Commissioning CMS online reconstruction with GPUs

Marc Huwiler on behalf of the CMS collaboration

E-mail: marc.huwiler@cern.ch

Abstract. Building on top of the multithreading functionality that was introduced in LHC Run-2, the CMS software framework (CMSSW) has been extended in LHC Run-3 to offload part of the physics reconstruction to NVIDIA GPUs. The first application of this new feature is the High Level Trigger (HLT): the new computing farm installed at the beginning of LHC Run-3 is composed of 200 nodes, and for the first time each one is equipped with two AMD Milan CPUs and two NVIDIA T4 GPUs. In order to guarantee that the HLT can LHC run on machines without any GPU accelerators - for example as part of the large scale Monte Carlo production running on the grid - the HLT reconstruction has been implemented both for NVIDIA GPUs and for traditional CPUs.

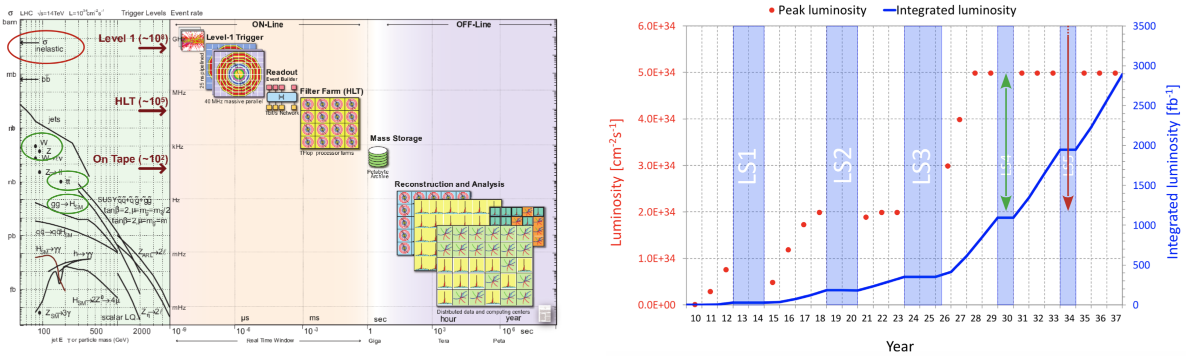
CMS has undertaken a comprehensive validation and commissioning activity to ensure the successful operations of the new HLT farm and the reproducibility of the physics results while using either of the two implementations: some have taken place offline, on dedicated Tier-2 centres equipped with NVIDIA GPUs; other activities ran online during the LHC commissioning period, after installing GPUs on few of the nodes from the LHC Run-2 HLT farm. The final steps were the optimisation of the HLT configuration, after the installation of the new HLT farm.

This contribution will describe the steps taken to validate the GPU-based reconstruction and commission the new HLT farm, leading to the successful data taking activities after the LHC Run-3 start up.

1. Introduction

The Large Hadron Collider (LHC) at CERN will undergo a major upgrade program, the High-Luminosity (HL)-LHC, aiming to increase by one order of magnitude the recorded luminosity. The beams circulating in opposite direction collide at a rate of 40 MHz, and the average number of collisions per bunch crossing (pileup) is expected to increase from the current average of 50 to about 140, or even 200 at the end of LHC Run-5. As a result, the instantaneous luminosity recorded by the CMS experiment [1] will be at least 2.5 times higher than in LHC Run-2, as shown in figure 1b.

In order to keep the same physics reach as in LHC Run-2, the subdetectors are being upgraded with faster readout electronics. To mitigate the higher pileup, the tracker, the endcap calorimeters as well as the muon system, will have increased granularity at hardware level, resulting in a larger number of readout channels. The combined effect of higher bunch crossing rate, higher pileup, and more readout channels would require a 30-fold increase in computing resources. As a result, the CMS trigger will undergo a major upgrade in order to cope efficiently with the new data taking conditions [2].



(a)

(b)

Figure 1: (a) Illustration of the data reduction performed in order to extract physics results from the collision data. (b) Projected luminosity increase for the LHC operation during the HL-LHC. The instantaneous luminosity (red) is expected to be a factor of 2.5 higher than in LHC Run-2 and LHC Run-3, for a total integrated luminosity (blue) increased by a factor of 10 by the end of data collection. [3, 2]

