PROCEEDINGS OF SCIENCE

PoS(BPU11)088

The CMS High Level Trigger System

Milos Dordevic on behalf of the CMS Collaboration[a,*]

[a] *Vinca Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade,*
Mike Petrovica Alasa 12-14, 11351 Vinca, Belgrade, Serbia

E-mail: milos.djordjevic@cern.ch

The CMS experiment at CERN uses a two-level triggering system that is composed of the Level-1 (L1), instrumented by custom-designed electronics with an output rate of 100 kHz, and the High Level Trigger (HLT), a streamlined version of the offline software reconstruction running on a computer farm, with around 1.5 kHz of physics rate stored for further analysis. New trigger algorithms and also new features, as well as an optimized trigger menu at the HLT, are essential in order to be able to successfully record events at higher data loads due to increasing luminosity and pileup at the LHC in Run 3 which has just started. Many measurements and searches will profit from the updates implemented in the CMS trigger. The highlights of Run 2 CMS trigger results will be presented in this proceedings, together with improvements for Run 3.

11th International Conference of the Balkan Physical Union (BPU11),
28 August - 1 September 2022
Belgrade, Serbia

[*] Speaker

1. Introduction

The Compact Muon Solenoid (CMS) is a general purpose experiment [1] at the Large Hadron Collider (LHC) at CERN. It has several sub-detector layers, the innermost being a silicon tracker, followed by electromagnetic calorimeter (ECAL) and a hadron calorimeter (HCAL). These three are immersed in a magnetic field of 3.8 T provided by a superconducting solenoid. Outside of the solenoid volume are muon sub-detectors stationed in a flux-return yoke. In the CMS experiment the Particle Flow (PF) algorithm [2] is used for event reconstruction, exploiting all sub-detector information.

The proton bunches which circulate in the LHC are time spaced with 25 ns intervals for the two beams running in opposite directions. The bunches collide at a 40 MHz rate. However, only around 1.5 kHz was kept for the data analysis at the end of Run 2, the data taking period at the LHC from 2015 to 2018. The limiting factors for recording more data online are the available storage and also computing resources allocated for the data processing. The CMS Trigger System performs a selective readout of data in real time, keeping only events of interest for physics, but at the same time having an efficient decision complying with the rate constraints of the CMS computing farm.

2. The CMS Trigger System design and implementation

The Trigger System of the CMS experiment is organized in two tiers, first being the Level-1 trigger (L1) [3] using custom-made electronics that is reducing the event rate to about 100 kHz with 3.8 μs latency, and the High Level Trigger (HLT) filtering the events with a software running on computing farm. The HLT is based on commercial CPUs, reducing the rate to about 1.5 kHz. Heterogeneous computing resources are used in Run 3, the data taking period at the LHC which started in 2022, with both CPU and GPU reconstruction.

3. Performance of the High Level Trigger System

The HLT system brings the L1 Trigger output rate of about 100 kHz down to around 1.5 kHz. The HLT uses offline reconstruction algorithms, optimized to be around a hundred times faster. At CMS the HLT selection is made with a trigger menu running of hundreds of HLT paths, targeting a broad range of physics signatures. The HLT path is composed of a sequence of reconstruction and filtering modules, arranged and executed in increasing complexity. If a particular event is rejected by a filter of HLT path, the subsequent modules in the same HLT path are not run.

The reconstruction of tracks in Run 3 at the HLT has been significantly revised with respect to Run 2 and is now performed using only a single global iteration, instead of several tracking iterations in Run 2, and presently seeded by the pixel tracks reconstructed by the Patatrack algorithm [4]. Figure 1 presents the tracking efficiency as a function of the simulated track p_T for the Run 2 HLT tracking and the Run 3 HLT single-iteration tracking, which clearly improved over the Run 2. The efficiency at large track p_T is lower due to the presence of a large number of tracks in high p_T jets, where the high track density and limitations of the silicon detector pitch lowers the capacity to disentangle hits from overlapping particles. Figure 2 presents the tracking fake rate as a function of the reconstructed track p_T. In the Run 3 HLT tracking no track with $p_T < 0.3$ GeV is reconstructed [4]. Hence, the total number of fake tracks in Run 2 HLT tracking is larger with respect to Run 3.

