

ML for event selection in data collection

Roope Niemi



NextGen
Next Generation Triggers

Triggers and Machine Learning

[Source](#)

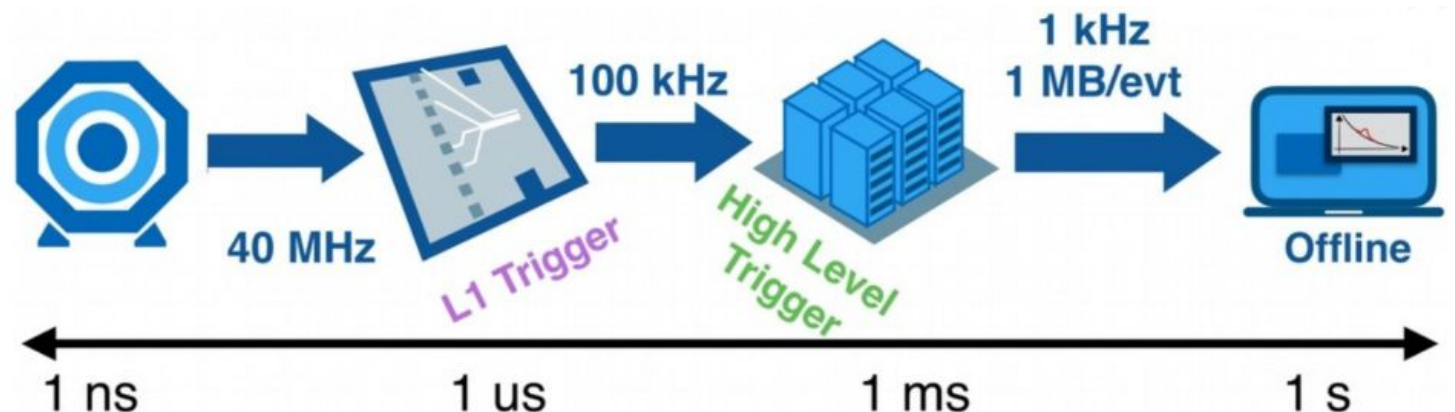
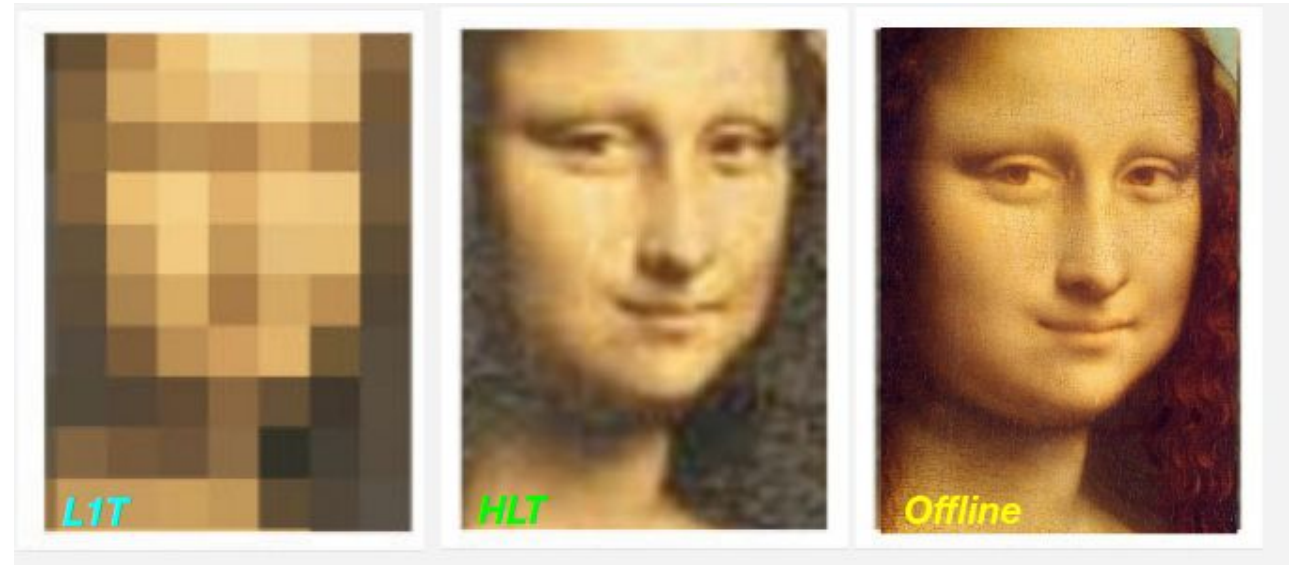
In LHC, collisions at **40MHz** frequency

Triggers are used to filter the data, reducing the data rate

99.75% of data filtered already in **L1T**. An event is lost forever if it is not found here!

Strict latency constraints

L1T is hardware based, makes decision using very coarse grained information

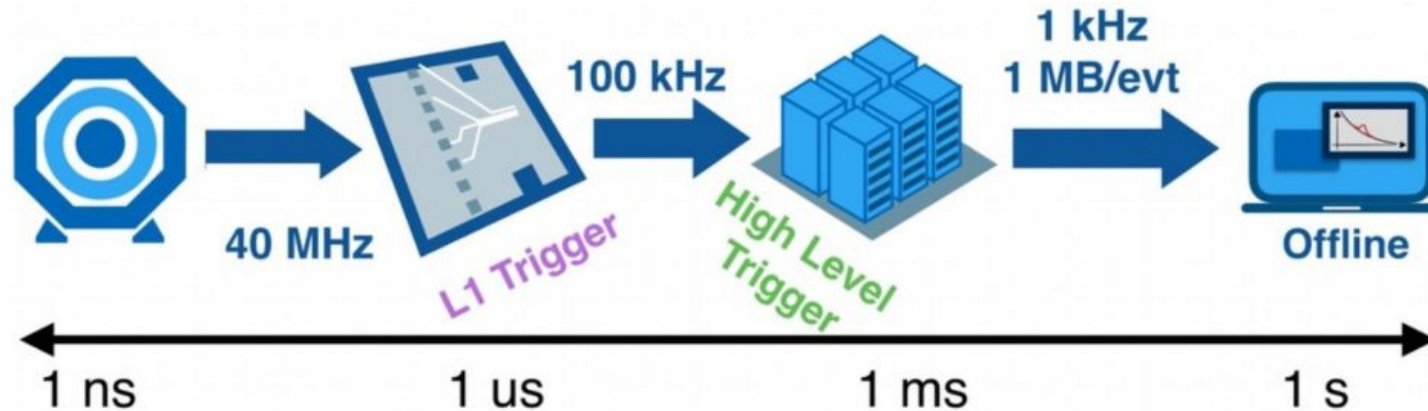


Triggers and Machine Learning

To improve pattern recognition and keep up with increasing data rates in the trigger,¹ use ML

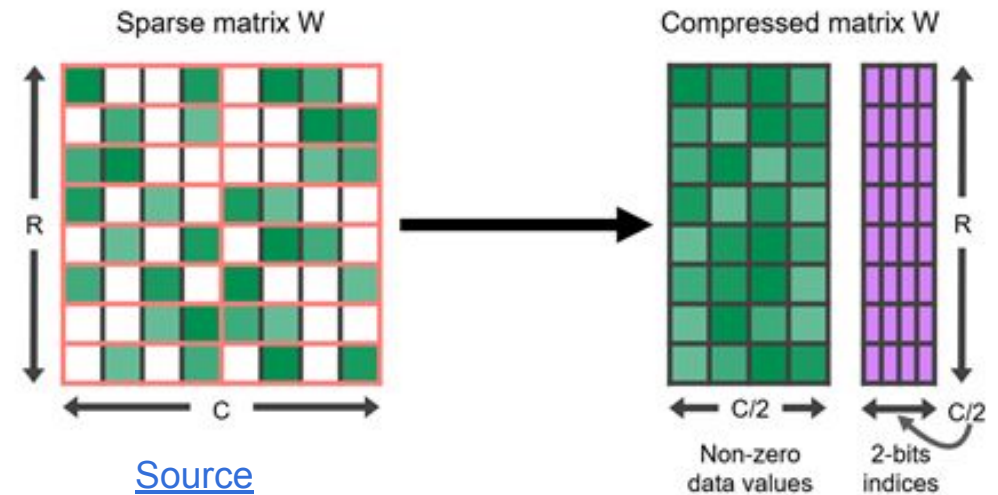
Deploy ML models on **FPGAs** to achieve nano-second inference latency

