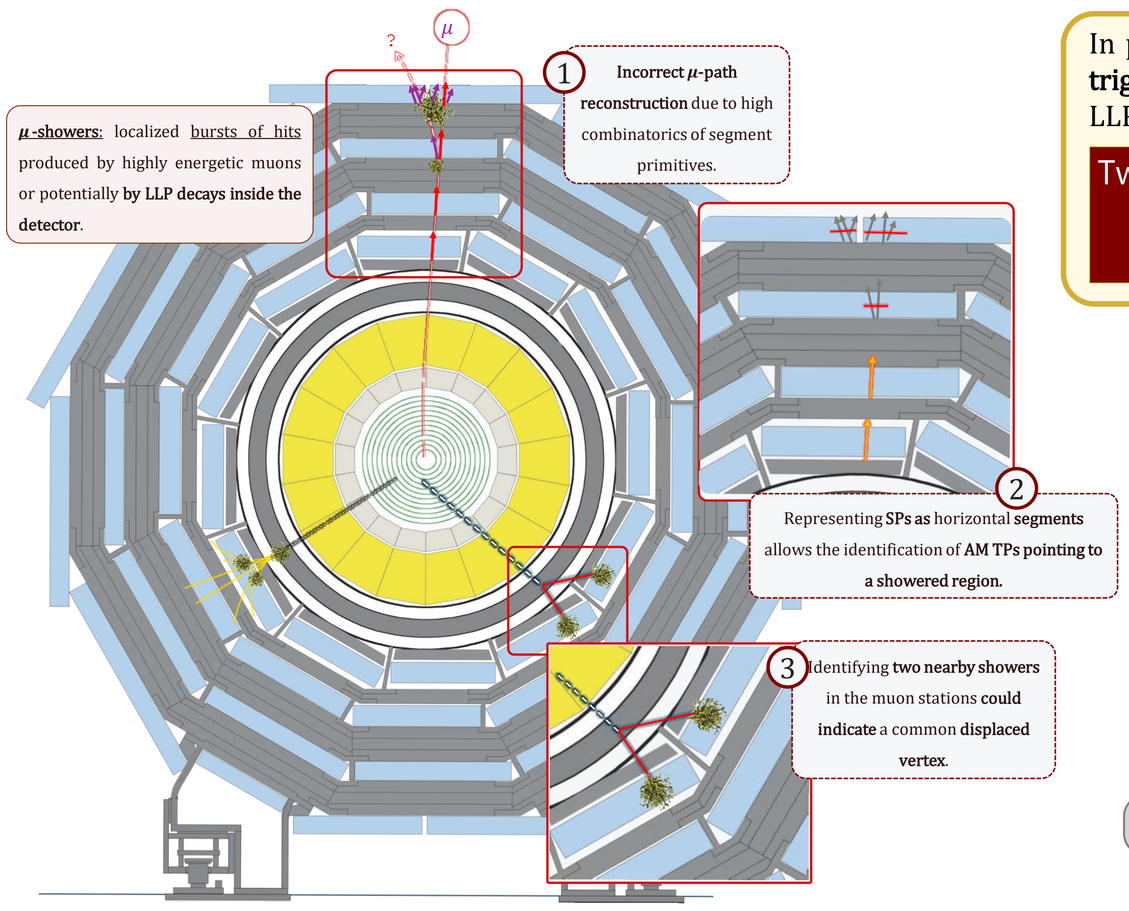
# FPGA-based muon shower identification and graph neural network tracking algorithms for HL-LHC triggers

E. Aller, A. Cardini, J. del Riego, D. Estrada, S. Folgueras, P. Leguina, J. Prado
Institute of Space Sciences and Technologies of Asturias (ICTEA), University of Oviedo, Spain.

International Workshop 27th iWoRID on Radiation Imaging Detectors
Universidad de Oviedo




In preparation of the HL-LHC, the CMS detector will entirely **upgrade the electronics for the hardware (L1) trigger** [1], allowing to **deploy complex algorithms** to identify and reconstruct rare signatures arising from LLPs with unprecedented precision.

Two complementary strategies are being developed:

- **Muon shower tagging** in the Barrel Muon Trigger.
- **GNN-based muon tracking** exploiting the full detector granularity for the overlap region [5–7].

