CMS HIGH LEVEL TRIGGER PERFORMANCE IN RUN 2

H. Sert[1]
on behalf of the CMS Collaboration

[1] *RWTH Aachen University, Experimental Physics Institute 3B, Aachen, Germany*

E-mail: hale.sert@cern.ch

The CMS experiment selects events with a two-level trigger system, the Level-1 (L1) trigger and the High Level trigger (HLT). The HLT is a farm of approximately 30K CPU cores that reduces the rate from 100 kHz to about 1 kHz. The HLT has access to the full detector readout and runs a streamlined version of the offline event reconstruction. In LHC Run 2 the peak instantaneous luminosity reached values around 2×10^{34} $cm^{-2}s^{-1}$, posing a challenge to the online event selection. An overview of the object reconstruction and trigger selections used in the 2016-2018 data-taking period will be presented. The performance of the main trigger paths and the lessons learned will be summarised, also in view of the coming LHC Run 3.

Keywords: LHC, HL-LHC, CMS, HLT

Hale Sert

1. Introduction

The Compact Muon Solenoid (CMS) experiment [1] is one of the multipurpose detectors of the CERN Large Hadron Collider (LHC). The LHC is a proton-proton collider with a design centre-of-mass energy of 14 TeV, instantaneous luminosity of 10^{34} $cm^{-2}s^{-1}$ and a bunch spacing of 25 ns resulting in 40 MHz interaction rate. However, it is not possible to store all events due to computing resources. The interesting events for the offline storage are chosen by a two-level trigger system in the CMS [2]. Level 1 (L1) triggers are hardware triggers taking the decision within a few microseconds by using the information from calorimeters and muon detector. It reduces the rate down to 100 kHz. High Level Triggers (HLT) in the CMS are software triggers running in a computing farm with approximately 30000 CPU cores. It exploits full detector information and reduces the core physics data rate down to approximately 1 kHz.

The timeline of the LHC (Figure 1) is scheduled to increase the centre-of-mass energy and luminosity in steps. This document focuses on Run 2 2016-2018 data-taking period, where the CMS recorded 146 fb^{-1} data at the centre-of-mass energy of 13 TeV. The instantaneous luminosity during Run 2 reached the peak value of more than 2×10^{34} cm^{-2} s^{-1}, while the number of overlapping proton-proton interactions (pileup) reached up to around 60 overlapping interactions. In Run 3, it is expected to run at 2×10^{34} cm^{-2} s^{-1} for a considerable portion of an LHC fill due to luminosity levelling which would result in collecting twice the data of Run 2. After Run 3, the LHC will be upgraded to high luminosity LHC (HL-LHC), Phase 2, which will result in almost 4 times more pileup. Moreover, the L1 output rate will increase to 750 kHz. These major increases will be challenging for the HLT at Phase 2.

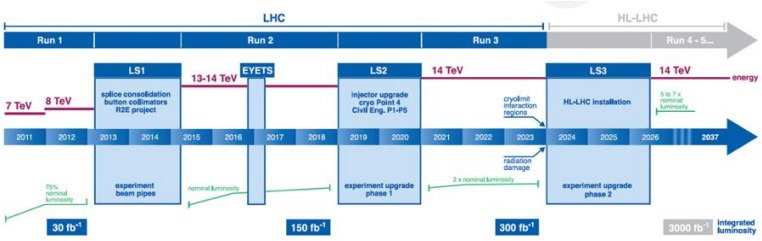

Figure 1. Timeline of the LHC baseline program and its upgrade phases, showing the energy of the collisions together with collected/expected to be collected integrated luminosities [3]

2. The High Level Trigger in CMS

The HLT triggers in CMS [2] are designed as a menu made of over 600 different paths targeting a broad range of physics signatures and purposes. Each HLT path consists of a sequence of reconstruction and filtering modules arranged in increasing complexity. The faster algorithms run first and the time consuming algorithms, which are mostly the ones for a reconstruction similar to offline (e.g. Particle Flow) as presented in Figure 2, are run at the end of the path. If a filter fails during reconstruction, the remaining part of the path is skipped in order to keep the CPU time under control.