FPGAs have limited resources. Model architectures and their deployment to FPGAs need to be heavily optimized with no room for redundancies. **Compression of ML models!**



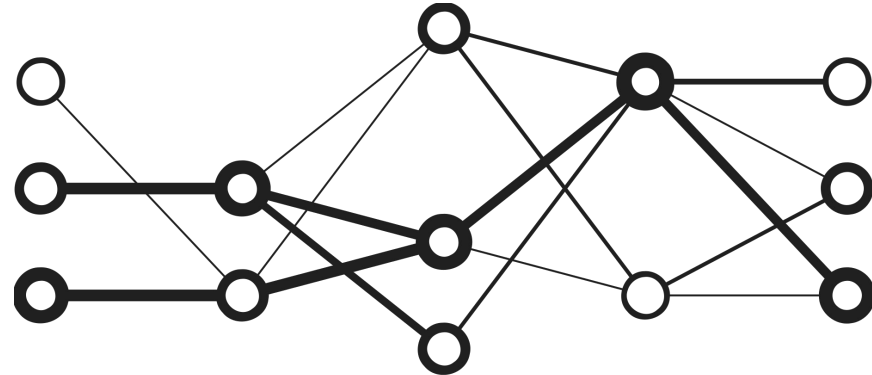
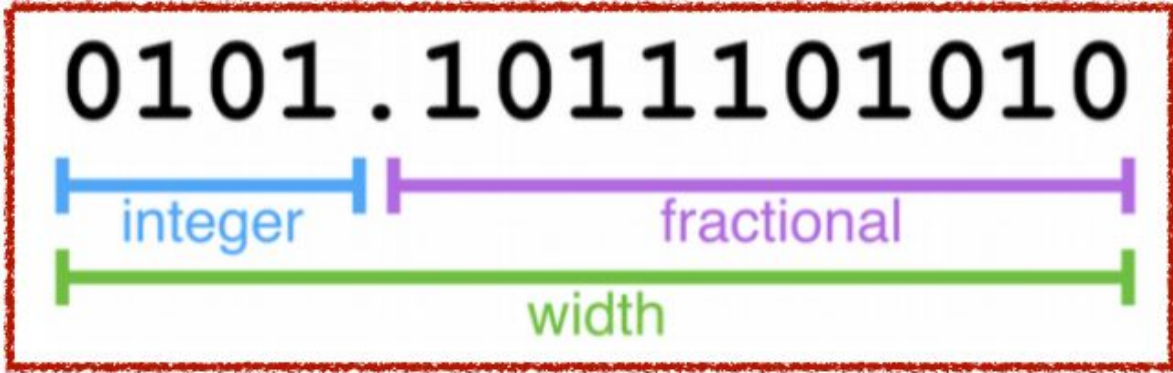
ML model compression

- An umbrella term for multiple techniques and their subcategories:
 - Quantization, pruning, distillation, low-rank approximation etc.
- During training, post-training
- Different granularities of compression:
 - Unstructured / N:M / structured pruning
 - Per-tensor, per-channel, per-weight quantization, HGQ
- Constraints by hardware



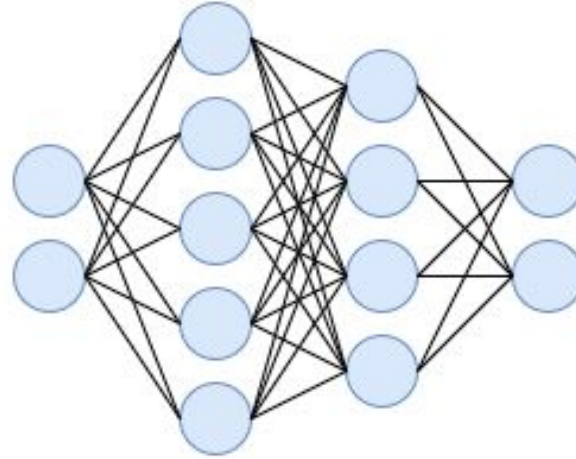
| W | | | $Q(W)$ | | |
|------|------|------|--------|---|-------|
| 0.1 | 0.01 | 0.21 | 0.125 | 0 | 0.25 |
| 0.02 | 0.01 | 0.83 | 0 | 0 | 0.875 |
| 0.77 | 0.03 | 0.01 | 0.75 | 0 | 0 |