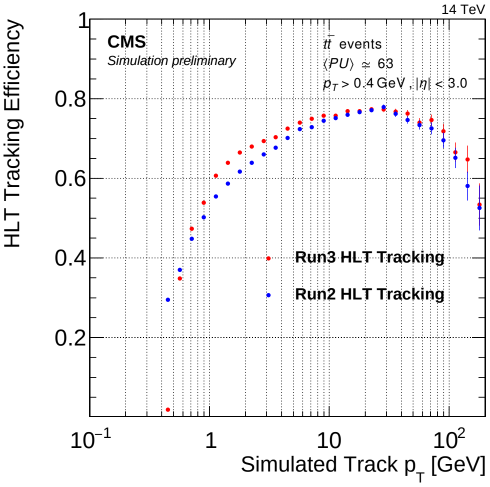

Figure 1: Tracking efficiency as a function of the simulated track p_T for the Run 2 HLT tracking (blue) and the Run 3 HLT single-iteration tracking (red)[4].

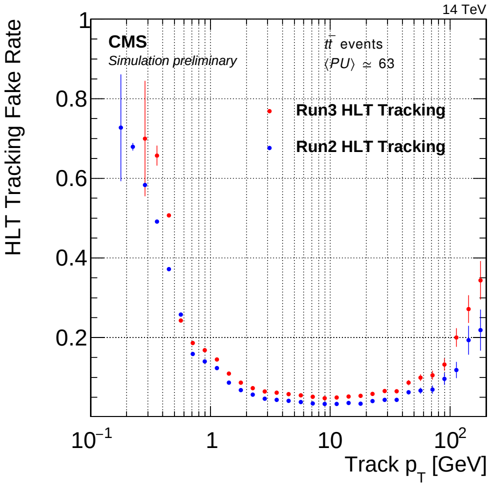

Figure 2: Tracking fake rate as a function of the reconstructed track p_T for the Run 2 (blue) and the Run 3 HLT single-iteration tracking (red)[4].

A neural network based classifier named Deep CSV (Combined Secondary Vertex) is used at CMS as of 2017 to identify the b-tagged jets [5]. Performance of b-tagged jets at the HLT was measured in di-lepton $t\bar{t}$ events, requiring one isolated electron and one isolated muon and two jets in the event. The Deep CSV discriminator for online PF jets is shown in Figure 3, with the individual contributions from different jet flavours. Figure 4 shows the performance of the online (red and blue) and offline (black) b-jet identification efficiency demonstrating the probability for a light-flavor jet to be misidentified as b-jet as a function of the efficiency to correctly identify a b-jet. The performance of the CSV (dashed) and DeepCSV (solid) algorithms are shown for comparison.

Performance of the online selection of electrons at the HLT is shown in Figure 5, for one characteristic HLT path, with respect to an offline reconstructed electron as a function of the electron p_T in different η regions [6]. The efficiency of HLT path that requires a photon with p_T higher than 200 GeV, used in Supersymmetry and other new physics searches, is shown in Figure 6 [7].

The efficiency of reconstructing jets at the HLT is measured using events collected by single muon trigger and it is shown in Figure 7 as a function of the offline reconstructed jets [8], shown for different HLT jet p_T thresholds. Both offline and online jets are reconstructed using the PF algorithm. The HLT jets were matched with the offline jets. The absolute value of the pseudorapidity for the jets is restricted to 2.4. The efficiency of HLT requiring an isolated single muon with p_T >24 GeV is shown in Figure 8 as a function of muon p_T [9], before and after the updated reconstruction in 2018. The efficiency is estimated with respect to the offline muon matched to L1 trigger object with $\Delta R < 0.3$ as well as passing tight ID and PF-based isolation requirements.

