

University of
Zurich^{UZH}

Low-latency AI for triggering on electrons at High Luminosity LHC with the CMS Level-1 hardware trigger

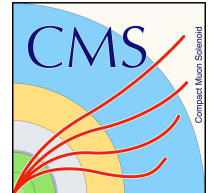
Piero Viscone^{1,2} on behalf of the **CMS Collaboration**.

¹CERN, ²University of Zürich

CHEP 2024 (Track 2)

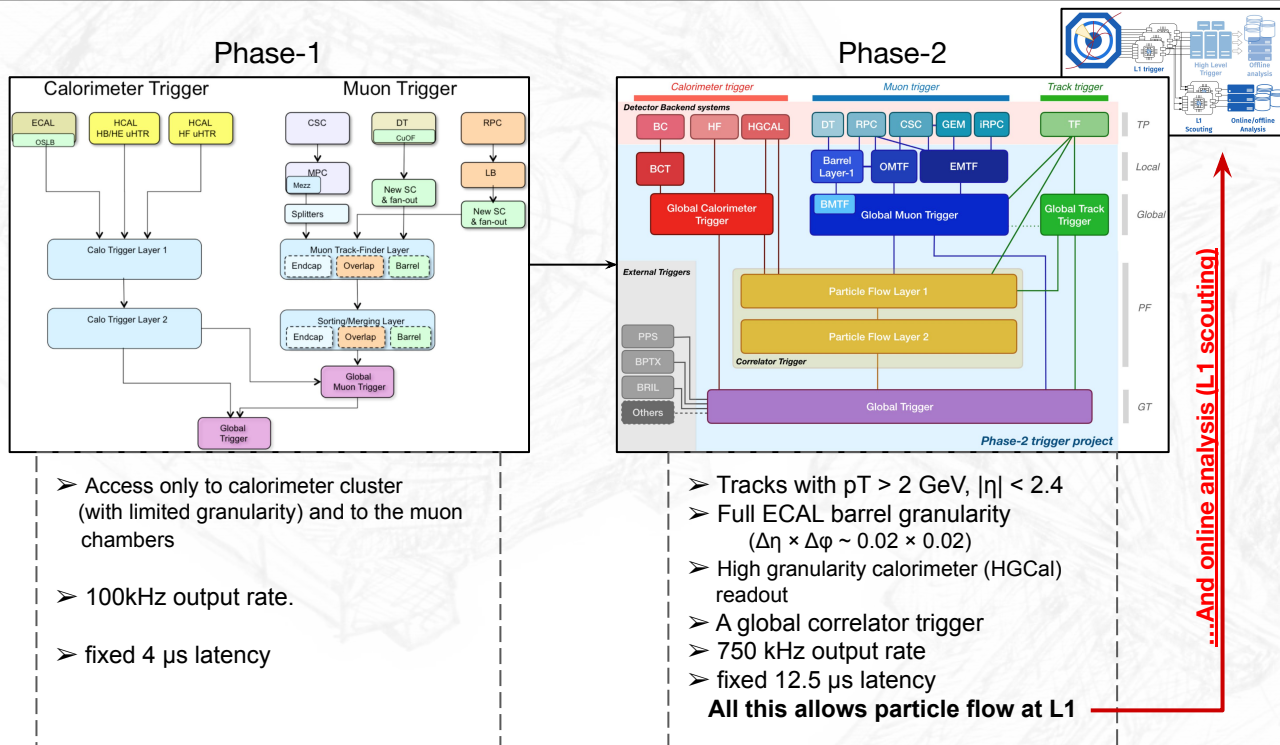
October 23rd, 2024

Kraków, Poland



The Phase-2 L1 trigger

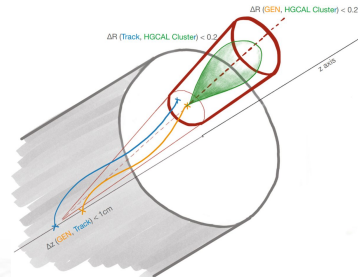
- The CMS Level-1 Trigger selects the most relevant collision events in real-time to reduce data for further processing
- Algorithms implemented in FPGAs and data transmitted through optical fibers for fast and low latency processing.



Tracks provide a nice handle to identify electrons and reduce the rate, extrapolating the tracks to the calorimeter and matching them to calorimeter clusters on the correlator trigger.

