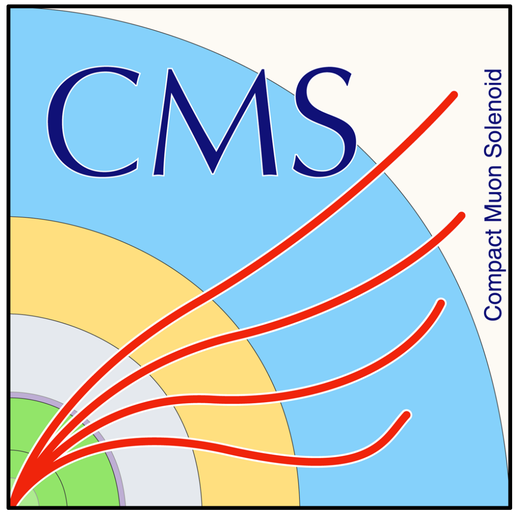


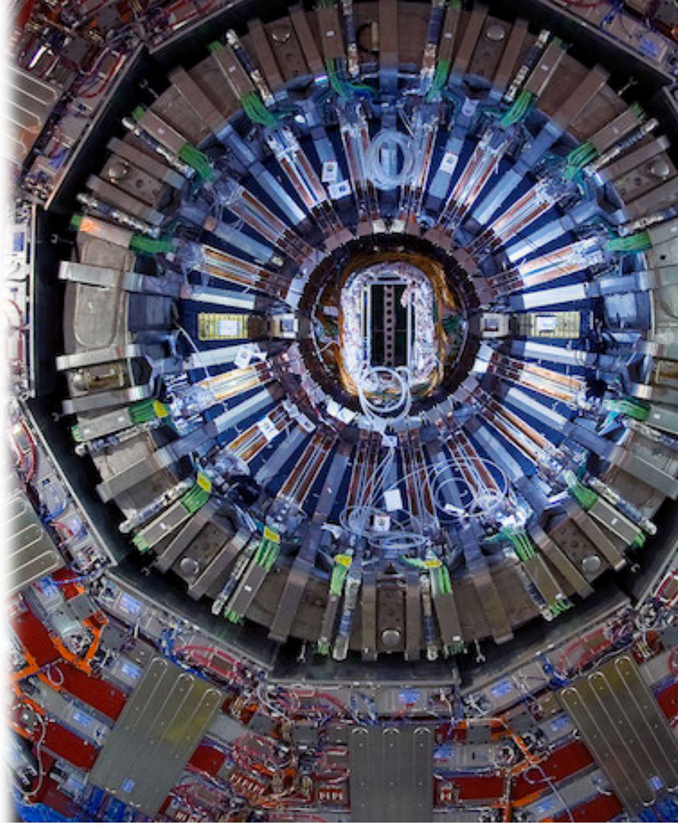
Real-time deep learning inference and FPGA based processing for level 1 trigger scouting at CMS

Thomas James *(CERN)*, Emilio Meschi, Rocco Ardino, Sabrina Giorgetti

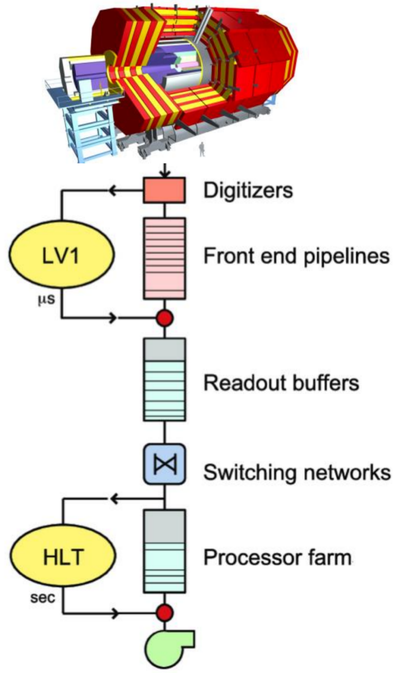
CERN Openlab workshop

16th March 2023

The CMS Detector at the LHC



- 2.4 billion collisions / second
  - In CMS ~ 100M sensors
  - Produce ~ 1.5 MB @ 40 MHz, ~500 Tb/s
  - Impossible to read out (or store) all data
  - Need fast ‘trigger’ to select *interesting* collisions for analysis

