The CMS Trigger System

The CMS Trigger System

Milos Dordevic on behalf of the CMS Collaboration

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

Abstract.

The CMS experiment at CERN uses a two-stage triggering system composed of the Level-1 (L1), instrumented with custom-designed hardware boards with an output rate of 100 kHz, and the High Level Trigger (HLT), streamlined version of the offline software reconstruction that runs on the computing farm, allowing to store around 1.5 kHz of rate. New trigger algorithms and new features, as well as optimized trigger menus at both L1 and HLT are mandatory in order to be able to successfully record the events at higher data loads due to increasing luminosity and pileup at the LHC in Run 3. 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, together with the improvements for Run 3.

1. Introduction

The Compact Muon Solenoid (CMS) is a general purpose experiment [1] built at the Large Hadron Collider (LHC) at CERN. It is made of several sub-detector layers, comprising a silicon tracker, a homogeneous electromagnetic calorimeter (ECAL) and a sampling hadron calorimeter (HCAL) being immersed in a magnetic field of 3.8 T produced by a superconducting solenoid. Outside of the solenoid volume there are muon sub-detectors positioned in a flux-return yoke. CMS uses Particle Flow algorithm [2] for event reconstruction with all sub-detector information.

The circulating proton bunches at the LHC are time spaced with 25 ns intervals in two beams running in opposite directions. These proton bunches collide at 40 MHz rate, but only about 1 kHz was kept for the data analysis at the end of Run 2. The limitation factors to record more data online are the available storage capacities and also computing resources for the data processing. The CMS Trigger System provides a selective readout of data in real time, keeping only events of interest for further physics analysis, aiming at an efficient decision within the rate constraints.

2. The CMS Trigger System design, architecture and implementation

The CMS Trigger System is organised in two tiers, the Level-1 trigger (L1) based on custom-made electronics that is reducing the event rate to about 100 kHz with 3.8 μs latency, and the High Level Trigger (HLT) that is filtering the events with a software running on computing farm based on the commercial CPUs, which further reduced the rate to about 1 kHz in the Run 2. Heterogeneous computing resources are allocated for Run 3 using CPU and GPU reconstruction.

Figure 1 shows a diagram of the upgraded CMS Level-1 trigger system during Run 2. At L1, each event is analysed by the Muon and Calorimeter trigger. The Muon trigger consists

of three muon detection systems used early in the processing chain of the trigger, in order to improve the efficiency and resolution, but also the trigger rate. The Calorimeter trigger is used for reconstructing electrons, photons, tau candidates, jets and energy sums. No tracking readout is used at L1 trigger stage. The Global Trigger combines the various objects provided by the Global Muon Trigger and Layer 2 Calorimeter Trigger. A large set of around 400 requirements is performed on the trigger objects, applied in a logical "OR", comprising the L1 trigger menu. At L1 trigger, the reconstruction of electrons and photons is performed using the cluster shape and electromagnetic fraction information to discriminate against jets. The reconstruction of jets at L1 trigger is done using a sliding window algorithm that looks for trigger tower seeds with an energy over a given threshold. The 9×9 trigger towers are summed in order to match the offline reconstructed jets of the cone size $\Delta R = 0.4$ after which the jets are pileup-subtracted and calibrated. The H_T variable is calculated by summing the jet energies with restriction to pseudorapidity η, while the missing E_T variable at L1 trigger is calculated by summing all trigger towers over the η and pileup dependent energy threshold E_T in the full η range. The muons are reconstructed using an extrapolation based track finding in the barrel region, and a pattern based track finding in the overlap and endcap region, where also the BDT regression algorithms are applied to improve the reconstruction [3].