BUT

Electrons are tricky to reconstruct due to Bremsstrahlung and pair conversions



Offline reconstruction: Use Gaussian Sum Filter (GSF) to take in account Bremsstrahlung for tracking

L1 reconstruction: Computing power is limited, we can't use GSF

- Bad track ϕ / p_T resolution for electrons
- Some electron tracks are not reconstructed or the track trigger reconstruct multiple tracks for a single electron

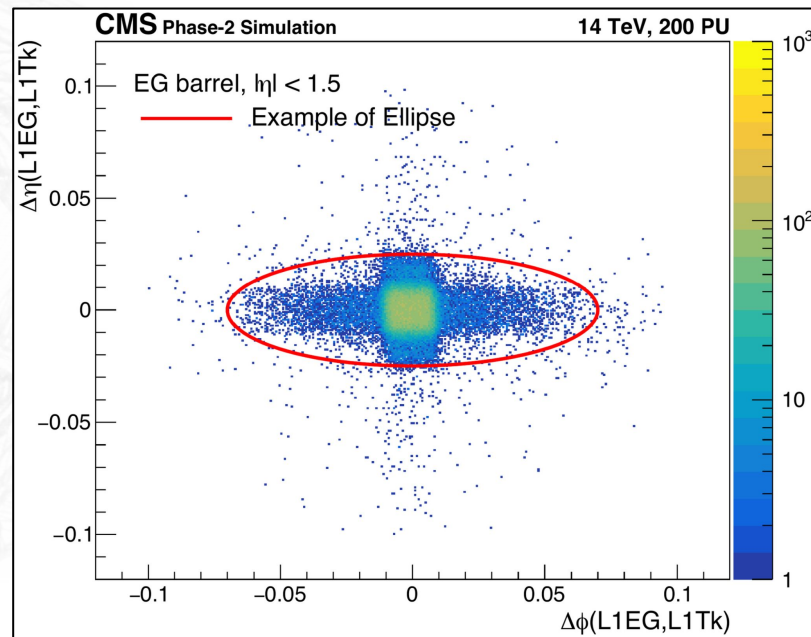
All of this at PU 200.

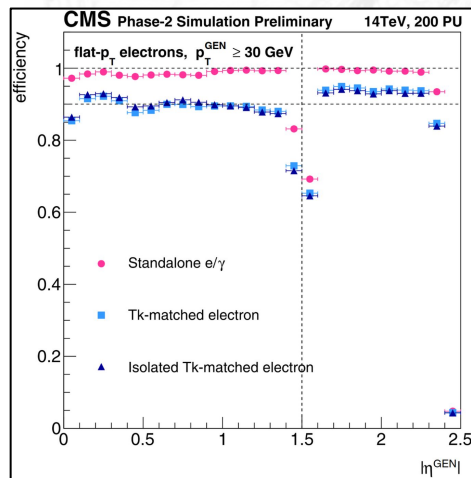
Given the high rate, the single lepton trigger for standalone objects has an high p_T threshold
Tk-matched electrons allow a significant lower p_T threshold.

The availability of the tracks @ L1 is exploited

Tracks with $p_T > 10$ GeV are matched to ECAL clusters through a tight elliptic matching in the η - ϕ plane

$$\Delta\eta_{\max} = \begin{cases} 0.025 & \text{for } |\eta| \leq 0.9 \\ 0.015 & \text{for } 0.9 < |\eta| \leq 1.479 \\ 0.0075 & \text{for } 1.479 < |\eta| \leq 2.4 \end{cases}$$
$$\Delta\phi_{\max} = 0.07$$

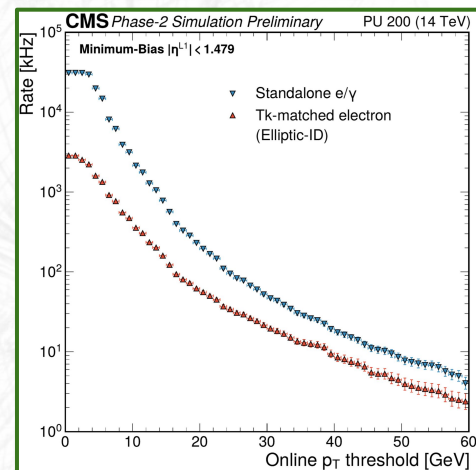
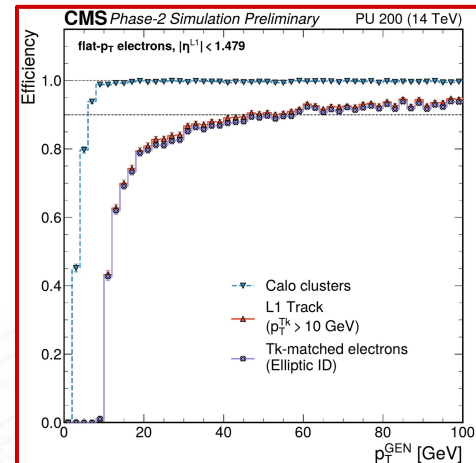




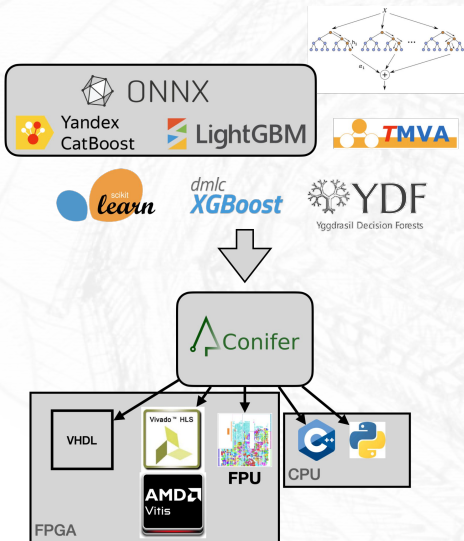
To keep the rate under control only tracks with $p_T > 10$ GeV are considered

- **Zero efficiency at $p_T < 10$ GeV**
- **Track matching allows us to keep the same threshold of the Run-2 menu while being at PU200**

But can we do better?