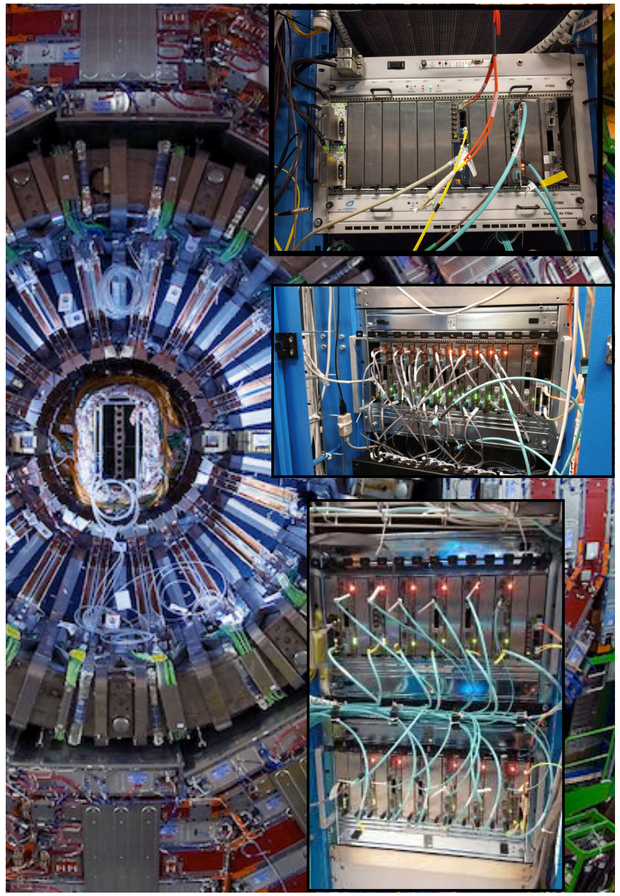


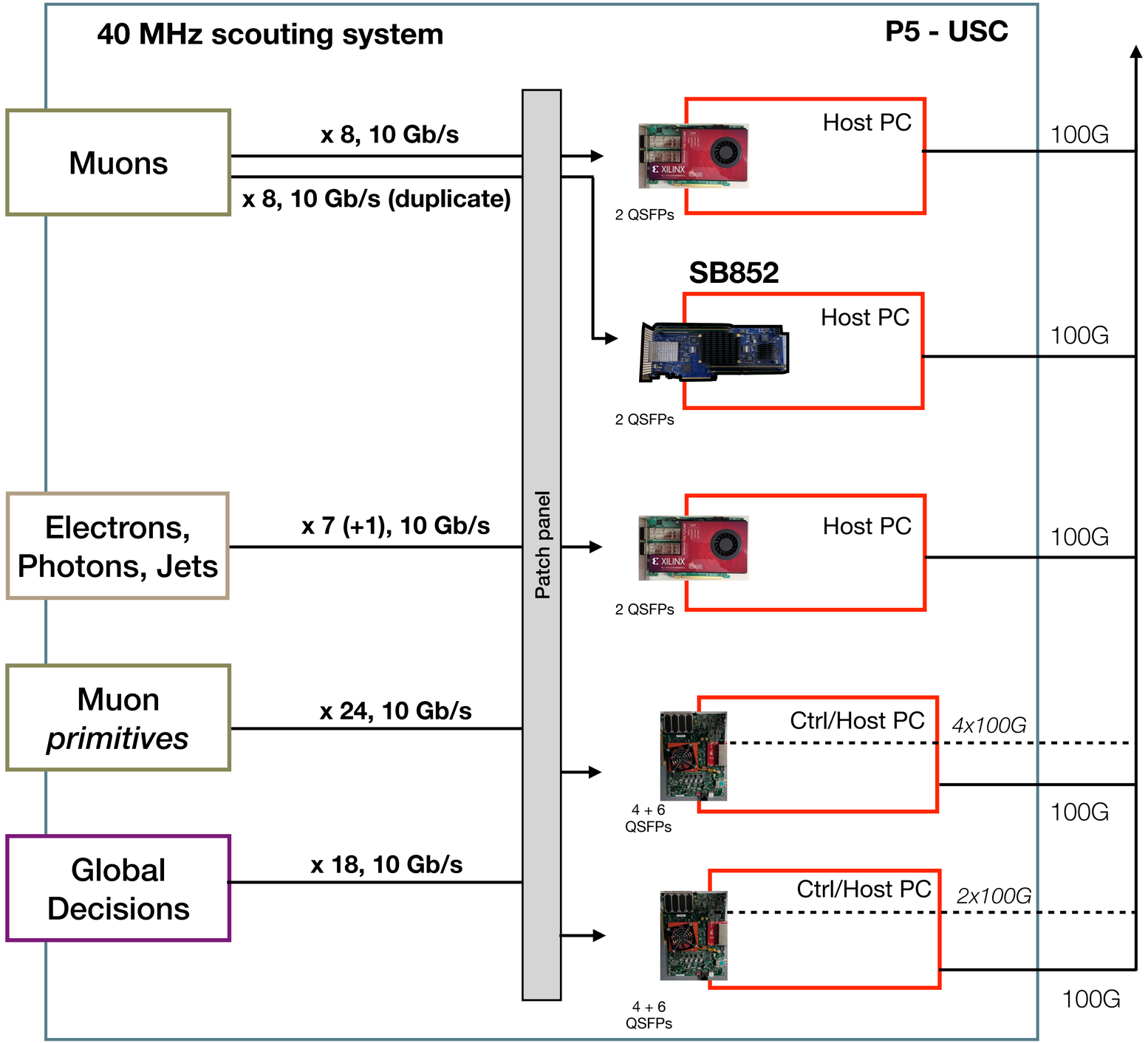
- Two layered:
  - Level 1: Fixed latency of 3.2 microseconds -> ASICs and FPGAs required
  - High Level Trigger: Flexible latency ~100 ms compute / event -> CPUs/GPUs

40 MHz Scouting: What does L1 trigger 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
- **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
- **Diagnostic and monitoring capabilities**
  - BX-to-BX correlations always available
  - Independent per-bunch lumi measurement