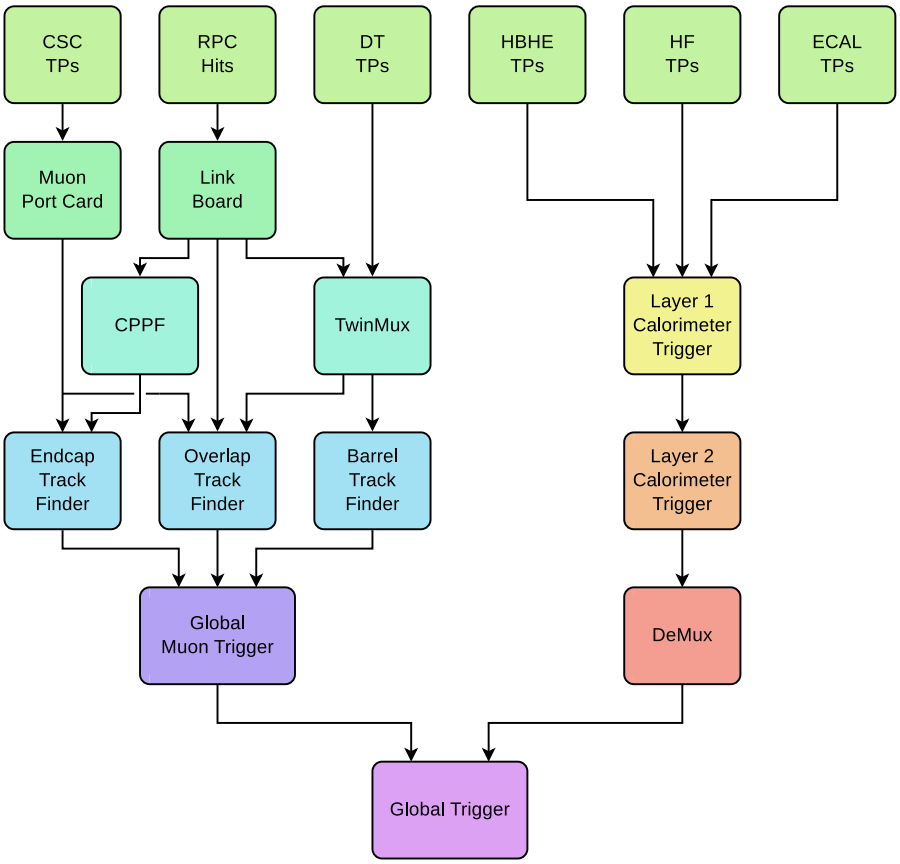

Figure 1. Diagram of the upgraded CMS Level-1 trigger system during Run 2 [3].

3. Performance of the Level-1 Trigger System

The selection of results demonstrating the performance of the L1 trigger in reconstructing of the most relevant physics objects is outlined here. Figure 2 (left) presents the efficiency to

reconstruct muons at L1 trigger, as a function of the offline reconstructed muon transverse momentum, shown for three different η ranges and also inclusive up to 2.4 [3]. The efficiency is sharpest in the barrel due to a better resolution for reconstruction of muons in the barrel region. Figure 2 (right) shows the efficiency as a function of the pseudorapidity of the offline reconstructed muon, that exhibits a moderate drop in the forward region. This is due to the fact that in the forward region, corresponding to the EMTF part of the L1 trigger, no detector redundancy is available, as only the Cathode Strip Chambers (CSC) out of the three muons systems of CMS are used. The assignment of muon transverse momentum (p_T), is more difficult due to reduced lever arm and more showering that occurs in the forward region of the detector. The L1 efficiency of the muon track finder is shown to have a small dependence on the number of offline vertices, presenting it to be stable against pileup, and also to have a flat distribution versus the azimuthal angle ϕ [3].

