Run-3 Commissioning of CMS Online HLT reconstruction using GPUs

Ganesh Parida on behalf of the CMS collaboration[1,*]

[1]University of Wisconsin-Madison

Abstract. The software-based High-Level Trigger (HLT) of CMS reduces the data readout rate from 100 kHz (obtained from Level 1 trigger) to around 5 kHz. It makes use of all detector subsystems and runs a streamlined version of CMS reconstruction. Run-2 of the LHC saw the reconstruction algorithms run on a CPU farm. However, the need to have increased computational power as we approach the high luminosity phase of LHC demands the use of Graphical Processing Units (GPUs) to reign in the cost, size and power consumption of the HLT farm. Parallelization of the reconstruction algorithms, on top of the multi-threading functionality introduced in Run-2, allowed parts of the Hadronic Calorimeter (HCAL), Electromagnetic Calorimeter (ECAL) and Pixel Tracker reconstruction to be offloaded to NVIDIA GPUs. In order to ensure the reproducibility of physics results on any machine, the HLT configuration was designed to run seamlessly with and without GPUs, that is, the algorithms were automatically offloaded to a GPU when one was available and otherwise fell back to running on the CPU. This contribution will describe the development of GPU-based algorithms for the HLT and the challenges they presented, along with the comprehensive validation and commissioning activity undertaken by CMS to ensure the successful operations of the new HLT farm.

1 Introduction

The Compact Muon Solenoid (CMS) detector [1] at the CERN LHC generates an enormous amount of data (several hundred gigabytes of information per second) with collisions happening at a rate of 30 MHz. To handle this vast data flow and select the most interesting physics events for further analysis, the experiment relies on a two-stage trigger system. The first is called the Level-1 Trigger (L1T) [2] implemented in custom hardware and FPGAs. It reduces the rate down to 100 kHz by using information from the muon detectors and the calorimeters with a reduced granularity. The second is the High-level Trigger (HLT) [3] which takes events selected by L1T as input and runs a streamlined version of CMS reconstruction software (CMSSW) using the full detector readout, with full granularity. It makes further selections on the physics objects reconstructed, thereby bringing the rate down to 5 kHz for Run-3.

In the High Luminosity phase of LHC (HL-LHC), the peak instantaneous luminosity is projected to be 2.5 times what it is for Run-3, as shown in Figure 1. Due to the increased event complexity, higher L1T rate and new detectors such as the High Granularity Calorimeter, the

*e-mail: ganesh.parida@cern.ch

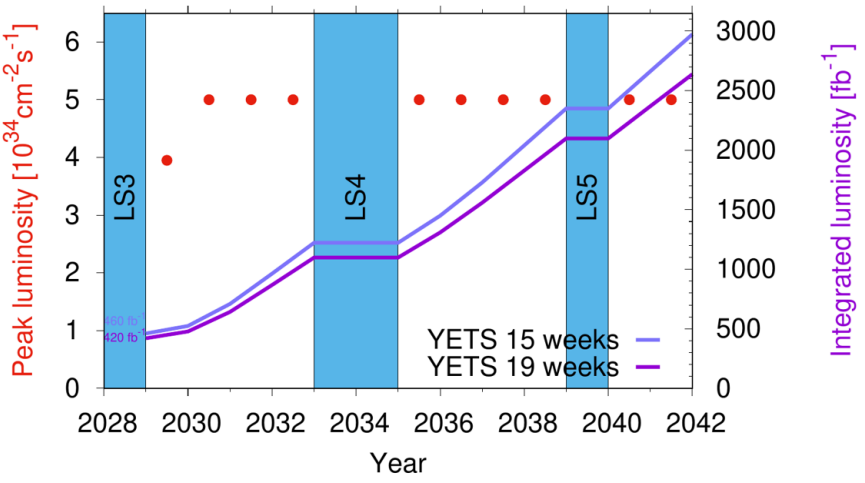


Figure 1: The projected peak luminosity during HL-LHC (marked in red dots) is two and a half times the peak luminosity during LHC-Run-3. The total integrated luminosity (in solid lines) is expected to increase by five times or more by the end of the HL-LHC data-taking period.

