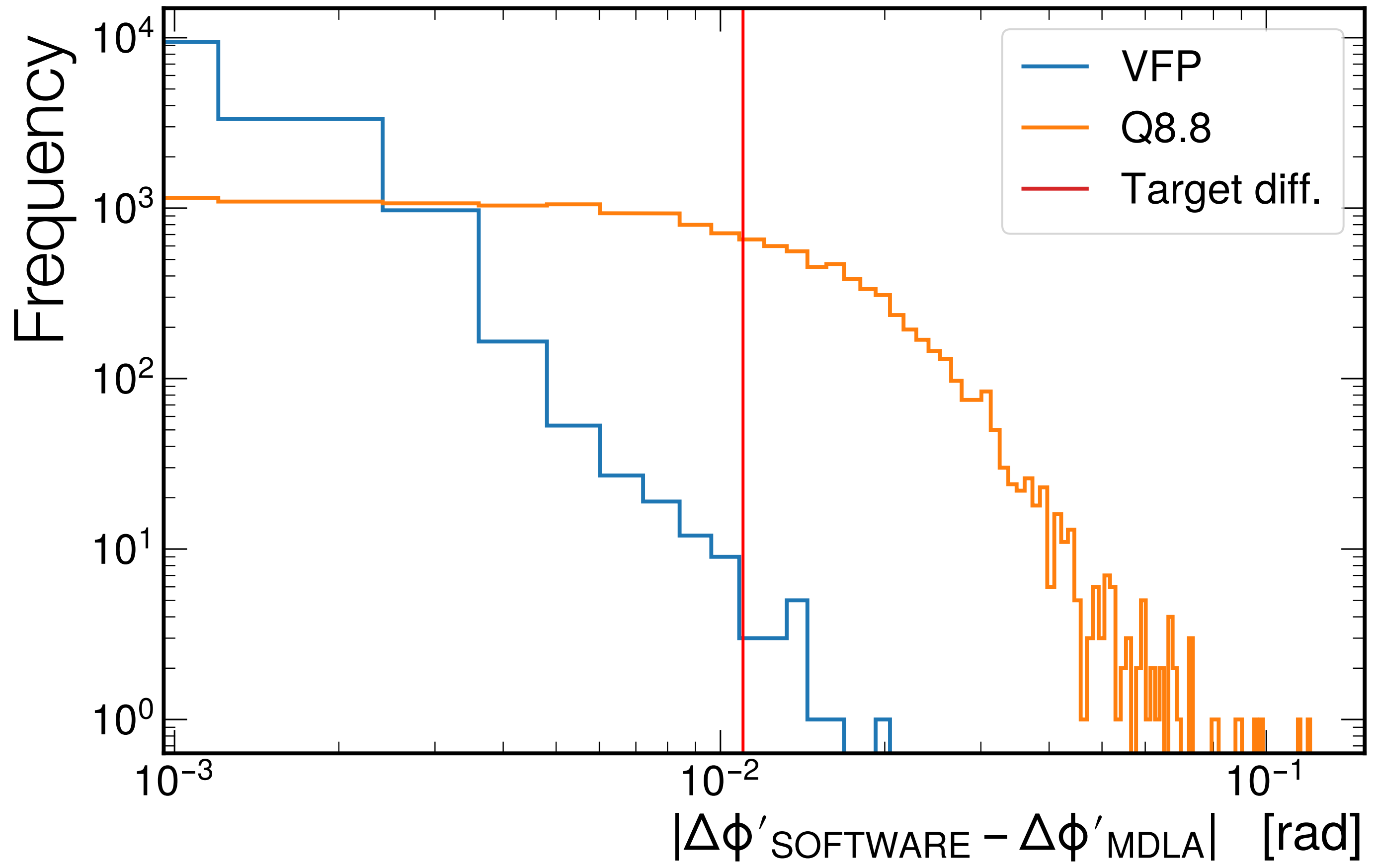
- Micron SB-852 for optical input -> DMA to PC
- Perform NN inference with Micron DLA after firmware ZS
- MDLA is embedded within the infrastructure & L1 scouting firmware

