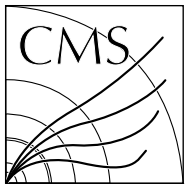
Available on CMS information server

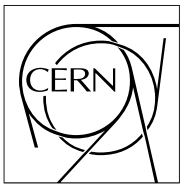
CMS DP -2023/079



The Compact Muon Solenoid Experiment

CMS Performance Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland



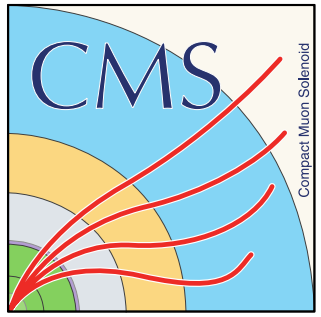
04 October 2023

Anomaly Detection in the CMS Global Trigger Test Crate for Run 3

CMS Collaboration

Abstract

This DPS note shows the design, implementation, and performance of an ML-based trigger algorithm, AXOL1TL, which selects anomalous events in real-time. The AXOL1TL algorithm was implemented in Level-1 Global Trigger Test Crate for 2023 data taking.



Anomaly Detection in the CMS Global Trigger Test Crate for Run 3

CMS Collaboration
cms-dpg-conveners-l1t@cern.ch and cms-conveners-ml@cern.ch

Glossary

- AXOL1TL – Anomaly eXtraction Online Level-1 Trigger Lightweight. Acronym for anomaly detection implementation at the global trigger level
- μGT – Level-1 global trigger system
- μGT test crate – Hardware system with identical inputs to production μGT that doesn't influence trigger decisions
- FPGA – Field Programmable Gate Array – reprogrammable digital chip used to implement Level-1 Trigger algorithms
- FPGA resources: LUTs (Look Up Tables) - general purpose logic; FFs (Flip Flops) - registers; DSPs (Digital Signal Processors) - multipliers; BRAMs (Block Random Access Memories) - small high speed memories
- MP7 – circuit board equipped with an FPGA and high speed optical fibres, on which the μGT is implemented
- Xilinx Virtex-7 – Model of FPGA chip on MP7 board
- QKeras – Deep learning library extension to TensorFlow API Keras. Used to quantize machine learning algorithms for improved performance when implemented on hardware
- VAE – Variational Autoencoder – Neural Network architecture encoding then decoding to/from a latent-space with probability distribution
- Latent space – Low-dimensional space into which network inputs are encoded
- KL-Divergence - Kullback–Leibler divergence, a statistical measure of distance between two distributions, used in the loss function to train the VAE
- μ – Mean value of distribution for each feature in latent space. μ^2 may refer to quadrature sum of individual μs
- VHDL - Very High-Speed Integrated Circuit Hardware Description Language. Language used to describe hardware, for example for FPGAs
- HLS - High-Level Synthesis – Process by which object-oriented code is synthesized into a lower-level description, in this case VHDL, for implementations on FPGAs or ASICs
- hls4ml – Python package for implementing machine learning algorithms in firmware via HLS
- ZeroBias / Ephemeral ZeroBias – Events recorded during data-taking at rate (before prescale) of LHC clock, where no signal is explicitly triggered on
- SWATCH – Control software for the CMS Level-1 Trigger hardware
- Prometheus – Monitoring software that queries SWATCH modules for real-time information on trigger systems
- ModelSim – Software used for simulating hardware description languages like VHDL
- Vivado – Software used for synthesis and analysis of FPGA designs
- L1_SingleMu22 – Unprescaled level-1 trigger seed requiring 1 muon with $p_T > 22$ GeV
- PUPPI - PileUp Per Particle Identification, pileup mitigation technique used in CMS

Introduction

- ML-based trigger algorithm, AXOL1TL, selects anomalous events in real-time
  - Signal-agnostic approach means sensitivity to wide variety of signals
- Variational autoencoder (VAE) trained on ZeroBias data to detect data outliers
- Information bottleneck in Gaussian-distributed latent space enforces efficient encoding ⇒ learning
- Calculated from μGT quantities
  - (p_T, η, ϕ) hardware integer inputs from: 1 MET, 4 e/γ, 4 μ, and 10 jets
  - The objects are p_T-sorted and the first N are selected
  - A higher number of objects is available (12 e/γ, 8 μ, 12 jets, 12 taus) and will be explored for future versions of the algorithm
- HLS implementation of μGT interface synthesizes VHDL for hardware
  - hls4ml-based CERN Gitlab repository for HLS dependencies
  - Generate bit files for MP7 global trigger boards
- Standalone HLS emulation of Anomaly Detection output for validating performance
- v1 of algorithm deployed on μGT Test Crate with L1 Menu containing 4 threshold triggers