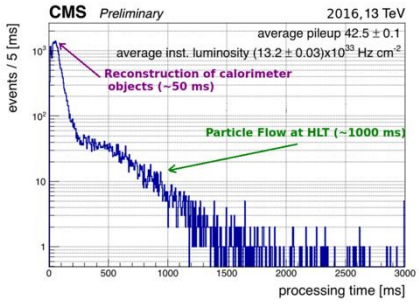

Figure 2. HLT processing time for an event with an average number of interactions per crossing of 42.5 for an average instantaneous luminosity of 13.2×10^{33} Hz cm^{-2} in 2016 [18]

Processing time of an HLT path is limited due to computing resources, therefore there are some simplifications applied in the online reconstruction. One of them is the intermediate selection steps before running CPU time consuming reconstruction parts. These steps use the information taken only from calorimeters, pixel tracks or muon detectors to filter events. In addition, tracking follows a simplified version of the offline tracking. Furthermore, reconstruction of many trigger objects is performed regionally in a specific region of the detector instead of global volume.

The main limitations on the value of the HLT menu rate are coming from the ability to promptly reconstruct the data at Tier-0 (T0) and from the limited disc space needed to store the reconstructed data. On the other hand, one needs to store events with high rate in some physics cases. There are two ways used in the CMS to increase the HLT rate: data scouting and data parking that are detailed in the following section.

2.1 Data Scouting and Parking

Data scouting [4] stores objects with online reconstruction at HLT only resulting in reduced event size by 100 or 1000 times. The raw data is not stored in scouting and the offline reconstruction is not applied. After the objects are reconstructed at HLT, looser selection compared to the one applied for normal triggers is applied and then the events are stored for offline analysis. The looser selection results in storing more events with higher rates.

Scouting has been used in the CMS since 2011. The first application was dijet resonance search performed by using data taken in 2011 [5]. Dimuon scouting triggers were introduced in Run 2 that covers the dimuon masses below 10 GeV which is not probable with the normal triggers that cover the dimuon masses between 10 GeV and 4.5 TeV. With the dimuon scouting triggers, CMS records events with two muons reconstructed in the CMS HLT system inclusively. Events are required to have at least two muons with $p_T > 3$ GeV and $|\eta| < 2.4$. They are required to pass a set of CMS muon L1 triggers. The invariant mass distribution of the dimuon system reconstructed using scouting triggers is presented in Figure 3 (left). The figure shows how good the resonances below 10 GeV can be reconstructed. The dimuon scouting triggers have already been used successfully to search for a narrow resonance decaying to a pair of muons [6].

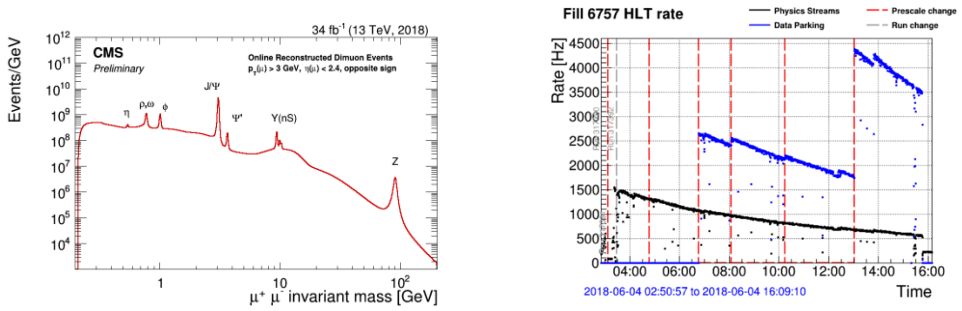

Figure 3. *Left:* Dimuon invariant mass spectrum using scouting dimuon triggers [18]. *Right:* HLT rates for an LHC fill in 2018 representing both physics streams with promptly reconstructed data and data parking with non-prompt offline reconstruction

Since the full event information is not stored with the scouting triggers, it is challenging to perform a detailed analysis in case of a potential signal is observed. For this purpose, the RAW data including full event information is parked to be reconstructed offline when it is needed after the data-taking period. Since there is no prompt offline reconstruction performed in parking, it allows to store more events. Data parking [7] is not only used for scouting triggers, but also for investigating the B physics anomalies that requires large number of $B\bar{B}$ events. One of the B meson is tagged by using a displaced muon trigger, while the other unbiased B mesons are collected to search for B anomalies. Using parking, one can achieve 3 - 5 times higher HLT rates as shown in the right plot of Figure 3.