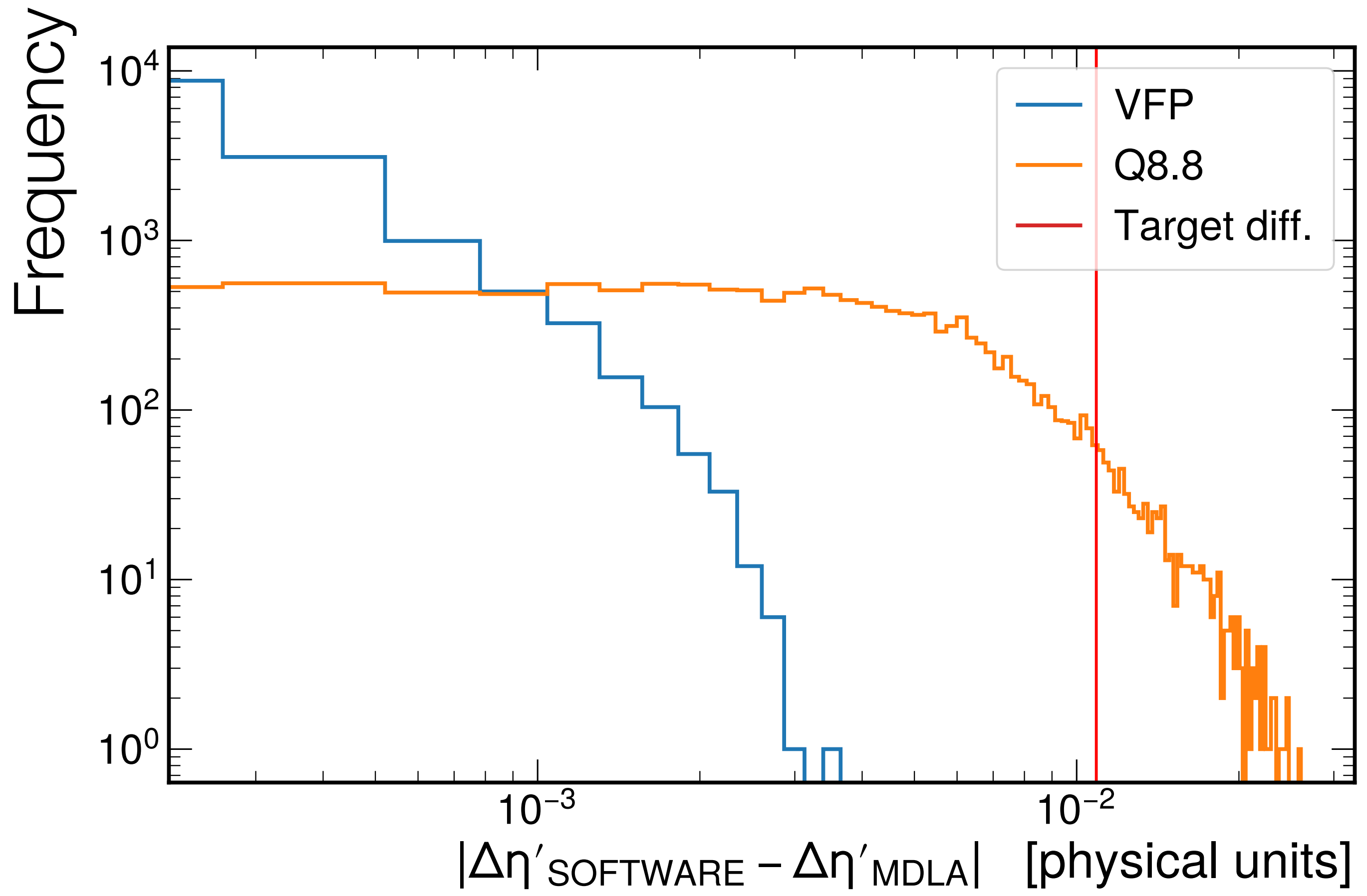


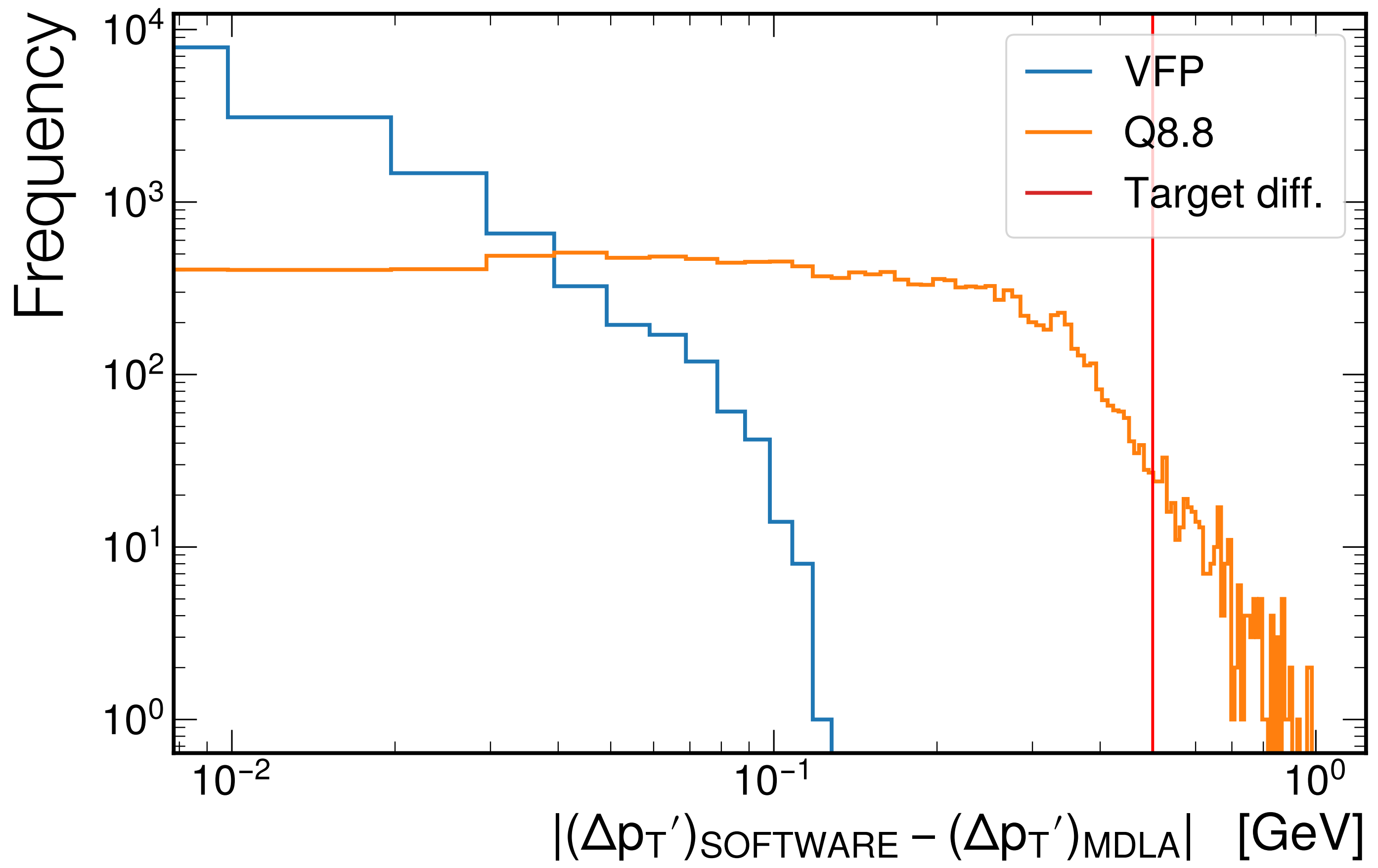
MDLA precision

- Three ways of running:
  - Full software e.g tensorflow, ONNX real-time
  - In the hardware SB-852
  - Micron-provided sw *emulator* (100% accurate!)
- To improve precision:
  - "Scaling" Integer inputs / 256
  - Batch normalisation
- **Q8.8** & **Variable Fixed Point (VFP)** modes available
- **Target precision** is to be < L1 object LSB step size of same variable e.g < 0.5 GeV p_T

Precision |hardware - tensorflow software|	Frac. Values < 1% diff
Model w/ integer inputs	99%