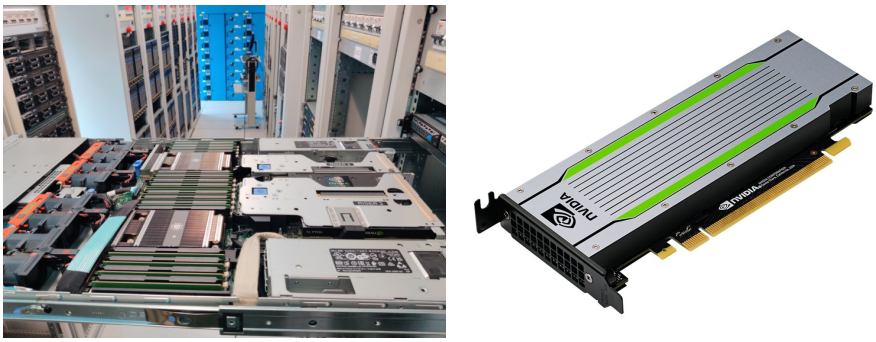
HLT will require a factor 30 increase in computing resources to achieve the same physics acceptance [4] as today. CMS plans to address this by further optimizing the online software and improving the reconstruction algorithms. Computing resources in the future are going to come more and more in the form of GPUs and other parallel processors. It is expected that they will provide higher computing performance at lower cost and power consumption. Adapting to use GPUs and exploiting them requires rethinking the reconstruction software in terms of extremely parallel algorithms.

2 High-Level Trigger Farm

The new HLT computing farm for Run-3 consists of 200 nodes. Each node (figure 2a) is equipped with two AMD EPYC "Milan" 7763 CPUs (128 cores, 256 threads), two NVIDIA T4 GPUs (figure 2b) and 256 GB of system memory. The NVIDIA GPU has 2560 CUDA "cores" running at 1.59 GHz, with 16 GB GDDR6DRAM and 6 MB L2 Cache. This setup runs on Red Hat Enterprise Linux 8. The details of the power consumption of the machines are as follows. The Thermal Design Power (TDP) of an AMD CPU is 280 W [5]. The system memory and an NVIDIA GPU have a TDP of 96 W [6] and 70 W [7] respectively. This adds up to a single node consuming 660 W without GPUs and about 800 W with GPUs. While this does not take into account other components of the machines such as motherboards, hard drives, network cards, etc, they are not expected to significantly increase power consumption.

3 Physics Improvements

The general idea of data processing at the HLT is to run multiple algorithms in increasing order of complexity that reconstruct physics objects (such as electrons, photons, muons, jets, etc) and filter events based on some thresholds motivated by physics. For 2022 data taking, CMS has offloaded serialized time-consuming processes for three sub-detector systems:



(a) (b)

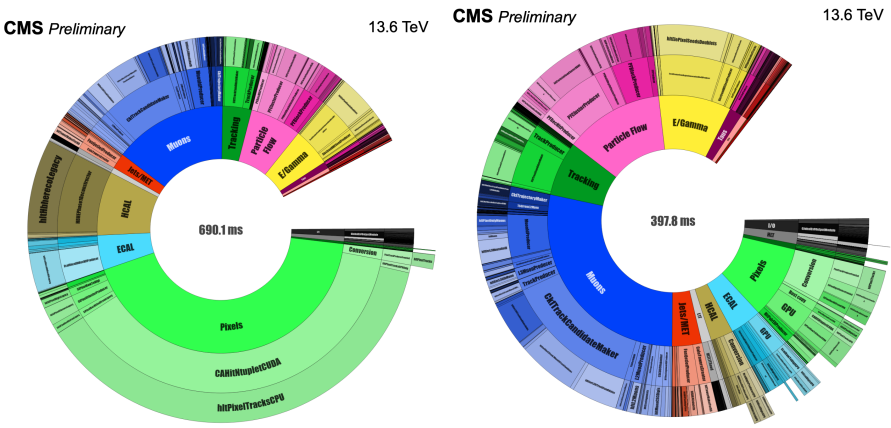
Figure 2: (a) Single HLT Node for Run-3, equipped with two AMD Milan 64-core CPUs and two NVIDIA T4 GPUs. The HLT Farm for Run-3 has 200 nodes. (b) NVIDIA T4 GPU.