2.2 HLT Object Reconstruction and Its Performance

Tracking at HLT [8] carries an important role in the reconstruction of many trigger objects, where a simplified version of offline tracking [9, 10] is used with reduced number of iterations and regional tracking in some of the iterations. The tracking algorithm starts the seeding with quadruple pixel hits after the phase 1 upgrade of the pixel detector, where the number of layers were increased by one layer in both barrel and endcap [11] and performed the reconstruction in three iterations. In mid-2017, an additional recovery sequence was introduced to overcome the reduced tracking efficiency due to inoperative pixel modules. Figure 4 shows the performance of HLT tracking with the doublet recovery sequence. The recovery sequence recovers most of the efficiency loss and brings the efficiency very close to the perfect detector case, while it does not increase the fake rate in the tracking reconstruction.

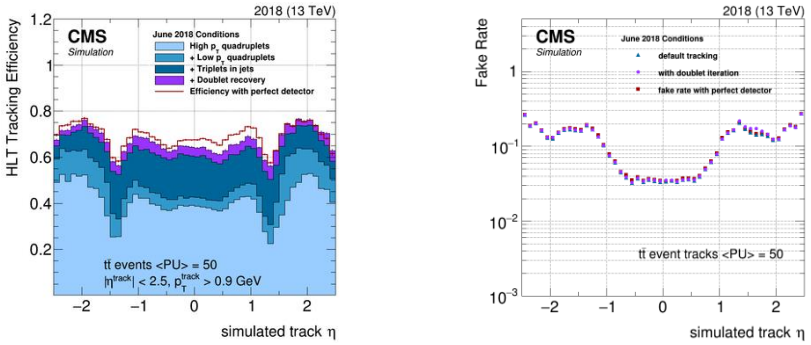

Figure 4. HLT tracking efficiencies (*left*) and fake rate (*right*) as a function of simulated track η for the original three tracking iterations and the doublet recovery compared with the perfect detector case, where there are no pixel modules considered broken [18]

Electrons at HLT are reconstructed starting from the reconstruction of the superclusters (SC), and their matching with the pixel hits, and continuing with the track reconstruction that is similar to the one used in offline [10]. The electron online reconstruction, updated after the phase 1 upgrade of the pixels, shows that the trigger efficiency of electrons is reduced with the new pixel detector, however it reduces the rate as well by 70% [12]. This significant reduction of the rate allowed to modify the working points to increase the efficiencies. The left plot in Figure 5 presents the performance of single electron trigger with $p_T = 32$ GeV for the last part of data taken in 2016 and 2017. The plot shows the gain in the endcap region from the modification of the working point, which makes the dependency on η flatter.

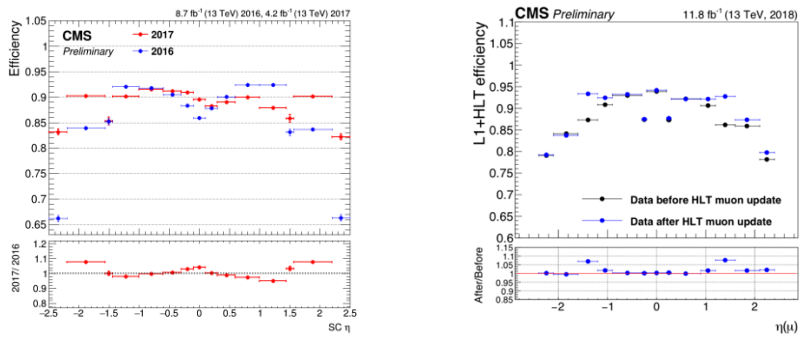

Figure 5. *Left:* Trigger efficiency of an electron with $p_T = 23$ GeV of a double electron trigger with $p_{T_1} = 23$ GeV and $p_{T_2} = 12$ GeV. *Right:* Trigger efficiency for a single isolated muon trigger with $p_T = 24$ GeV [18]