2. The CMS DAQ system

The physics processes studied at the LHC are several orders of magnitude lower than the inelastic scattering cross section (left part of figure 1a). Both online (middle of the plot) and offline (right part of the plot) filtering are necessary to extract the results for publication. The online filtering (real-time window) is done by the trigger, and brings the event rate down to a level that can reasonably be recorded for further analysis. The CMS experiment uses a two-level trigger system. A first stage called the level 1 (L1) trigger consists of custom electronics and FPGAs, which read out the muon system and the calorimeters. The L1 trigger reduces the rate to 100kHz (700kHz at HL-LHC) with a latency of 3.8μs (12.5μs). The second stage of the trigger system, the High Level Trigger (HLT) features a streamlined version of the reconstruction software, running on a computing farm located at the experiment cavern, near the CMS detector. It builds higher level objects, such as tracks, jets, photons and electrons, and applies looser analysis level cuts, while fitting into a time budget of O(100ms). The HLT farm consists of 200 servers (figure 2a) with 2 sockets, equipped with AMD EPYC 7763 ”Milan” 64-core processors, for a total of 128 physical cores and 256 hardware cores per machine. In addition, two low profile NVIDIA T4 GPUs (figure 2b) are equipping each machine. The NVIDIA Tesla T4 offers 2560 processing cores running at 1.59 GHz, 16GB GDDR6 DRAM and 6MB L2 cache. Their low-profile form-factor allowed fitting them into the existing server racks. The new heterogeneous HLT farm has been in service since the start of LHC Run-3, on 4 July 2022.

3. Heterogeneous reconstruction at the CMS HLT

A heterogeneous computing model, featuring GPU accelerators, has been chosen and validated for the HLT farm [2, 4]. Reconstruction of detector signals is a task that can be parallelized and that is suitable for GPUs, since a large number of independent objects are computed. Time consuming processes are offloaded to GPUs, starting with the pixel track reconstruction [4], and including the current Electromagnetic Calorimeter (ECAL) and Hadron Calorimeter (HCAL) local reconstruction, as well as the reconstruction for the future HGCAL detector [5]. The prototype was successfully integrated into the CMS software and the promising throughput increase enabled the deployment of GPU accelerators and their usage in the online event reconstruction for the LHC Run-3 already.

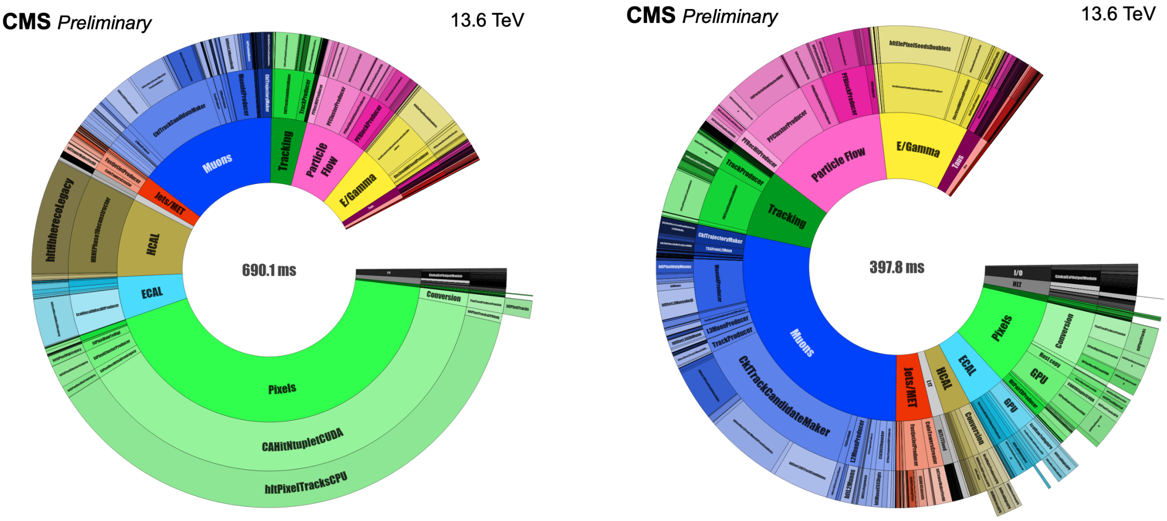
(a)

(b)

Figure 2: (a) Server of the CMS HLT farm. (b) NVIDIA Tesla T4 GPU, equipping the HLT machines (2 per server) since the start of LHC Run-3.

4. Timing and throughput

The average processing time per event at the HLT was reduced from 690.1ms using the CPU-only workflow to 397.8ms with the new heterogeneous workflow using GPUs (figure 3), corresponding to a speedup factor around 1.7. As shown in figure 3b, a significant fraction of time is spent in conversion between the new heterogeneous data format (SoA), and the legacy data format still used by downstream algorithms. Adapting all algorithms to run on the SoA format therefore leaves room for further improvement in timing. Efforts in this direction are already ongoing for the vertex reconstruction [6], and with the development of a universal SoA dataformat [7].