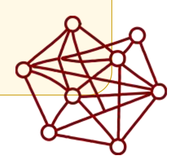
Raw hits

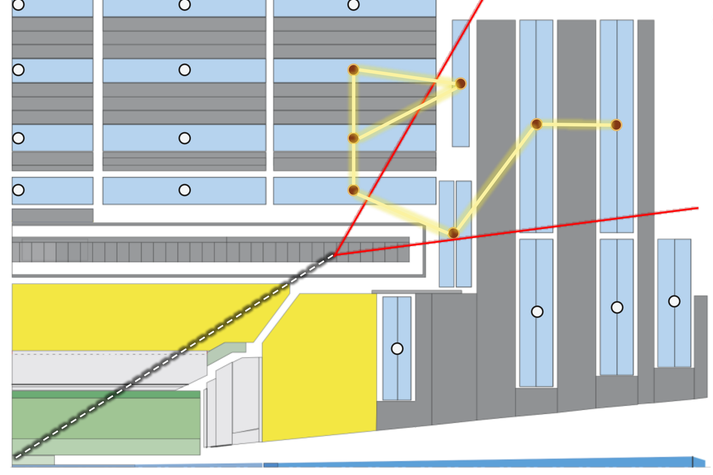
Endcap Stubs

Barrel Stubs

Detector **stubs** (OMTF Hybrid stubs) can be represented as **Graph Nodes** and then connected through **geometric compatibility** as **Graph Edges**.

**Edges attributes** : Layer, $\eta, \phi, \Delta\eta, \Delta\phi$. (normalized)



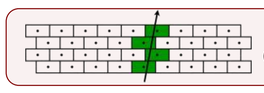


DT primitives

Graph

## BMTL1 Barrel Layer-1

### Analytical method



Optimized algorithm to track individual $\mu$ paths [3]

### Shower algorithm

Monitors **hit multiplicities** within a 400 ns time window (to consider the drift time) and generates compact **shower primitives** (SP) encoding timing and spatial information when predefined threshold is exceeded (Fig. 1) [4].

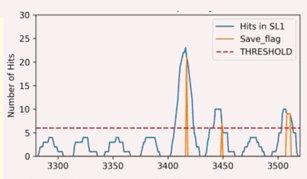
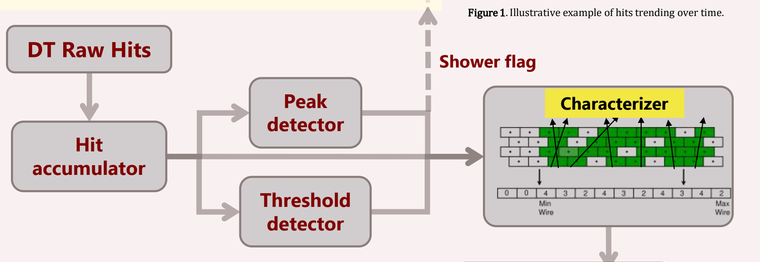



**Figure 1**. Illustrative example of hits trending over time.




| LUTs | FF | Latency |
|---|---|---|
| 2.5 k | 2.5 k | 20 ns |

- **Ultra-low resources** consumption.
- Algorithm **latency sub-BX**.
- Firmware-emulator **agreement >95%**.
- Simulated data based studies (**preliminary**) shown approximately **80% of accuracy** and almost **10 times less local trigger rate** than the expected for the **AM algorithm.** Current studies prioritise recall (Fig 2).

| Accuracy | Precision | Recall | F1 score |
|---|---|---|---|
| 0.77 | 0.49 | 1.00 | 0.65 |

**Figure 2.** Performance metrics estimations.

Algorithm was **deployed** this year during **slice tests** to evaluate performance **during data-taking** (under study).

* Metrics are defined such as:
- Accuracy $= (TP + TN) / (TP + TN + FP + FN)$ – Overall correctness
- Precision $= TP / (TP + FP)$ – of all predicted positive, how many correct?
- Recall $= TP / (TP + FN)$ – of all actual positive, how many found?
- F1 Score $= 2 \cdot (\text{Precision} \cdot \text{Recall}) / (P + R)$ – harmonic mean · balances P and R

TP, FP, TN, and FN means **true/false positive/negative** cases respectively

## BMTL1 Barrel Filter

- Primitives can be **analyzed and correlated** here with **neighboring sector-stations**
- **Several algorithms**, each **targeting different signatures**, can be implemented.