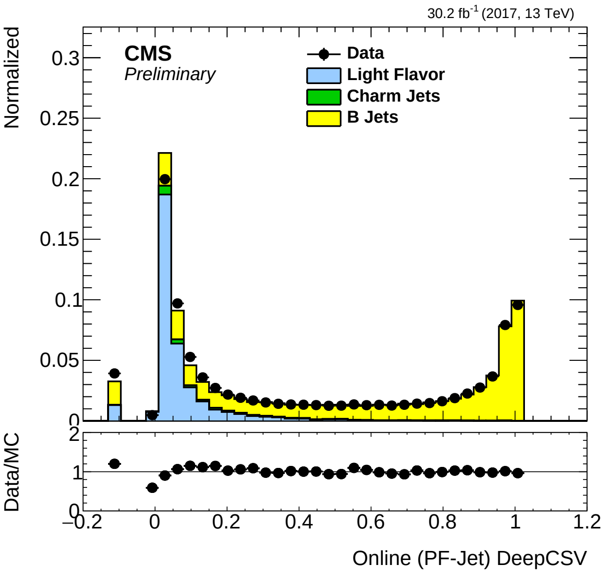

Figure 3: The DeepCSV discriminator distribution for online (PF-Jets). Different colours show the contributions in simulations arising from different jet flavors[5].

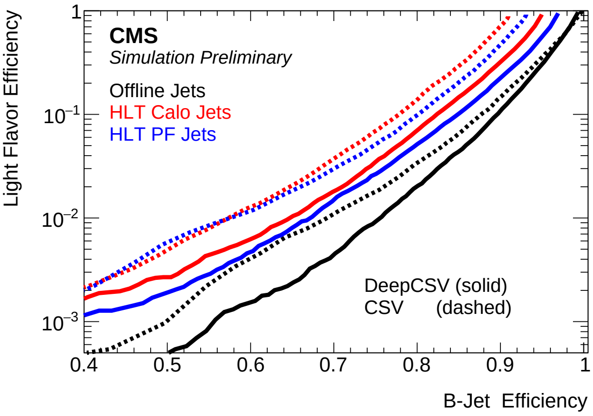

Figure 4: Performance of the online (red and blue) and offline (black) b-jet identification efficiency: a probability for a light-flavor jet to be misidentified as b-jet as a function of efficiency to identify b-jet[5].

4. Trigger Menus for Data Selection

The trigger menu represents a large set of selection criteria in the online event selection, enabling to fulfil the broad physics program of the CMS experiment. Each trigger menu consists of general or multi-purpose triggers, but also some specific triggers used only in particular studies and searches, as well as backup triggers. The trigger menu used at the HLT has about 600 items.

Figure 9 shows the HLT rate per physics group in 2018, presenting the total, shared and pure rates [10]. The total rate is the rate from an event assigned to all groups that trigger the event. The shared rate is the rate from an event shared equally among all groups that trigger the event. The

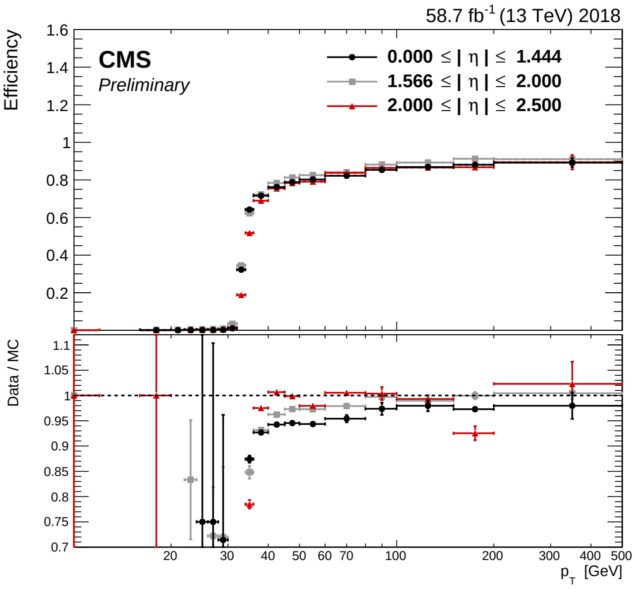

Figure 5: Efficiency of a single electron trigger with p_T >32 GeV, with respect to an offline reconstructed electron as a function of the electron p_T, obtained for different η regions with full 2018 dataset [6].

pure rate is the rate from an event assigned to a given group if it is the only one that triggers the event. The rates were evaluated by running the HLT menu on a commissioning data set. The latter is made of a fraction of events passing the L1 trigger, without any further HLT requirement. The average recorded instantaneous luminosity in the input data was $1.2 \times 10^{34} cm^2 s^{-1}$. All rates were scaled to $2.0 \times 10^{34}\mathrm{cm}^2\ \mathrm{s}^{-1}$.