The improved computing power of the **correlator trigger** and the flexibility of the High Level Synthesis (HLS) allow us to deploy machine learning models on programmable logic.



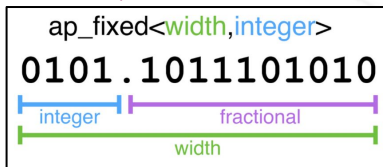
FPGAs allows low-latency inference but impose constraints on the size and complexity of the models.

Models must fit in a limited latency frame and in the limited availability of Look Up Tables (LUTs), Digital Signal Processors (DSPs) and Flip Flops (FFs)

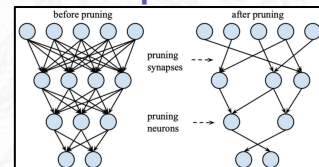
There are different techniques to reduce the used FPGA resources by a **NN/BDT**



Quantization



Compression



Electron identification in the Barrel region

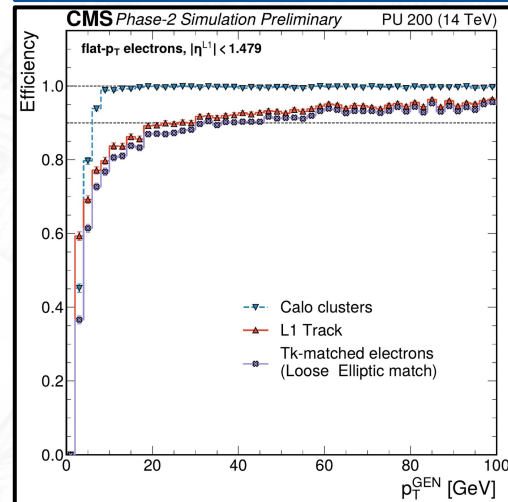
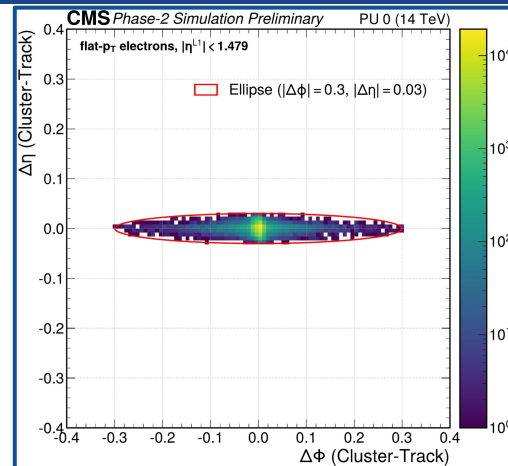
Step 1:

Tracks are matched to calorimeter clusters with a looser elliptic matching with minimal requirements on the track

Step 2:

Use a machine learning model to classify the track-calorimeter cluster pairs

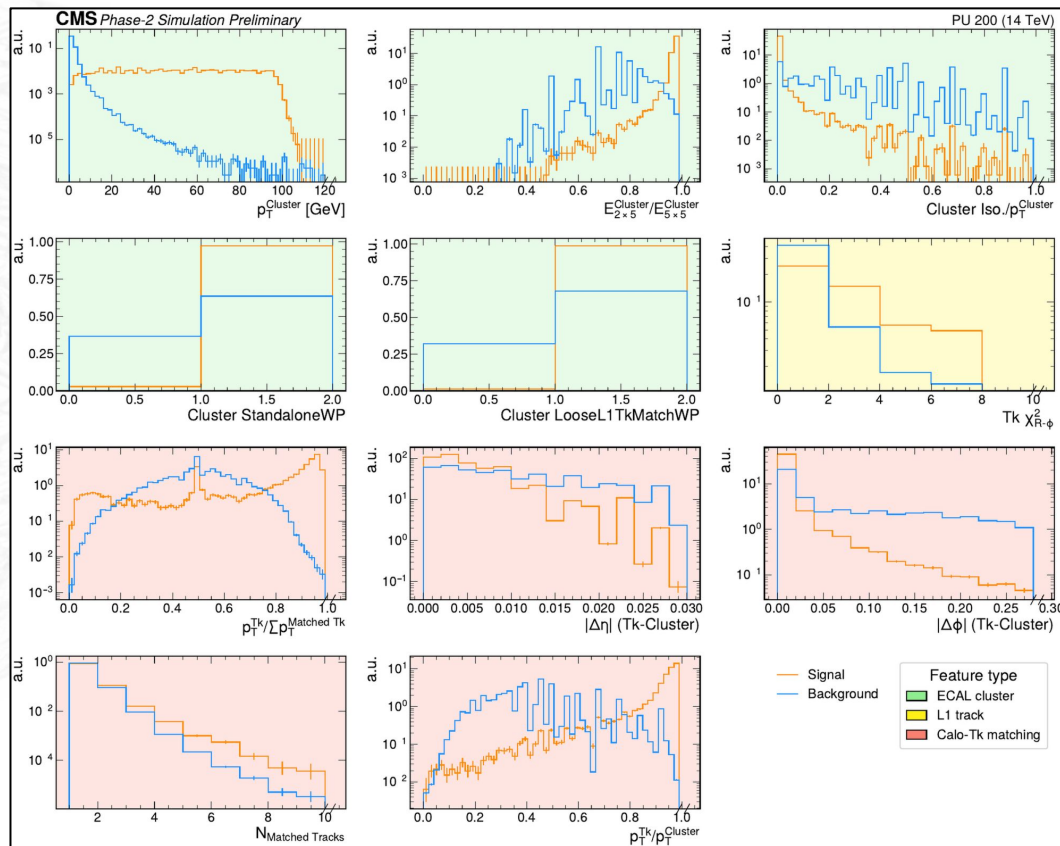
Goal: improving the electron identification, allowing a significant reduction of the rate