### High-$p_T$ tagging

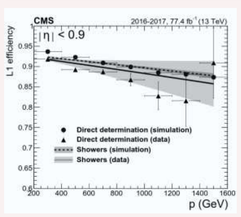
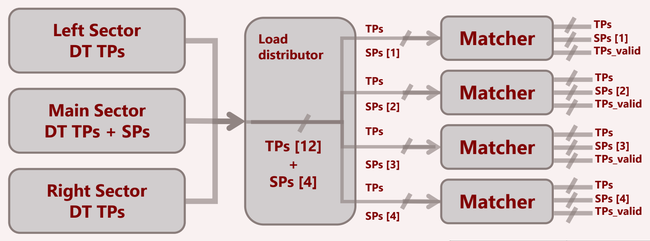



**Figure 3**. Plot showing how efficiency decreases with increasing muon momentum.

L1 trigger **efficiency losses** (Fig. 3) result from **faulty $\mu$-path reconstruction** (1 see ), which degrades reconstruction of kinematics. Consequently [2], **early identification of high–$p_T$ muon** could provide valuable tracking data, potentially preventing misreconstruction.

**To achieve this...**

Performs a **geometrical matching** between **AM primitives** and **SP segment** representations (see 2). This approach **flags muon** primitives **pointing toward shower regions**, providing a tag that downstream **track finders can utilize** to handle potential track loss [4, 8].




- Also **sub-BX latency. Pipeline able process 48 TPs per clk @480 MHz**

| LUTs | FF |
|---|---|
| 4.3 k | 5.6 k |

- Simulation results using only $\phi$ data shown that the proposed matching algorithm **successfully flags showering muons** with an efficiency **greater than 80%**. Further studies on the longitudinal axis are still pending ($\theta$).

### LLPs tagging

Under development... See 3

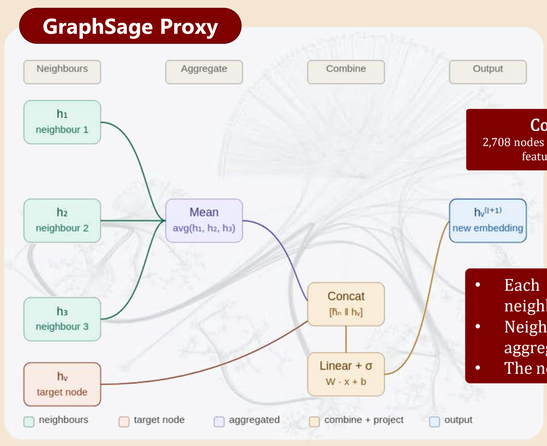
### Others...

## Can a GNN fit in the L1 Trigger?

FPGA deployment tested with a **two-layer GraphSAGE** model before testing more complex algorithms.

- Used a **Cora dataset** to test a hardware-friendly implementation
- Investigated weight quantization techniques to maximize algorithmic complexity while complying with low latency

**GraphSage Proxy**

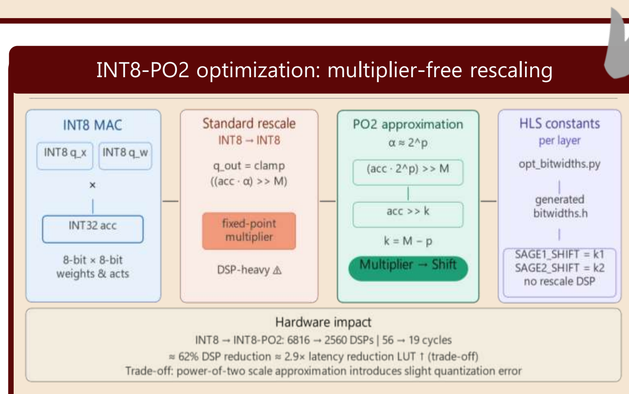



Tests with different weight quantization techniques found **best performance** with an **optimized version of INT8 with power of 2 bit-shift** [7, 9].

| Design | Acc. | Type | Latency (cyc) | Latency (ns) | DSP | FF | LUT | Freq. |
|---|---|---|---|---|---|---|---|---|
| Float32 | 78.0% | FP32 | 603 | 1,670 | 34,880 (283%) | 4.45M (128%) | 2.14M (123%) | — |
| Fixed-PTQ | 75.7% | Fixed | 77 | 213 | 6,976 (56%) | 262k (7%) | 110k (6%) | 397 MHz |
| INT8-PO2 | 75.0% | INT8 | 28 | 77.6 | 5,320 (43%) | 254k (7%) | 188k (10%) | 450 MHz |
| QAT-v2 | 76.8% | INT8 | 55 | 152.3 | 6,760 (55%) | 561k (16%) | 282k (16%) | 493 MHz |
| ★INT8-PO2 Opt. best | 75.0% | INT8 | 19 | 52.6 | 2,560 (20%) | 232k (6%) | 477k (27%) | 502 MHz |

This **approach best meets L1-trigger latency and resource budgets** by minimizing DSP usage with negligible accuracy loss.