Muon online reconstruction consists of two steps: in the first step the muons are reconstructed using the information only from muon detectors, and in the second one the reconstruction is performed

by exploiting the full detector information. The second part of the reconstruction algorithm underwent an important upgrade in the beginning of 2017, where the two different algorithms [13], cascade and tracker muon, were combined into a single algorithm [14]. For further improvements, additional updates were performed in 2018, such as adding one more iteration with doublet hits and adding a simple identification to keep the high purity with lower rate. The improvement obtained with these updates is shown as a function of η of offline muons in the right plot of Figure 5.

Jet and missing transverse energy (MET) trigger objects are reconstructed by using particle flow (PF) algorithm [10], which is a time consuming reconstruction algorithm as shown in Figure 2. Therefore, jets and METs are reconstructed by using only the calorimeter information as a first step, and then the PF reconstruction is performed by exploiting the full detector information. The jetMET HLT paths provided consistent results with high performance during Run 2 [18].

For the identification of b-jets at HLT, there are two b-jet tagging algorithms used in Run 2: Combined Secondary Vertex (CSVv2) [15] used up to 2018 and DeepCSV [15] started to be used in 2018. The b-jet tagging algorithm performs the tracking as in the PF sequence. Alternatively, it can be also performed regionally around the leading calorimeter jets. The regional tracking reconstruction around calorimeter jets reduces the computing time by approximately 75% [18].

Tau leptons decaying into hadrons, τ_h, are reconstructed at HLT using the PF algorithm, globally or regionally, depending on the type of tau lepton triggers. Since the reconstruction of di-τ_h triggers are CPU consuming processes, they use regional PF reconstruction around the L1 τ_h candidate, while $e\tau_h$ and $\mu\tau_h$ triggers are reconstructed globally. In the case of di-τ_h triggers, for the same purpose two more filters by using the jets reconstructed from the calorimeter information only and by using a track based isolation are applied. The approximate processing time of a di-τ_h trigger even after mentioned special treatments is around 50 ms, while this is around 10 ms for lepton+τ_h triggers for an average pileup of $< PU > = 50$. In the final step, tau leptons were reconstructed by using the cone-based algorithm until 2018, where there is no separation between decay modes. The reconstruction was updated to hadron-plus-strip (HPS) algorithm [16] in 2018 that makes the separate decay mode reconstruction possible in online and which aligns it with the offline reconstruction. The HPS-based algorithm provides better p_T resolution as seen from Figure 6 (left). The middle and right plots of Figure 6 represent the comparison of two different tau reconstruction algorithms for $\mu\tau_h$ and di-τ_h triggers, where they provide approximately similar performances, while the HPS-based algorithm reduces the rate of tau lepton trigger by 10% per tau-leg [17].

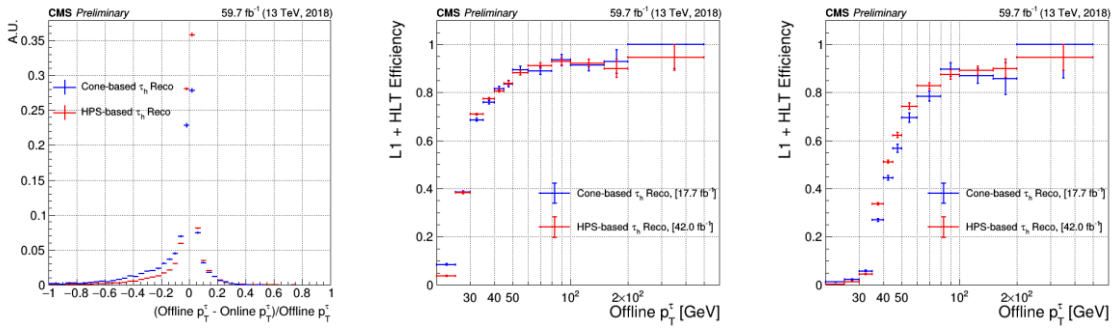

Figure 6. *Left:* p_T resolution of two different reconstruction algorithms of tau leptons used in 2018 data-taking. *Middle and Right*: Performance comparison of tau reconstruction algorithms used in 2018 data-taking [18]