Anomaly Detection Neural Network

The AXOL1TL anomaly detection uses a Variational Autoencoder (VAE). A dense feed-forward neural network reads in (p_T, η, ϕ) hardware inputs of 19 L1 objects. The encoder network computes a latent space vector of Gaussian probability distributions, $N(\mu_8, \sigma_8)$. The decoder network reconstructs the original input from the latent space.

$$\text{Loss} = (1-\beta)\left\|x-\hat{x}\right\|^2 + \beta\frac{1}{2}(\mu^2+\sigma^2-1-\log\sigma^2)$$

Reconstruction term — Full regularization term

Equation: VAE loss function. The reconstruction term is computed from the difference between the input (x) and output ($\hat{x}$) of the VAE. The second, full regularization term, is the Kullback–Leibler divergence (KL-divergence) between the latent space distribution and a standard normal distribution with mean μ and standard deviation σ. The parameter β can be tuned to balance the reconstruction performance with more efficient latent space encoding. At inference time, the loss is approximated by the mean-squared term $\Sigma\mu_i^2$ of the KL-divergence for latency considerations. This approximation has no impact on performance.

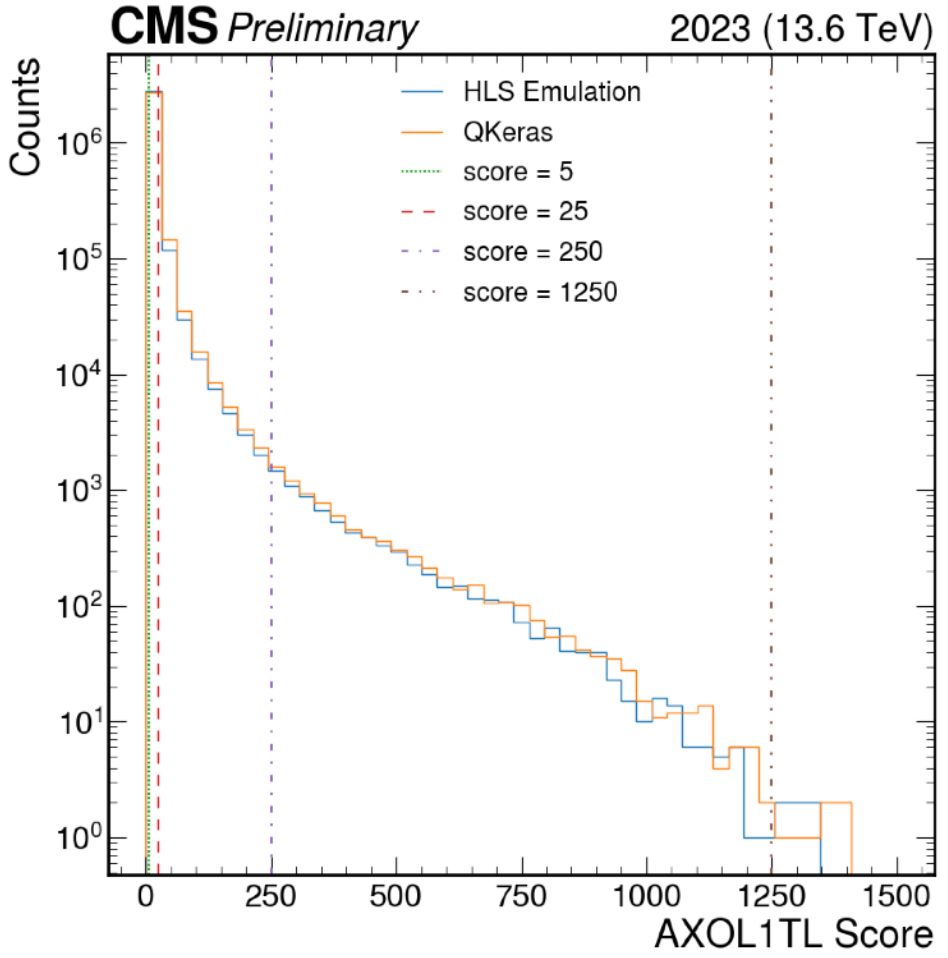
Dataset

- This algorithm has been trained using data collected by the CMS experiment in 2023 at a center-of-mass energy of $\sqrt{s}$=13.6 TeV, where lumi-leveling at pile-up 62 was used.