### INT8-PO2 optimization: multiplier-free rescaling




Standard INT8 inference requires fixed-point rescaling after each INT32 accumulation. In the INT8-PO2 variant, each **rescale factor is approximated by a power of two, replacing multiplications by arithmetic shifts**.

Shift constants are emitted to the HLS header and used by the FPGA kernel, **reducing DSP usage and latency** at **the cost** of **additional LUT** logic and a **small accuracy loss**.

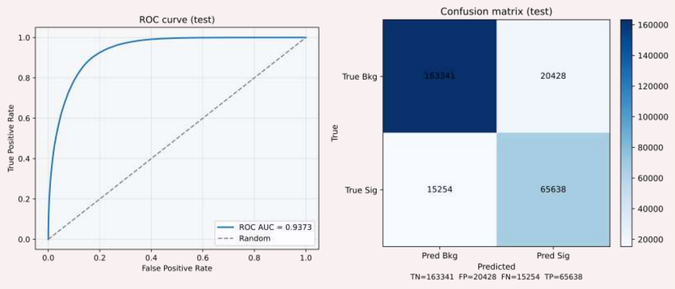
**DSP reduction is a consequence of implementing multipliers in logic.**

## X2O OMTF

- GNN architectures for CMS track finding (**primarily for displaced muons**) were evaluated in software as a baseline before FPGA deployment.
- Heterogeneous **training datasets** cover key physical scenarios: high-pileup, prompt and displaced (isolated and non-isolated) candidates.

### TransformerConv

Three-layer **Graph Transformer** model that combines source, target nodes, and edge attributes to build edge embeddings. Binary **edge classification** is performed by an MLP.

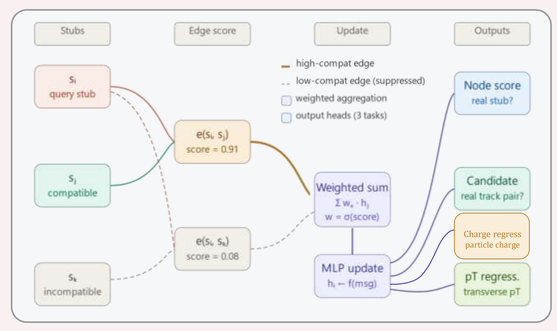



| Accuracy | Precision | Recall | F1 score |
|---|---|---|---|
| 0.86 | 0.76 | 0.81 | 0.79 |

**Figure 4.** Performance evaluation plots and metrics estimations.

### EdgeCompact

**GNN model** that treats detector stubs as nodes and learns a **compatibility score for pairs of stubs**.




**More about INTREPID**

**References Here!**

**Other stubs + standalone μ's**

**Better stubs + standalone μ's**

Global Muon Trigger (+BMTF)

Correlators

Global Trigger

New potential BSM or Phase-1 lost events for analysis

## Conclusions

- A **low resources and latency** Shower algorithm was **successfully designed and integrated** into CMS data acquisition **during slice tests.**
- **High-$p_T$ tagging algorithm** is currently being optimized by **integrating $\theta$** primitives.
- **Firmware implementation** of the **high-$p_T$ tagging algorithm** is **still at an early stage**, and the **data pipeline needs to be adapted** to support data from up to five CMS wheels.
- Barrel Filter **algorithms targeting LLP signatures** are currently **under initial study** (see 3 ).
- Initial FPGA deployments prove that **GNN's can fit** within strict **L1 trigger** latency and resource limits.
- Proxy model studies show that **quantization with INT8-PO2 optimization** provides the **best trade-off** between accuracy and resource usage.
- More complex architectures, such as **TransformerConv** and **EdgeCompact**, are under evaluation.
- This framework opens a new avenue for LLPs searches at the HL-LHC.

Contact Information: Daniel Estrada Acevedo, Department of Physics, University of Oviedo, Email: destrada@cern.ch