

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

Cristina Botta¹ , Gianluca Cerminara¹ , Kyungmin Park² , and Piero Viscone^{1,3} 

¹CERN, 1211 Genève 23, Switzerland

²Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States

³University of Zürich, Rämistrasse 71, CH-8006 Zürich, Switzerland

Abstract. In preparation for the High Luminosity LHC (HL-LHC) run, the CMS collaboration is working on an ambitious upgrade project for the first stage of its online selection system: the Level-1 Trigger. The upgraded system will use powerful field-programmable gate arrays (FPGA) processors connected by a high-bandwidth network of optical fibers. The new system will access highly granular calorimeter information and online tracking: their combination for identifying physics objects is a key asset to cope with the harsh HL-LHC environment without compromising physics acceptance. The track matching is particularly relevant for identifying calorimeter deposits originating from electron particles. Traditional identification techniques rely on several independent selection stages applied to the calorimeter and track primitives, followed by an angular matching procedure. A new machine learning approach is presented, combining track and calorimeter information into a single identification and matching step. The new algorithm leverages new technologies for running fast inference on FPGA.

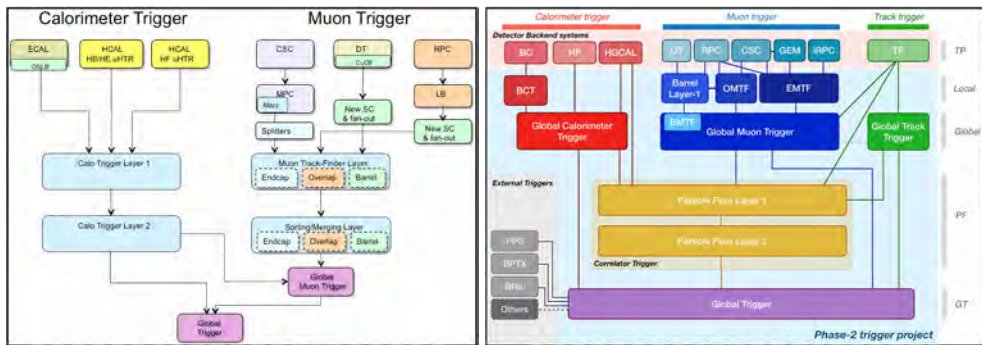
1 Introduction

The High Luminosity Large Hadron Collider (HL-LHC) presents a challenging environment for real-time event selection due to the high multiplicity of simultaneous interactions, also known as pileup (PU). To address these challenges, the CMS [1] Phase-2 Level-1 Trigger (L1T) [2] will be significantly upgraded. The new architecture will open the way to new identification techniques for the online electron reconstruction. After an overview of the L1T system upgrade in Section 2, the document focuses on novel identification algorithms exploiting Machine-Learning inference on FPGAs, enhancing the identification efficiency for low transverse momentum electrons. Section 3 presents an overview of the advantages and challenges of implementing machine learning models on FPGAs. Section 4 focuses on the online identification of clusters originating from electromagnetic energy deposits in the endcap calorimeter, while Section 5 discusses electron identification exploiting online tracking and calorimeter information. The new approach and its physics performance are compared to the baseline algorithm described in the TDR for both cluster classification and electron identification use cases.

2 The upgrade of the Level-1 Trigger system

The Phase-1 L1T system, depicted in Fig. 1, primarily relied on a combination of calorimeter and muon trigger information, with limited granularity in the calorimeter data. However, the Phase-2 upgrade introduces significant improvements by incorporating tracking information for tracks with transverse momentum $p_T > 2$ GeV and pseudorapidity $|\eta| < 2.4$ and providing full calorimeter barrel granularity along with three-dimensional High Granularity Calorimeter (HGCAL) information in the forward region.

Key improvements also include an increased output rate, from 100 to 750 kHz, and an increased fixed latency from 4 to 12.5 μs , providing additional time for more sophisticated algorithms. These enhancements, together with the Correlator Trigger that combines the information of all subsystems, enable full Particle Flow (PF) [3] reconstruction at L1T, providing a more complete event description in real-time at 40 MHz.