5. GPU-based acceleration at High Level Trigger

In the CMS experiment, a heterogeneous architecture is used in the online reconstruction in Run 3, comprising CPUs and GPUs [11]. The online reconstruction of pixel-based tracks and vertices is now performed also on GPUs. The ECAL and HCAL local reconstruction were ported to GPUs as well and more reconstruction code is foreseen to be ported to GPUs in the near future. The pie-charts in Figures 10 and 11 show the distribution of CPU-only and GPU-enabled time, respectively, spent in different instances of CMS software framework (CMSSW) modules (outermost ring), their corresponding C++ class (one level inner) and grouped by physics object or detector (innermost ring). The empty slice indicates the time spent outside of the individual algorithms. The timing is measured on a pileup of 50 events from 2018, running 4 jobs in parallel, with 32 threads each, on a node (2x AMD "Rome" 7502) with SMT enabled, and, in the case of a heterogeneous architecture, additionally equipped with NVIDIA T4 GPU. Around 25% of the CPU time is offloaded to GPU.

6. Data Scouting and Data Parking

There are alternative strategies with respect to the standard online data taking, in order to surpass the limited storage and computer processing resources. One option is called Data Scouting,

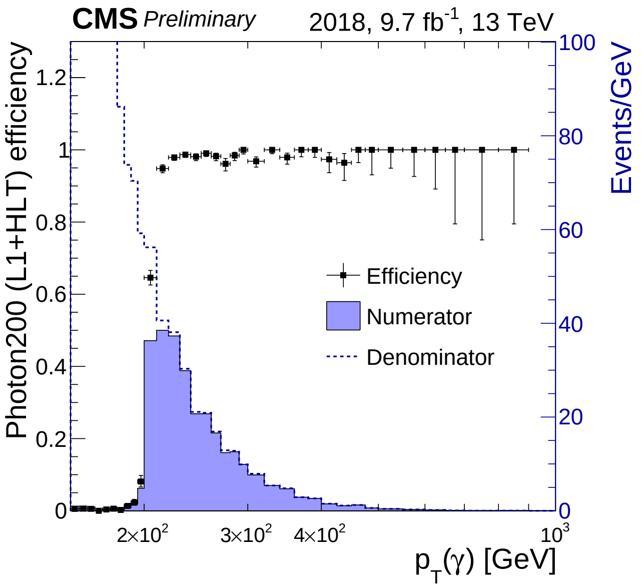

Figure 6: Efficiency of a single electron trigger with p_T >32 GeV, with respect to an offline reconstructed electron as a function of the electron p_T, obtained for different η regions with full 2018 dataset [6].

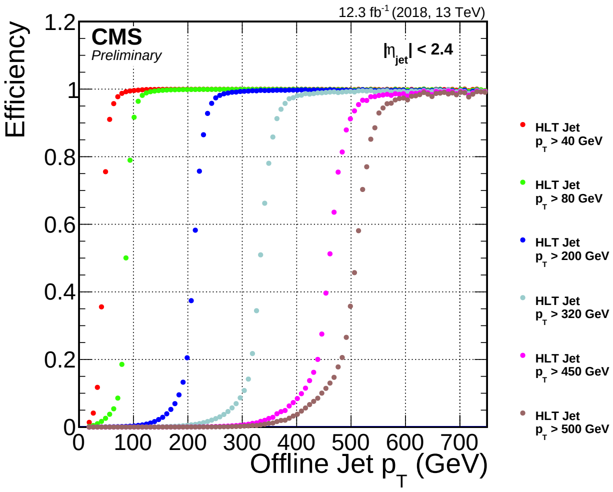

Figure 7: Efficiency to reconstruct jets at the HLT as a function of the offline reconstructed jet p_T[8]. The different colours correspond to different HLT jet p_T thresholds applied.