Pixel [8], Electromagnetic Calorimeter (ECAL) [9] and Hadronic Calorimeter (HCAL) [10] onto GPUs (CUDA Toolkit [11] was used to target NVIDIA GPUs). For the pixel, GPUs are used to reconstruct the hits, combine the hits to form tracks (which seed the full tracking or are used standalone for data scouting which is described later) and use the tracks to reconstruct vertices. Adopting new reconstruction algorithms requires re-engineering the existing code which can in turn bring about gains in physics performance. More computing power allows CMS to invest in improved and accurate methods of reconstruction, such as tracking [12] done in a single iteration, dedicated tracking for Long-Lived Particles (LLP) and reconstruction of low transverse momentum electrons.

Data Scouting [13] is a trigger strategy pioneered by CMS to investigate physics processes that have a rate too large for available computing resources (for example, search for di-jet resonances). In scouting, higher acceptance is achieved (by lowering thresholds and increasing event rate) and the event size is reduced. No RAW data is stored and the analysis is performed with objects reconstructed at the HLT instead of offline reconstruction(prompt or otherwise). The increased computing power due to the use of GPUs has allowed for the storing of new high-level physics objects (electrons, photons and pixel-only tracks) in addition to the objects stored during the Run-2 (muons, jets and jet constituents). At the end of 2022 data taking, the data scouting ran at a rate of up to 30 kHz (30% of L1T).

4 HLT Timing and Throughput

Timing measurement was performed on a fully loaded machine running 8 concurrent jobs, each launched on 32 threads (8×32) with 24 concurrent events which is also the configuration that is used in production for 2022 data taking. The average time spent per event at the HLT reduces from 690.1 ms in CPU-only configuration to 397.8 ms with a GPU offloading-enabled configuration. The pie charts in figure 3 show the time slices of pixel, ECAL and HCAL shrinking significantly upon running on GPU. This reduction in timing translates to increased throughput at the HLT. Figure 4 shows the HLT throughput as a function of the number of CPU threads launched per job. The blue curve shows the throughput measured by CPU only. The red and the green curves show the throughput measured with GPU offloading enabled, using native GPU sharing and improved sharing via the NVIDIA MPS [14] service, respectively. Each job requires one copy of the detector conditions and algorithm parameters



(a) (b)

Figure 3: Measurement of average HLT reconstruction time spent per event with 8 concurrent jobs, each running on 32 threads and 24 concurrent events for two different configurations, (a) CPU-only configuration (b) CPU + GPU configuration without NVIDIA MPS. HLT timing improves by 40% with GPU offloading. This measurement was performed on (2x AMD 7763 CPU and 2x NVIDIA T4 GPU) with pp collisions data collected in 2022 at an average pileup 56 with the HLT menu used in 2022 data taking

in memory. A smaller number of jobs, therefore, reduced the CPU and GPU memory usage, which is the reason for sticking to the 8×32 (instead of 16×16) configuration for production. The throughput increased by a factor of 80%. Equipping a singular HLT node with GPUs does increase the power consumption from 660W to 800W (as described earlier in section 2), however, the throughput increase means that the overall HLT farm size (required for Run-3 physics goals) is smaller compared to a "CPU Only" farm. This brings about a reduction in power consumption (per throughput) by 30%.