4 HGCAL cluster classification

The Phase-2 HGCAL [8] spans the pseudorapidity range of $1.52 < |\eta| < 3$, providing highly granular information about the particle showering in its volume. In the TDR baseline approach shown in Fig. 2, the identification of electromagnetic-like deposits in the HGCAL follows a two-stage BDT-based strategy. Since observables related to electron and photon identification evolve rapidly with η , two models are implemented: one for the lower η region ($1.5 < |\eta| \leq 2.7$) and another for the higher $|\eta|$ region. The input features include five longitudinal and four lateral shower-shape variables. The longitudinal variables encompass the cluster length, the position of shower onset, and the energy-weighted RMS of the z-coordinates of the cluster components. The first step involves a BDT that classifies clusters as either originating from pileup (PU) or non-PU. Clusters classified as non-PU undergo further classification through two additional BDTs. One BDT distinguishes charged hadrons (π) from electromagnetic particles (e/γ) for PF reconstruction, while the other BDT refines the identification of e/γ candidates for lepton triggers.

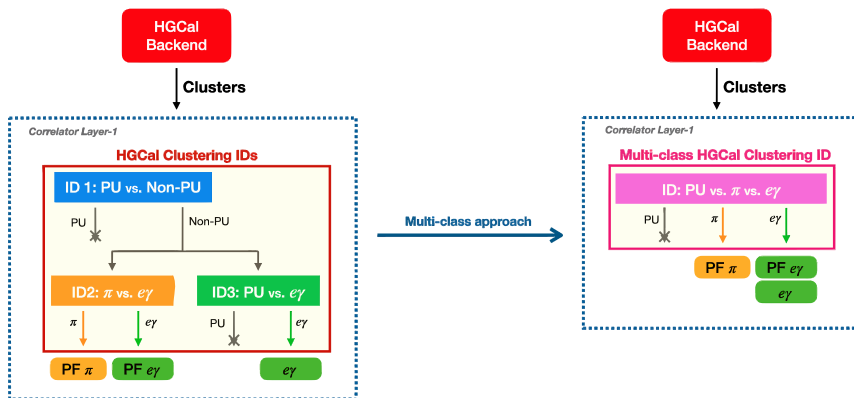


Figure 2. Comparison between the baseline double stage BDT approach on the left and the new Multiclass-ID on the right.

4.1 Multi-class cluster identification

A new multi-class BDT model [9] is introduced, differing from the baseline multi-stage binary classification approach by directly classifying clusters into multiple categories (PU, hadronic, or electromagnetic) within a single inference step. Training datasets are derived from detailed Geant-4 based detector simulations with 200 PU, incorporating minimum-bias samples for PU classification, $t\bar{t}$ events for charged hadrons (π), and flat transverse momentum electron/photon samples for e/γ identification. A $p_T > 5$ GeV and $|\eta| < 2.4$ selection is applied to all clusters, with additional generator-level matching $\Delta R(\text{Gen} - \text{HGCAL cluster}) < 0.1$ required for π and e/γ clusters. Eight cluster features are used as inputs, the number of shower layers, the number of continuous shower layers, and the fraction of energy deposited in the electromagnetic layers of the calorimeter. Other features include shower shape variables such as $\sigma_{\eta\eta}$, $\sigma_{\phi\phi}$, σ_{zz} , which are computed as the energy-weighted RMS over cluster constituents, as well as the absolute cluster pseudorapidity, and the energy-weighted barycenter in the longitudinal coordinate. To normalize the output scores and interpret them as probabilities, a softmax function is applied to the output nodes. Two different schemas of working

points (WP) were studied: an exclusive WP that assigns the class with the highest probability to each cluster and a non-exclusive WP that defines a minimum threshold for each class, allowing clusters to be categorized into multiple classes. The confusion matrices for both WPs are shown in Fig. 3. The non-exclusive WP was optimized for the best efficiency and purity of e/γ clusters and to preserve the performance of the PF reconstruction of the missing transverse energy and was chosen for its better ID efficiency and downstream interpretability. Unlike the baseline method, all e/γ clusters, whether used for PF reconstruction or trigger-