(a) Timing for CPU reconstruction

(b) Timing for GPU reconstruction

Figure 3: Average processing time per event (a) running on CPUs only (b) running with GPUs. The measurement was performed with 8 concurrent jobs, each running 32 CPU threads and 24 concurrent events. In (b) a single GPU, without NVIDIA MPS, was in use. [8]

The speedup per event translates into an increased throughput of the HLT farm, as shown in figure 4, where different configurations of the HLT workflow are compared. Always 256 threads are in use, split into 16, 32, 64 or 128 threads per job (resulting in respectively 16, 8, 4 and 2 concurrently running jobs). In blue, jobs are running on CPU only; in green, part of the computations are offloaded to the GPU, and in red, part of the computations are offloaded to the GPU using the NVIDIA Multi-Process Service (MPS) to improve GPU sharing.

The configuration of the HLT currently in use during datataking has 8 concurrent jobs, each

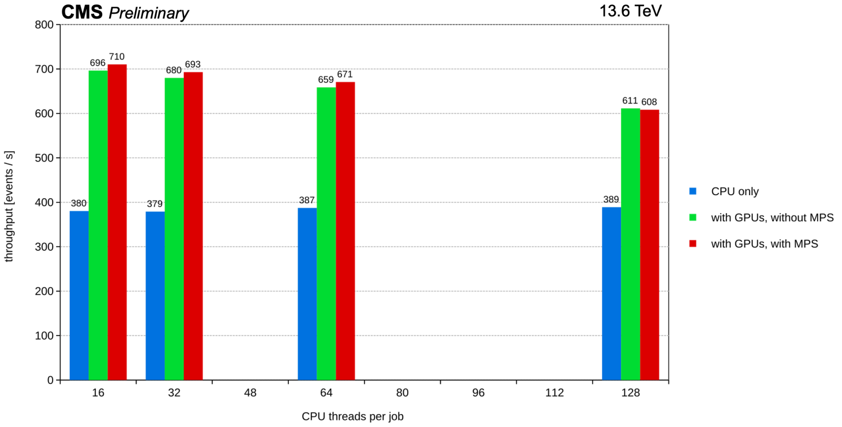
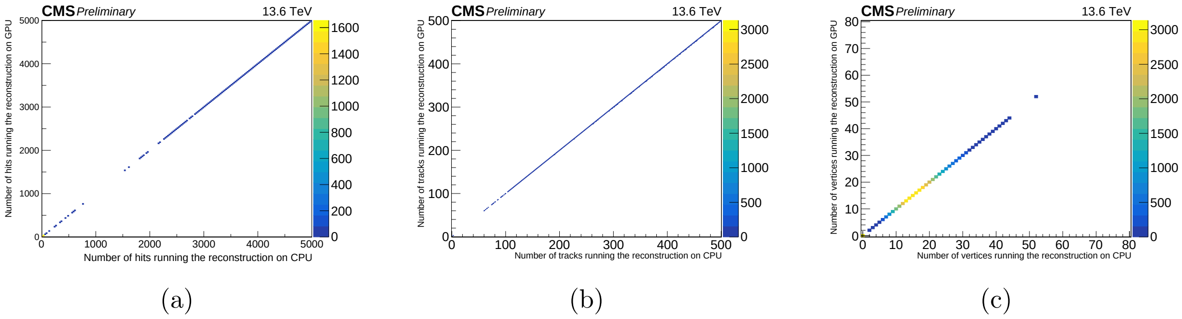


Figure 4: Throughput in events/s under different conditions. The total number of threads is always 256, split in 16, 32, 64 and 128 threads per job. The throughput for CPU only (blue), GPU without NVIDIA MPS (green) and GPU with NVIDIA MPS (red) workflows is shown. [8]

running with 32 CPU threads and 24 concurrent events. MPS has recently been enabled in the HLT workflow, but was not used at the time of the conference. The results are obtained from *pp* collisions at 13.6 TeV recorded in October 2022, with average pileup 55.

5. GPU vs CPU reconstruction validation

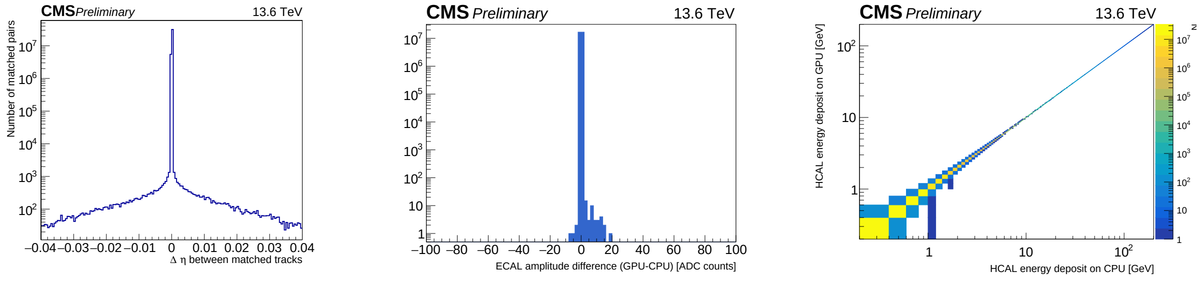
The new GPU reconstruction was validated in a series of studies, in order to make sure it does not introduce any regression. A few machines of the old HLT farm were equipped with GPUs to take data with cosmics in 2021/2022 and during the 900GeV run in May/June 2022. A pilot submission, to validate the latest pre-release of the reconstruction software with simulated benchmark datasets, was launched on GPU machines on the LHC Computing Grid. Event by event comparisons were implemented in the online Data Quality Monitoring (DQM) software. A set of plots with events recorded from the proton-proton collision run at 13.6 TeV on 2nd of October 2022 is shown in figures 5, and 6.



