PROCEEDINGS OF SCIENCE

CMS operations for Run II: preparation and commissioning of the offline infrastructure

Gianluca Cerminara* **for the CMS Collaboration**
CERN
E-mail: gianluca.cerminara@cern.ch

The restart of the LHC coincided with an intense activity for the CMS experiment. Both at the beginning of Run II in 2015 and the restart of operations in 2016, the collaboration was engaged in an extensive re-commissioning of the CMS data-taking operations. After the long stop, the detector was fully aligned and calibrated. Data streams were redesigned, to fit the priorities dictated by the physics program for 2015 and 2016. A new reconstruction software (both online and offline) was commissioned with early collisions and further developed during the year. A massive campaign of Monte Carlo production was launched, to assist physics analyses. This presentation reviews the main events of this commissioning journey and describes the status of CMS physics performances for 2016.

38th International Conference on High Energy Physics
3-10 August 2016
Chicago, USA

*Speaker.

PoS(ICHEP2016)169

1. Introduction

The design of the CMS detector [1] is based on a large super-conducting solenoid providing an intense (3.8 T) magnetic field, a high-precision silicon tracking system composed of about 76 millions channels including pixel and strips and hermetic calorimetry including an homogeneous Electromagnetic Calorimeter (ECAL) consisting in about 76000 $PbWO_4$ scintillating crystals. The return yoke of the solenoid houses a muon spectrometer used both for trigger and for tracking purposes. The high level of complexity and the large number of detector channels reflect in an elaborated structure for the offline workflows meant to prepare the data for physics analysis. This document will cover a selection of aspects of this preparation work focusing, in particular, on the changes that have been implemented in preparation of the LHC Run II. The topic is inherently broad the main goal is to provide the reader with an overview of the main areas of development and a set of references for more in-depth material.

2. Optimization of the reconstruction algorithms

In preparation of the LHC Run II, all the offline workflows of the experiments had to be optimized to withstand harsher running conditions. The accelerator raised the center of mass energy of the collisions from 8 to 13 TeV, increased the instantaneous luminosity and shortened the time separation of the proton bunches from 50 to 25 ns. This translated in higher pile-up of collision events, both because of the multiple interactions in the same crossing and because of the contributions from adjacent bunches (a.k.a out-of-time pile-up). At the same time, not to compromise on the physics reach, the experiment decided to increase the rate of events written to disk, imposing stronger constraints on the computing resources to effectively process and analyze them offline.

To cope with these new challenges, the algorithms for the reconstruction and identification of all physics objects have been subject of an intense development and consolidation phase mainly aiming at mitigating the effects of the pile-up both on the physics and computational performance. The key areas for this development work have been the reconstruction in the calorimeters and the tracking algorithms. In fact, the hadronic (HCAL) and electromagnetic (ECAL) calorimeters, due to the long signal integration time, are inherently more sensitive to overlapping signals from different bunch crossings, while the tracking has to cope with the higher occupancy of the silicon tracker related to the increased luminosity.

In the following section we will concentrate on the development for the ECAL reconstruction algorithm.

2.1 Out-of-time pile-up reduction in the ECAL energy reconstruction

The electric signal from the photodetectors of the CMS ECAL [2] is digitized by a 12 bit ADC running at 40 MHz. The data read out consists of a series of consecutive digitizations and a set of 10 consecutive samplings is used to reconstruct the signal amplitude.

During the LHC Run I a digital filtering algorithm was used for the estimation of the signal amplitude through a linear combination of the 10 samples S_i weighted to minimize the variance of the amplitude. In the Run II LHC running conditions, this method does not ensure the needed robustness against out-of-time contributions. Several alternative methods have been evaluated in

preparation to the 2015 data-taking and the final choice has been for a template fit with multiple components [3], called "muli-fit" in the CMS jargon. This algorithm estimates the in-time signal amplitude together with the amplitude of the 9 out-of-time contributing signals via the minimization of the χ^2 defined as:

$$\chi^2 = \sum_{i=1}^{N=10} \frac{(\sum_{j=1}^{M=10} \mathscr{A}_j p_{ij} - S_i)^2}{\sigma_{S_i}^2} \tag{2.1}$$

where $\mathscr{A}_j$ are the amplitudes of the $M = 10$ interactions. The templates for the pulse shapes $\mathbf{p}$ of each channel have been measured on low pile-up pp collision data recorded in 2013 and they are assumed to have the same shape for the in time (BX=0) and out-of-time (BX -5 to +4) bunch-corssing (j), just shifted in time by multiples of 25 ns for the latter. The electronic noise and its associated covariance matrix σ_{S_i}, are measured from dedicated pedestal runs, which measure the noise in the absence of signal pulses.