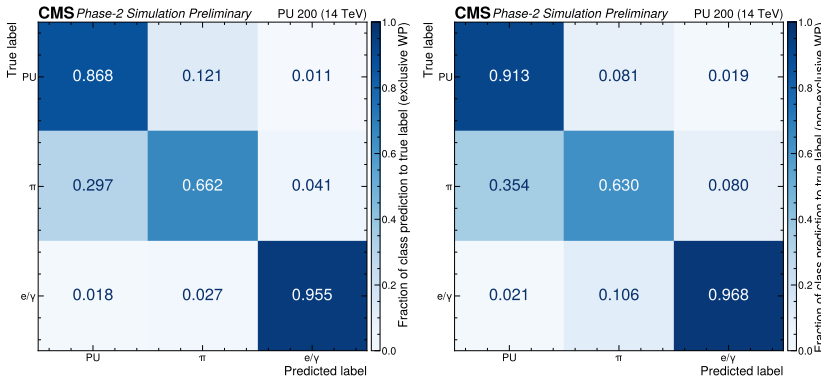


Figure 3. Confusion matrices for exclusive WP (left) and non-exclusive WP (right).

ing, originate from the same output node. The differentiation between clusters used in PF reconstruction and those used for constructing e/γ triggers is achieved by applying different thresholds to the same output by employing the non-exclusive WP. This unified approach ensures consistency in classification while allowing flexibility in optimizing selection criteria for both purposes. As shown in Fig. 4, the efficiency at the low p_T region is significantly improved with the new multi-class ID, with an overall 35% improvement in efficiency for $5 < p_T < 10$ GeV. While only standalone electromagnetic objects with $p_T > 25$ GeV are used in the TDR Trigger Menu [2], efficient identification of low- p_T clusters can be exploited in building track-matched electrons for multi-object triggers and in the L1T Scouting system [2] to target signals with low- p_T electrons. The model was synthesized for FPGA implementation using the Conifer library and Vivado HLS quantizing the input features to fixed point precision of 9 integer and 10 decimal bits, achieving a latency of 13.89 ns (5 clock cycles at 2.78 ns per clock) on Xilinx Virtex UltraScale+VU13P hardware and using 6% of the available LUT on the board. Further optimization of the model size is possible.

5 Electron identification

Since at PU 200 the single-lepton trigger for standalone objects has a high p_T threshold, the selection of lower-energy electrons can be obtained only through track-matched electrons. Tracks provide a crucial handle to identify electrons and reduce the trigger rate by extrapolating them to the calorimeter and matching them to calorimeter clusters in the Correlator Trigger. However, electron tracks are particularly challenging to reconstruct due to effects such as Bremsstrahlung radiation and pair conversions. While offline reconstruction employs Gaussian Sum Filters (GSF) to account for Bremsstrahlung in tracking [10], the limited computing power available at L1T prevents the use of GSF, requiring reliance on a simpler Kalman filter instead. This constraint results in poor track ϕ and p_T resolution for electrons

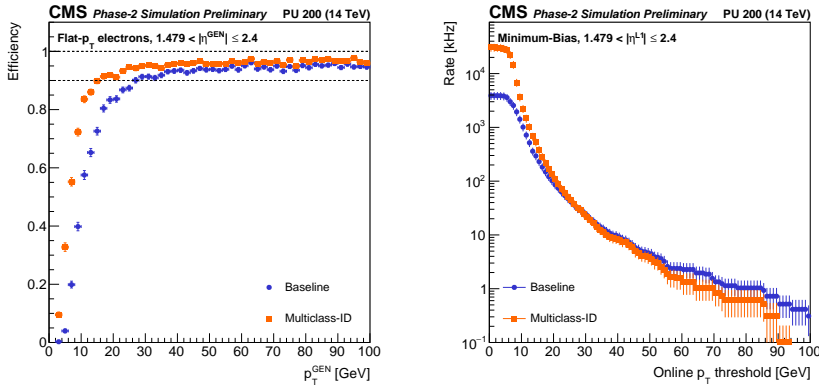


Figure 4. The left plot shows the efficiency of reconstructing a genuine electron at L1T for the baseline and the new multiclass e/γ IDs using the non-exclusive WP for standalone electromagnetic objects on a flat- p_T electron sample, while the right plot shows the rates of standalone electromagnetic objects measured on minimum bias samples, comparing the current baseline and new multi-class clustering ID. 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).

and worse track reconstruction efficiency. The baseline algorithm for electron identification in the Phase-2 L1T, as described in the TDR [2], is based on an elliptic matching approach that consists of matching tracks to electromagnetic calorimeter (ECAL) clusters through a tight elliptic matching in the $\eta - \phi$ plane, shown in Fig. 5.