SB-852 resource utilisation & throughput

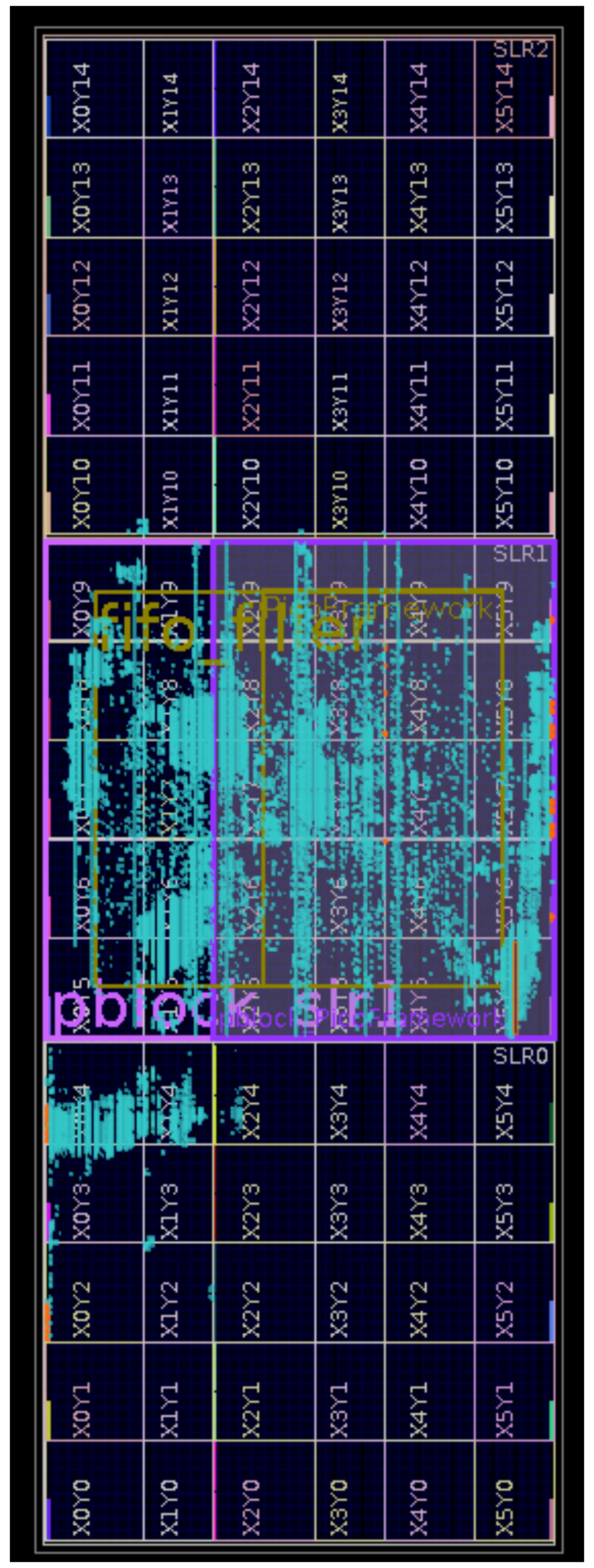
VU9P - MDLA w/ VFP

N DLA clusters	LUTs [%]	BRAM [%]	URAM [%]	DSP [%]
0	2.72	28.10 ← Extra readout buffers needed w/o DLA	0.21	0
1	21.61	28.96	6.88	16.10
2	29.95	43.70	13.33	32.02

N DLA clusters	Inference rate	Average latency / muon inference	Encoding
4 cluster	5.2 MHz	192 ns	Q8.8
2 cluster	2.6 MHz	385 ns	Variable Fixed Point (VFP)