BDT trained on 11 input features
(max depth=12, 15 boosting round)

- Signal:
Tk-Cluster pairs matched to a generated electron in a PU200 flat- p_T electron sample
- Background:
All the Tk-Cluster pairs in a PU200 minimum bias sample

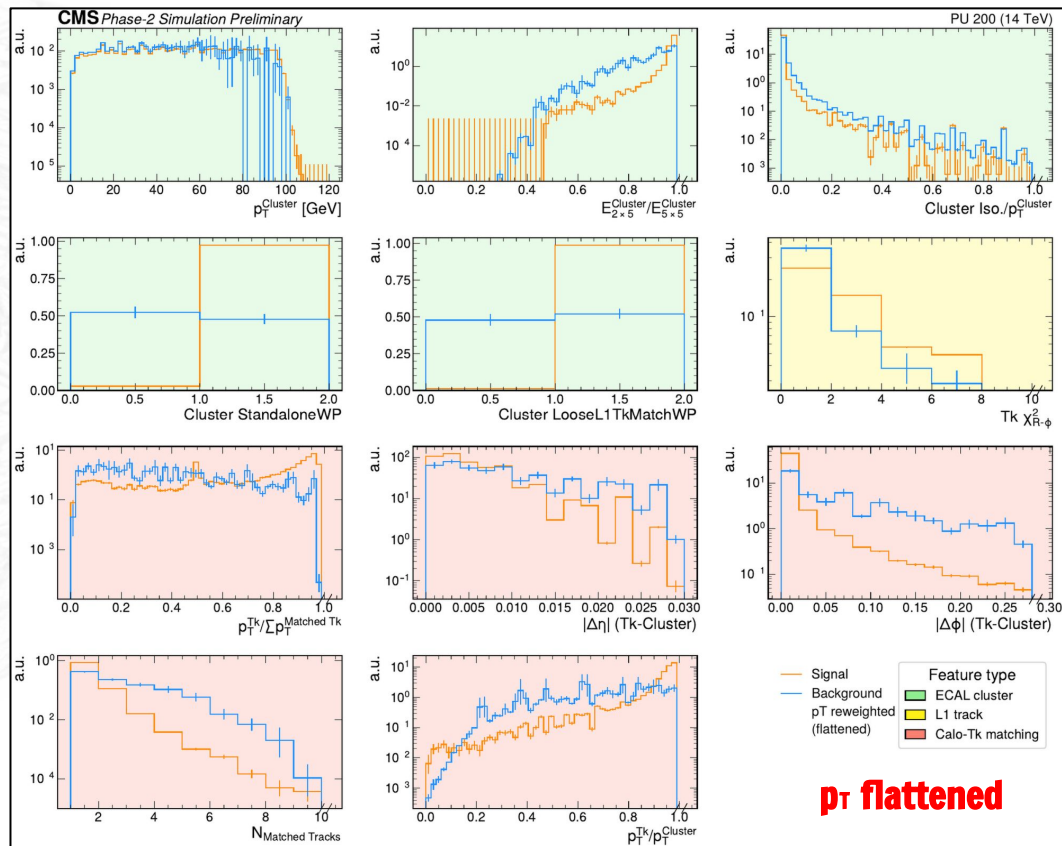
To prevent the model from imposing a tight cut on the cluster p_T , the features of all the background candidates used for training were reweighted to flatten the cluster p_T distribution