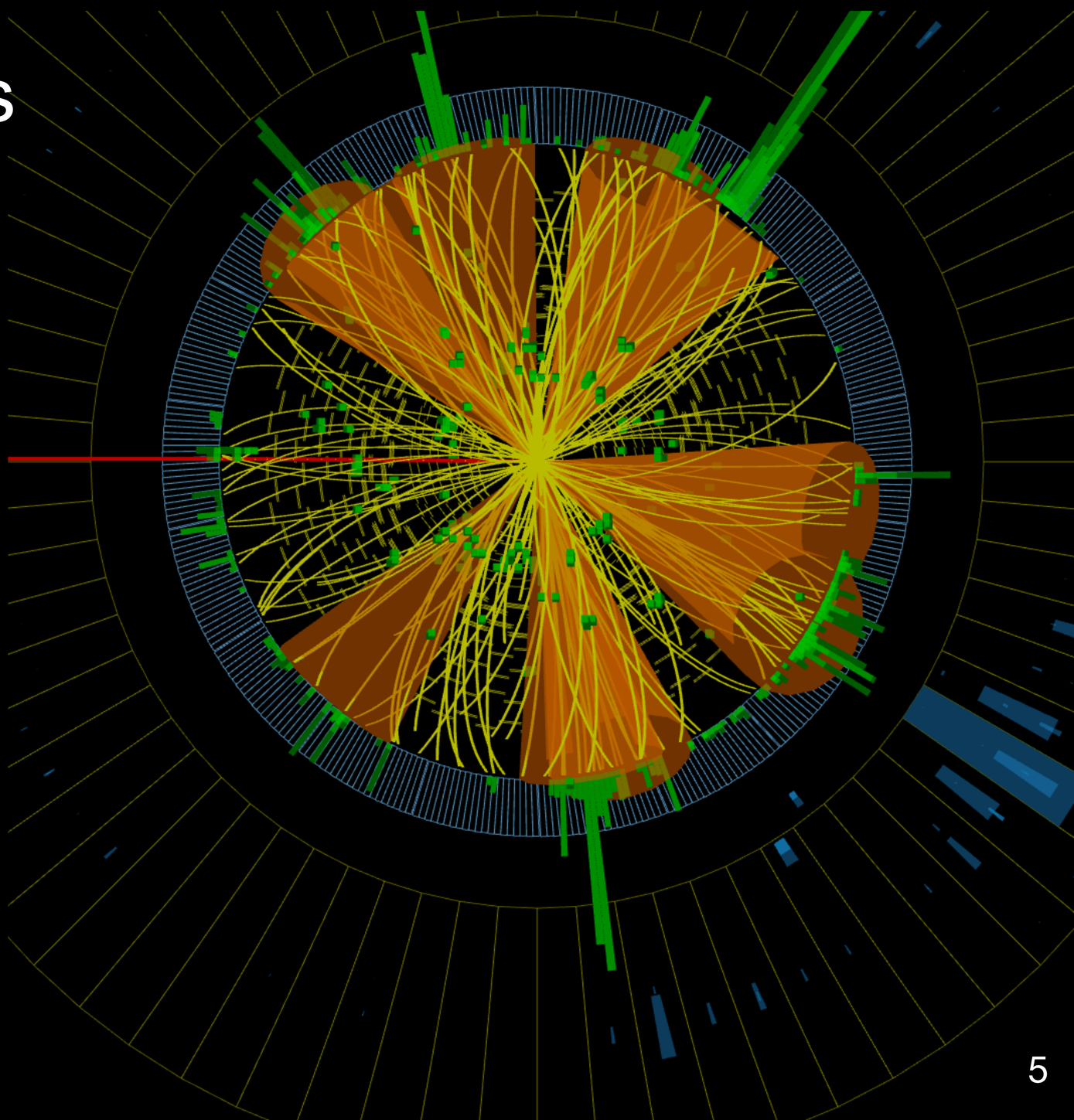
L1 Scouting Demonstrator





Fast Inference in FPGAs

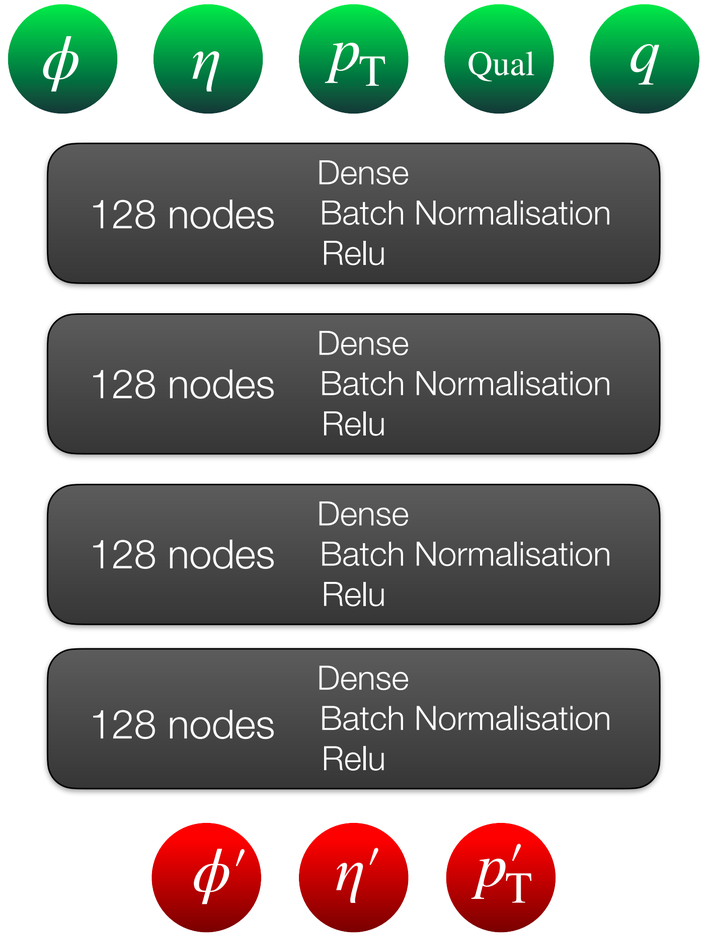
- Machine learning being exploited across particle physics
- Particles of interest rare
- Need to detect them with high efficiency, low fake rate
- Can train a network to achieve performance of complex algorithms in much lower latency
- FPGAs provide ultra-low latency, high bandwidth, and fixed timing



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

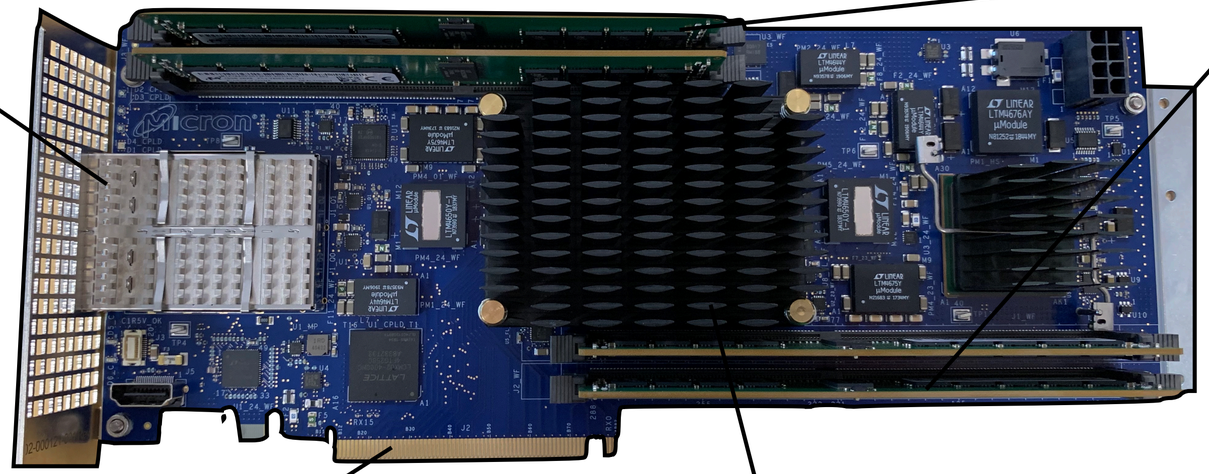
- **Inputs** - L1 objects e.g muons:
- **Target** - Offline fully reconstructed objects
- Use of classical ***fully connected*** neural networks to ‘recalibrate’ L1 information to improve their utility for an online analysis