- A total of 10.5 million events are used. ~50% is used for the training itself and the other 50% are used for setting the anomaly thresholds based on the anomaly score distribution of those events

Anomaly Score and Thresholds

Anomaly score distributions for 2023 Ephemeral ZeroBias events. Individual event scores/losses for the QKeras model in Python (orange) and standalone HLS emulator (blue). Dotted lines represent scores that correspond to trigger paths in the μGT test crate.



Physics Performance

Efficiency improvement of AXOL1TL trigger bits to 2023 L1 Menu with respect to multiple SM and BSM signals. The model used is trained on Run 3 ZeroBias events. Efficiency gains for the BSM signal of Higgs decaying to two (pseudo)scalars of mass 15 GeV, where the (pseudo)scalars decay to bottom quark pairs, from AXOL1TL at various triggering rates are shown in the table. We also observe a significant improvement for several other signal models.

AXOL1TL Rate	1 kHz	5 kHz	10 kHz
Signal Efficiency Gain	46%	100%	133%

Firmware Validation

	Latency	LUTs	FFs	DSPs	BRAMs
AXOL1TL	2 ticks 50 ns	2.1%	~0	0	0

Vivado latency and resource utilization report for Anomaly Detection trigger on Xilinx Virtex-7 FPGA. Results show that firmware build meets L1 latency requirements, fitting within 2 clock cycles at 40 MHz. The resource utilization of the module is also relatively small.

The FPGA resources used are:

- LUTs (Look Up Tables) - general purpose logic
- FFs (Flip Flops) - registers
- DSPs (Digital Signal Processors) - multipliers
- BRAMs (Block Random Access Memories) - small high speed memories

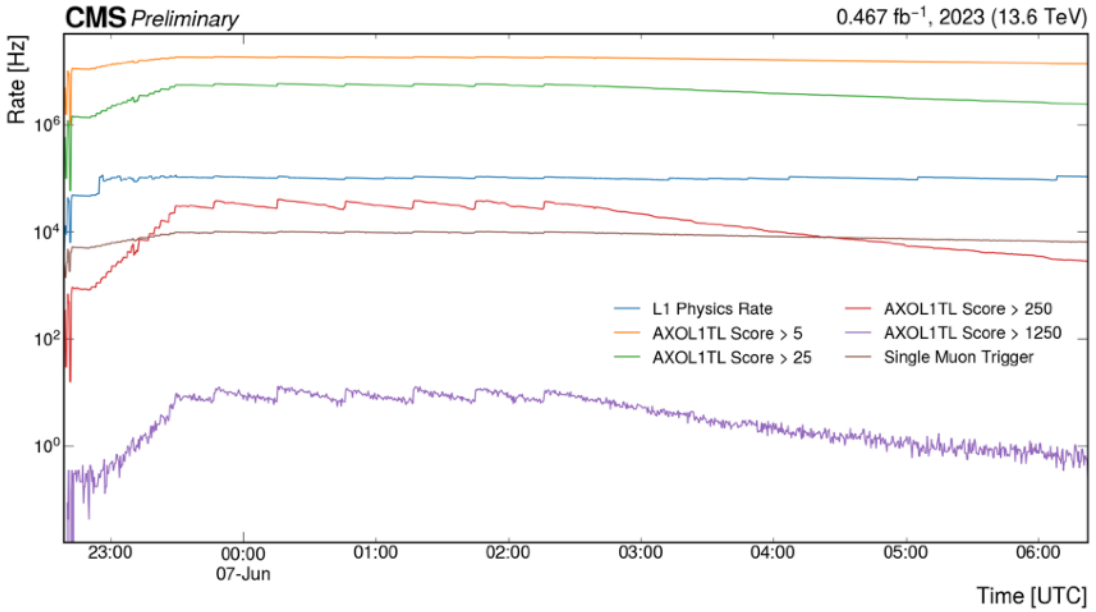
Firmware Validation

Test Crate firmware validation. The table shows trigger bits for the L1 menu including 4 anomaly detection thresholds: scores >1250, >250, >25, and >5 from top to bottom. Test vector column is generated from inference results of standalone emulator and HW count comes from standard global trigger firmware simulation workflow using ModelSim. Perfect bit agreement is observed.