Quantization and pruning

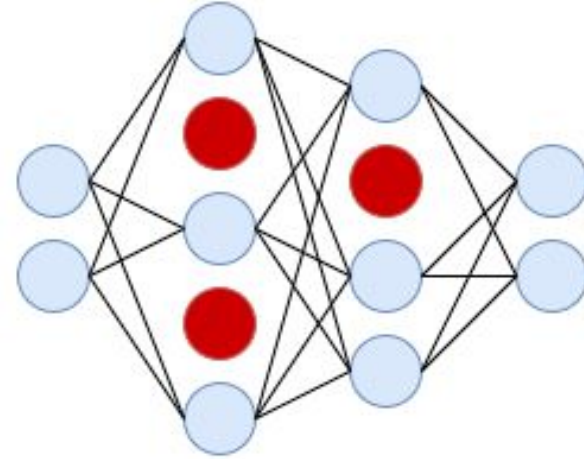


HGQ II

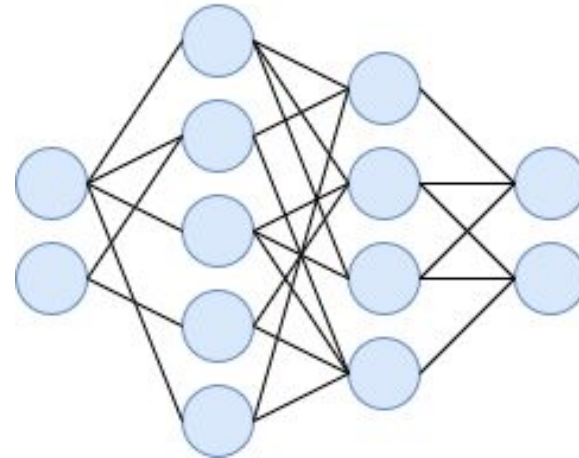
Dense model



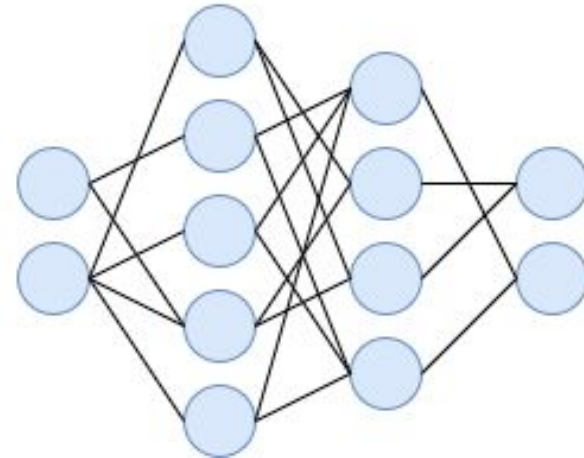
Structured pruning



Unstructured pruning

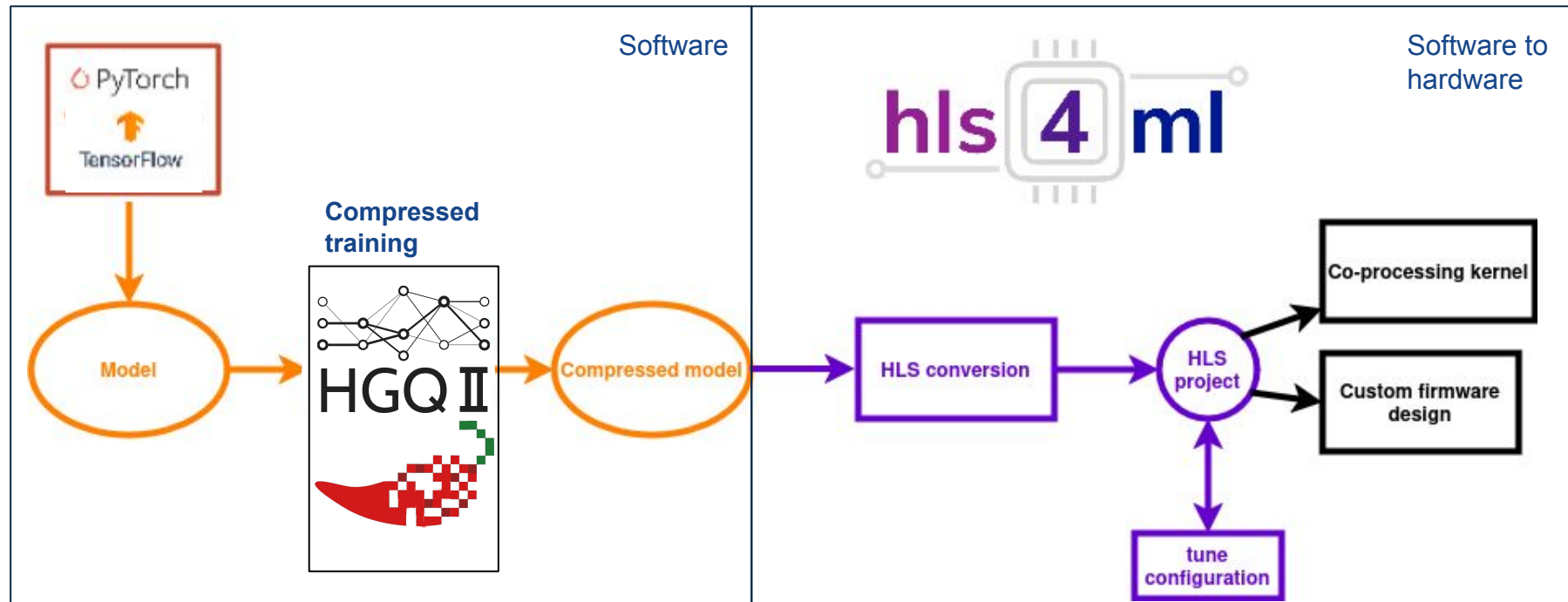


2:4 pruning



Model to FPGA

- [HGQ](#): A quantization library that learns individual, per-weight / per-activation bit-widths. Can learn a bit-width of 0 -> prune
- [PQuantML](#): brings various pruning and quantization methods, including HGQ, under one library. Allows both the individual use of different methods and the combined use of different methods



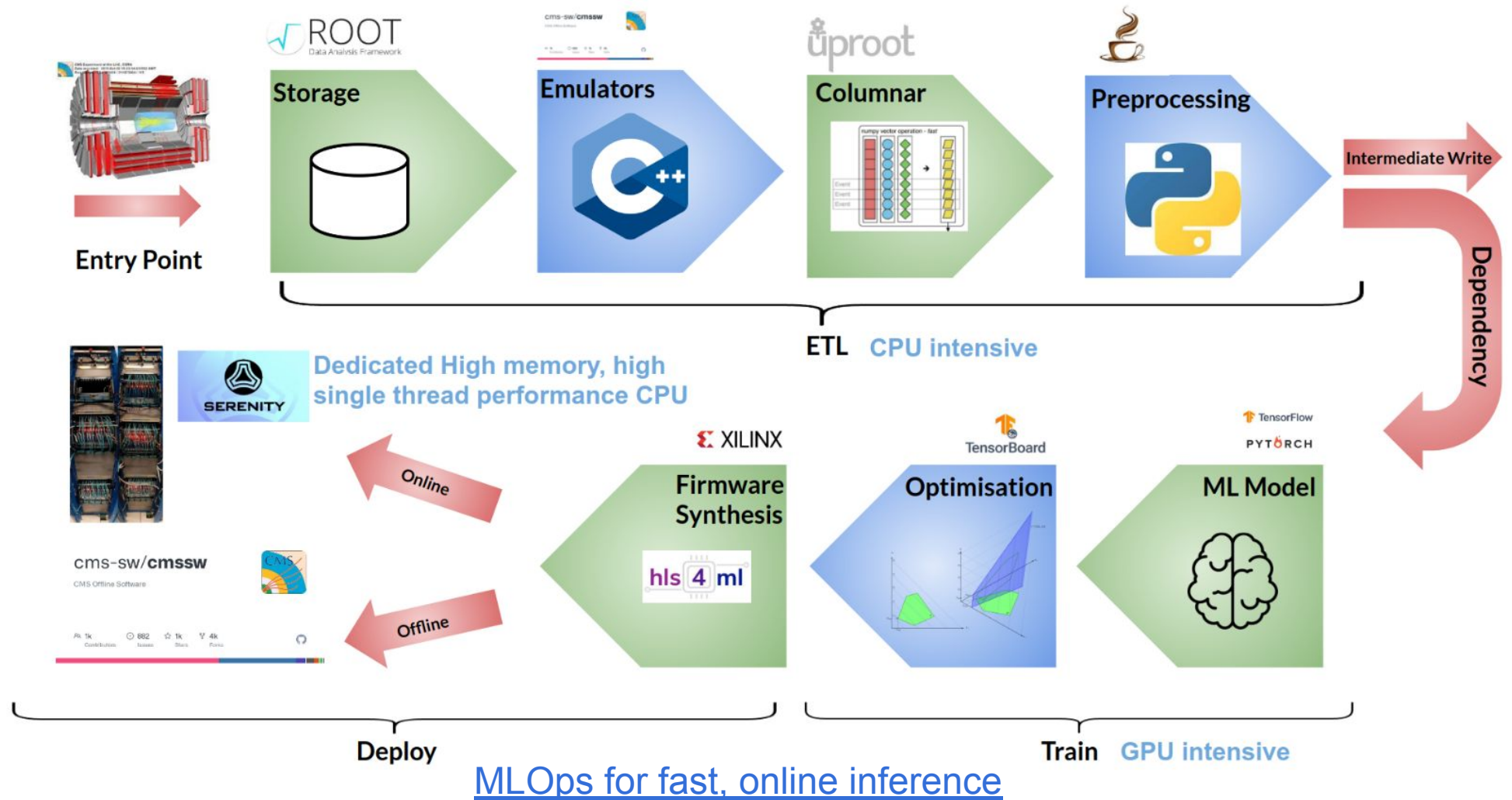
[hls4ml](#): Used to translate trained ML models to optimized HLS code. Synthesize into FPGA bit stream, run on FPGAs