2.3 Prospects for Run 3 and Phase 2 of the LHC

LHC Phase 2 will have 7.5 times more the number of events to process with 4 times more pileup. Therefore, it is extremely challenging to perform the HLT reconstruction solely on CPUs. Heterogeneous computing farm with the GPUs for the HLT reconstruction is considered to overcome this challenge [19]. However, using GPUs already in Run 3, which is under evaluation, would give valuable experience running the HLT reconstruction code in a heterogeneous environment. With the heterogeneous HLT farm reconstructions that consume more CPU time like pixel tracking, ECAL,

HCAL local reconstructions could be run in GPUs. A study with the pixel tracking using GPUs showed that GPU provides better p_T resolution, higher efficiency and lower fake rate as seen in Figure 7.

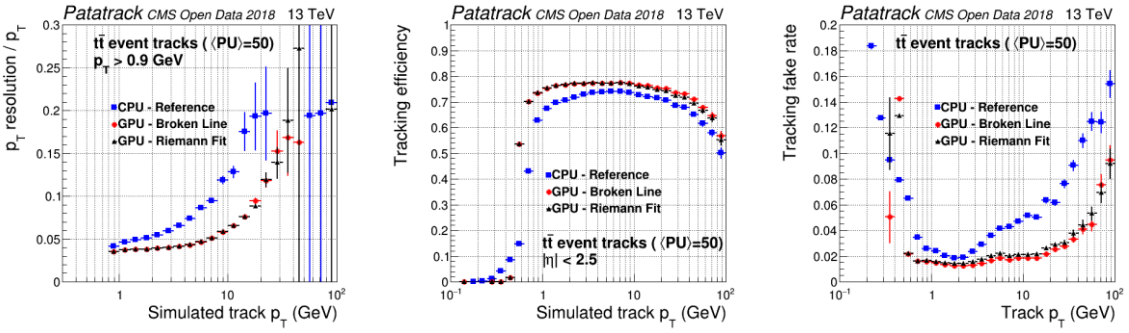

Figure 7. Physics performance comparison of the pixel-only track reconstruction for running HLT in GPUs with two different fitting algorithms and CPUs as reference. *Left*: p_T resolution, *Middle*: tracking efficiency, and *Right*: fake rate of the track reconstruction are shown as a function of track p_T. The results are obtained by using simulated $t\bar{t}$ events with average pileup of 50 [19]

Figure 8 shows the throughput for the pixel reconstruction for different architectures. The first and the second blocks correspond to the results obtained by running a single job on two different GPU accelerators, while the third one shows the results when two concurrent jobs running on a Tesla T4. The blue lines show the throughput when one uses CPUs. The first column of each block represents the throughput when copying the raw data to GPU and run the reconstruction algorithms there and leaving the results on the GPU, while the second column shows the throughput when the results are copied back to the CPU but keeping the data format as it was, and the third one is the throughput when all data converted to legacy data formats. The figure presents that copying the final products of the reconstruction from GPU to CPU cause a significant reduction of the throughput. Converting the data format reduces the throughput even more. On the other hand, running two jobs on a GPU accelerator improves the performance. The HLT reconstruction is considered to be run as much as possible in GPUs such that one needs to copy as less information as possible back to the CPU.