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

$$\Delta\phi_{\text{max}} = 0.07$$

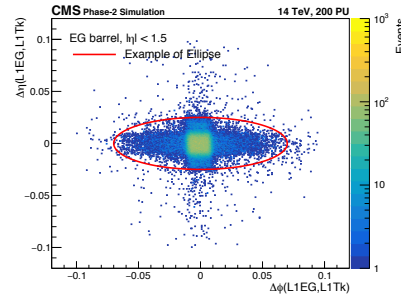


Figure 5. $\Delta\eta$ vs $\Delta\phi$ distances between calorimeter clusters and the closest $p_T > 10$ GeV L1 track in the $\eta - \phi$ plane. The plots are computed for single electron events at PU 200 in the barrel region. The elliptic cuts used in the baseline algorithm to select electron candidates are shown. [2]

To keep the trigger rate under control, only tracks with $p_T > 10$ GeV are considered. While the elliptic matching technique allowed the Phase-2 L1T to retain the same trigger threshold as the Run-2 menu, even under high pileup conditions at PU 200, the efficiency in the $p_T < 10$ GeV region is small. A more sophisticated approach, called "Composite-ID", integrates both track and cluster information within a unified ML model. This method significantly enhances electron identification efficiency even in the low- p_T region and further

reduces the trigger rate. While this approach is implemented also for the endcap [11], here we concentrate on the barrel.

5.1 Composite model for track-calorimeter cluster matching

In the novel approach, candidate track-cluster pairs are initially selected based on loose elliptic matching constraints ($\Delta\phi < 0.3$, $\Delta\eta < 0.03$). The baseline requirement on the track $p_T > 10$ GeV is dropped to maximize signal retention also in the low- p_T region, as shown in Fig. 6.

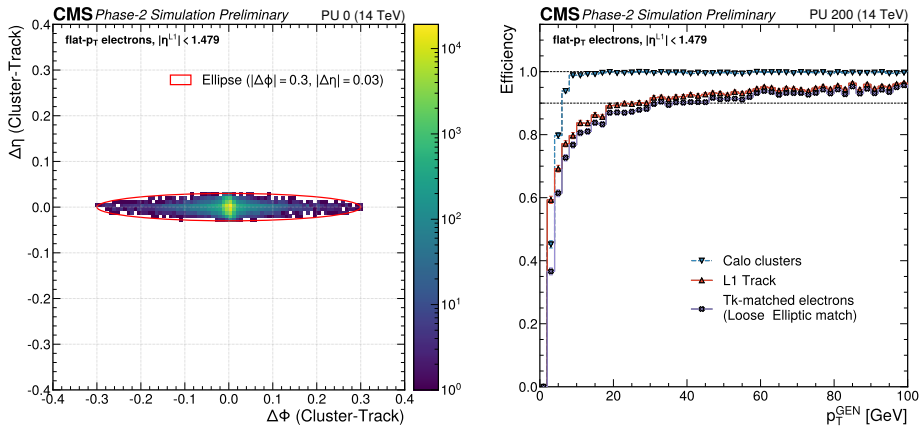


Figure 6. The left plot shows the Track-Cluster $\Delta\eta$ vs $\Delta\phi$ 2D distribution in a flat- p_T single electron 0 PU sample, while the right plot shows the efficiencies of the standalone object, L1 tracks, and Tk-matched candidates as a function of the generated electron p_T using a flat- p_T single electron 200 PU sample.