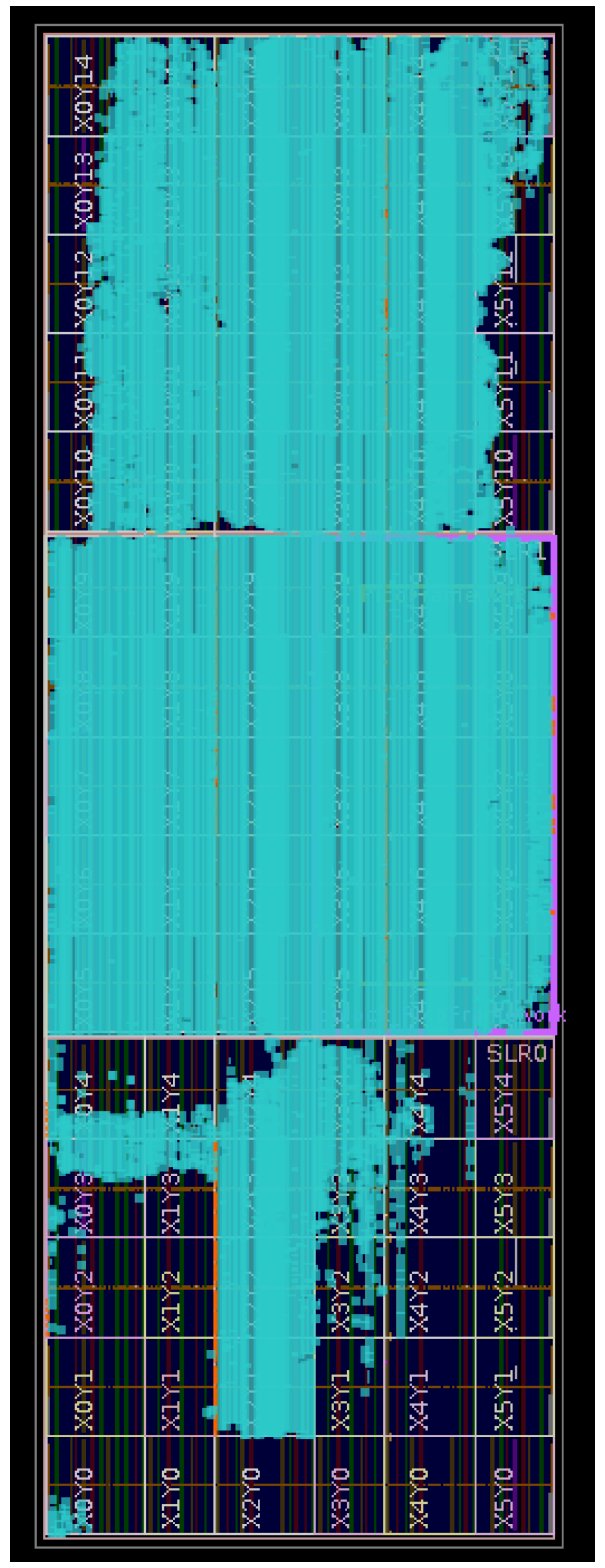
- Not yet able to fit 4 clusters w/ VFP

SB-852 infrastructure + L1 scouting firmware



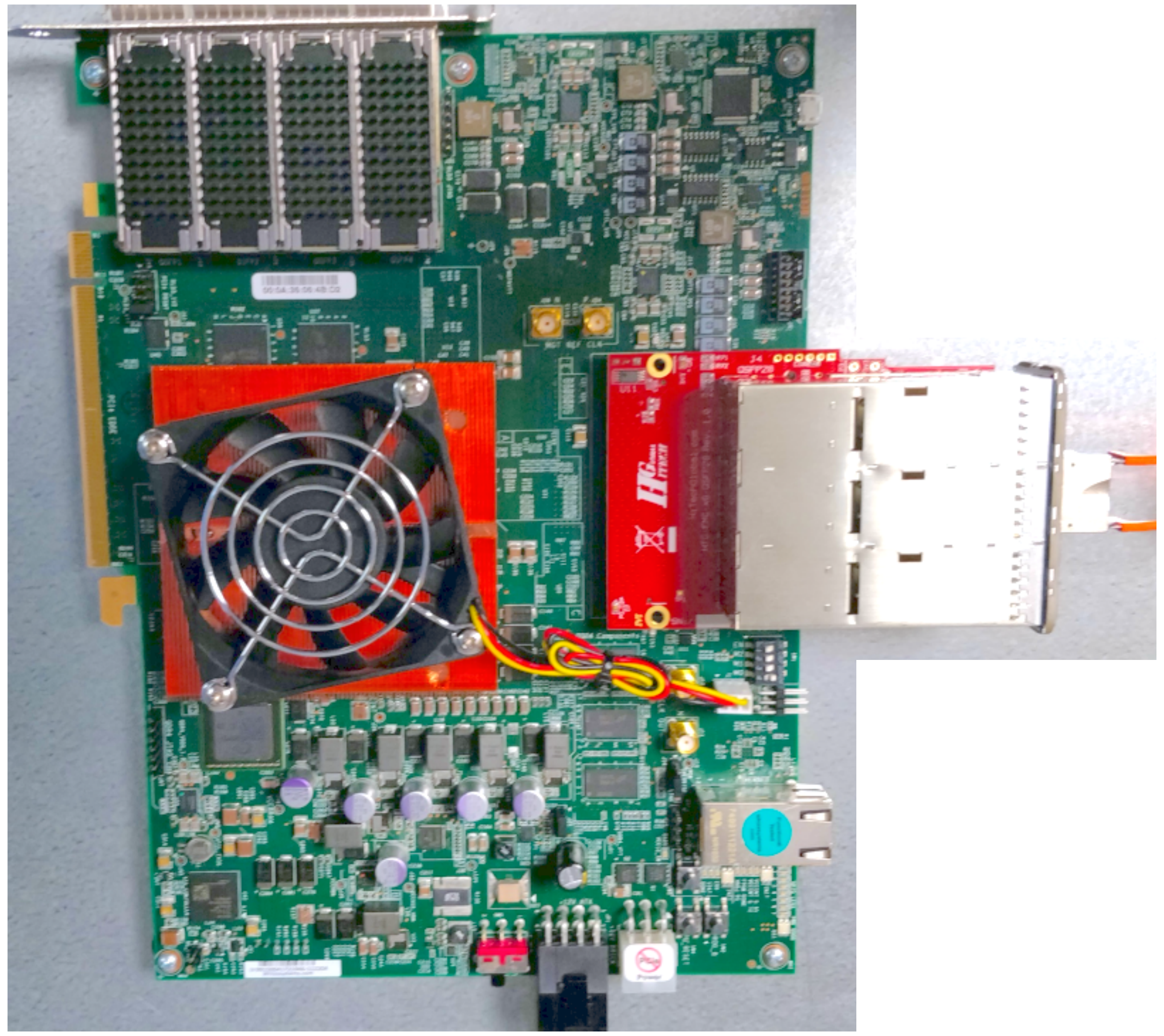
Mostly confined to SLR1

SB-852 infrastructure + L1 scouting firmware + 2 clusters of MDLA

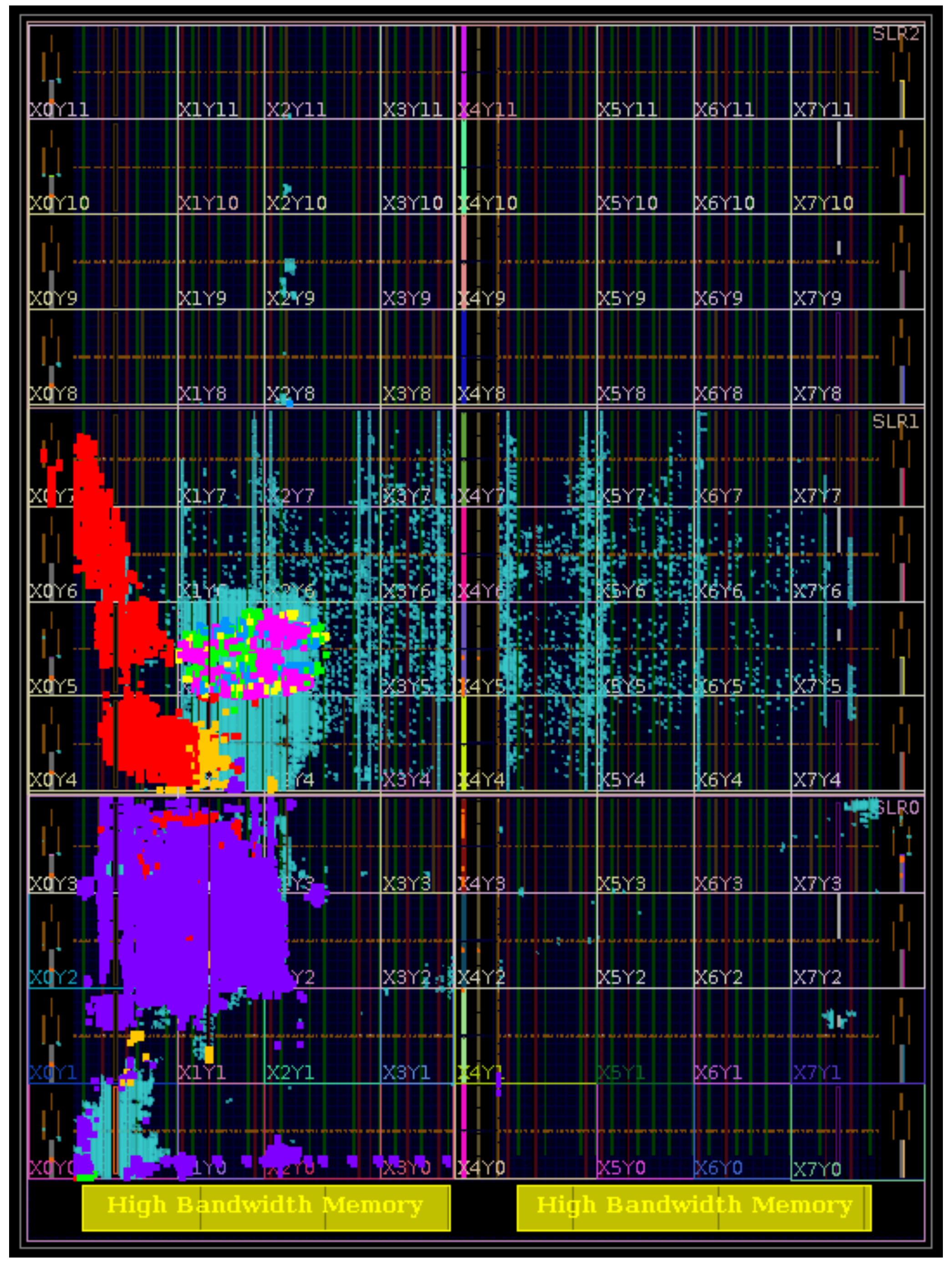


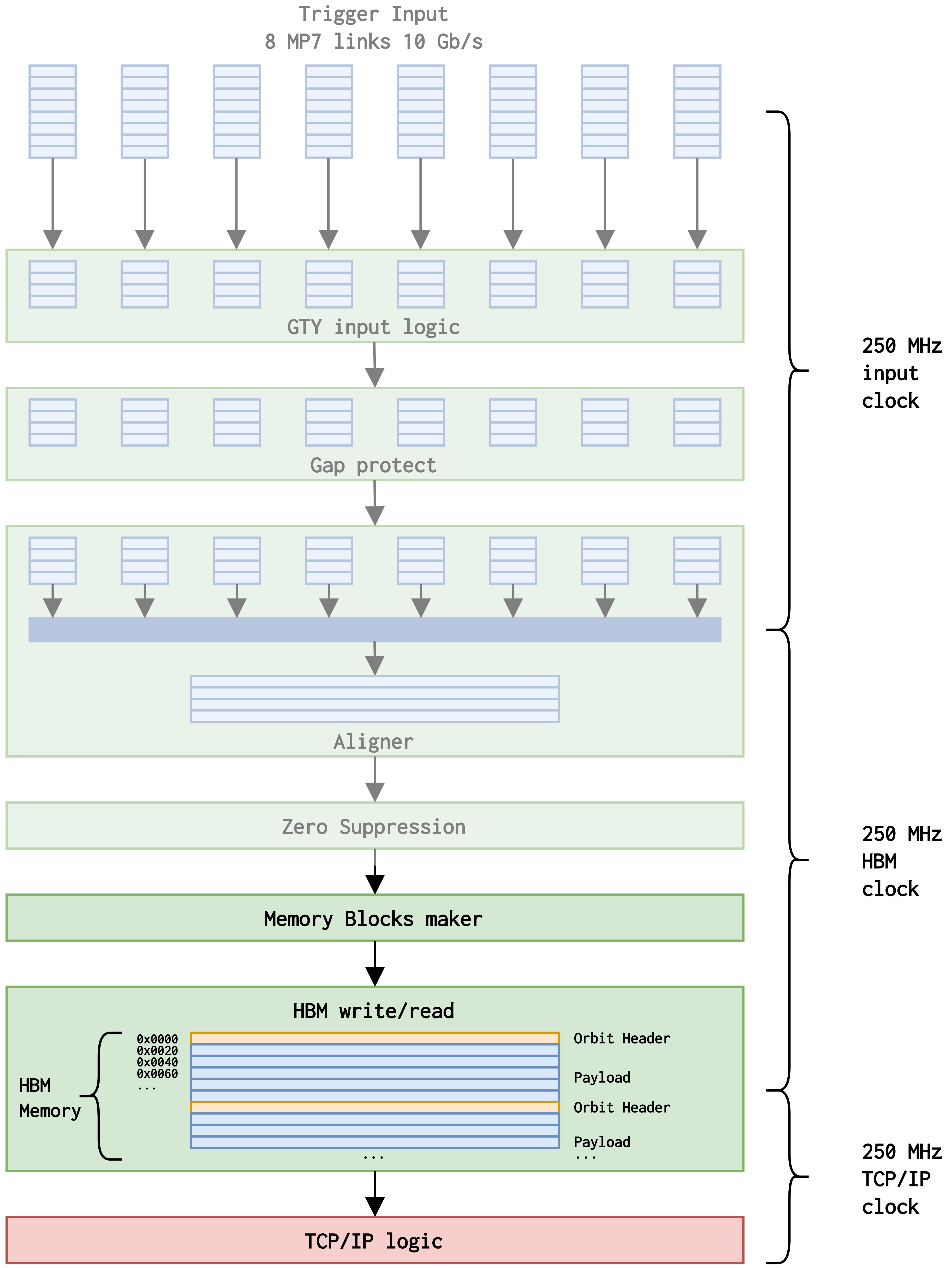
40 MHz scouting w/ VCU128

- (4 + 6 w/ mezzanine) QSFPs & HBM
- Replace DMA w/ TCP/IP to surface
- Replace FIFO chain w/ HBM
- DMA data-taking also supported



Legend:
- GTY input emulation logic
- Gap protect module
- Aligner module
- Zero Suppression module
- Memory blocks for HBM maker module
- HBM write/read logic
- TCP/IP engine