Deployment pipeline

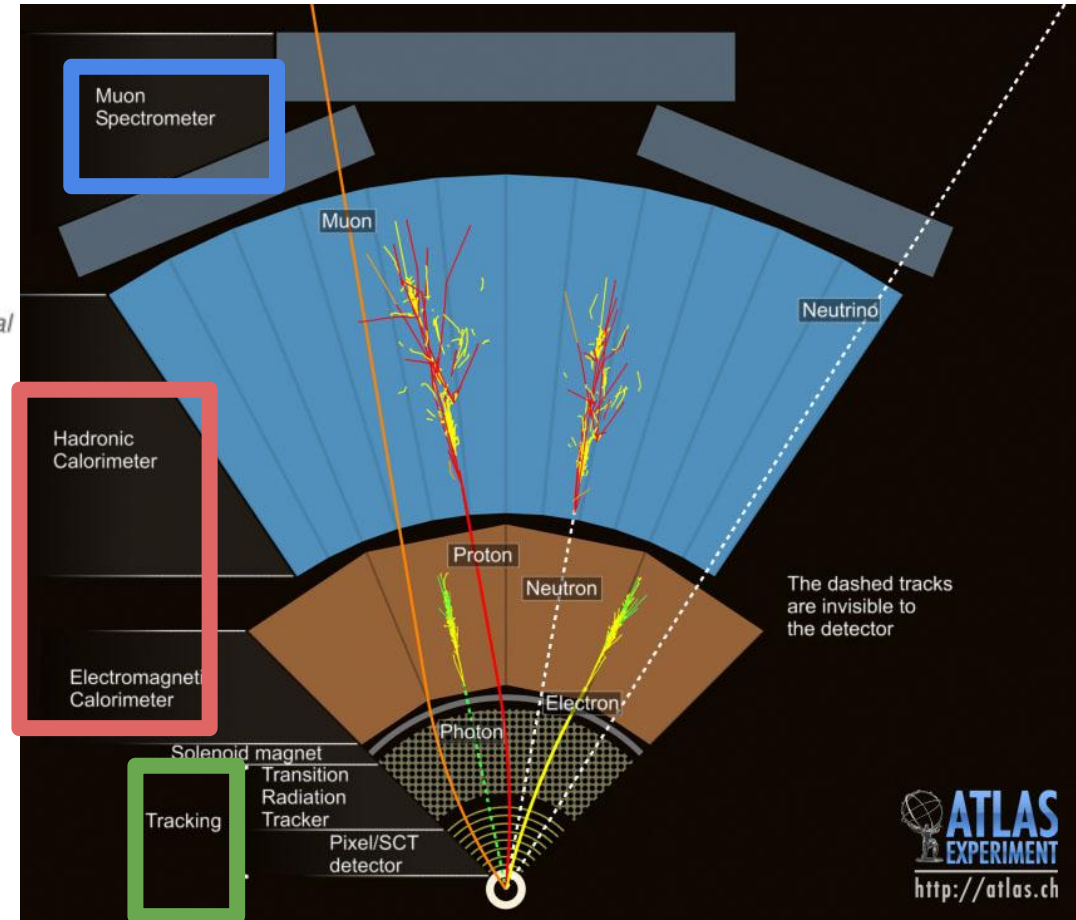
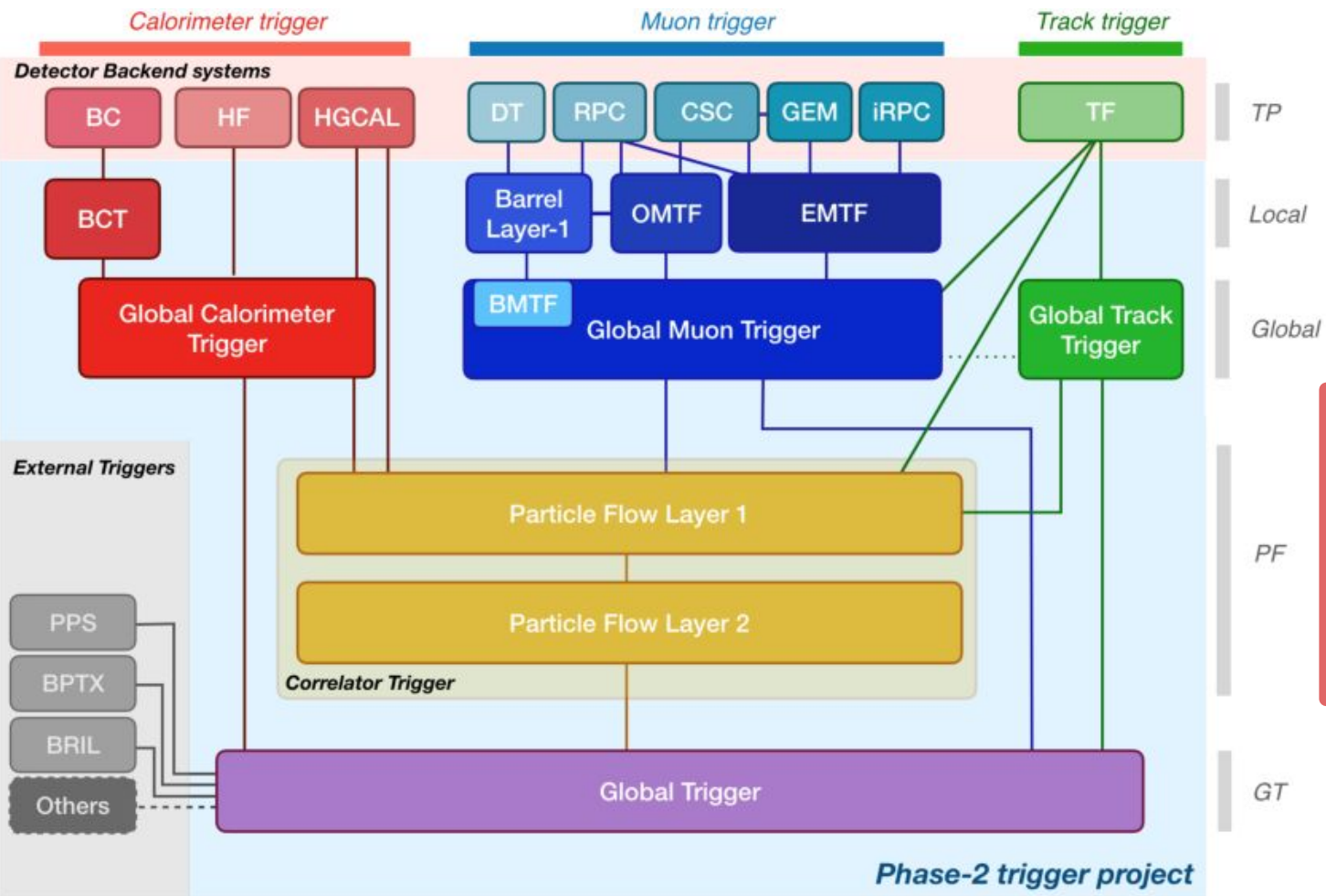
Data format: h5, Parquet, Numpy, Torch/TF Tensor, ROOT, pattern file (raw binary)

FW synthesis and validation

- Synthesise FW
- Validate HLS standalone
- Validate SW emulator
- Validate FW with test vectors