5 Commissioning Timeline

The GPU reconstruction code was integrated into CMSSW during 2020-21. To ensure successful data taking using GPUs at HLT, the commissioning was done in stages. During 2021-22, a few machines of the 2018 HLT farm were equipped with GPUs to take data with cosmics as a proof of concept (which worked successfully). The GPU menu was then integrated into the central HLT menu in December of 2021. To validate the physics results using GPUs, jobs with few Monte Carlo samples representative of phase-1 data-taking conditions were launched with the last pre-release of the software on GPU-equipped machines on the Worldwide LHC Computing Grid(WLCG). Post-installation of the new farm, there was a commissioning run of 900 GeV pp collisions so that the farm could be tested in real data-taking conditions. The GPU-equipped HLT farm was fully commissioned in time for the start of Run-3 data taking, on the 4th of July, 2022.

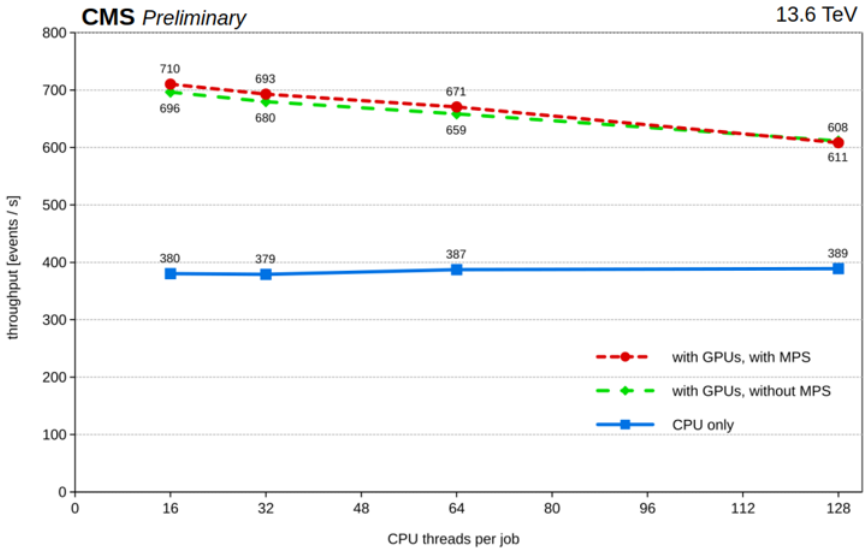


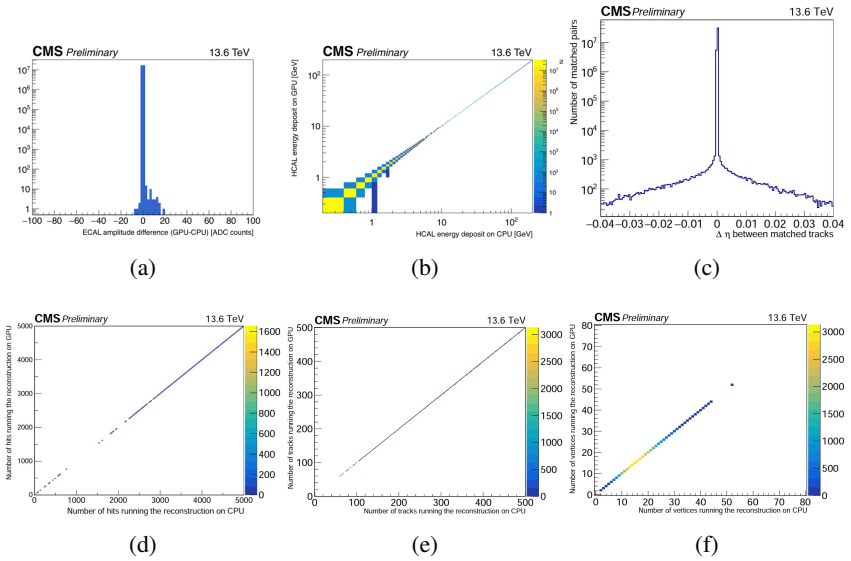
Figure 4: HLT Throughput for a whole machine as a function of the number of threads launched per job for CPU only (blue), CPU + GPU (dashed green without MPS and dashed red with MPS). The 32 threads per Job configuration is used in production(at the end of 2022 data taking), and the offloading onto the GPUs increases the throughput by a factor of 1.8. This measurement was performed on (2x AMD 7763 CPU and 2x NVIDIA T4 GPU) with pp collisions data collected in 2022 at an average pileup 56 with the HLT menu used in 2022 data taking