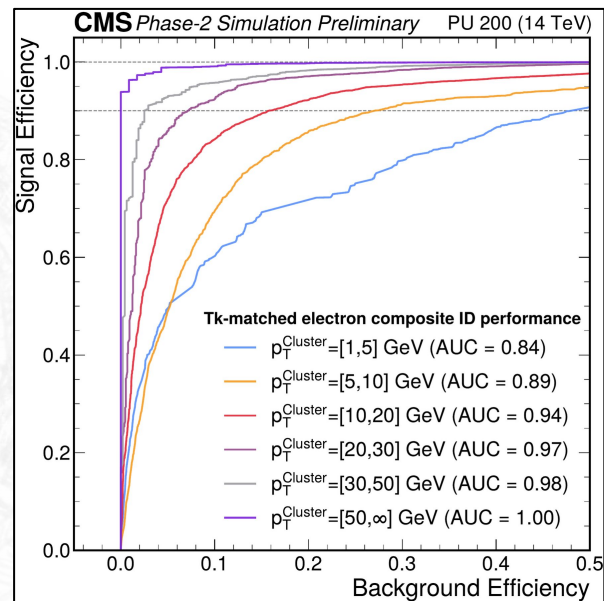
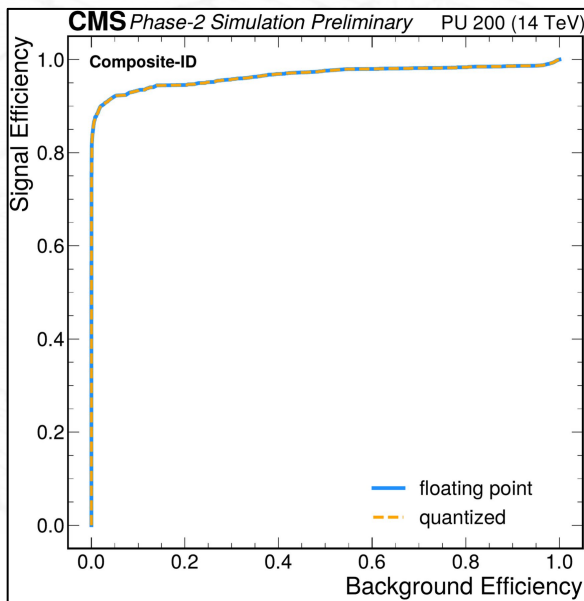


BDT trained on 11 input features
(max depth=12, 15 boosting round)

- Signal:
Tk-Cluster pairs matched to a generated electron in a PU200 flat- p_T electron sample
- Background:
All the Tk-Cluster pairs in a PU200 minimum bias sample

To prevent the model from imposing a tight cut on the cluster p_T , the features of all the background candidates used for training were reweighted to flatten the cluster p_T distribution



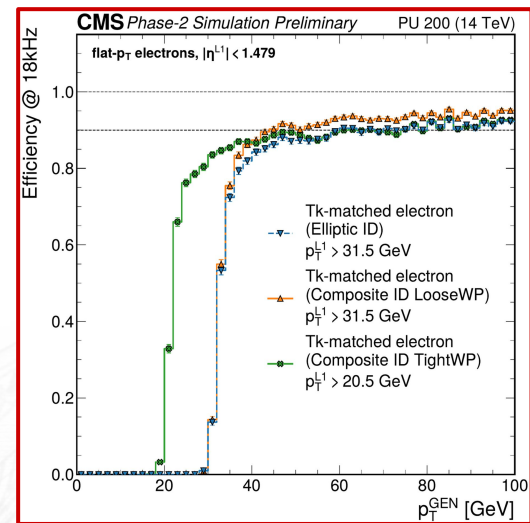
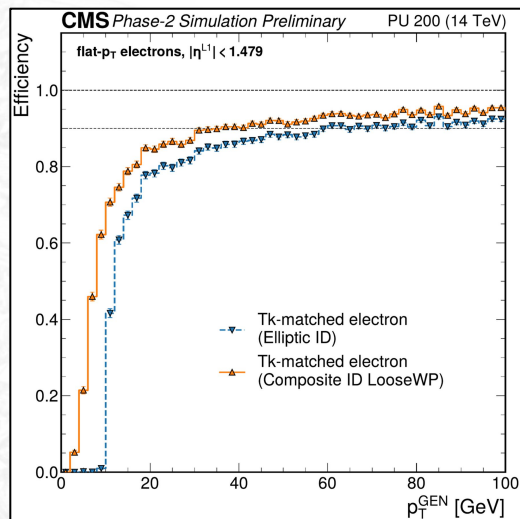
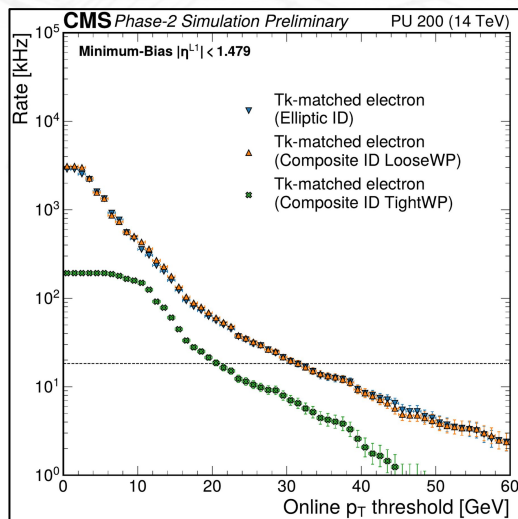


The input variable were quantized with fixed point precision (**9 integer and 15 decimal bits**).