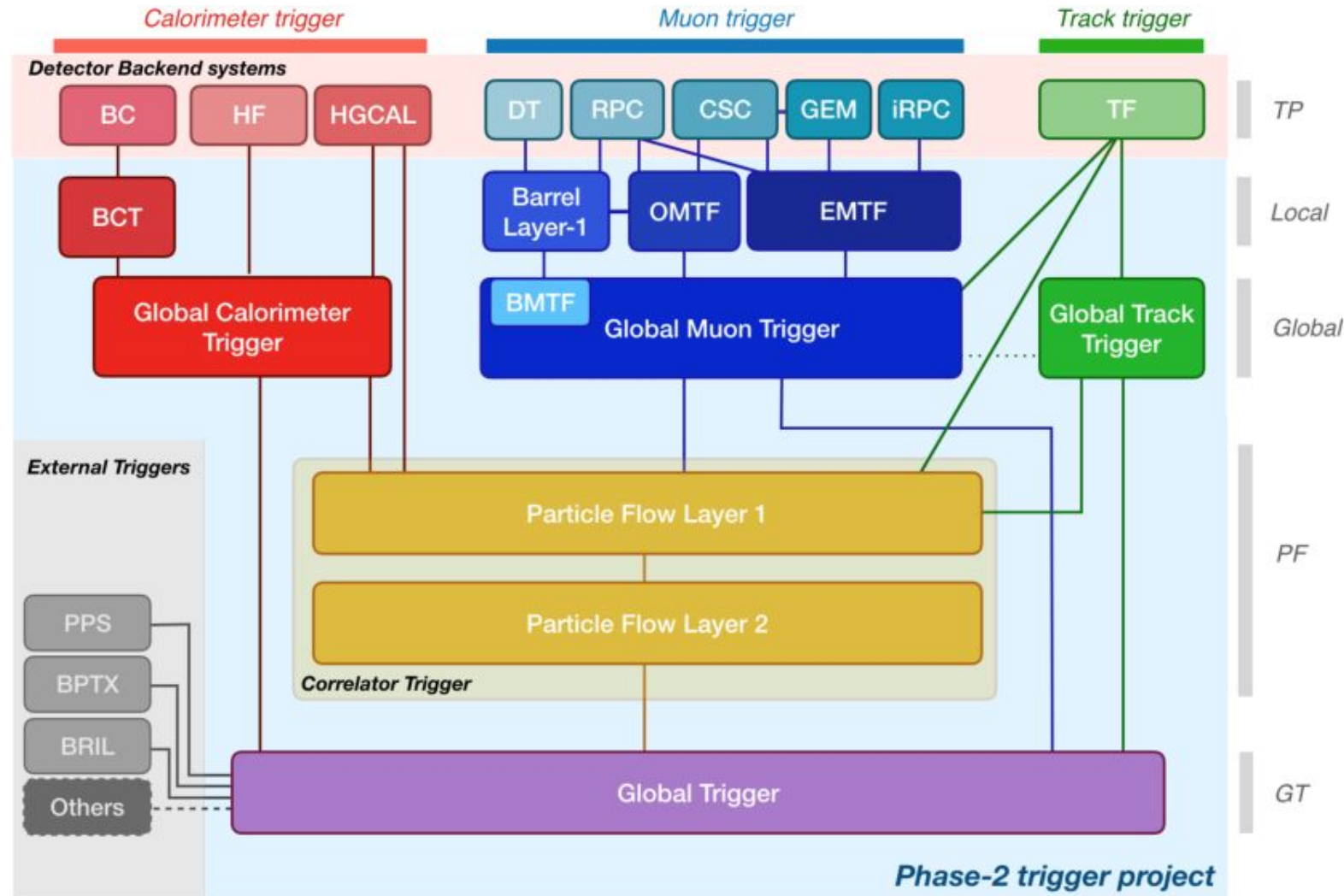
L1T



L1T

- The L1T receives condensed event information in the form of trigger primitives.
- Dedicated subsystem modules reconstruct physics objects from varying detectors and/or regions.
- L1T is an all FPGA design hosted on custom boards interconnected via GB/s optical links.
- The final L1T decision (L1T accept/reject) is propagated to the Data Acquisition System (DAQ).

[Source](#)