SB-852 PCIe board

2x QSFP 100G

64G DDR4



PCIe Gen 3 interface / form factor

Xilinx VU9P FPGA

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

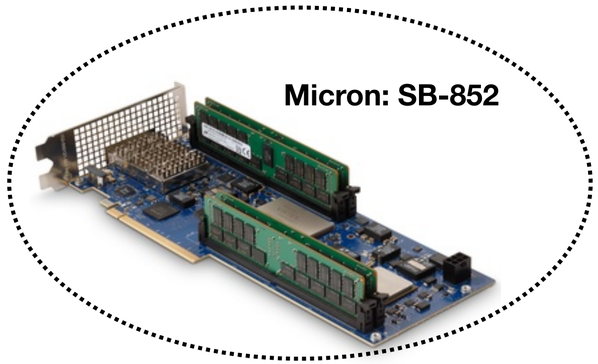
TensorFlow

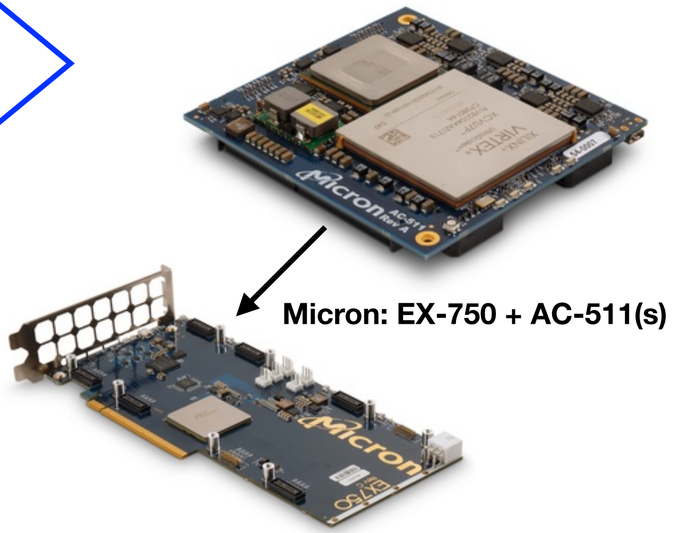
Caffe2

PyTorch

ONNX

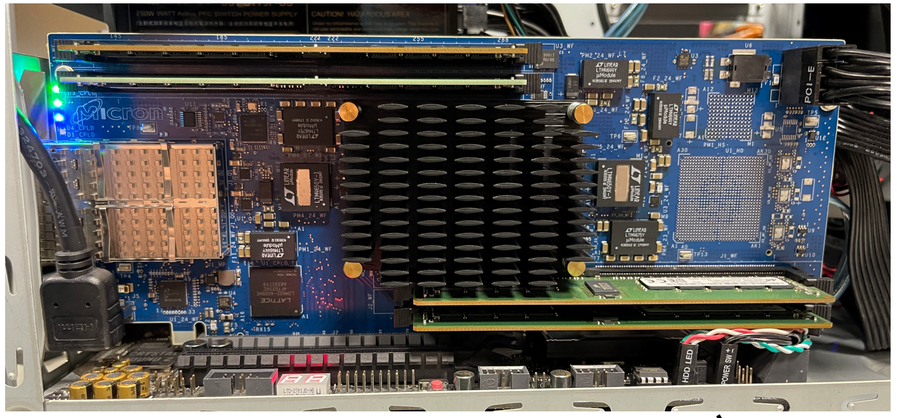
Micron Deep Learning Complier



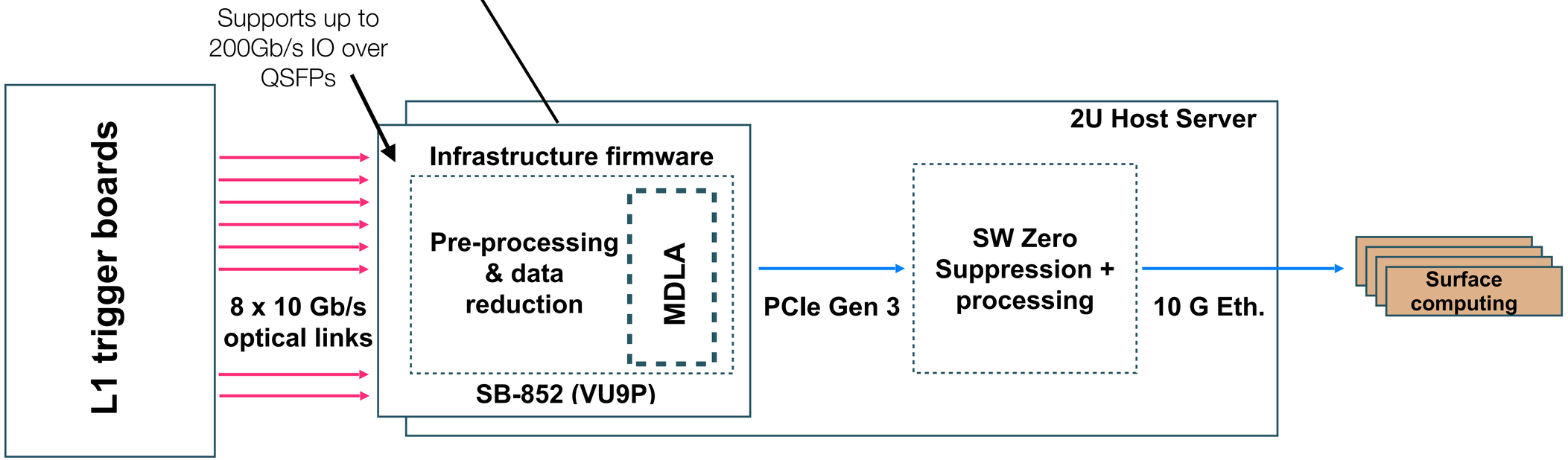


User friendly API reports diagnostics of interest: latency, precision, bandwidth

CMS 40 MHz Scouting with Micron™ SB-852

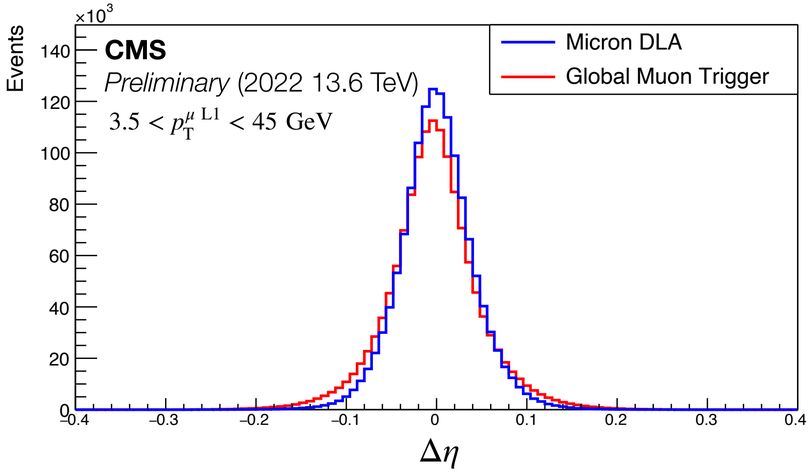


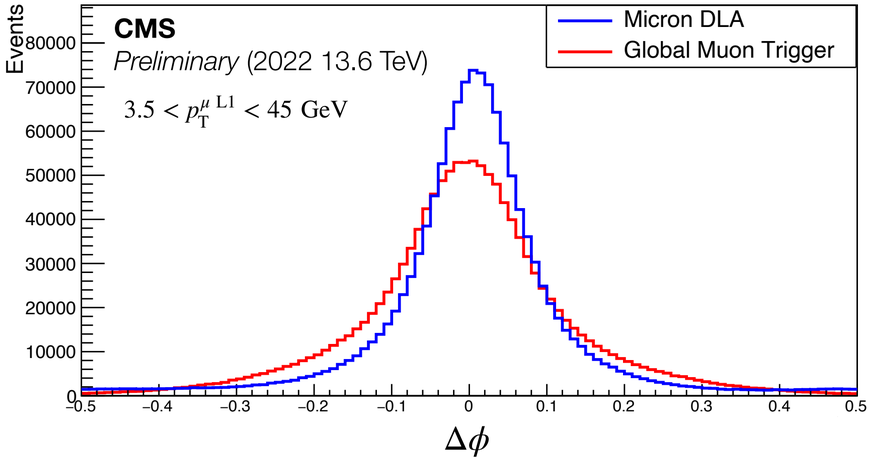
- Micron SB-852 for optical input -> DMA to PC
- Perform NN inference with Micron DLA after firmware data reduction / zero suppression
- MDLA is embedded within the infrastructure & L1 scouting firmware