The output scores are obtained from the bit-wise emulator using Conifer.

The quantization procedure do not cause any noticeable loss of performance compared to the floating point precision

At low p_T is more difficult to distinguish genuine electrons from PU particles but, as expected, the model is not imposing an hard cut on the cluster p_T due to reweighting



To compare the Composite-ID with the Elliptic-ID, two working points (WPs) applied in different p_T bins were created: a tight WP that matches the efficiency of the Elliptic-ID and a loose WP that matches the rate of the elliptic-ID.

Significant gain in efficiency at same rate and significant rate reduction at same efficiency

p_T threshold can be lowered by more than 10 GeV on a single electron trigger with a fixed rate of 18kHz in the barrel (Run3 like threshold)

The model has been synthesized in firmware using the Conifer library and Vivado HLS.

The model is synthesized for a clock frequency of 180MHz, matching the implementation of the e/y reconstruction in the Correlator Layer-1 boards
(Xilinx Virtex UltraScale+ VU13P FPGA)

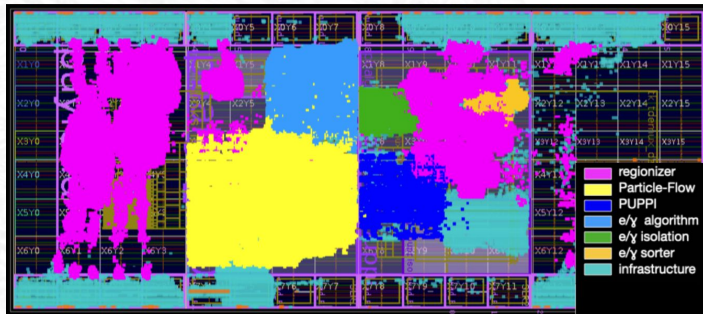
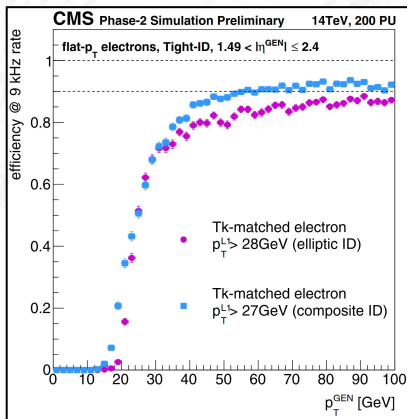
The model use just 1.6% of the Look Up Table used

	BRAM	DSP	FF	LUT
SLR	0.0%	0.0%	0.12%	6.5%
Total	0.0%	0.0%	0.02%	1.6%

	Clock cycles	Latency
Clock (5.56 ns)	5	27.8 ns

Background rejection in the Endcap region

A similar approach was employed also in the endcap.