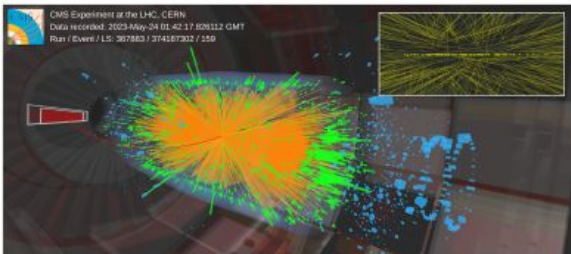


L1T

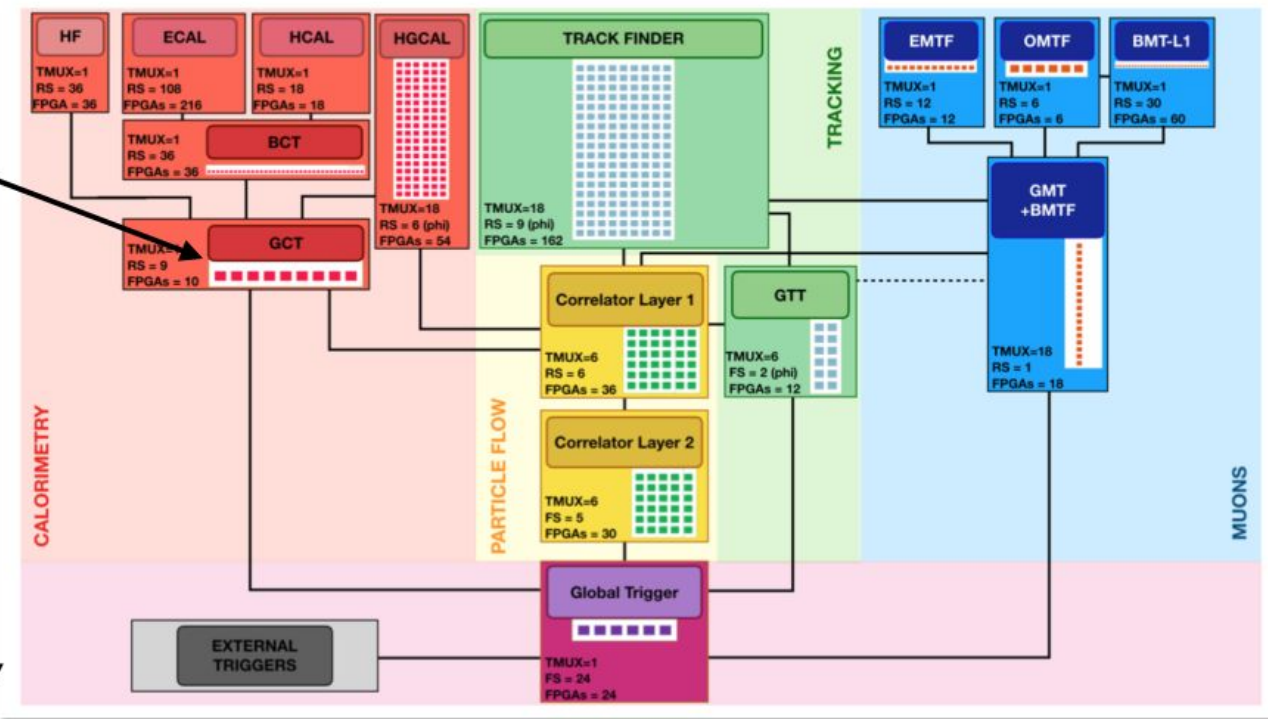
$t = 0$



1 small box = 1 FPGA board with AMD VU13P FPGA



$t < 12.5 \mu\text{s}$



0 μs

Detector hits

5 μs

Clusters & Tracks

6 μs

Particles

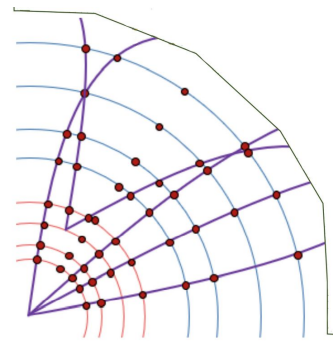
7 μs

Event Categorisation

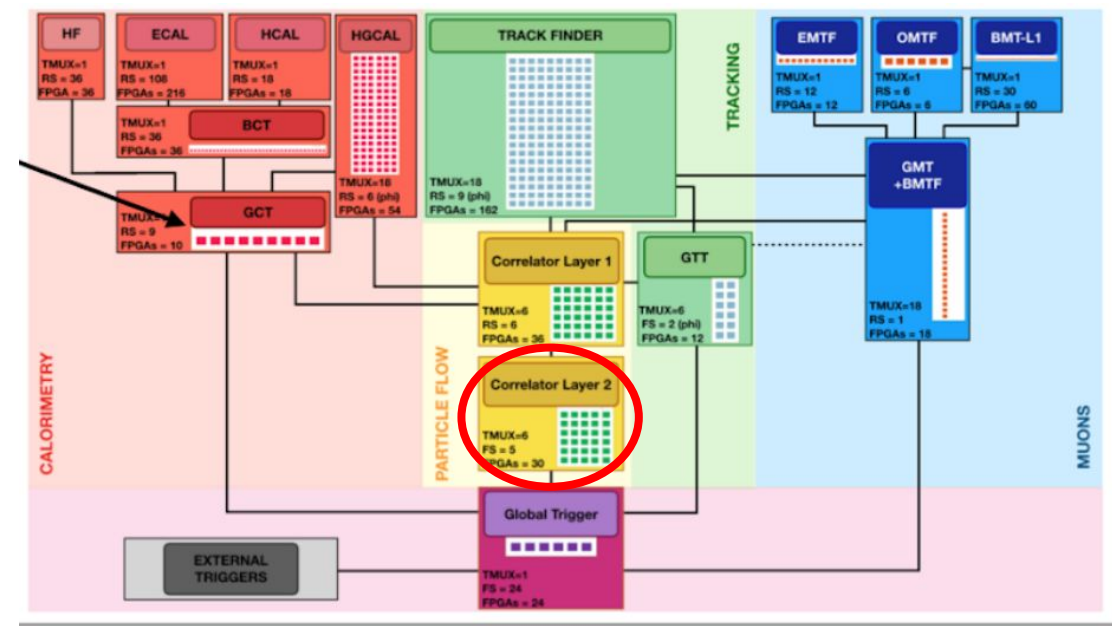
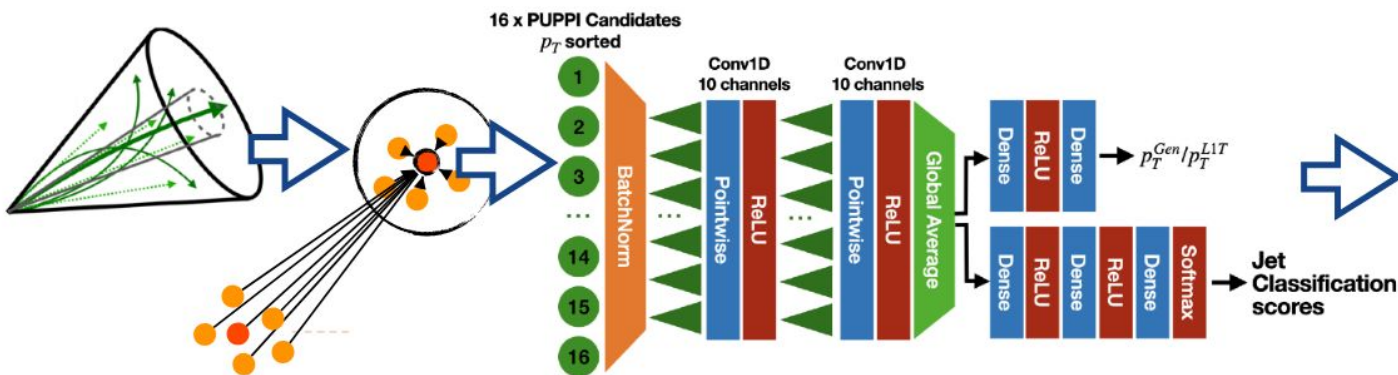
8 μs

1 bit: keep / discard

Jet Tagging

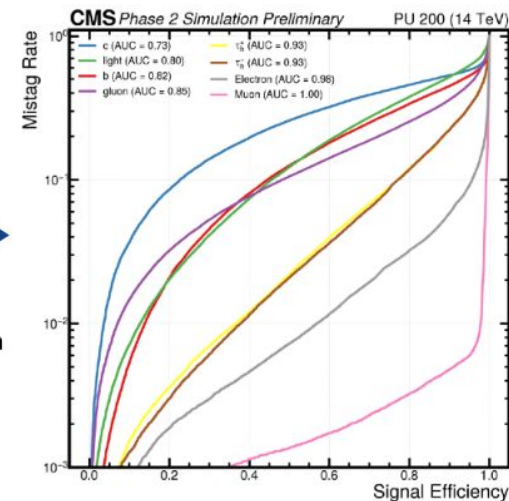


- Inputs from PF + PUPPI, some preprocessing
- Identify type of quark / gluon that initiated the jet
- Final-state particles do not have a natural order. Artificial ordering may affect model performance ([source](#))
- The model architecture used, Deep Sets, is permutation-invariant
- Send all category scores to the Global Trigger



| Component | Latency (μs) |
|----------------------------|---------------------------|
| Preprocessing (incl. sort) | 0.10 |
| Jet Tag NN | 0.19 |
| Jet Reco | 0.74 |
| Total System | 1.01 |

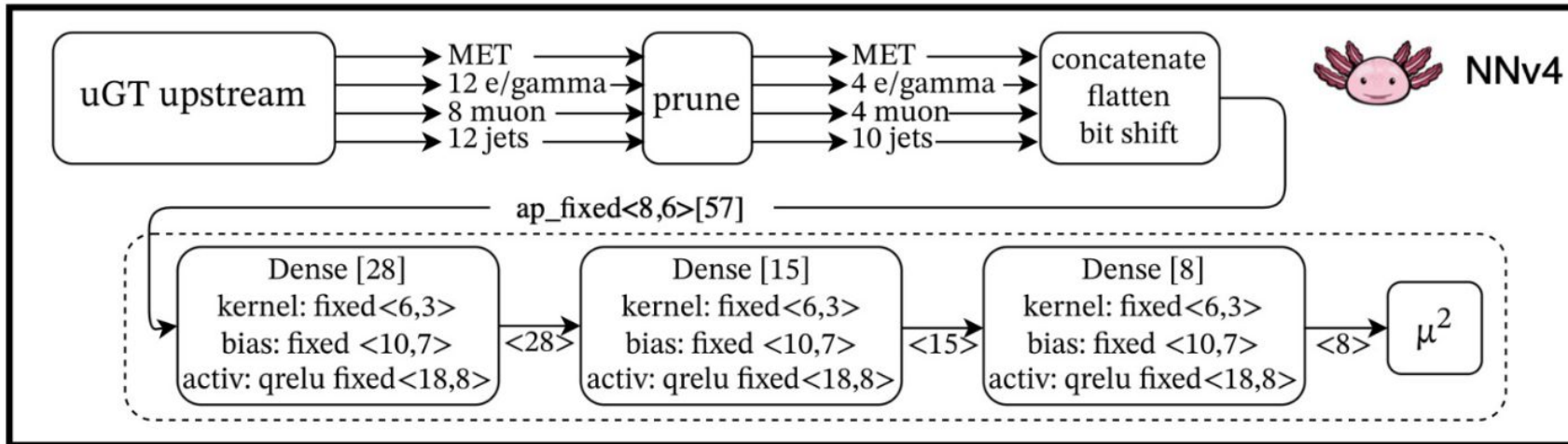
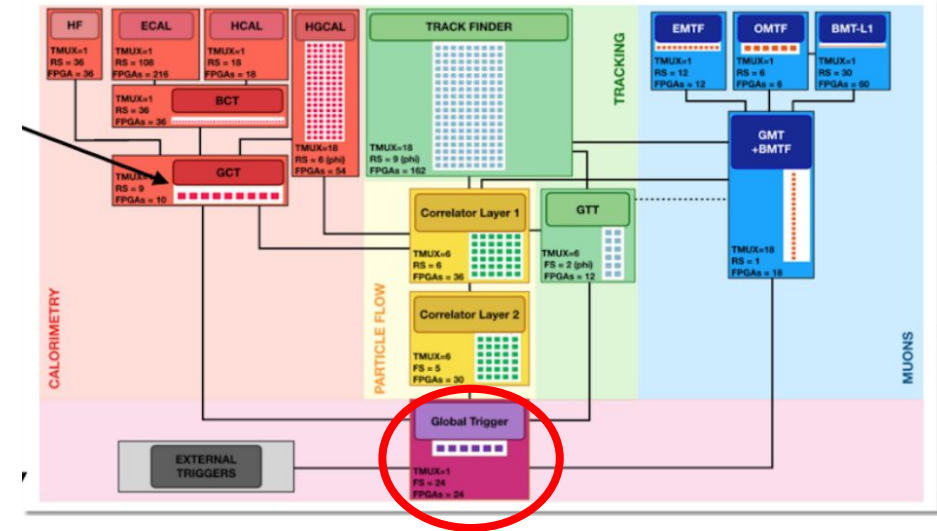
[Source](#)