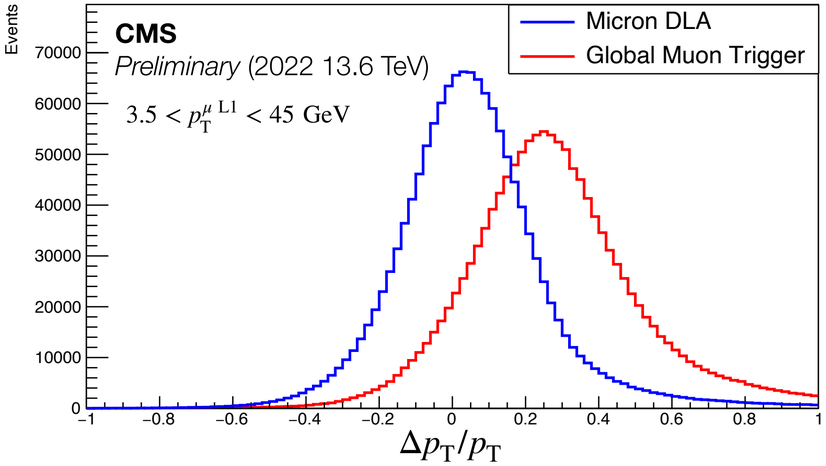


Muon re-calibration with Neural Network

- Trained with **2022 data** *Zero-bias* dataset
- $\Delta\eta$, $\Delta\varphi$, Δp_T is the difference between the prediction (or muon trigger) values, and the offline reconstructed muon tracks for matched muons ($\Delta R<0.1$ at 2nd muon station)