6 Reconstruction Validation

Consistency checks of CPU-GPU reconstruction were implemented in the Online Data Quality Monitoring (DQM) system of CMS for the three sub-detectors to ensure that the physics is unaffected by the use of GPUs. These checks were performed on a small fraction of events (one in every 3000) in real time. Overall, excellent agreement was seen between the CPU and the GPU results. Differences in the pulse amplitude of the ECAL barrel when run on CPU and GPU are shown in Figure 5a, where only a fraction of pulses of the order of 10^{-6} show a difference. Similarly, Figure 5b, shows excellent correlation in the response for the same energy deposit (in GeV) reconstructed on CPU and GPU for HCAL (barrel and endcap). The small discrepancies can be explained by the slightly different numerical approximations used by CPUs and GPUs. Finally, excellent agreement was also seen between the number of hits, tracks and vertices reconstructed on CPU and GPU in Figures 5(d-f), respectively. A closer look at the $\Delta\eta$ distributions of a CPU track and its closest matched GPU track in Figure 5c shows a sharp peak around zero and the bins above $|\Delta\eta| > 0.01$ are of the order 10^{-5} lower than the central peak. The random order of completion of GPU tasks running in parallel for pixel tracking gives rise to these small differences with respect to the CPU reconstruction.

7 Effect on Trigger Results

To study the GPU effects on trigger results, an event-by-event comparison was also performed for all the paths in the HLT menu (~700 paths) offline. Upon running the same menu once fully on CPU and once allowing for GPU offloading on the same input dataset (pp collision



(a) (b) (c)

(d) (e) (f)

Figure 5: Series of plots from the Online CPUvsGPU Data Quality Monitoring system of CMS which validates in real time the CPU and GPU reconstruction for HCAL, ECAL and Pixels on a fraction of events (1 out of 3000). (a) The difference in pulse amplitude between CPU and GPU in ECAL barrel. (b) Response of the same energy deposit (in GeV) reconstructed on CPU and GPU for both barrel and endcap. (c) The difference in η of a CPU track and the closest matched GPU track. (d) Correlation between the number of reconstructed pixel hits on a CPU and a GPU. (e) Correlation between number of tracks reconstructed on CPU and GPU. (f) Correlation between the number of vertices reconstructed on CPU and GPU.

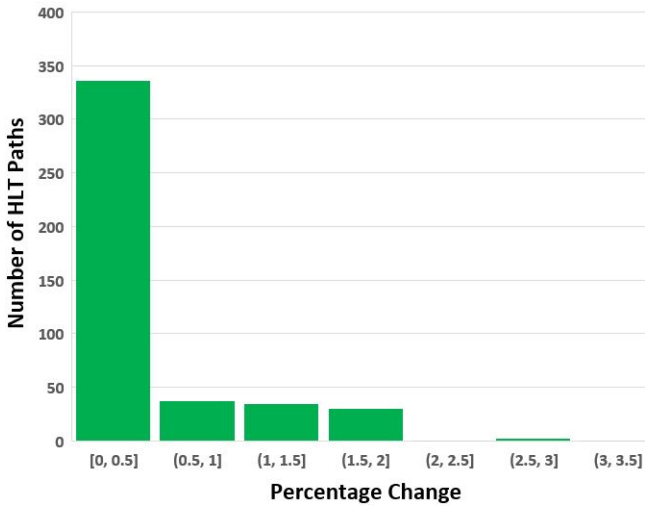


Figure 6: Event by event difference in results of HLT paths (that select at least 100 events) with and without GPU offloading. 99% of paths have differences less than 2%.