Idx	L1 Menu Algorithm Name	Test Vector Count	HW Count	Agreement
94	L1_ADT_20000	0	0	✓
95	L1_ADT_4000	29	29	✓
103	L1_ADT_400	2618	2618	✓
108	L1_ADT_80	3331	3331	✓

Test vectors generated from Run 3 data

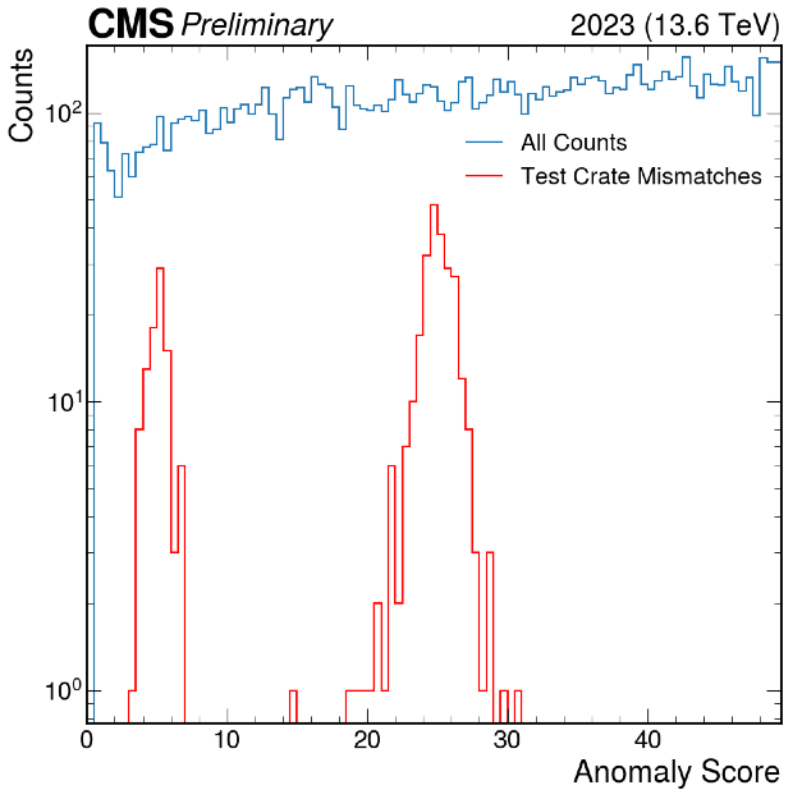
Test Crate Monitoring



Test crate rate monitoring time series. L1 trigger rates shown for 4 Anomaly Detection threshold triggers, overall L1 physics rate, and the L1_SingleMu22 un-prescaled single muon reference trigger. Time-averaged rates are read from μGT test crate SWATCH cell via Prometheus server at ~20s buffer rate during good data-taking conditions in 2023. AXOL1TL model is trained on 2018 ZeroBias data and thresholds are chosen to test possible range of accessible trigger rate. **Thresholds are not meant to model realistic trigger rates**. Consistent performance is shown over the course of partial fill-cycle. The turn-on corresponds to the beginning of an LHC fill and the sawtooth pattern corresponds to luminosity levelling.

Test Crate Validation

L1 Menu Algorithm Name	Test Crate Count	Standalone Emulator Count	Mismatches
L1_ADT_20000	1	1	0
L1_ADT_4000	742	741	19
L1_ADT_400	21236	21229	253
L1_ADT_80	25468	25481	93