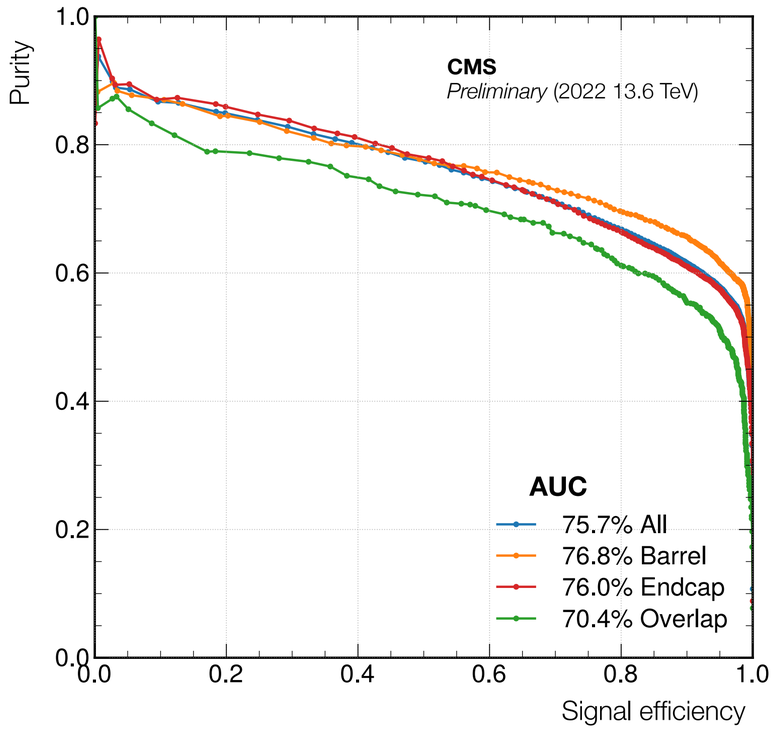


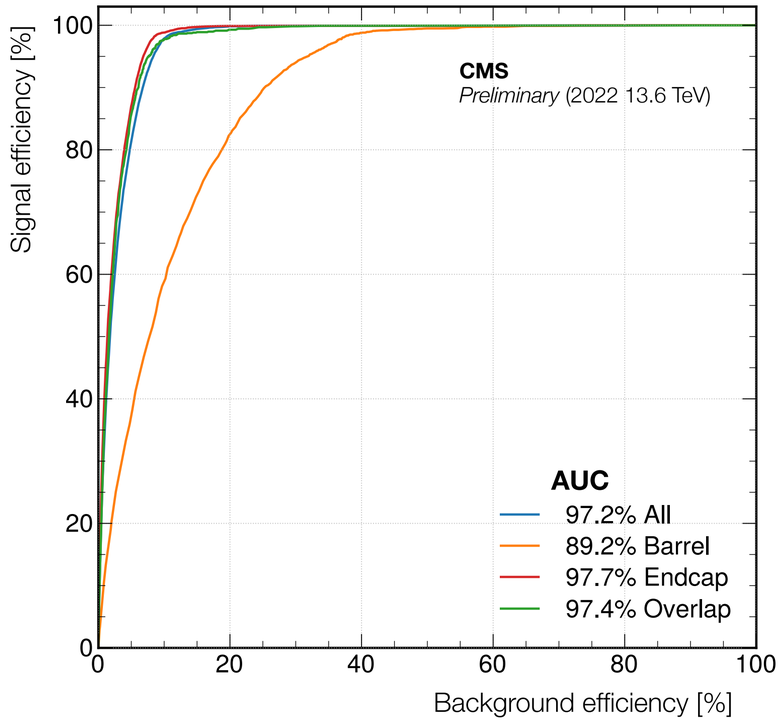


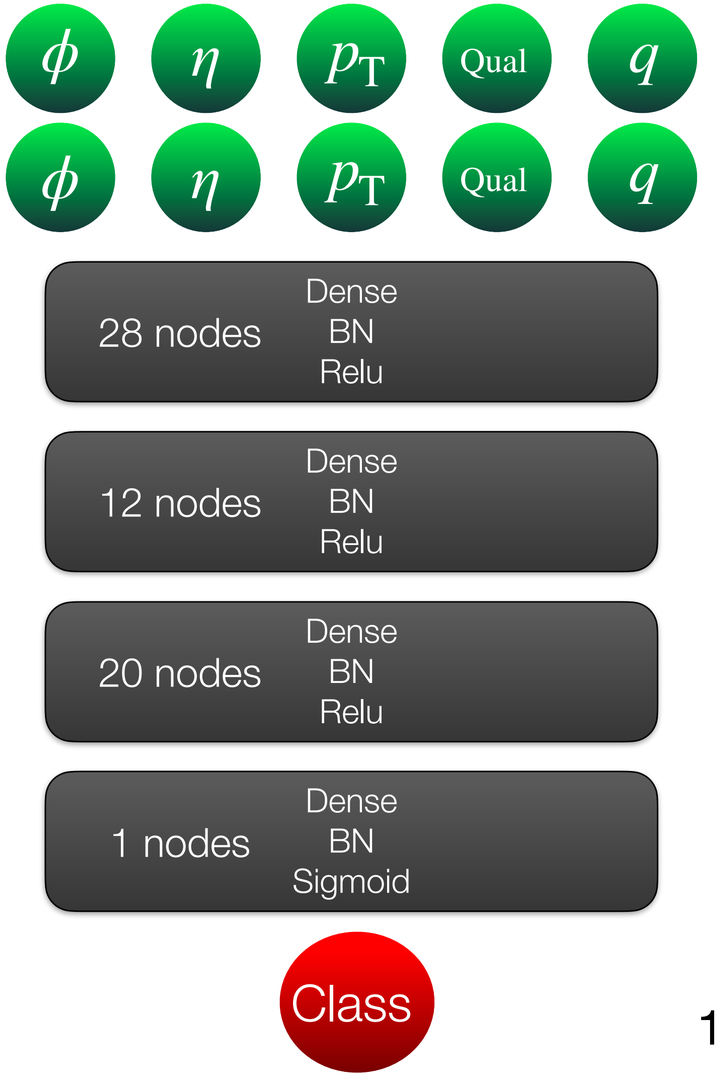
- Target: centred on zero & higher peak
- NN shows improvement for all variables

Fake muon pair classifier

- Acts on muon pairs: rejects misconstructed pairs of L1 muons not found in offline reco i.e L1 duplicates
- Trained/tested with 2022 data - 305k pairs (identical pre-removed)
- Huge gains in purity for small efficiency cost



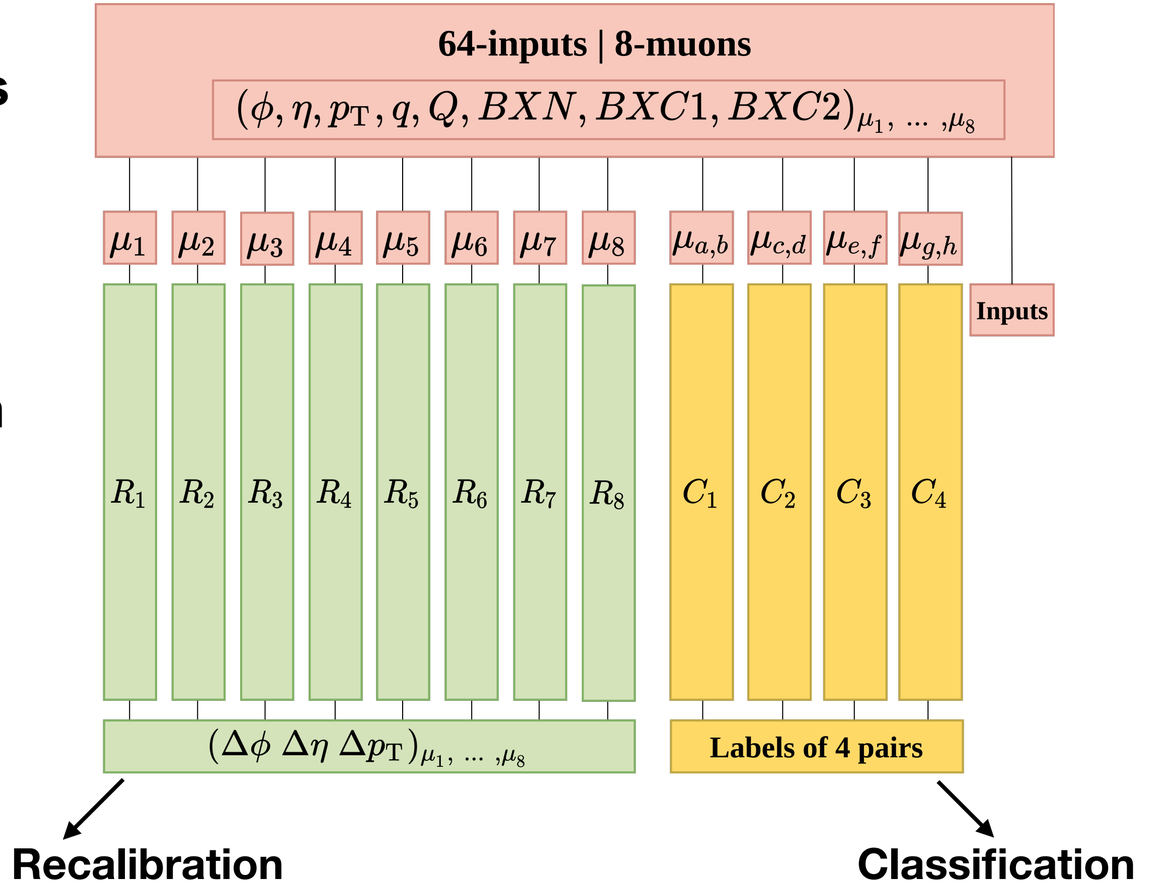




Network architecture improvements

- **Re-shaped network into convolutional structure** to **maximise efficiency** of MDLA resources
  - i.e max parallelism, 8 muons (full BX) at once
- **8 recalibration branches & 4 classification branches**
- One recalibration & one classification branch trained
  - *Mega-network* constructed by duplication of weights
- Copy of inputs forwarded to output for later matching on orbit number, bunch crossing

→ Throughput up $\times 2$ & combined w/ classification since last year



SB-852 resource utilisation & throughput

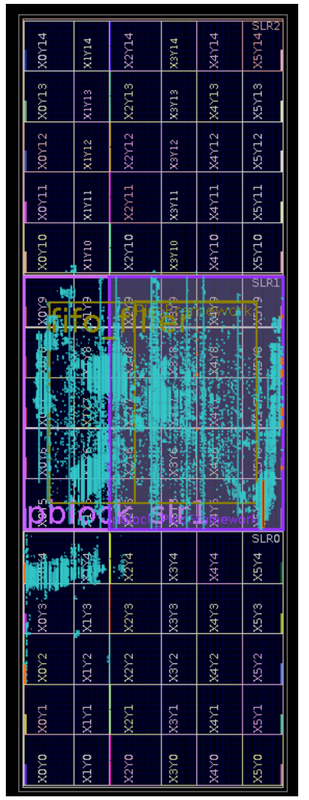
VU9P - MDLA w/ VFP

N DLA clusters	LUTs [%]	BRAM [%]	URAM [%]	DSP [%]
0	2.72	28.10	0.21	0
1	21.61	28.96	6.88	16.10
2	29.95	43.70	13.33	32.02
3	38.08	53.24	20.00	47.94