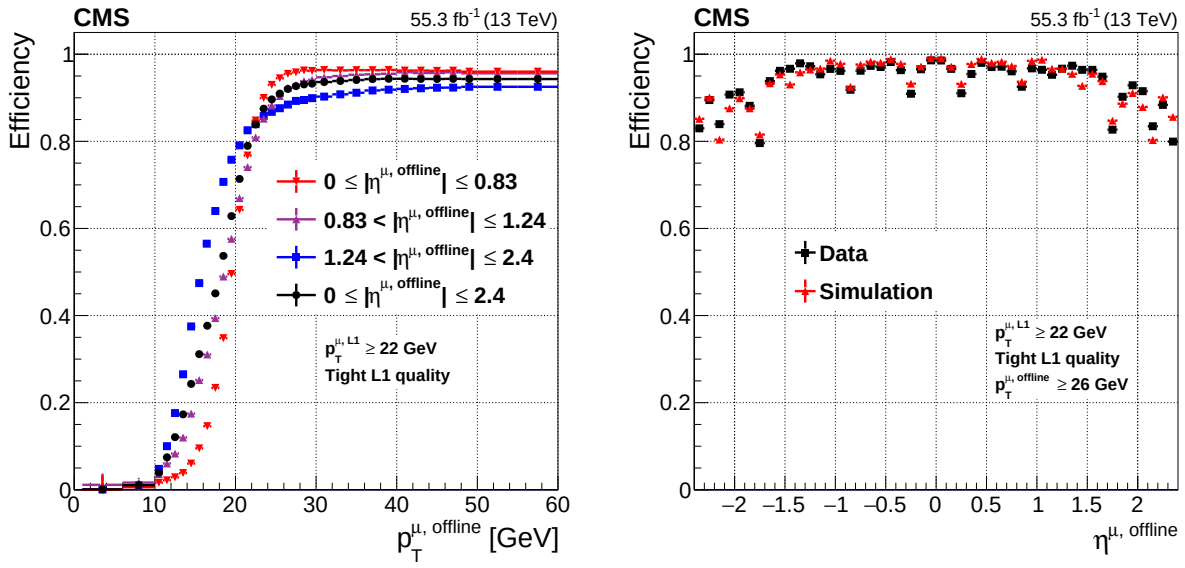

Figure 2. The left plot shows the L1 muon trigger efficiency for data as a function of the offline reconstructed muon $p_T^{\mu,offline}$ for each η region: barrel (red), overlap (violet), endcap (blue) and the total (black) [16]. The right plot shows the L1 muon efficiency for data and simulation as a function of the offline reconstructed muon pseudorapidity [16].

Figure 3 presents the sharp L1 trigger efficiency for reconstructing electrons or photons over given online threshold of 30 and 40 GeV, with respect to the offline reconstructed electron transverse energy [3]. Figure 4 shows only a small dependence of the electron and photon reconstruction and isolation requirements at the L1 with respect to the number of vertices (pileup) [3].

In Figure 5 the single tau object trigger efficiency is presented for the different working points of 26, 30 and 34 GeV for the inclusive tau trigger at the L1 [3]. The energy of the tau leptons is calibrated at the L1 as a function of energy and pileup. The resolution of the tau lepton transverse momentum reconstructed at the L1 trigger is at about 20 to 25 % in the range of 30 to 50 GeV. Figure 6 presents the efficiency curves for the L1 jet triggers for several characteristic thresholds of 35, 90, 120 and 180 GeV in the barrel pseudorapidity region ($|\eta| < 3$) [3].

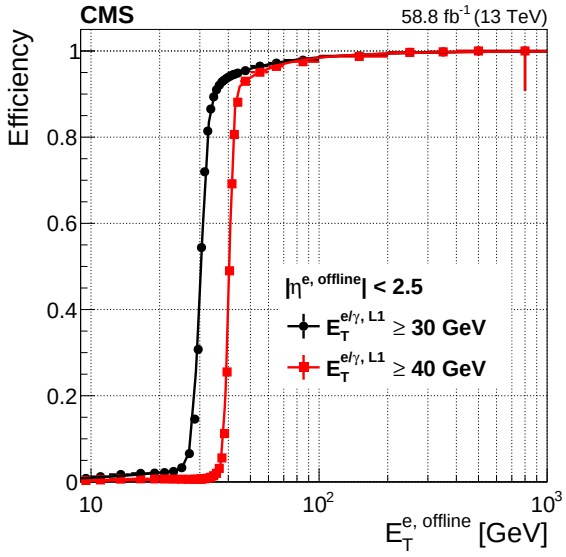

Figure 3. The L1 e/γ trigger efficiency as a function of the offline reconstructed electron E_T for thresholds of 30 and 40 GeV[3].

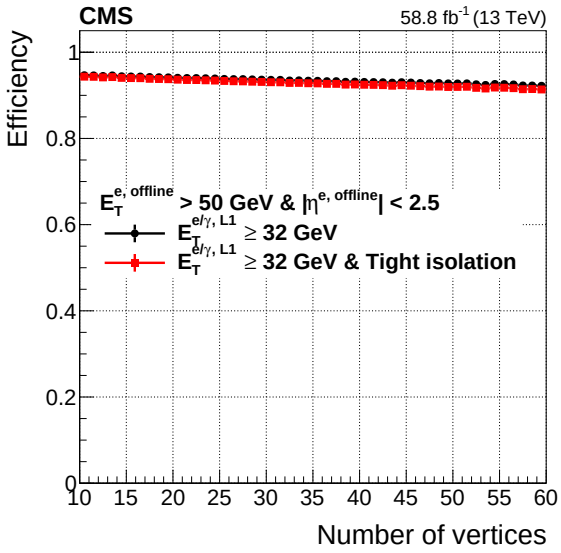

Figure 4. The L1 e/γ isolated trigger efficiency, presented as a function of the offline reconstructed vertices[3].