where only a small summary of the reconstructed event quantities is saved and not all the raw data, thus reducing the event size, to be able to record the events at a higher rate [12]. An example of the usage of Data Scouting is presented in Figure 12 where the di-muon events are recorded largely unconstrained by requirements on muon kinematics otherwise imposed by CMS data acquisition and event reconstruction workflows. For low invariant mass dimuon events the data taking efficiency is improved by one or two orders of magnitude. Another option is Data Parking, where the data is parked on tape, skipping the prompt reconstruction, thus reducing the required computing resources, and reconstructing the data later, in the shutdown period when the experiment is not taking data[13].

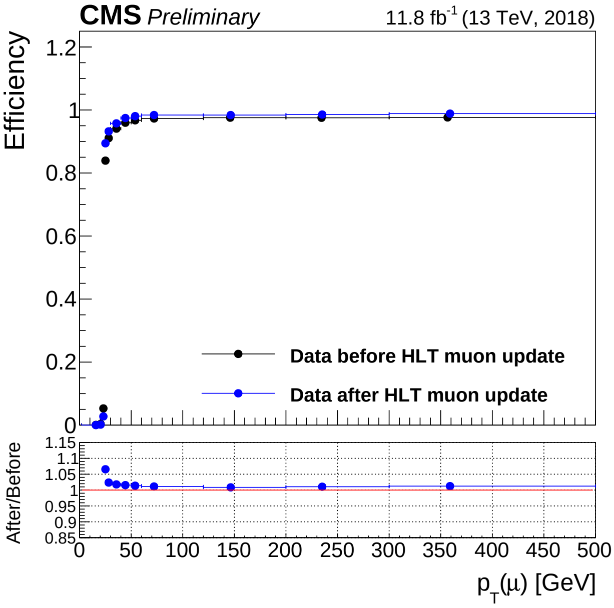

Figure 8: HLT efficiency requiring isolated single muon with p_T > 24 GeV, calculated as a function of muon p_T [9].

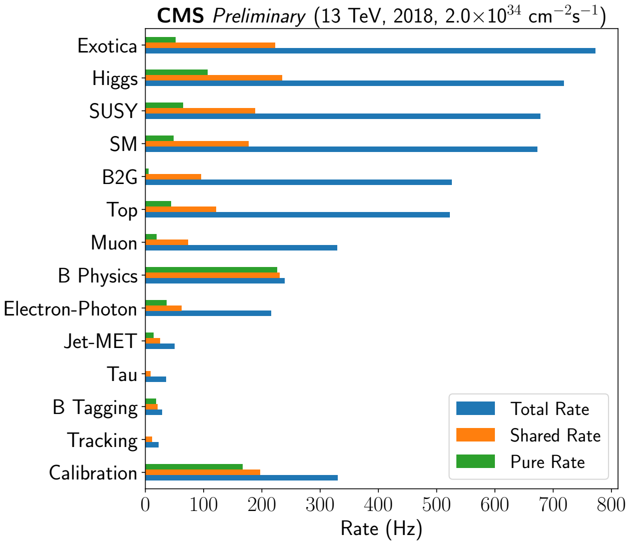

Figure 9: The HLT rates consumed by each CMS physics group during 2018 data taking [10].

7. Summary and Outlook for Run 3

The CMS High Level Trigger System has proved to be robust, flexible and effective during Run 1 and Run 2. It has also shown to be able to deal with a large number of events required to fulfil the CMS physics goals. Excellent performance was obtained in LHC Run 2, from sharp efficiency curves to only a moderate pileup dependence. In Run 2 and during the Long ShutDown 2 (LS2), many new technologies were integrated into the CMS HLT System. Also, many trigger algorithms

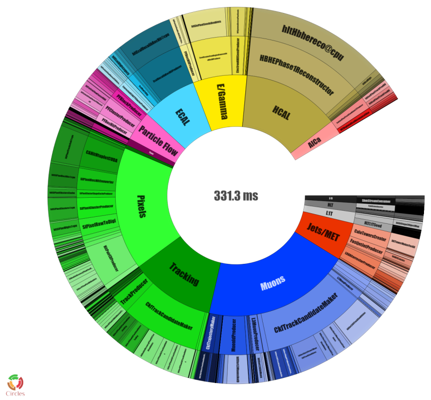

Figure 10: Pie-chart shows the distribution of CPU-only time in different instances of CMSSW modules [11].