Extra readout buffers needed w/o DLA

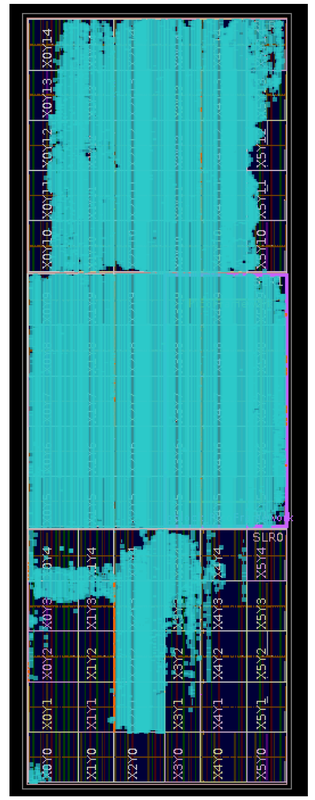
N DLA clusters	Inference rate	Average latency / muon inference	Encoding
4 cluster	5.2 MHz	192 ns	Q8.8
2 cluster	2.6 MHz	385 ns	Variable Fixed Point (VFP)

SB-852 infrastructure + L1 scouting firmware



Mostly confined to SLR1

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

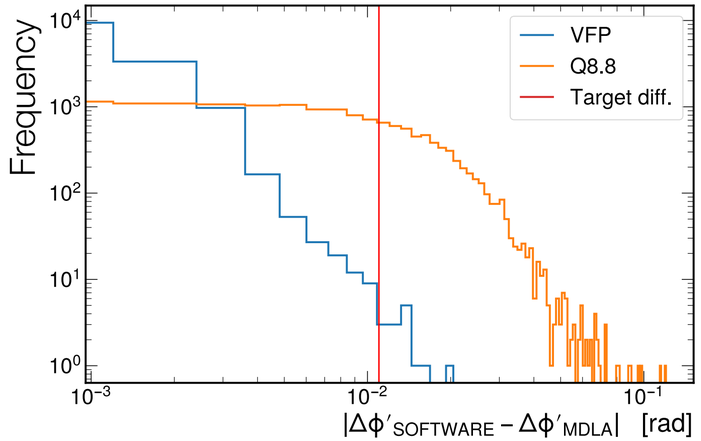


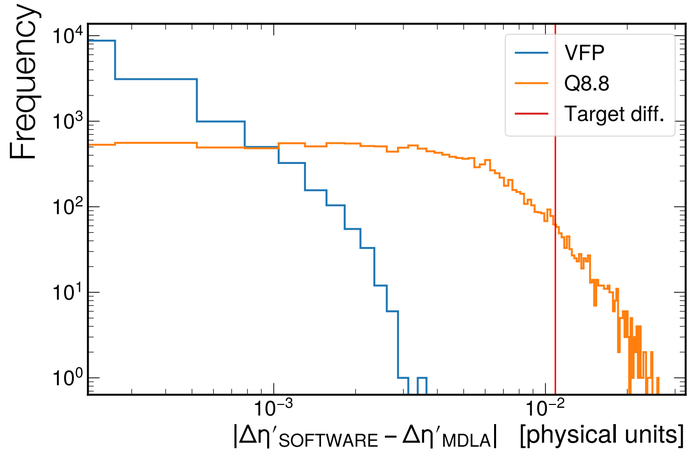
- Not able to fit 4 clusters w/ VFP
- **Throughput up x2 from previous year** as a result of increased efficiency i.e convolutional shape

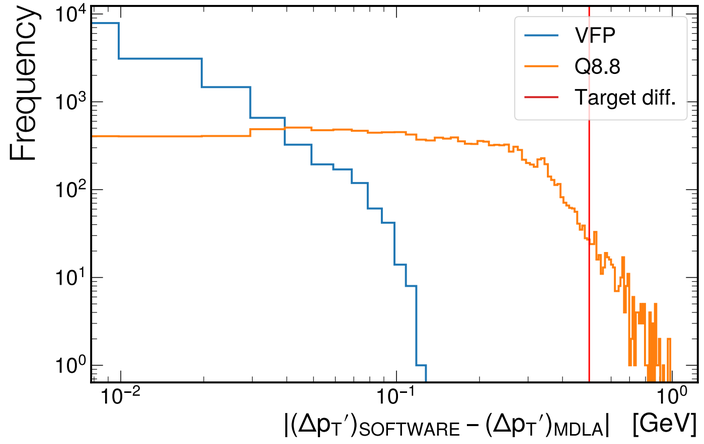
MDLA precision

- 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 / 64
  - 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%