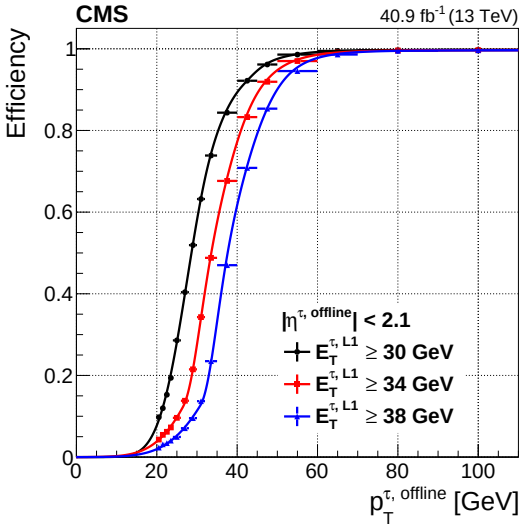

Figure 5. The L1 τ trigger efficiency, as a function of the offline reconstructed τ lepton p_T, for typical thresholds of 30, 34, and 38 GeV [3].

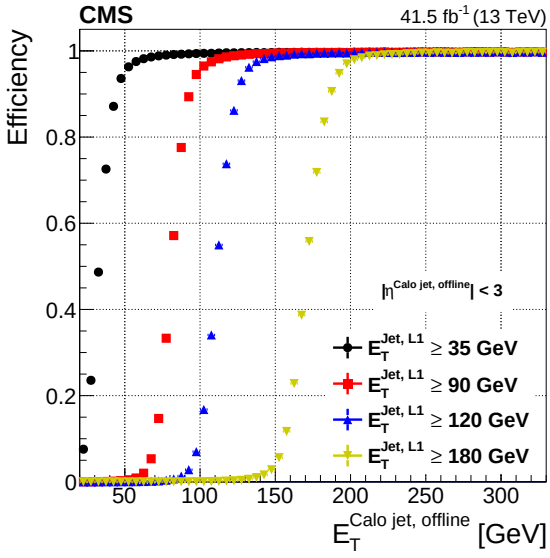

Figure 6. Efficiency curves for the Level-1 jet trigger for the barrel pseudorapidity range [3].

4. Performance of the High Level Trigger System

The High Level Trigger system reduces the L1 Trigger output rate of about 100 kHz further down to around 1 kHz at the end of Run 2, with an estimated HLT rate at around 1.5 kHz for the Run 3. The HLT uses offline reconstruction algorithms, but optimised to run around

hundred times faster. It is composed of hundreds of HLT paths, targeting a broad range of event topologies, that consist of reconstruction and filtering modules executed in sequences. When an event is rejected by a filter, the subsequent modules within the same path are not run.

Figure 7 presents the tracking efficiency as a function of the simulated track p_T for the Run-2 HLT tracking, given in blue, and the Run-3 HLT single-iteration tracking, shown in red [4]. The Run-3 tracking clearly improved over the previously used for the Run-2. The efficiency at large track p_T is reduced due to the presence of a large number of tracks in the core of high p_T jets. Figure 8 shows the tracking fake rate as a function of the reconstructed track p_T. In the Run-3 HLT tracking, unlike in Run-2, no track with $p_T < 0.3$ GeV is reconstructed [4]. As a consequence, the total amount of fake tracks in Run-2 HLT tracking is sensibly larger with respect to Run-3.

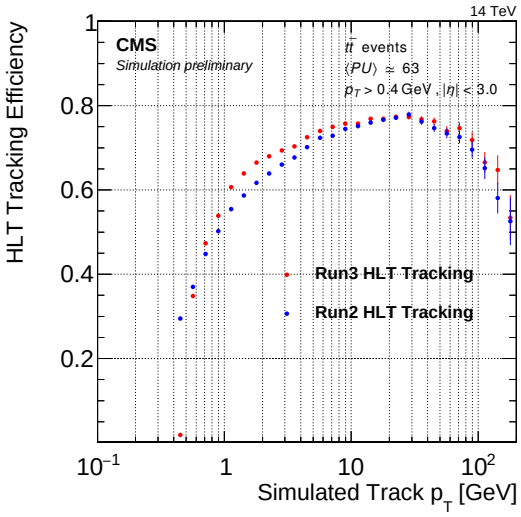

Figure 7. The tracking efficiency is shown as a function of the simulated track p_T for the Run-2 HLT tracking (blue) and the Run-3 HLT single-iteration tracking (red)[4].

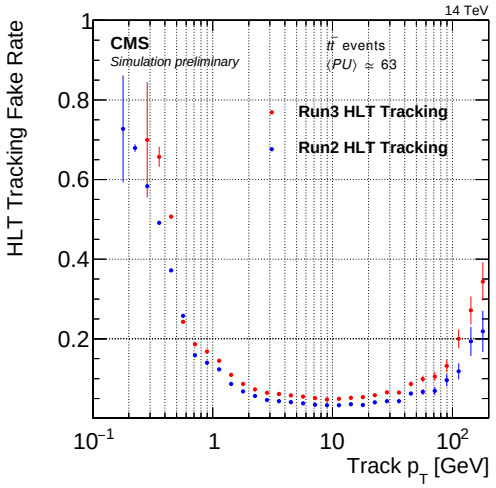

Figure 8. The tracking fake rate is shown as a function of the reconstructed track p_T for the Run-2 HLT tracking (blue) and the Run-3 HLT single-iteration tracking (red)[4].