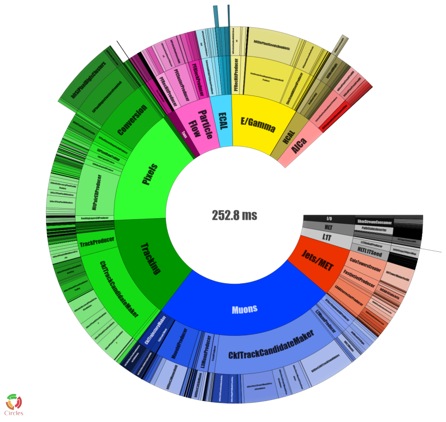

Figure 11: Pie-chart shows the distribution of CPU time with part of the reconstruction offloaded to GPU, in different CMSSW modules [11].

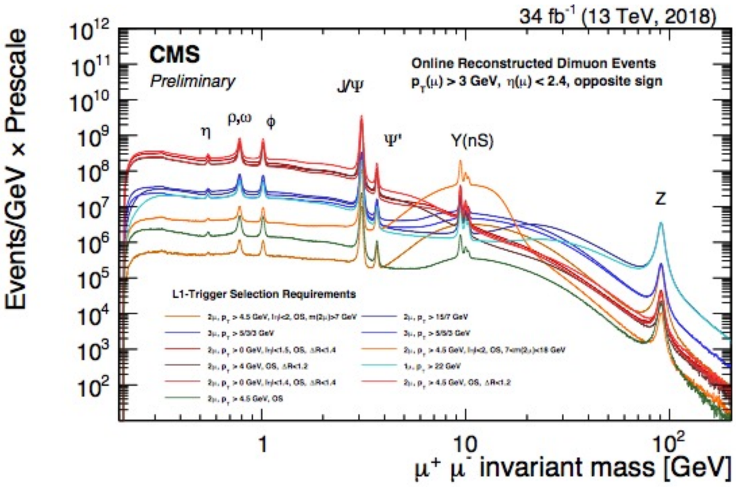

Figure 12: Dimuon invariant mass spectra reconstructed in the High Level Trigger system of the CMS detector for various muon Level-1 trigger requirements deployed by the CMS collaboration in 2017. For a subset of Level-1 requirements CMS recorded only a fraction of the data [12].

were improved and innovated [14]. The HLT reconstruction was improved following the Phase-I Pixel upgrade [15] from 2016 to 2017. HCAL has undergone the Phase-I Upgrade in the endcap and barrel, during 2018 and 2019, respectively [16]. One particular improvement is introduced for Run 3, referred to as the heterogeneous reconstruction comprising CPU and GPU processors.

8. Acknowledgement

The research was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (grant number 451-03-47/2023-01/ 200017).

References

[1] The CMS Collaboration, JINST 3 (2008) S08004.

[2] The CMS Collaboration, JINST 12 (2017) P10003.

[3] The CMS Collaboration, JINST 15 (2020) P10017.

[4] The CMS Collaboration, CMS-DP-2022/014.

[5] The CMS Collaboration, CMS-DP-2019/042.

[6] The CMS Collaboration, CMS-DP-2020/016.

[7] The CMS Collaboration, CMS-DP-2018/049.

[8] The CMS Collaboration, CMS-DP-2018/037.

[9] The CMS Collaboration, CMS-DP-2018/034.

[10] The CMS Collaboration, CMS-DP-2018/057.

[11] The CMS Collaboration, CMS-DP-2021/013.

[12] The CMS Collaboration, CMS-DP-2018/055.

[13] The CMS Collaboration, CMS-DP-2019/043.

[14] https://home.cern/press/2022/CMS-upgrades-LS2.

[15] The CMS Collaboration, CERN-LHCC-2013-011; CMS-TDR-12.

[16] The CMS Collaboration, CERN-LHCC-2012-016 ; CMS-TDR-11.