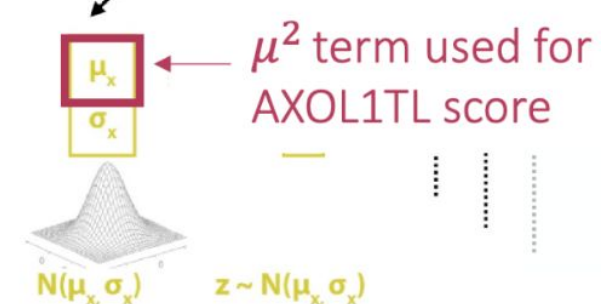
Anomaly detection in GT

AXOL1TL:

- Event level trigger that uses GT input objects
- Latency requirement of 50ns
- L2-norm of the latent vector (size 8) used as the anomaly score, deploy only the encoder



Regularize latent space to avoid overfitting

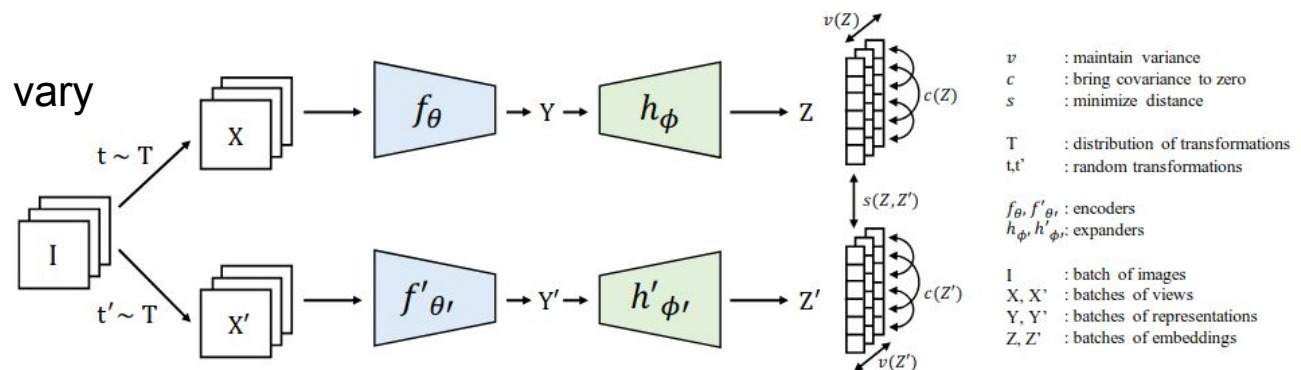
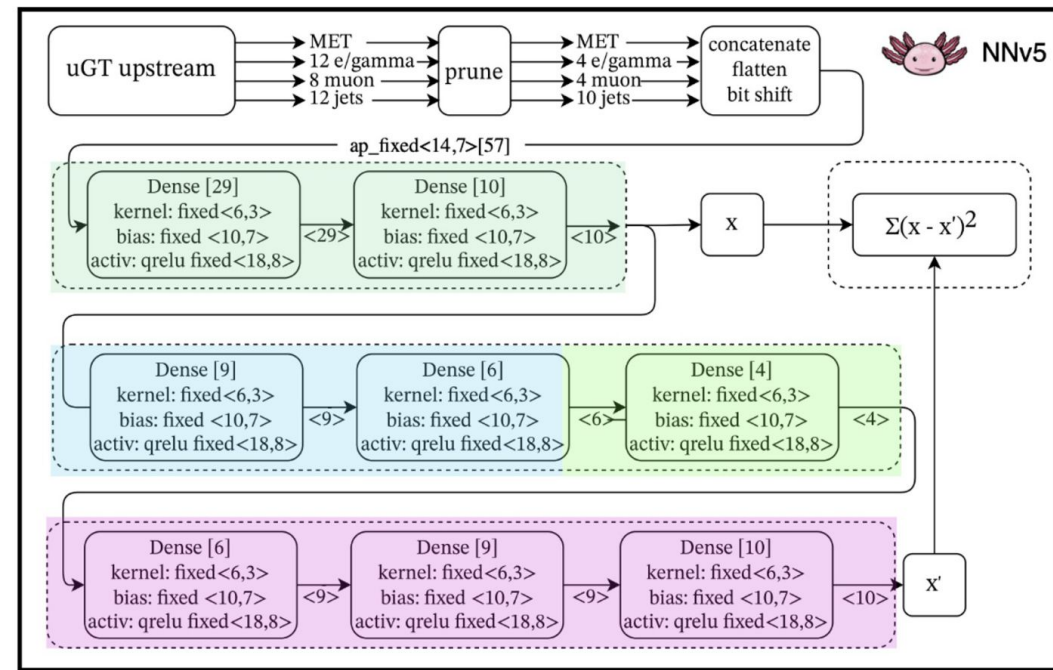