Figure 9 shows the efficiency to identify b-jets online, compared to the corresponding offline reconstruction [5]. A neural network based classifier called Deep CSV (Combined Secondary Vertex) is used at CMS since 2017 for the identification of b-tagged jets. The improvement over the previously used CSV algorithms ranges from 5 to 15 % at the same light flavor efficiency. The distribution of the Deep CSV discriminator for online PF jets is shown in Figure 10, where different colours present the individual contributions that are coming from different jet flavors.

The performance of the online selection of the electrons at the High Level Trigger is shown in Figure 11, for one characteristic HLT path, with respect to an offline reconstructed electron as a function of the electron p_T, obtained for different η regions using the full 2018 dataset [6]. The efficiency of a characteristic HLT path that requires a photon with the p_T higher than 200 GeV, used for example in Supersymmetry and other new physics searches, is in Figure 12 [7].

The efficiency to reconstruct jets at the HLT is measured in single muon events and shown in Figure 13 as a function of the offline reconstructed jets [8]. 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 high-level

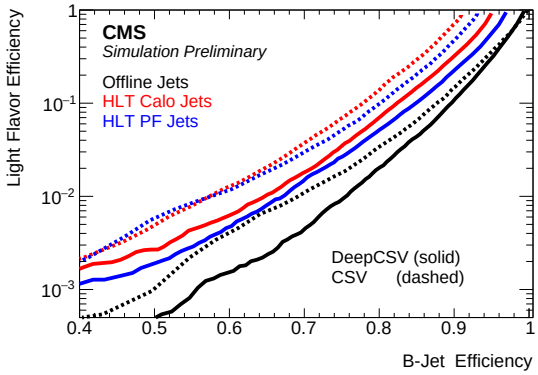

Figure 9. Performance of the online (red and blue) and offline (black) b-jet identification efficiency demonstrating the probability for a light-flavor jet to be misidentified as b-jet as a function of the efficiency to correctly identify a b-jet[5].

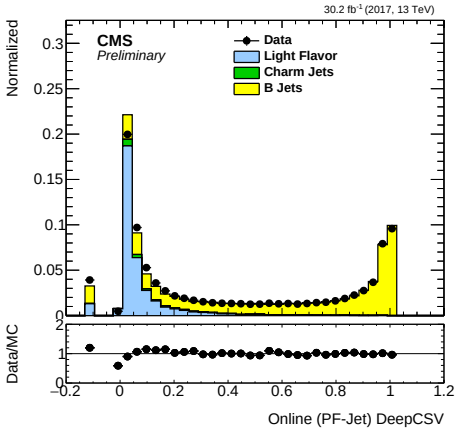

Figure 10. The DeepCSV discriminator distribution for online (PF-Jets). Different colours show the contributions in simulations from different jet flavors[5].

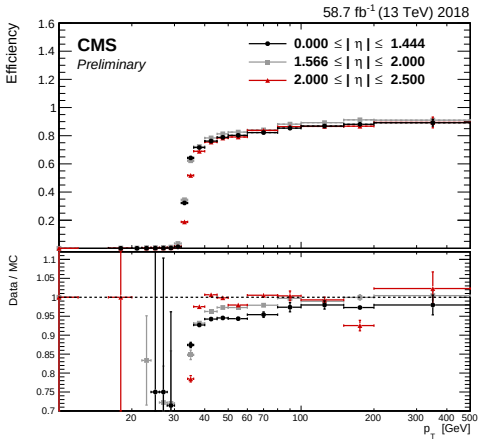

Figure 11. The efficiency of a single electron trigger with p_T >32 GeV, with respect to an offline reconstructed electron as a function of the electron p_T, obtained for different ηregions using the full 2018 dataset [6].