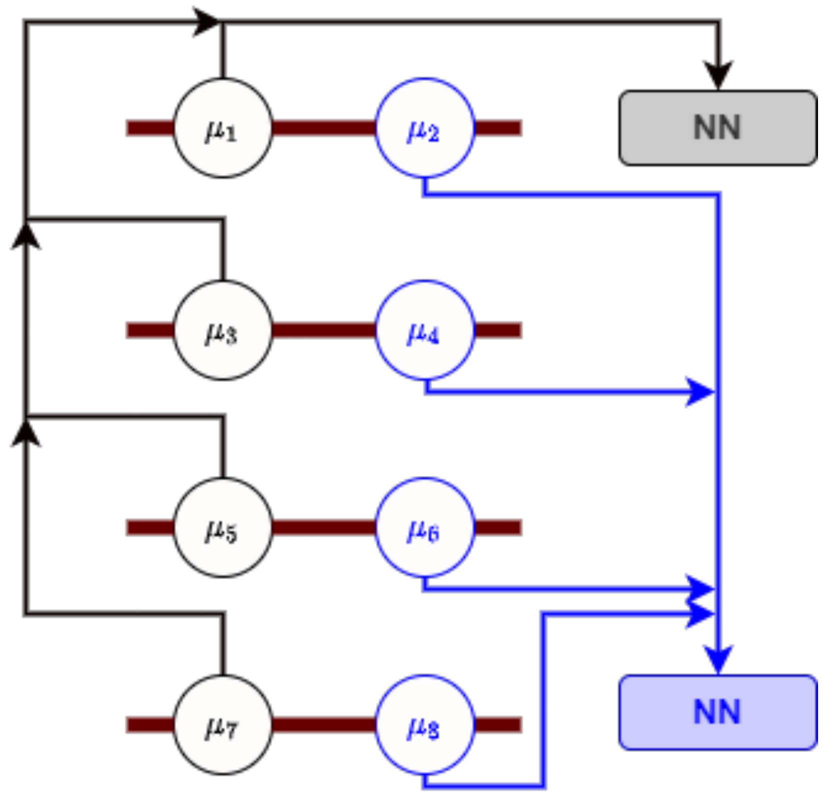


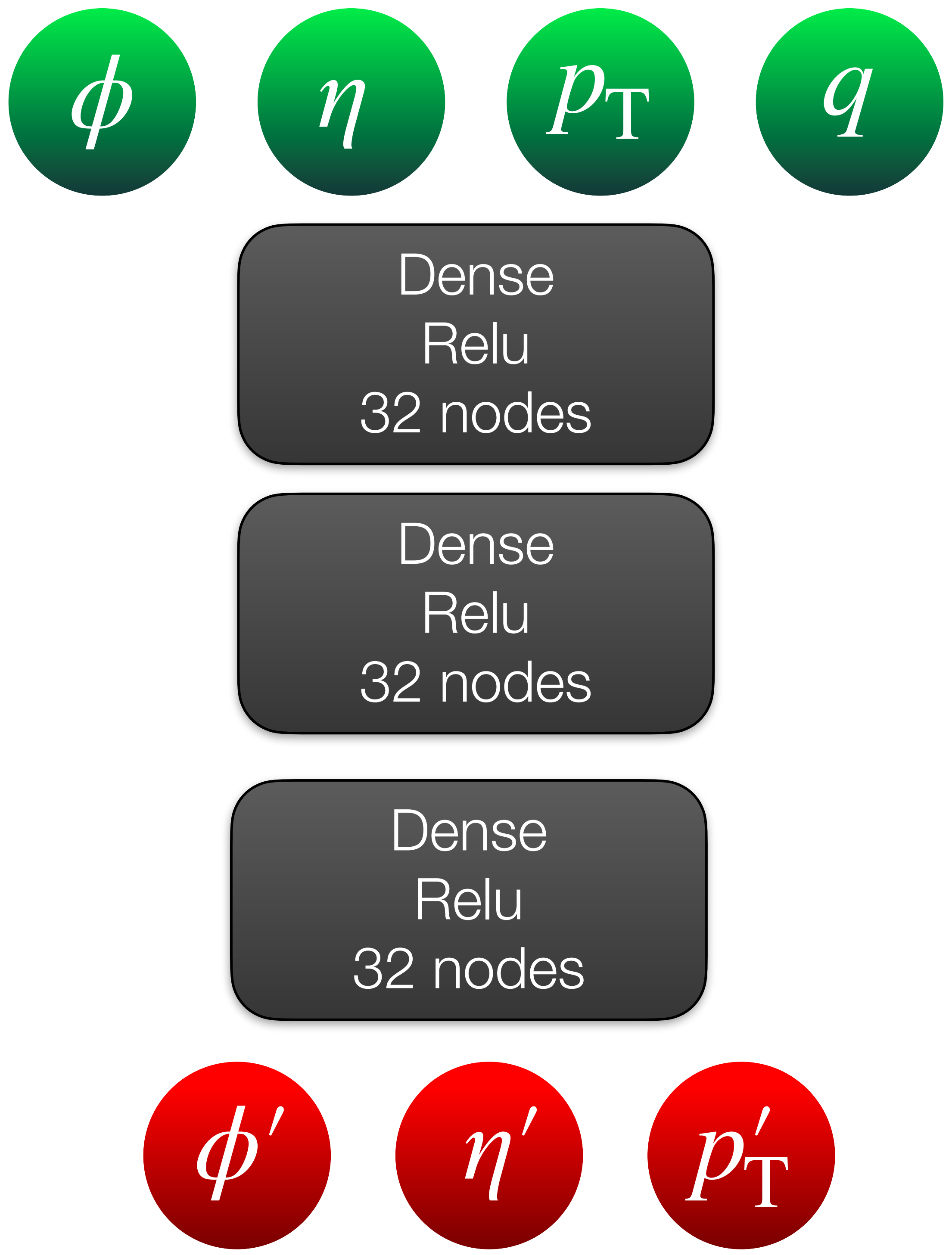


VCU128 - NN w/ hls4ml

- Integrated NN for muon recalibration generated w/ HLS4ML*
- Q6.12 precision, pruning factor 0.5
- 2 NN each process 4 muons / BX
- Latency $\lesssim$ 100 ns FIFO latency, can accept 2 muons / clock

	VU37P
LUTs [%]	2.72
BRAM [%]	21.61
DSP [%]	29.95

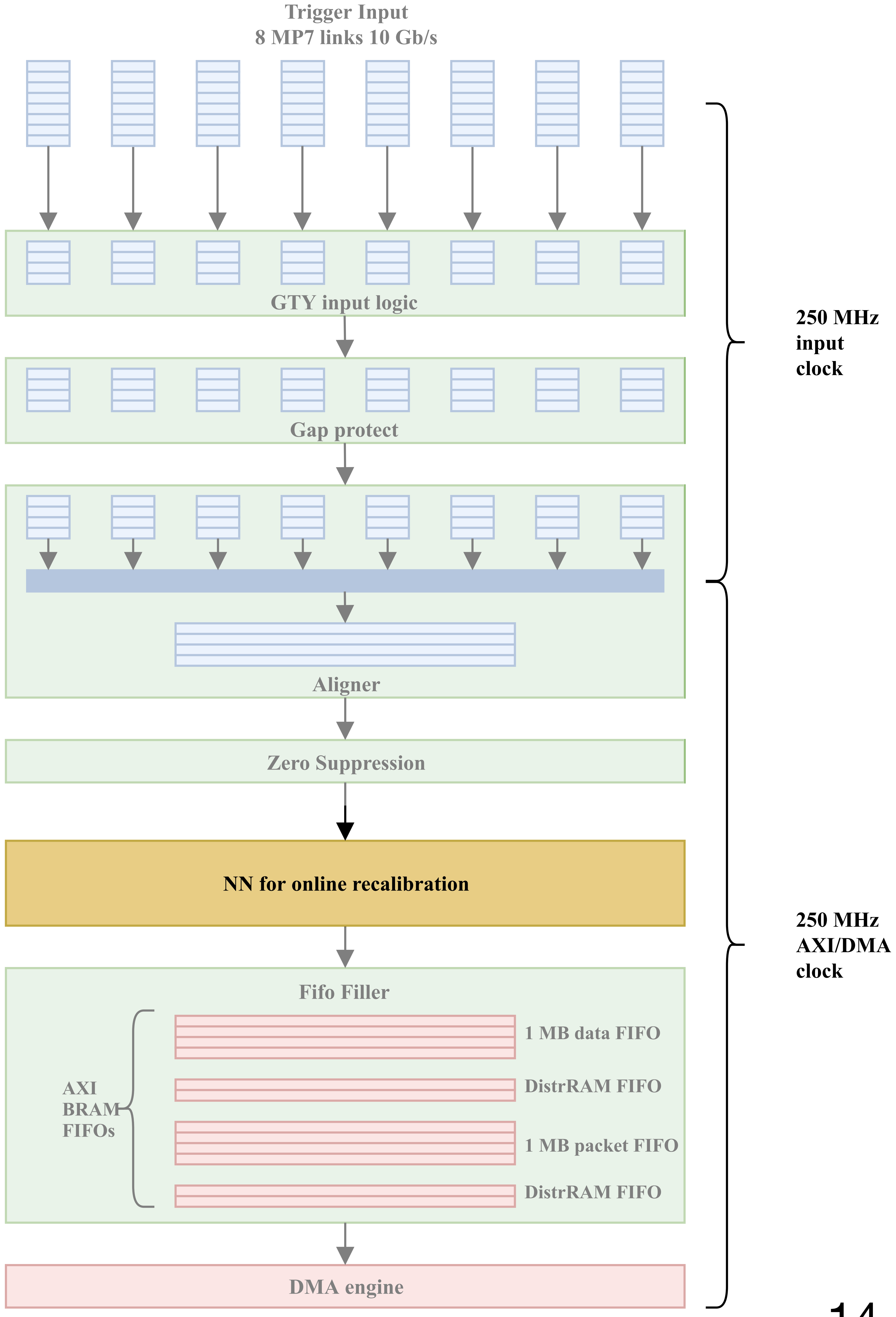




*Python API & command line tool that translates trained NNs to synthesizable FPGA firmware