Examples of one fit for a hit in the barrel and a hit in the endcap are shown in Fig. 1.

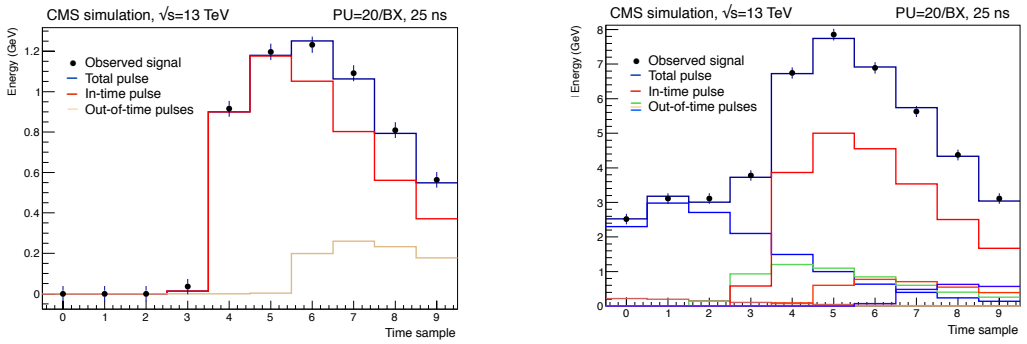

Figure 1: Examples of fitted pulses for simulated events with 20 average pile-up interactions and 25 ns bunch spacing, for a signal in the barrel (a) and in the endcaps (b). The dots represent the 10 digitized samples, the red distribution represents the in-time (BX=0) pulse while the other light colors are the fitted pulses for the out-of-time contributions with positive amplitude. The dark-blue line represents the sum of all the fitted contributions [3]

The new algorithm has been successfully employed for the reconstruction of the 2015 and 2016 data. The improvement in energy resolution with respect the Run I reconstruction algorithm for collisions with 25 ns bunch spacing is substantial especially for low p_T photons and electrons, given the larger contribution of pile-up to the total energy estimate, and is still significant for those at high p_T ($p_T > 50$ GeV). The new algorithm thus reduces the pile-up dependence in the electromagnetic components of reconstructed jets and $\not{E}_T$.

3. Automated low-latency alignment and calibration workflows

Having the best possible calibration and alignment of all sub-detectors available with short turnaround is a key asset to keep the online event selection in the High Level Trigger (HLT) fully efficient and to produce reconstructed datasets ready for physics analysis already within a few hours

after the acquisition of the data. This strategy also limits the need for re-reconstruction passes allowing for higher acquisition rate and therefore for a broader physics reach of the experiment given the budget of available computing resources. For this purpose, low-latency automatic calibration workflows [4] have been implemented to monitor and update crucial conditions which vary with very short time granularity and have been already successfully utilized during the data taking of the LHC Run I. In preparation to the Run II they went through a general consolidation effort and new calibration algorithms have been integrated. Currently the following workflows are operated in this framework:

- three-dimensional track based fit of the beam-spot position and width;
- identification of transient problematic channels in the SiStrip tracker for event-by-event optimization of the tracking reconstruction;
- determination of gains of the silicon strip tracker sensors to correct for radiation induced effects;
- track-based alignment of high-level structure of the pixel detector;
- monitoring and correction for the radiation damage of the ECAL crystals.

The strategy for running low-latency calibration workflows is based on the delay between the reconstruction of a selection of the data feeding the calibration algorithms and the reconstruction of the bulk of the data for physics analysis. This is implemented having dedicated data streams produced by the HLT and reconstructed with different latency on the Tier-0 processing farm at CERN: the average data-collection rate during a LHC fill is about 1 kHz in Run II. Only a limited bandwidth, corresponding to about 30-40 Hz, is allocated for the express processing in order to guarantee a fast reconstruction time. In the following paragraph we will describe the workflow for the track-based alignment of the pixel detector which has been operated in production for the first time during the LHC Run II.

3.1 Track-based alignment of the pixel tracker high-level structures

It has been observed that the high-level structures of the pixel detector (the two half-shells which compose the barrel and the 4 half disks composing the endcaps) move w.r.t to each other especially in relation to thermal and magnetic field changes [5]. An automated workflow has been setup to detect and, when needed, correct for these movements during the execution of the reconstruction algorithms. The plot in Fig. 2 shows, for example, the movements w.r.t to the nominal position in one of the measured coordinates over a period of about one months of data-taking. Typical movements during magnet-off periods are smaller than 50 μm in x and y, and smaller than 150 μm in z. These movements are mostly recovered after a magnet cycle.