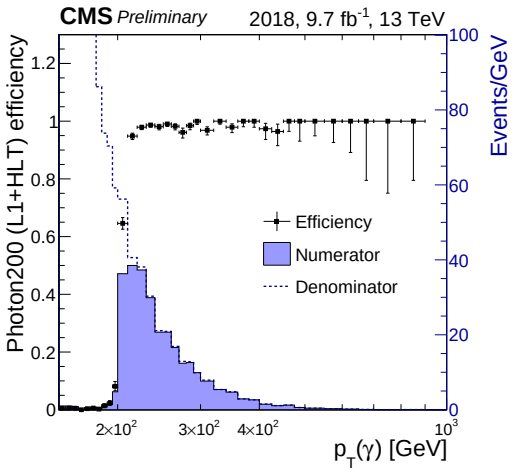

Figure 12. The efficiency of a High-Level Trigger (HLT) algorithm that requires a photon with a transverse momentum p_T greater than 200 GeV[6].

trigger requiring isolated single muon with p_T >24 GeV is shown in Figure 14 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.

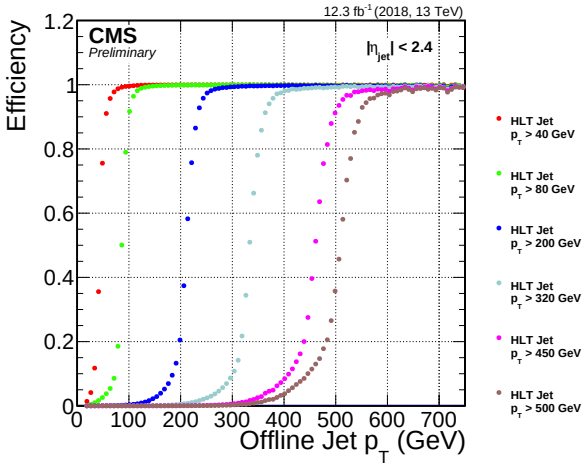

Figure 13. The efficiency to reconstruct jets at the HLT as a function of the offline reconstructed jet p_T[8].

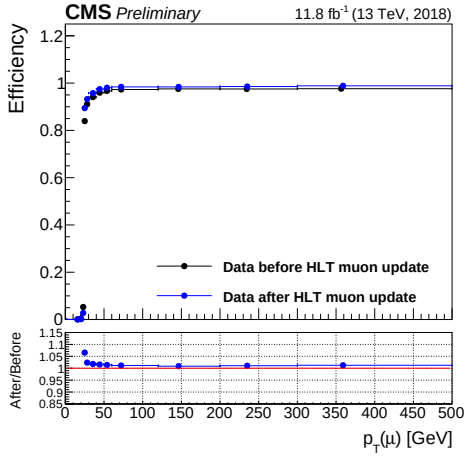

Figure 14. The efficiency of high-level trigger requiring isolated single muon with p_T >24 GeV as a function of muon p_T [9].

5. Trigger Menus for Data Selection

The trigger menu represents a large set of selection criteria that is imposed during 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 very specific triggers used only in particular studies and searches, as well as backup triggers. There are separate trigger menus used at the L1 and at the HLT, with about 400 and 600 items, respectively. Figure 15 presents fractions of the total L1 rate (100 kHz) allocation for single- and multi-object triggers and cross triggers in a typical CMS physics menu during Run 2 [3]. The dominant fraction of the total L1 rate is taken by the single and multi objects, with less rate given to the cross triggers.

Figure 16 shows the HLT rate per physics group during the data taking 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, while the 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 consists in a fraction of events passing the Level-1 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 $2.0 \times 10^{34} cm^{-2} s^{-1}$.

6. GPU-based acceleration at High Level Trigger

At CMS experiment, a heterogeneous architecture will be used in the online reconstruction in the Run 3, comprising CPUs and GPUs [11]. The pixel and pixel-based tracking, as well as the ECAL and HCAL local reconstruction have already been ported to GPUs and more reconstruction code is foreseen to be ported to GPUs in the near future. The pie-charts in Figures 17 and 18 represent the distribution of CPU and GPU time, respectively, spent in different instances of CMSSW modules (outermost ring), their corresponding C++ class (one level inner) and grouped by physics object or detector (innermost ring), while the empty slice indicates the time spent outside of the individual algorithms. The timing is measured on pileup 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 case of heterogeneous architecture, additionally equipped with