https://fastmachinelearning.org/hls4ml/

https://arxiv.org/abs/1804.06913

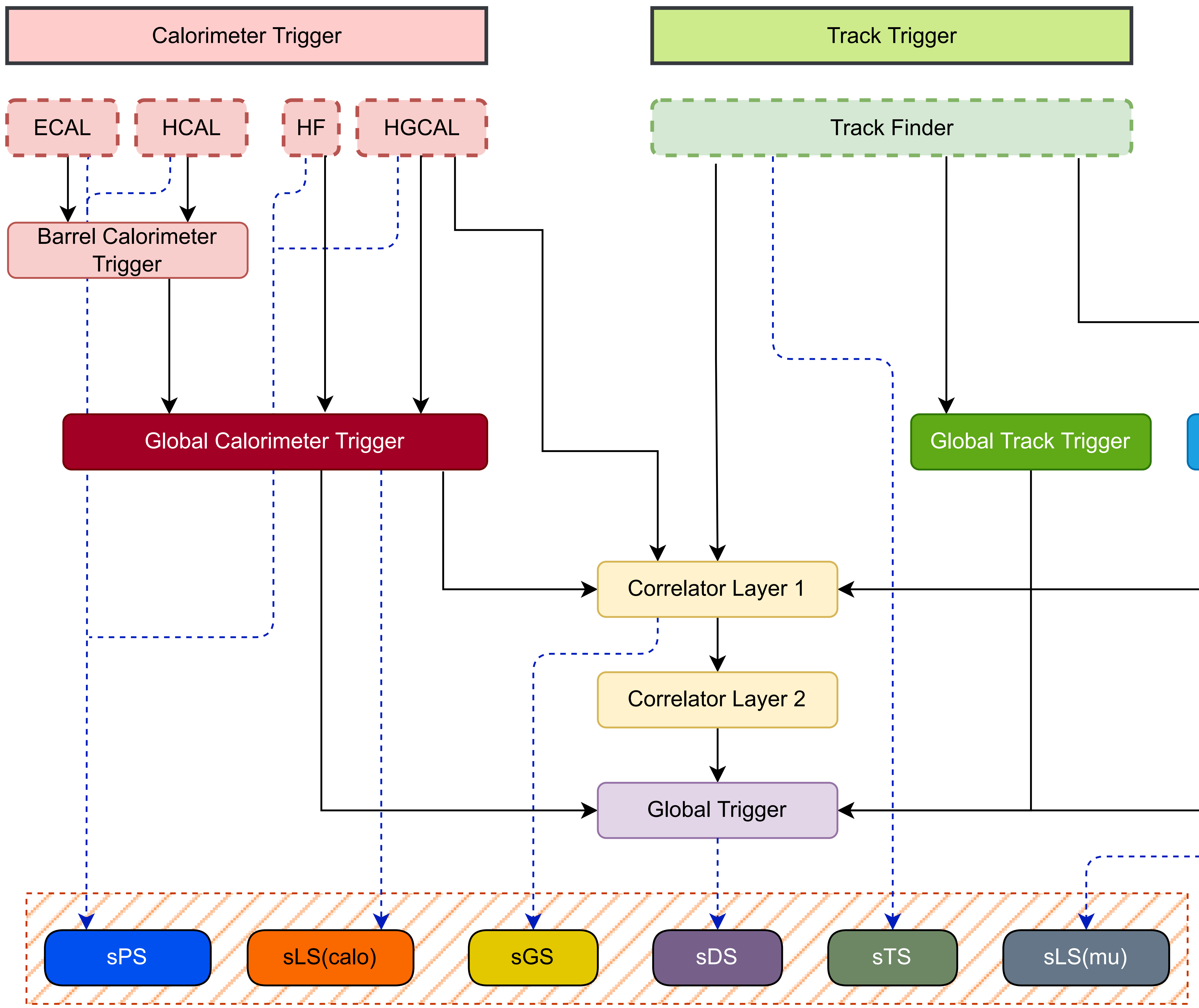


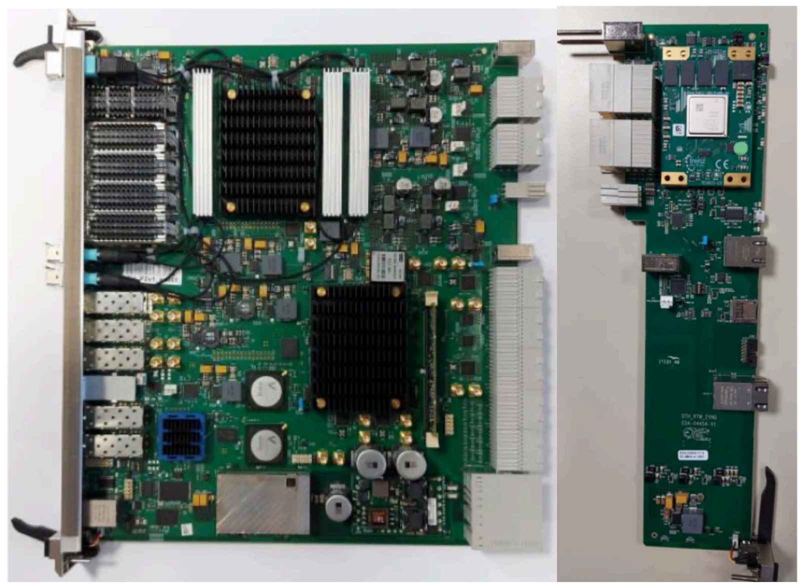
Plans for CMS Phase 2

New L1 trigger for CMS at HL-LHC

L1 scouting will have stageable architecture

1. **GT inputs & Outputs (sDS)**
2. **Calo & Muon local reco (sLS)**
3. **Tracker tracks (sTS)**
4. **Calo primitives (sPS)**