data at 13.6 TeV), it was observed that 400 out of 700 paths did not exhibit any difference. For the paths that select at least 100 events (out of the 1.28 M events processed), 99.9% of the paths had differences lower than 2% as shown in Figure 6. These differences arose from slightly varying inputs (such as the Pixel, HCAL, ECAL) to Particle Flow (PF). A trigger path (HLT_PFJet40_GPUvsCPU_v1) was introduced to monitor this behavior. This was designed to fire for events with different CPU-GPU results. During a pp collisions run at 13.6 TeV on October 13, 2022, this path recorded 5316 events at 0.18 Hz. Whereas the corresponding GPU trigger recorded 2,312,690 events. Thus the discrepancy arose for quite a small fraction ($\sim$ 0.22%) of events, further highlighting the excellent consistency between the CPU and GPU trigger results.

8 Conclusion

This contribution described the various activities that led to the successful deployment of GPUs at HLT for Run-3. This includes the installation of a new HLT farm with each machine equipped with 2 NVIDIA T4 GPUs. For the 2022 data taking, reconstruction for the HCAL, ECAL, Pixel Local Reconstruction, Pixel Only Track (used to seed the full tracking and standalone for scouting) and Vertex Reconstruction have been offloaded to GPUs. As a result, the HLT timing and throughput improved by 40% and 80%, respectively. Power consumption is also reduced by 30%. Extensive online reconstruction validation and offline comparison of CPU-GPU results showed no significant discrepancy and the residual differences are being investigated. To round up this discussion, it is important also to highlight ongoing GPU developmental efforts on multiple fronts, such as migration from traditional CMS data formats to Structure of Arrays (SOAs) for better utilization of CPUs and GPUs, rewriting other algorithms (e.g. Particle Flow) to run on GPUs and porting of Heterogeneous code to Alpaka performance portability library [15][16] to reduce code duplication and dependency on a particular architecture.

References

[1] The CMS collaboration, JINST **3** S08004 (2008)

[2] The CMS collaboration, The European Physical Journal C-Particles and Fields **34** s151–s159 (2004)

[3] The CMS collaboration, The European Physical Journal C-Particles and Fields **46** 3 (2006)

[4] The CMS collaboration, The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger, **CERN-LHCC-2021-007, CMS-TDR-022** (2021)

[5] AMD EPYC 7763, 2021, Advanced Micro Devices, Inc, https://www.amd.com/en/product/10906

[6] Micron Technology, 2019, https://uk.crucial.com/support/articles-faq-memory/how-much-power-does-memory-use

[7] NVIDIA Turing GPU Architecture, 2018, NVIDIA Corporation, https://images.nvidia.com/aem-dam/Solutions/design-visualization/technologies/turing-architecture/NVIDIA-Turing-Architecture-Whitepaper.pdf

[8] Bocci A, Innocente V, Kortelainen M, Pantaleo F and Rovere M, Frontiers in big Data **3** 601728 (2020)

[9] Reis T, Journal of Physics: Conference Series **2438** 012027 (2023)

[10] Massironi A, Khristenko V, DAlfonso M, Journal of Physics: Conference Series **1525** 012040 (2020)

[11] Kirk, David and others, ISMM, **7**, 103–104 (2007)
[12] Tosi M, Nuclear and particle physics proceedings, **273**, 2494–2496 (2016)
[13] Anderson D[The CMS collaboration], PoS ICHEP, **2016**, 190(2016)
[14] NVIDIA MPS Service, https://docs.nvidia.com/deploy/mps/index.html
[15] Zenker, Erik and Worpitz, Benjamin and Widera, René and Huebl, Axel and Juckeland, Guido and Knüpfer, Andreas and Nagel, Wolfgang E and Bussmann, Michael, IEEE, **2016**, 631–640(2016)
[16] Bocci A, Czirkos A, Di Pilato A, Pantaleo F, Hugo G, Matti K, Redjeb W Journal of Physics: Conference Series **2438** 012058 (2023)