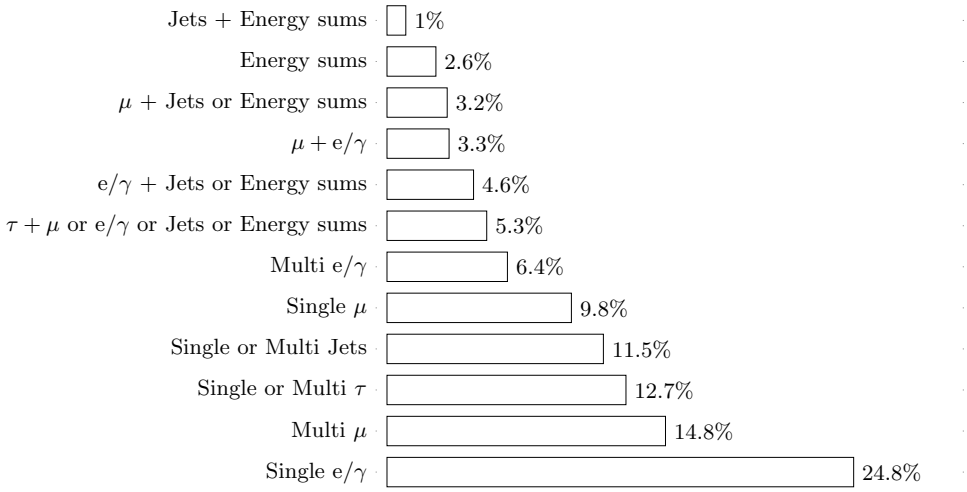

Figure 15. Fractions of the 100 kHz rate allocation for single- and multi-object triggers and cross triggers in a typical CMS physics menu during Run 2 [3].

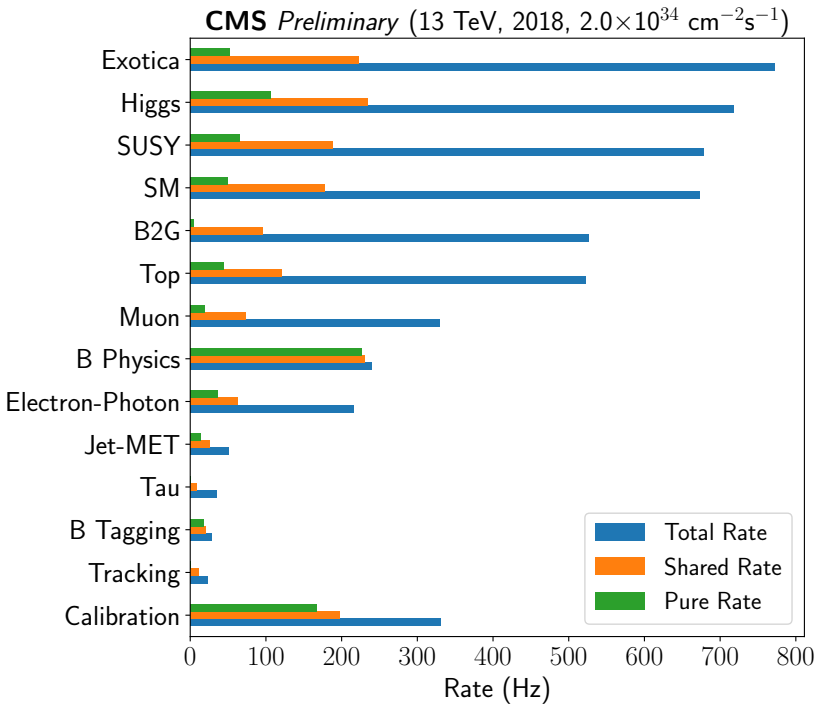

Figure 16. The HLT rates consumed by each CMS physics group during 2018 data taking [10].

NVIDIA T4 GPU. Around 25% of the CPU time is offloaded to GPU at present.

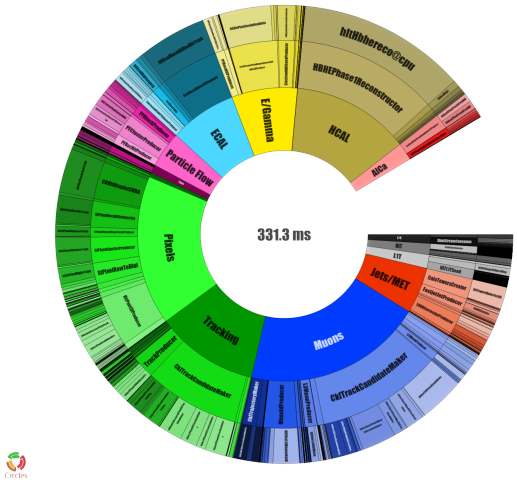

Figure 17. The pie-chart shows the distribution of CPU time in different instances of CMSSW modules [11].