Next steps: online processing

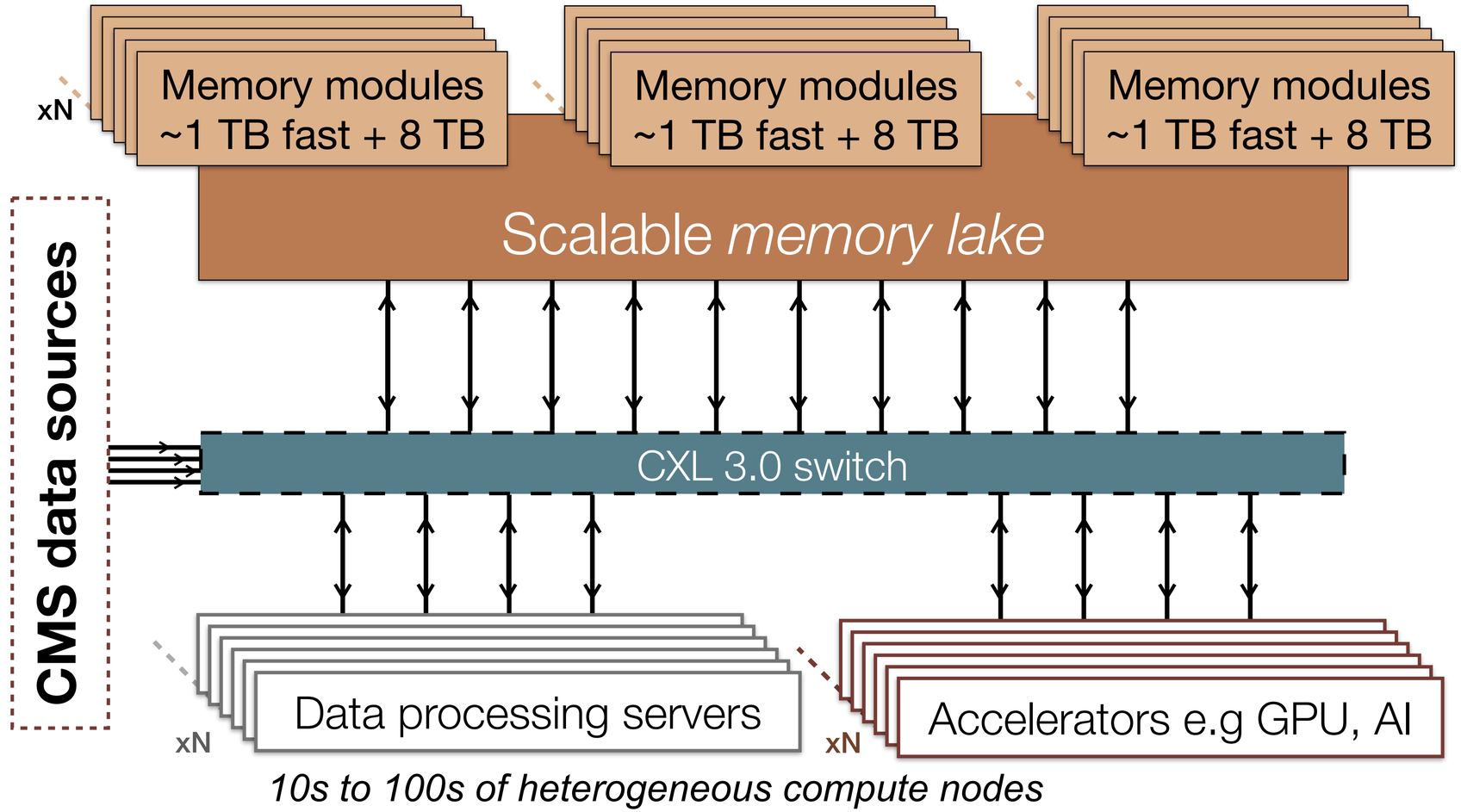
CERN data centre, 2016

Online processing

- Would like to perform real-time data analysis on incoming data streams
- Data reduction by e.g extracting topological information, fake/zero rejection, invariant mass, histogramming - cpus and co-processors - store only analysis products

Input
O(few Tb/s)

Output
O(few Gb/s max)



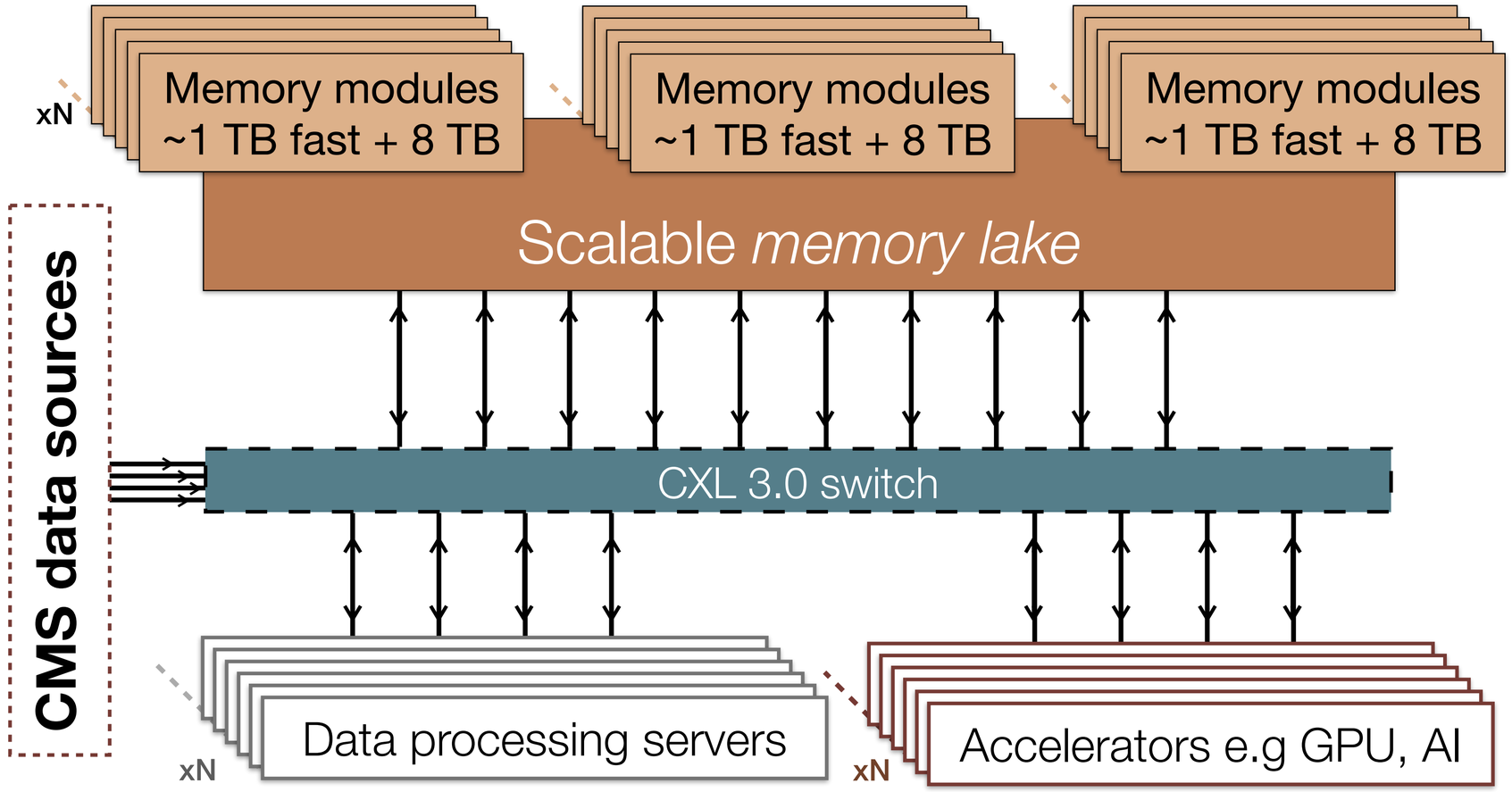
10s to 100s of heterogeneous compute nodes

Online processing

- Take advantage of next generation of Micron memory
- CXL 3.0 interconnect: builds upon PCIe, memory coherency between CPU & devices

Input
O(few Tb/s)

Output
O(few Gb/s max)



Summary

- L1 Scouting demonstrator system in operation, taking data from a multitude of L1 systems
- Fast machine inference for re-calibrations and fake reduction with MDLA and SB-852 achieved

- **Next steps**: look at back-end system
  - perform real-time data analysis on incoming data streams
  - *memory lake* concept w/ CXL 3.0 interconnect