The Composite-ID then refines selection through a 15 boosting round BDT model trained on 11 features, including cluster shower shape parameters (as energy fraction in 2×5 vs. 5×5 crystal arrays and isolation properties), track quality variables (as the χ^2 obtained during the track fitting procedure) and cluster-track matching features (as the number of matched tracks per cluster, $|\eta|$ and $|\phi|$ distances between the matched track and cluster, and the ratio between the cluster and track p_T). Signal candidates are selected based on the geometrical ($\Delta R < 0.1$) matching of the cluster with a generated electron particle in a flat- p_T single electron sample at PU 200, while background candidates are selected among all track-matched candidates in a Minimum-bias sample at PU 200. Since the p_T distribution of clusters originating from PU particles is mostly at low- p_T and is exponentially falling, 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. As can be noticed in Fig. 7, while at low- p_T it is more difficult to distinguish genuine electrons from PU particles, the reweighting prevents the model from imposing a hard cut on the cluster p_T . To compare the Composite-ID with the Elliptic-ID, two WPs were applied in different p_T bins: a tight WP that matches the efficiency of the baseline Elliptic-ID and a loose WP that matches the rate of the baseline Elliptic-ID. The tight WP allows a significant rate reduction compared to the baseline elliptic ID implying a significantly looser p_T threshold for the single electron trigger. Considering a fixed rate of 18 kHz, corresponding to the barrel bandwidth for Run-3-like thresholds, the tight WP allows to move the threshold from 31.5 GeV to 20.5 GeV, as shown in Fig. 8.

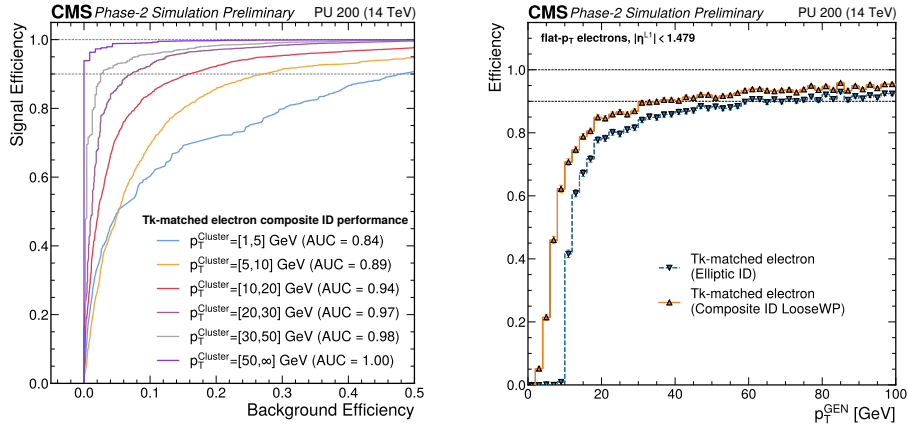


Figure 7. The left plot shows the receiver operating characteristic curve (ROC) curve in different p_T bins of the Composite ID model, while the right plot illustrates the efficiency in function of the p_T of the generated electron for the baseline elliptic ID and the LooseWP of the Composite ID, showing a substantial efficiency gain compared to the baseline elliptic-ID at parity of rate that is crucial for the Double Electron seed of the baseline L1 menu, which has a threshold of 25 and 12 GeV.

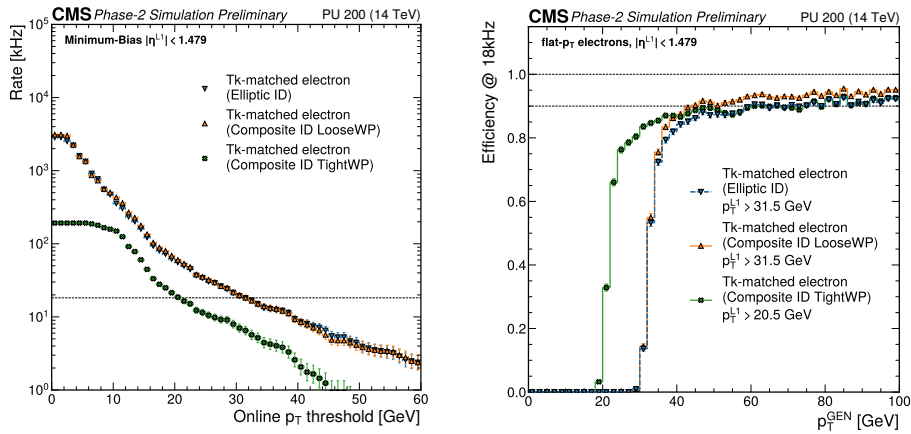


Figure 8. The left plot illustrates the rate as a function of the cluster p_T . The right plot compares the efficiency as a function of the p_T of the generated electron for a fixed 18 kHz rate, corresponding to the barrel bandwidth for Run-3-like thresholds.