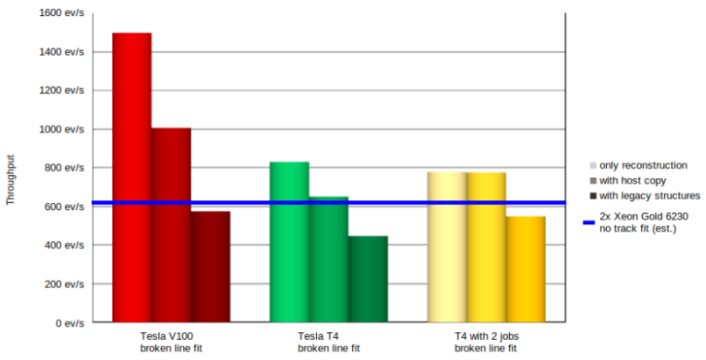

Figure 8. Throughput for the pixel reconstruction for different architectures: *Left:* NVIDIA Tesla V100, *Middle*: Tesla T4, *Right*: Tesla T4 with two concurrent jobs. The blue line shows the estimated throughput for a machine equipped with a pair of Intel Xeon Gold 6230 with 40 cores [19]

The studies showed that using the heterogeneous HLT farm with the GPUs in LHC Run 3 would improve the physics performance as well as it would bring the experience in running in a heterogeneous farm, commissioning and operating it. Usage of GPUs in LHC Run 3 does not preclude some other accelerator technology being used in the future.

3. Summary & Outlook

The HLT in the CMS is run in a computing farm with approximately 30000 CPUs in Run 2 data-taking period. The HLTs performed well and maintained high performance in Run 2. Many developments were performed to improve the reconstruction of HLT objects and also to mitigate the experienced issues during data-taking. However, phase 2 will be challenging for the HLT due to high pileup and input rate. Heterogenous computing farm will probably be the solution that is thought to be necessary to meet the high luminosity LHC needs. Deployment of a prototype already in Run 3 will provide the experience needed for phase 2 and would also improve the physics performance.

References

[1] CMS Collaboration, "The CMS experiment at the CERN LHC", JINST 3 S08004 (2008).

[2] CMS Collaboration, "The CMS trigger system", JINST 12 (2017) P01020.

[3] "Timeline of the LHC baseline program" https://cds.cern.ch/record/2199189/files/English.pdf.

[4] S. Mukherjee, "Data Scouting: A New Trigger Paradigm", CMS-CR-2017-194.

[5] CMS Collaboration, "Search for Narrow Resonances using the Dijet Mass Spectrum in pp Collisions at sqrt s of 7 TeV", CMS-PAS-EXO-11-094.

[6] CMS Collaboration, "Search for a narrow resonance decaying to a pair of muons in proton-proton collisions at 13 TeV", CMS-PAS-EXO-19-018.

[7] D. Trocino, "The CMS High Level Trigger", Journal of Physics: Conference Series 513 (2014) 012036.

[8] M. Tosi, "Tracking at High Level Trigger in CMS", Nuclear and Particle Physics Proceedings Volumes 273-275 (2016) Pages 2494-2496.

[9] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", 2014 JINST9 P10009.

[10] Sirunyan, A. M. and others, "Particle-flow reconstruction and global event description with the CMS detector", CMS-JINST-12-2017-10-P10003.

[11] CMS Collaboration, "CMS Technical Design Report for the Pixel Detector Upgrade", CMS-TDR-11, CERN-LHCC-2012-016 (2012).

[12] CMS Collaboration, "Electron trigger performance in CMS with the full 2017 data sample", CMS DP-2018/030.

[13] P. Verwilligen, "Muons in the CMS High Level Trigger System", Nuclear and Particle Physics Proceedings Volumes 273-275 (2016) Pages 2509-2511.

[14] K. Lee, "Muon Performance with CMS detector in Run2 of LHC", PoS ICHEP2018 (2019) 690.

[15] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", JINST 13 (2018) P05011.

[16] CMS Collaboration, "Performance of τ-lepton reconstruction and identification in CMS", JINST 7 (2012) P01001.

[17] CMS Collaboration, "2018 Tau Trigger Reconstruction Comparison", CMS DP-2018/035.

[18] CMS Collaboration, "High Level Trigger Run 2 Results",

https://twiki.cern.ch/twiki/bin/view/CMSPublic/HighLevelTriggerRunIIResults.

[19] CMS Collaboration, "Patatrack Heterogeneous Computing 2018 Demonstrator: Pixel Tracks", https://cds.cern.ch/record/2646774/files/DP2018_059.pdf .