$$\text{loss} = ||x - \hat{x}||^2 + \text{KL}[N(\mu_x, \sigma_x), N(0, I)]$$

Anomaly detection in GT

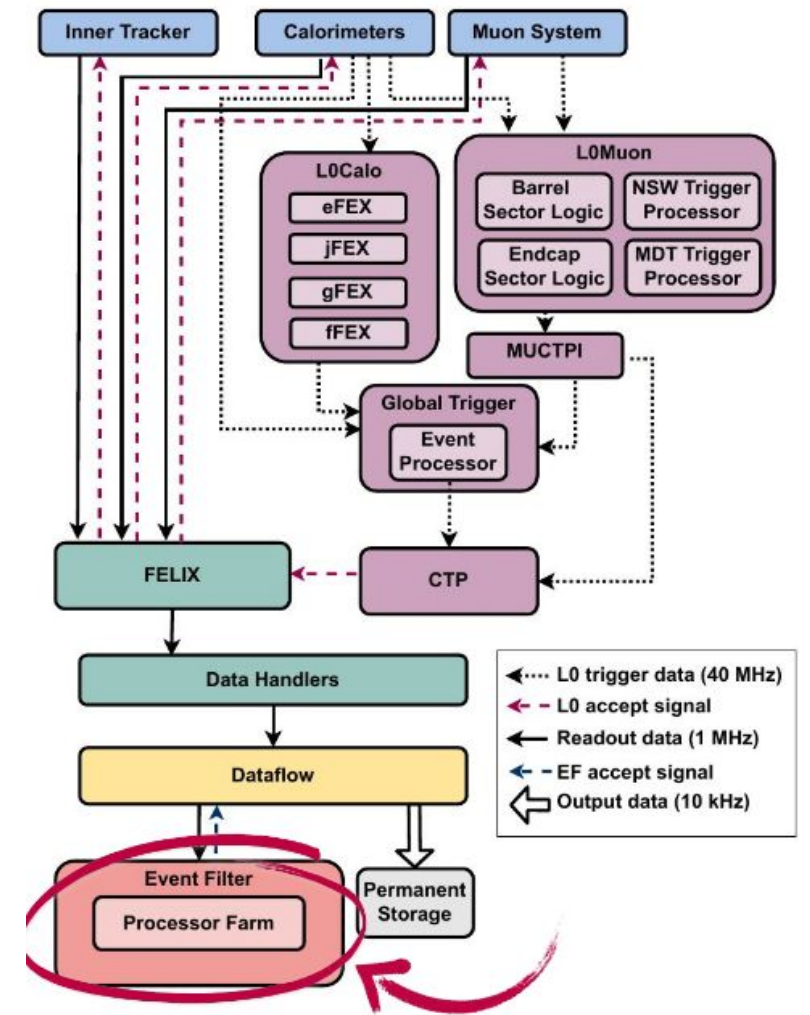
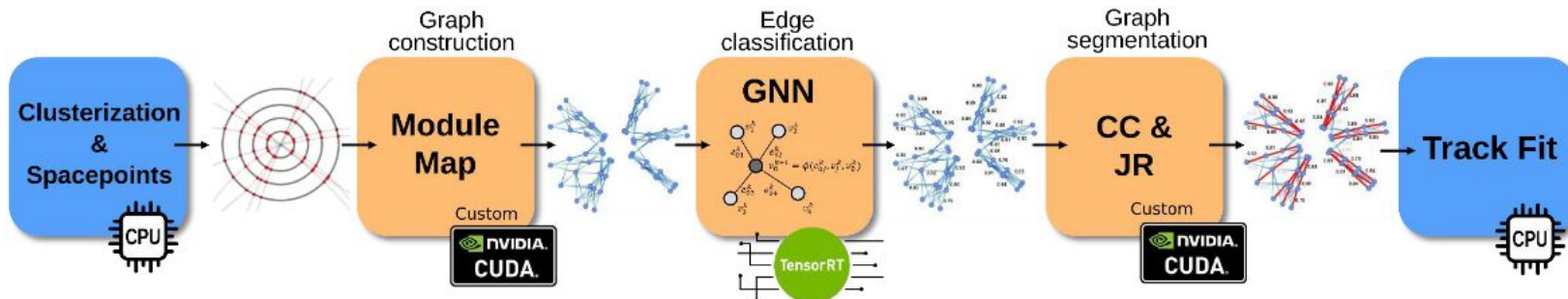
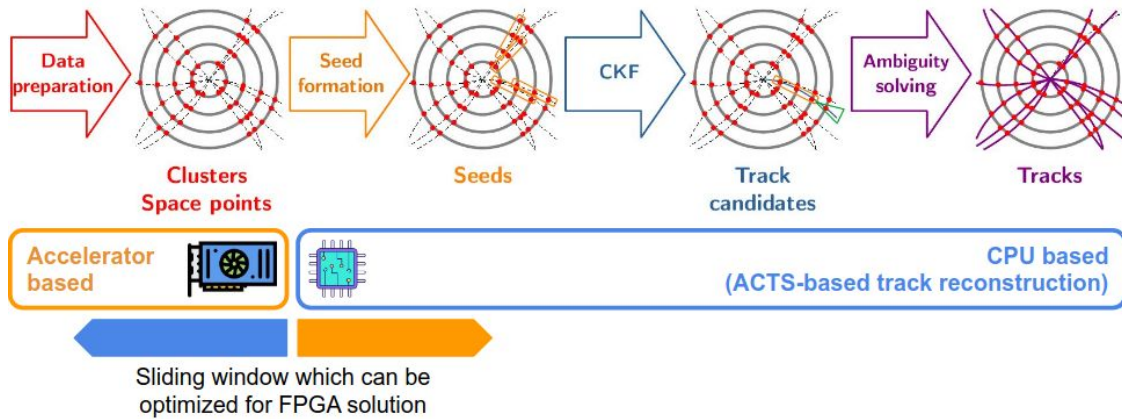
AXOL1TL V5:

- Stack a VICReg trained feature extractor on top of the VAE.
- **VICReg loss:**
 - **Invariance:** MSE between embedding vectors
 - **Variance:** a hinge loss to maintain the standard deviation (over a batch) of each variable of the embedding above a given threshold. This term forces the embedding vectors of samples within a batch to be different.
 - **Covariance:** a term that attracts the covariances (over a batch) between every pair of (centered) embedding variables towards zero. This term decorrelates the variables of each embedding and prevents an informational collapse in which the variables would vary together or be highly correlated.



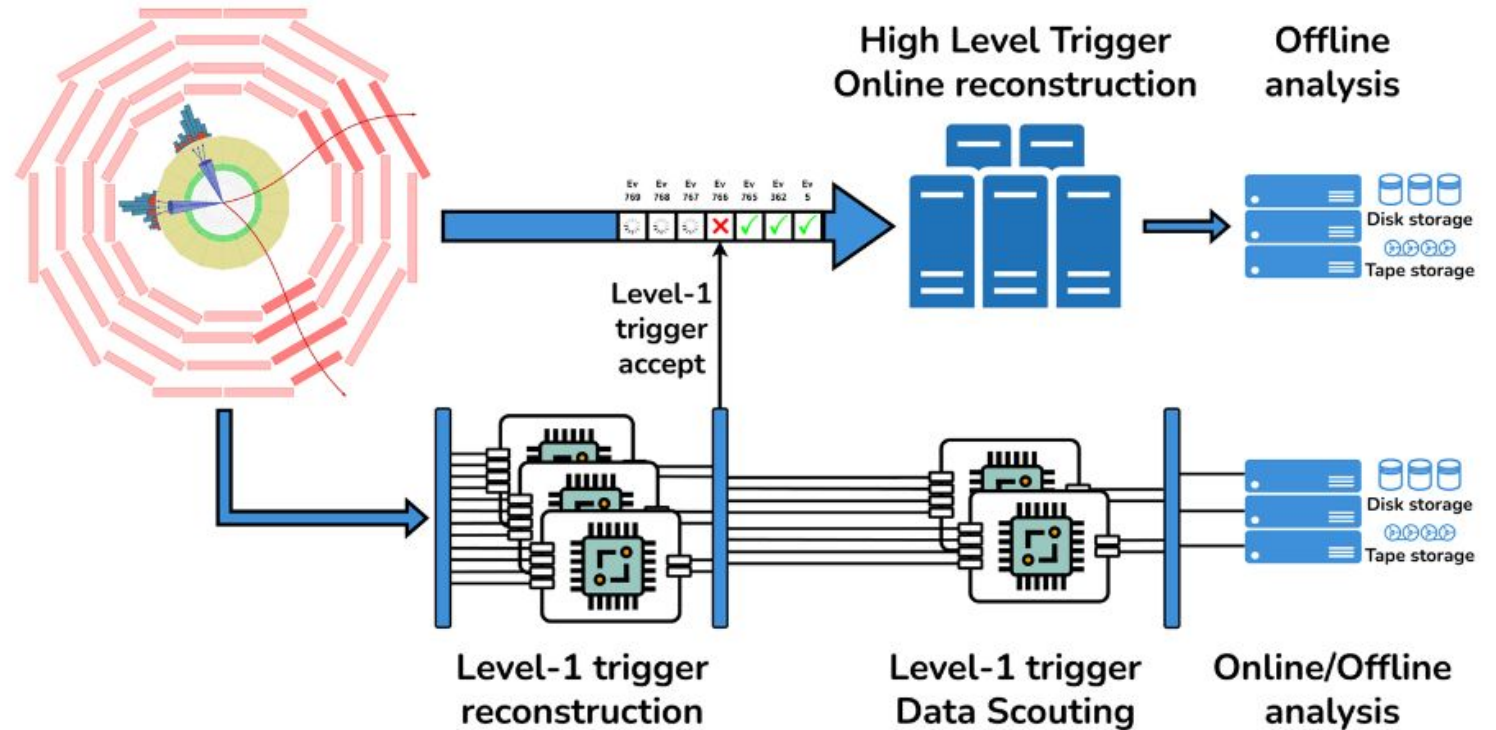
Event Filter Tracking

- ITk data preparation, track seeding and pattern finding, track fitting and ambiguity solving



L1T-Scouting

- **Capture the objects reconstructed by L1T, run physics analysis at 40MHz**
 - Large “irreducible” backgrounds
 - Identification can’t fit the L1 fixed latency or resource constraints
 - Identification requires data from multiple LHC bunch crossings
- **Study signatures evaded by the L1T**



[Source](#)