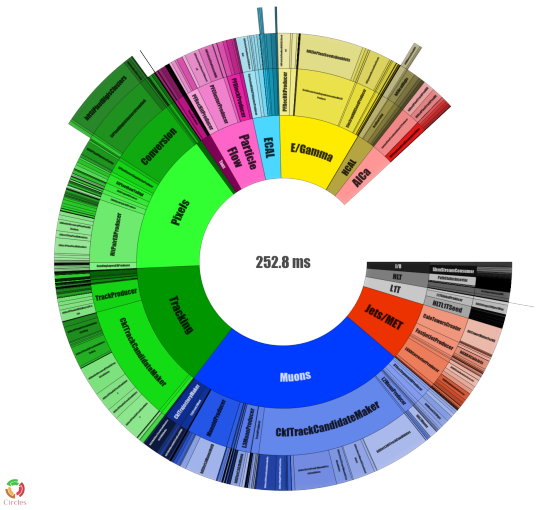

Figure 18. The pie-chart shows the distribution of CPU and GPU time in different instances of CMSSW modules [11].

7. Data Scouting and Data Parking

Alternative strategies exist with respect to the standard online data taking, as described above, in order to surpass the limited storage and computer processing resources. One option is referred as Data Scouting, where only a small summary of the reconstructed event quantities is saved and not all the raw data, thus reducing the event size, in order to be able to record the events with higher rate [12]. An example of the usage of Data Scouting is presented in Figures 19 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 shutdown period when the experiment is not taking data [13].

8. Summary and Outlook for Run 3

The CMS Trigger System has manifested its robustness and flexibility and has proven during the Run 1 and Run 2 to be able to deal with a large number of events to fulfil the CMS physics goals. Excellent performance was obtained in Run 2, from sharp efficiency curves to only a moderate pileup dependence. During the course of Run 2 and in the Long ShutDown 2 (LS2), many new technologies were integrated in the CMS Trigger System, and many trigger algorithms were improved and innovated [14]. The Phase-I L1 trigger upgrade had occurred in 2016, introducing finer calorimeter granularity, resulting in improved energy and position resolution, while remaining within the rate constraints. The pileup subtraction algorithms were implemented in the L1 trigger during the Run 2, at a large benefit of the L1 performance [3]. The High Level Trigger reconstruction was improved following the Phase-I Pixel upgrade [15] from 2016 to 2017. HCAL had the Phase-I Upgrade in endcap and barrel, during 2018 and 2019, respectively [16].

A particular improvement is introduced for the Run 3, referred to as the heterogeneous

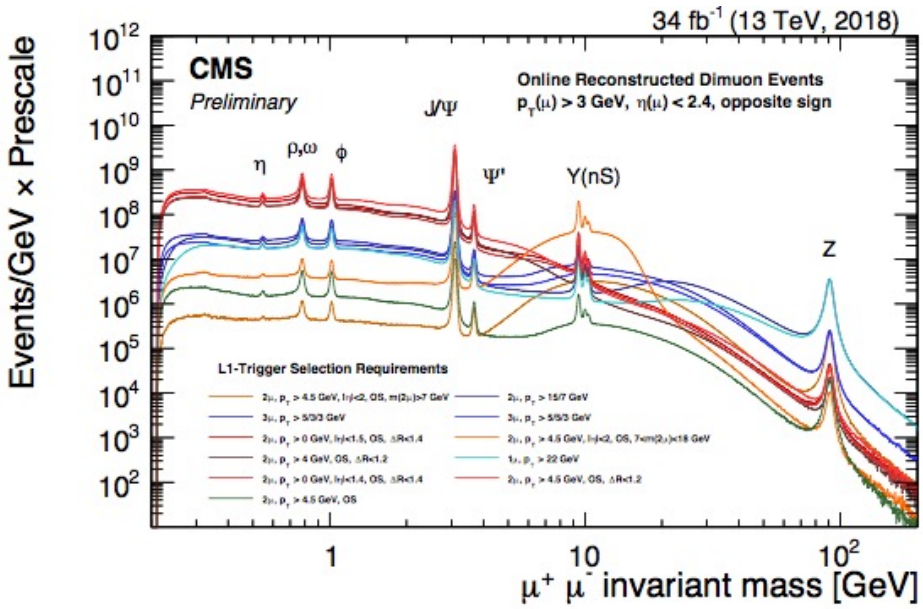

Figure 19. Dimuon invariant mass spectra reconstructed in the High Level Trigger system of the CMS detector for various muon Level-1 trigger requirements deployed by the CMS collaboration in 2017. For a subset of Level-1 requirements CMS recorded only a fraction of the data [12].

reconstruction comprising CPU and GPU. In the Run 3, it is planned to expand the reach to high rates and more exotic phase spaces, like developing special triggers for long-lived particles [17].

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.
[17] The CMS Collaboration, Phys. Rev. Lett. 127 (2021) 261804.