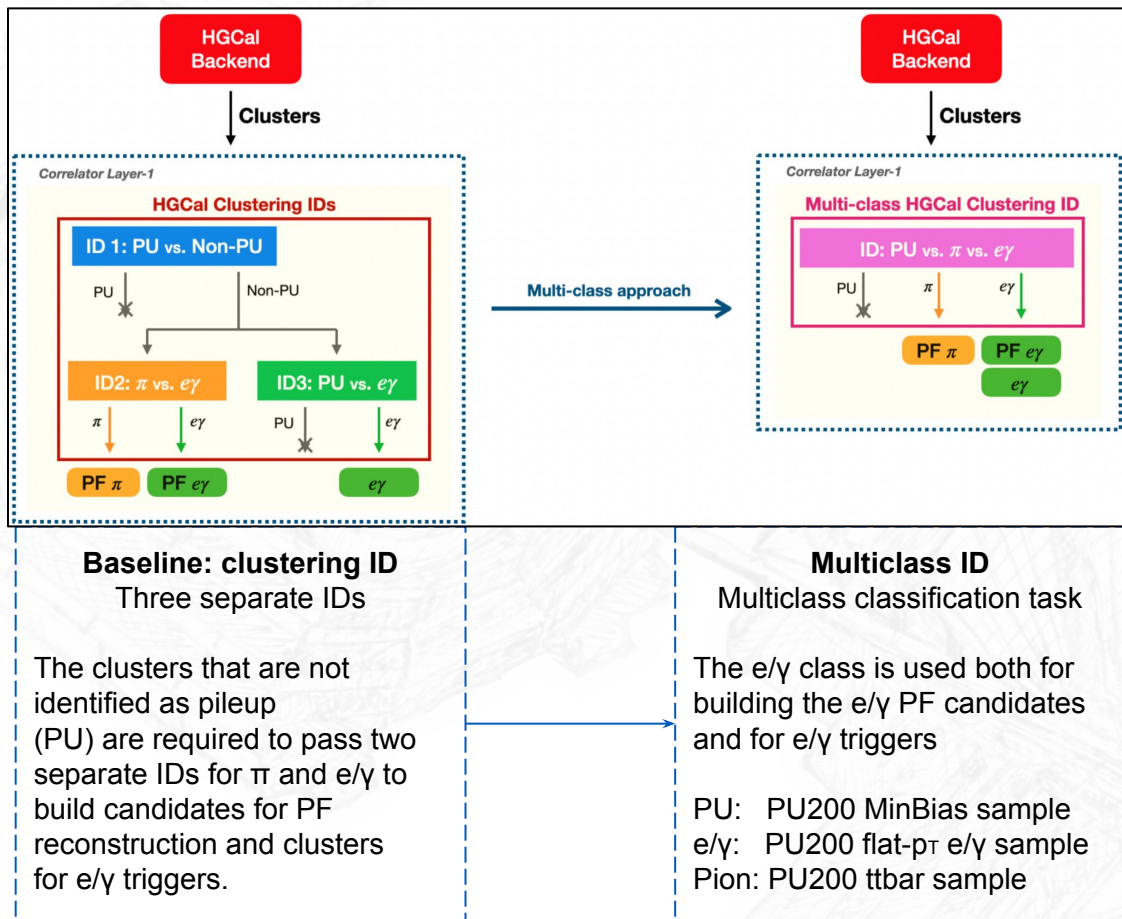
Full algorithm

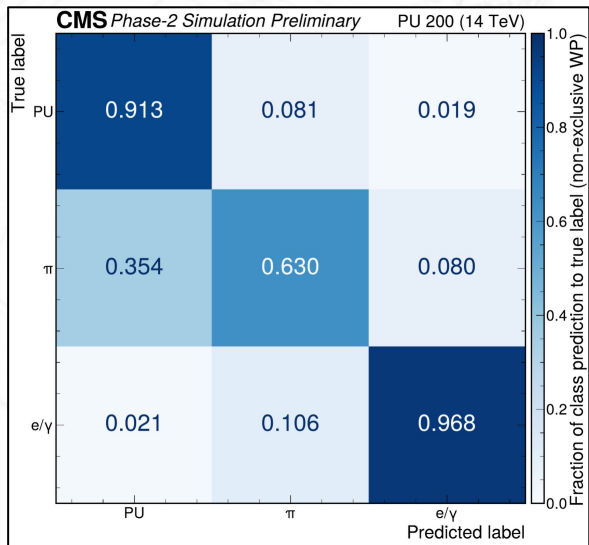
	LUT	FF	BRAM	DSP
e/ γ IP	3.1%	0.4%	0.0%	1.6%
Total	24.4%	17.6%	29.5%	14.3%

BUT

- HGCal give us a lot more information compared to the ECAL barrel

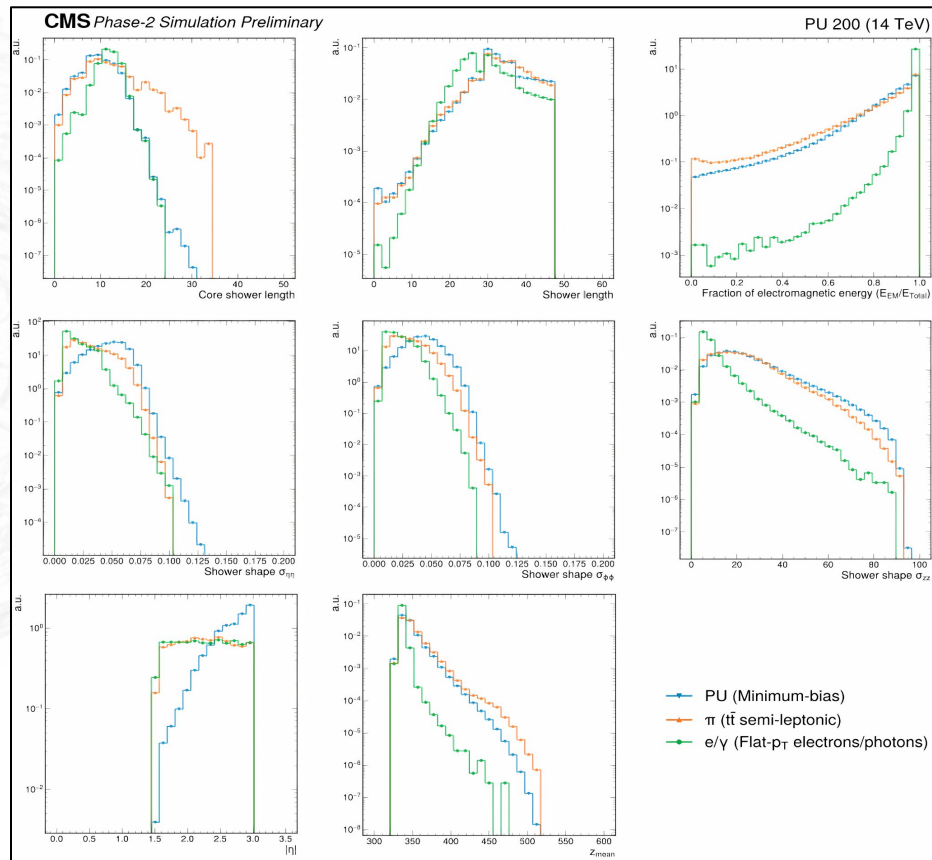
We can exploit the high granularity and the 3 dimensional information to preselect the calorimeter clusters