(a) (b) (c)

Figure 5: Tracking validation, (a) comparison of the number of rechits in the pixel detector per event, (b) comparison of the number of tracks per event, reconstructed within the pixel detector only, (c) comparison of the number of vertices per event, reconstructed from tracks in the pixel detector. [8]

In figure 5, correlations between the number of reconstructed hits (rechits) (5a), the number of tracks (5b), and the number of vertices (5c) reconstructed in the same event with the GPU and CPU workflows are shown. Overall, excellent agreement is observed, with a mismatch in the number of tracks of about 0.1%. The same fraction of disagreement is observed in figure 6a, where the difference in pseudorapitity between tracks reconstructed on GPU and on CPU is shown.



(a)

(b)

(c)

Figure 6: (a) Difference in η of a track reconstructed on CPU with a track reconstructed on GPU, matched within a geometrical acceptance of $\Delta R < 0.2$, (b) difference of amplitude in ADC counts of the same pulse in the ECAL barrel, when the fit is run on GPU and on CPU, (c) energy response of the same energy deposit in the HCAL (barrel + endcap), reconstructed on GPU and CPU. [8]

In figure 6b the difference of pulse amplitude in the ECAL barrel is shown, when the fit is run over the same pulse on GPU and on CPU. The difference is obtained by subtracting the CPU fit result from the GPU fit result. A fraction of the order of 10^{-6} of the pulses fit in this sample show a difference, and is often associated to cases where the fit takes longer to converge.

A correlation between the energy of the same HCAL energy deposit reconstructed on GPU and on CPU is shown in figure 6c. Excellent correlation is observed; the fraction of off-diagonal elements is 5 orders of magnitude lower than the diagonal.

6. Effect on trigger paths

The trigger menu was run with, and without GPU reconstruction on the same events, and the yields of each trigger path were compared. From the ∼700 trigger paths, about 400 showed no difference in yield at all. Considering paths that accepted more than 100 events (out of more than 1 million), 99% have a yield difference lower than 2%.

A dedicated trigger path was added to monitor differences in yield between CPU and GPU workflows after the full ParticleFlow reconstruction [9] is run. Events which are triggered by the CPU or GPU workflow alone are stored for further investigation. During a run of *pp* collision data at 13.6TeV on 13 October, this trigger path recorded 5316 events at 0.18Hz, while the corresponding GPU trigger recorded 2’312’690 events. The fraction of events recorded only in one of the two workflows is 0.22%.

7. Conclusion

The CMS HLT has been upgraded with two NVIDIA T4 GPUs per server, in view of the HL-LHC conditions. The successful integration of the heterogeneous reconstruction algorithms into the experiment’s software enabled them to go to production in LHC Run-3 already, and to commission the new GPU reconstruction at the HLT. The validation before and during datataking shows no significant discrepancy between GPU and CPU reconstruction, and residual differences are being investigated. The throughput has already increased by a factor 1.7, and the use of dedicated data structures throughout the reconstruction is expected to bring further speedup by avoiding unnecessary conversions. Additional algorithms are being ported to GPU reconstruction, in particular the ParticleFlow [10]. Efforts to reduce dependency on a particular architecture are also ongoing, in particular by moving to the Alpaka library [11].

References

[1] The CMS collaboration 2008 *JINST 3 S08004*

[2] The CMS collaboration 2021 *CERN-LHCC-2021-007; CMS-TDR-022*

[3] Bocci A 2016 *3rd Plenary KSETA Workshop*

[4] Bocci A, Innocente V, Kortelainen M, Pantaleo F and Rovere M 2020 *Frontiers in Big Data* **3** 601728

[5] Pantaleo F and Rovere M 2022 *CMS-CR-2022-037*

[6] Florio A D 2023 *J. Phys.: Conf. Ser. Submitted*

[7] Cano E 2023 *J. Phys.: Conf. Ser. Submitted*

[8] The CMS collaboration 2022 *CERN-CMS-DP-2023-004*

[9] The CMS collaboration 2017 *JINST 12 P10003*

[10] Pantaleo F 2023 *J. Phys.: Conf. Ser. Submitted*

[11] Bocci A 2023 *J. Phys.: Conf. Ser. Submitted*