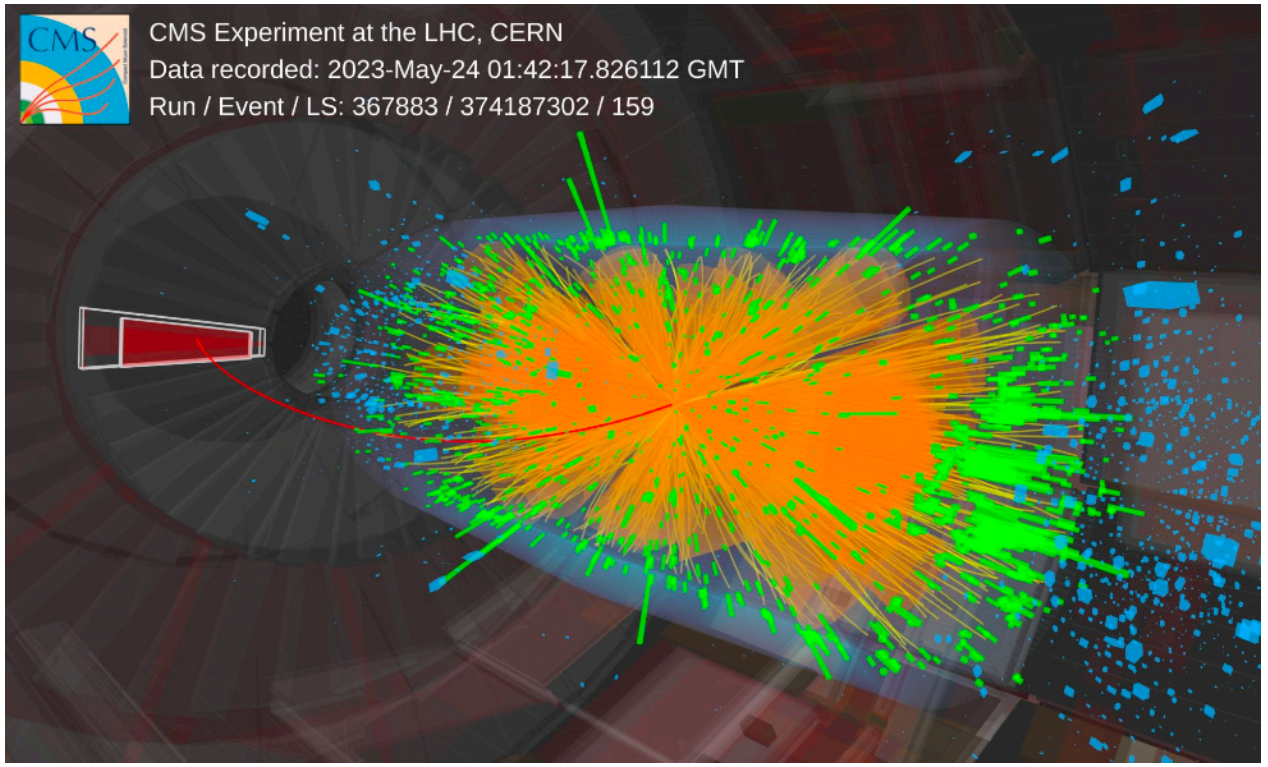
Anomaly Detection hardware vs. emulation trigger mismatches. Events from promptly reconstructed 2023 Ephemeral ZeroBias data where hardware bits are recorded from configured μGT test crate. In table (left), Test Crate Count shows events triggered in hardware and read out into data and Standalone Emulator Count is evaluated via offline inference with L1 objects. Anomaly score distribution of all events (right): red segments represent mismatches between hardware and emulation. Clustering near decision boundaries implies issue is due to precision/rounding problem. Minimal mismatches in hardware vs. emulation (≤ 1%) observed.

Event Display

Event display of the highest anomaly score event that is not selected by the normal L1T menu, from Ephemeral Zero Bias 2023 Run 367883.

This event features the maximal number of L1 jets (12), out of which 11 have $E_T > 20$ GeV. It also features a 3 GeV L1 muon. The offline reconstruction identifies 7 jets (reconstructed with the PUPPI algorithm) with $p_T > 15$ GeV, and 1 muon.

The event is also characterized by a very unlikely large number of reconstructed vertices (75), given the pile up profile of the data taken in Run 2 and Run 3.



References

- CMS Collaboration. "CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter". CERN-LHCC-2012-015, CMS-TDR-10 (2012). https://cds.cern.ch/record/1481837.
- M. Jeitler, et al. "The level-1 global trigger for the CMS experiment at LHC". JINST 2, P01006 (2007). https://doi.org/10.1088/1748-0221/2/01/P01006.
- E. Govorkova, et al. "Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider". Nat. Mach Intell. 4, 154 (2022). https://doi.org/10.1038/s42256-022-00441-3.
- FastML Team. hls4ml (Version v0.7.1) [Computer software]. https://doi.org/10.5281/zenodo.1201549.
- J. Duarte, et al. "Fast inference of deep neural networks in FPGAs for particle physics". JINST 13, P07027 (2018). https://doi.org/10.1088/1748-0221/13/07/P07027.
- Xilinx Virtex-7 FPGA. https://www.xilinx.com/products/silicon-devices/fpga/virtex-7.html.
- L1 Menu Repository. https://github.com/herbberg/l1menus.