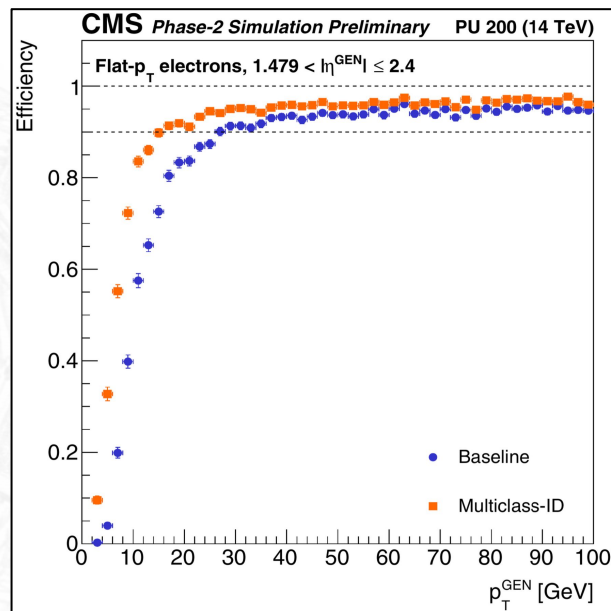
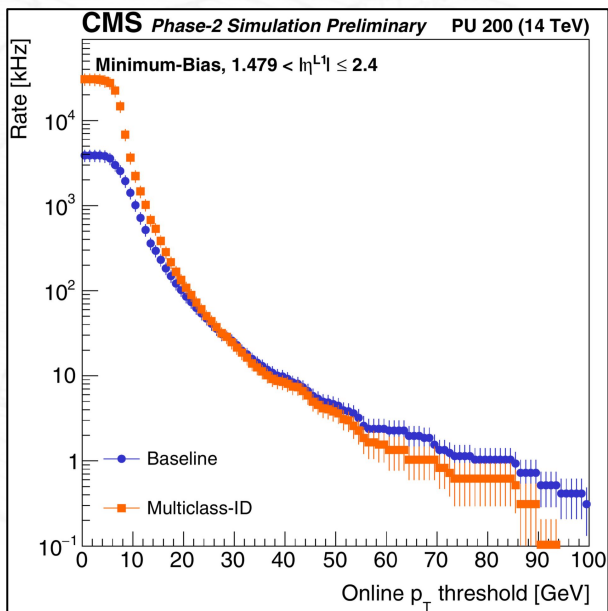




Non-exclusive WP:

Cut on the 3 scores in order to optimize the e/γ purity and efficiency.





- Efficient identification of low- p_T clusters can be exploited in building track-matched electrons and in the L1T Scouting system to target signals with low- p_T electrons.
- The multi-class ID increases the rate in low p_T regions but lowers it at the typical thresholds used in the TDR baseline Trigger Menu ($p_T > 25$ GeV)

The model has been synthesized in firmware using the Conifer library and Vivado HLS.

Floating value input features are quantized to fixed point precision of 7 integer and 10 decimal bits

The model is synthesized for a clock frequency of 360MHz on a **Xilinx Virtex UltraScale+ VU13P FPGA**

6% of the Look Up Table used

	BRAM	DSP	FF	LUT
SLR	0.0%	0.0%	1.0%	25.0%
Total	0.0%	0.0%	0.0%	6.0%

	Latency (cycles)	Latency (absolute)
Clock (2.78 ns)	5	13.890 ns

- Exploit the improved efficiency for low- p_T signature in the L1 scouting for signature that can not be triggered in the standards L1 menu
- Review the endcap composite ID model profiting from the new multiclass ID, improving the efficiency at low- p_T
- p_T regression to improve the electron p_T resolution
- Integrate the firmware of the barrel model into the correlator layer-1 design



CMS Experiment at the LHC, CERN

Data recorded: 2012-May-13 20:08:14.621490 GMT

Run/Event: 194108 / 564224000

BACKUP

A 3D visualization of a particle collision event. A central point of interaction is shown with numerous lines radiating outwards, representing the paths of particles. The lines are colored in shades of blue, orange, and green. The background is dark, with a large, semi-transparent blue volume surrounding the collision point. Two prominent green lines extend from the collision point towards the top and bottom right corners of the frame. The word 'BACKUP' is overlaid in large white letters across the center of the image.