It is important to stress that the loose and tight WP were defined solely for comparison between the Composite-ID model and the baseline Elliptic-ID. A working point positioned between these two extremes would likely be more beneficial in balancing efficiency and rate, maximizing the overall performance of the trigger.

The Composite-ID model has been bit-wise emulated and synthesized in firmware using the Conifer library and Vivado HLS. The input features are quantized with 9 bits for the integer part and 15 bits for the decimal precision, while the raw scores are quantized using 4 integer and 12 decimal bits. The model is synthesized for a frequency of 180 MHz, align-

ing with the implementation of the e/γ reconstruction in the Correlator boards. Resource estimates after synthesis on a Xilinx Virtex UltraScale+VU13P FPGA show that the model occupies 1.6% of the available LUTs, with a latency of 27.8 ns (5 clock cycles at 5.56 ns per clock) with minimal performance loss compared to the non-quantized version of the model.

6 Conclusion

The integration of ML-driven electron identification in the CMS Phase-2 L1T system represents a significant advancement in real-time event selection. The composite-ID and multiclass-ID methods demonstrate clear advantages over traditional approaches, particularly for low- p_T electrons. Additionally, their implementation was successfully emulated in CMSSW, ensuring bit-wise accuracy and seamless integration within the CMS software framework. Future work will focus on several key areas: exploiting the improved efficiency for low- p_T signatures in L1 scouting, allowing triggering on signatures that cannot be captured in the standard L1 menu; optimizing the endcap Composite-ID model using the new Multiclass-ID to enhance low- p_T efficiency; implementing a p_T regression to improve the electron p_T resolution; and integrating the firmware into the Correlator Layer-1 boards. These advancements will contribute to enhancing the physics discovery potential at the HL-LHC.

Acknowledgment This work has been [partially] funded by the Eric & Wendy Schmidt Fund for Strategic Innovation through the CERN Next Generation Triggers project under grant agreement number SIF-2023-004.

References

- [1] CMS Collaboration, The CMS experiment at the CERN LHC, Journal of Instrumentation (2008), S08004, <https://doi.org/10.1088/1748-0221/3/08/s08004>
- [2] CMS Collaboration, The Phase-2 Upgrade of the CMS Level-1 Trigger, CMS-TDR-021, (2020), <https://cds.cern.ch/record/2714892>.
- [3] CMS Collaboration, Particle-flow reconstruction and global event description with the CMS detector, JINST **12** (2017) no. 10, P10003, <https://doi.org/10.1088/1748-0221/12/10/P10003>
- [4] XGBoost, <https://xgboost.readthedocs.io/en/stable/>.
- [5] Conifer, <https://github.com/thesps/conifer>.
- [6] CMSSW, <https://cms-sw.github.io/>.
- [7] HLS4ML, <https://github.com/fastmachinelearning/hls4ml>.
- [8] CMS Collaboration, The Phase-2 Upgrade of the CMS Endcap Calorimeter (2017), <https://cds.cern.ch/record/2293646>
- [9] Electron Reconstruction and Identification in the CMS Phase-2 Level-1 Trigger, CMS-DP-2024-098, <https://cds.cern.ch/record/2916193/>.
- [10] CMS Collaboration, Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC, JINST **16** (2021) no. 05, P05014, <https://doi.org/10.1088/1748-0221/16/05/P05014>.
- [11] CMS Collaboration, Electron Reconstruction and Identification in the CMS Phase-2 Level-1 Trigger, CMS-DP-2023-047, <https://cds.cern.ch/record/2868782>.