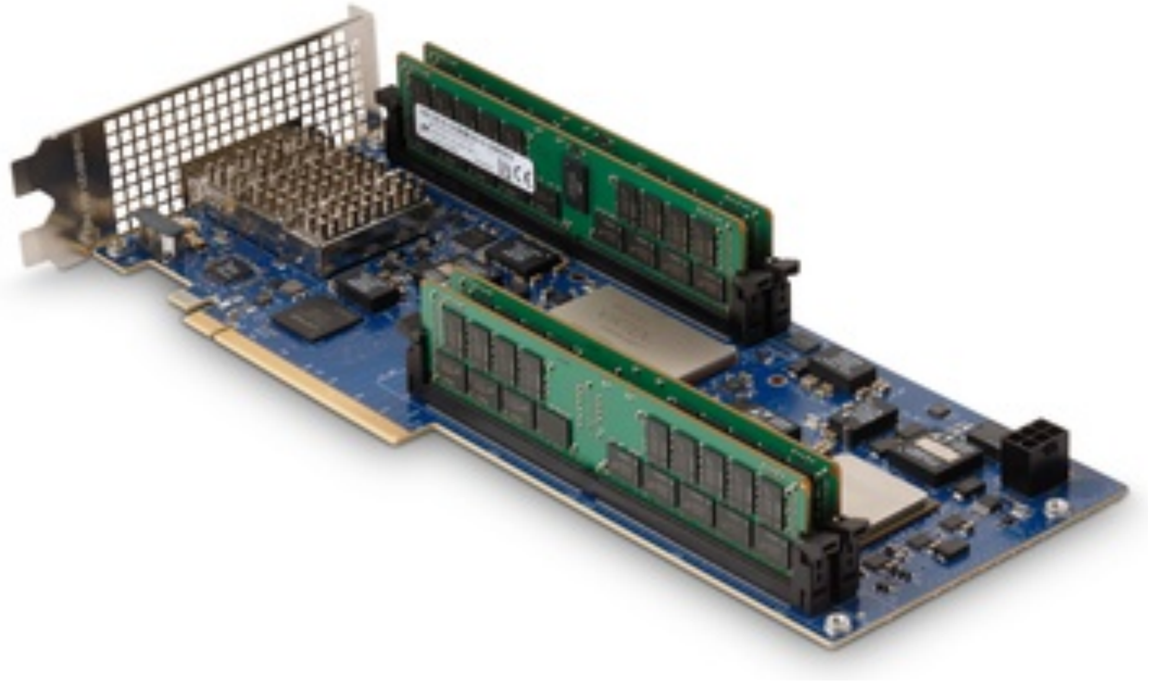
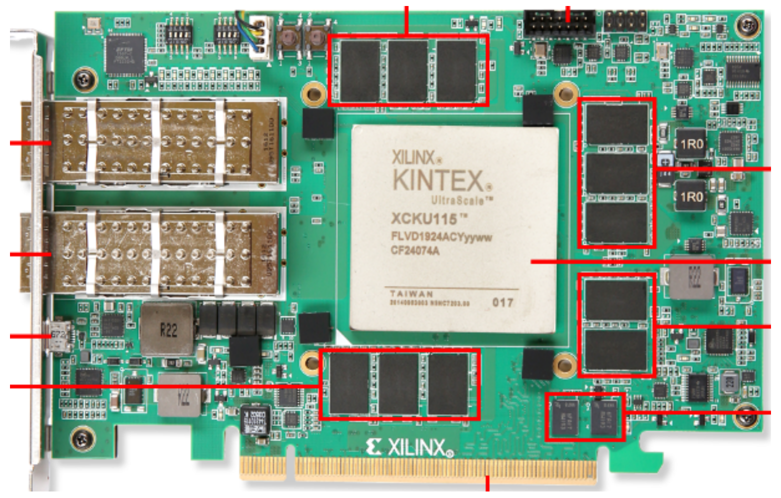
DAQ400 prototype-2 + RTM

Hardware = *DAQ800*

- CMS DAQ Ph2 readout board
- 2x VU35P FPGAs
- 6x4 FireFly inputs / FPGA
- 5 QSFP outputs / FPGA

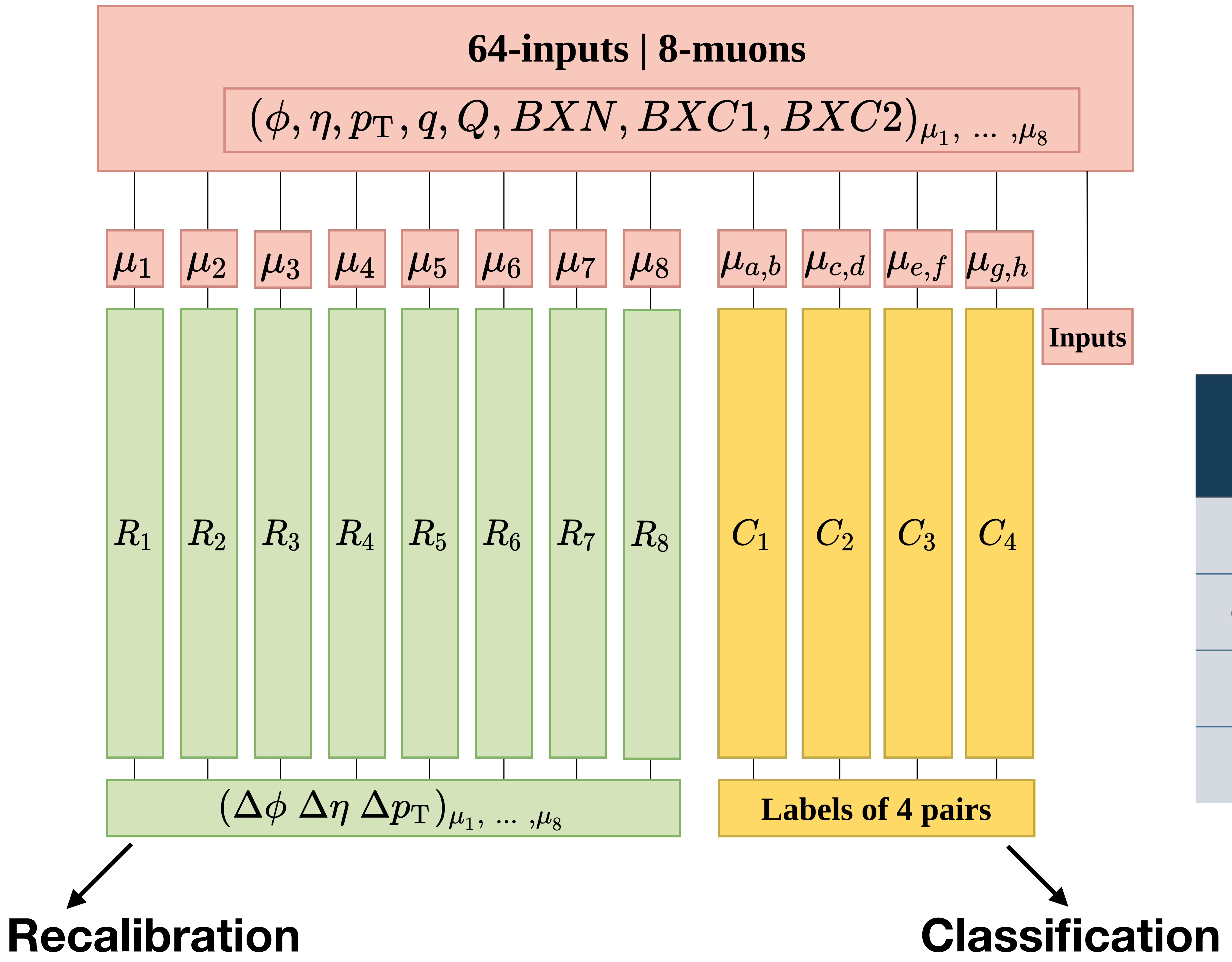
Summary

- L1 Scouting demonstrator system in operation, taking data from $\mu\mathrm{GMT}$ and CALO trigger Layer 2
- Three FPGA boards: Xilinx KCU1500, VCU128 & Micron SB-852
- Applying ML inference w/ help of Micron DLA framework and/or HLS4ML
  - for re-calibration of parameters and
  - fake detection
  - w/ real performance gains
- Full system in development w/ DAQ800 board for CMS at HL-LHC


Backup: Fake muon pair classifier

- Network consists of **8 recalibration branches & 4 classification branches**
- Trained/tested with Run 3 Zero-bias data



True positive/false positive: area under curve	
Barrel - only	89.2%
Overlap - only	97.4%
Endcap - only	97.7%
All	97.2%