4. Optimization of the analysis model

The analysis of the CMS data collected during the LHC Run I was based upon the Analysis Object Data (AOD) format, of average size per event of about 400 kilobytes. At the end of 2013, the total size of AOD stored was about 20 petabytes. On the top of the AOD datasets, the physics analysis groups used to generate intermediate datasets, called ntuples in the CMS jargon, that had more specialized formats intended for particular analysis needs. In preparation of the Run II, the collaboration optimized this strategy both in terms of storage footprint of the analysis data and in

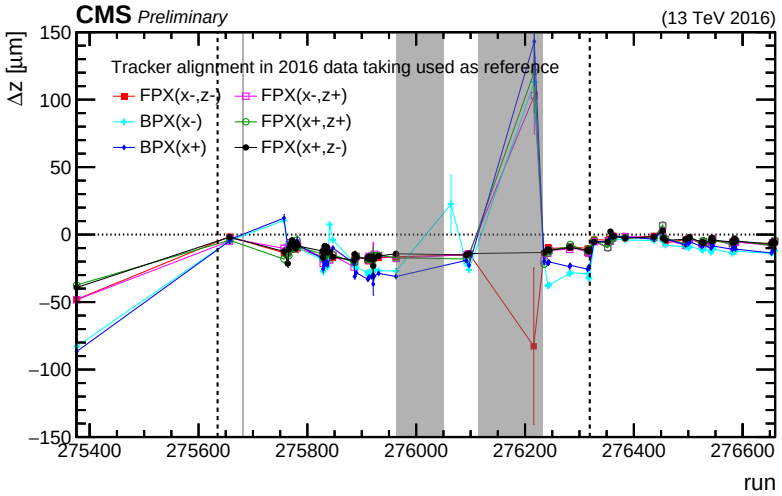

Figure 2: Observed movements of the high-level structures of the pixel detector in the longitudinal plane monitored as a function of run number. The run range covers the time from June 21 to July 12 2016, and the recorded data corresponds to an integrated luminosity of 7 fb^{-1}. Error bars represent the statistical uncertainties of the measurement. Grey bands represent runs during which CMS magnet was not at 3.8 T. Vertical dashed lines illustrate updates of the pixel high-level structure reference geometry, after which the mis-alignment is cured.

terms of common processing of the physics objects. This lead to the development of a new format, the so-called MiniAOD [6], produced centrally by the CMS computing group and providing a common foundation for CMS physics analyses. Its compressed format is one tenth the size of AOD, and it provides the solution for handling the larger data volumes of Run II.

The main principles driving the design of the new data-format have been, on one side the quest for minimal size dropping unnecessary data from existing data formats, and on the other side the need for flexibility. To achieve these goals the following information is stored in the MiniAOD format:

- high-level physics objects: leptons, photons, jets, and $\not{E}_T$;
- basic kinematic information for all particle candidates as produced by the CMS Particle Flow (PF) reconstruction algorithm [7, 8]. Their presence allows analysts to re-reconstruct physics objects with new techniques;
- information about the trigger decision to enable the computation of trigger efficiencies;
- information about final state simulated particles, generated jets and reference information;
- miscellaneous information like the interaction vertices and $\not{E}_T$ cleaning filters.

This allowed the new format to be adopted by the vast majority of physics analysis, enabling the development of novel analysis techniques and re-tuning of algorithms and calibrations while these were refined during the data-taking.

5. Data Quality Monitoring and Data Certification for analysis

Another crucial aspects of the offline workflows concerns the complex infrastructure for mon-

itoring the quality of the data [9] and aiming at certifying the datasets as usable for the physics analysis. The creation of datasets with uniform quality and not affected by any detector problem has been a key asset to exploit the physics potential of the detector already during Run I. The CMS collaboration developed a complex and powerful framework for achieving this goal, based on dedicated monitoring applications running both online, in parallel to the data-taking, and offline during the event reconstruction. This infrastructure has been further consolidated in preparation of the LHC Run II and allowed to achieve high efficiency in exploiting the acquired data for analysis. A team of detector and physics object experts checks continuously the plots automatically produced and published by this Data Quality Monitoring framework and looks for unexpected effects that could affect the analysis level physics quantities. The list of lumi-sections (quanta of data corresponding to about 23 s of data-taking and considered atomic for luminosity bookkeeping purposes) usable by all CMS physics analysis is then distributed weekly. Figure 3 shows the performance of the data certification on the datataset used for the ICHEP analysis. The overall fraction of integrated luminosity certified as usable is about 93% of what the experiment collected over the first months of LHC run. This fraction increased up to about 96% over the full 2016 dataset. The losses being manly due to isolated incidents to the detector, without recurring systematic problems.

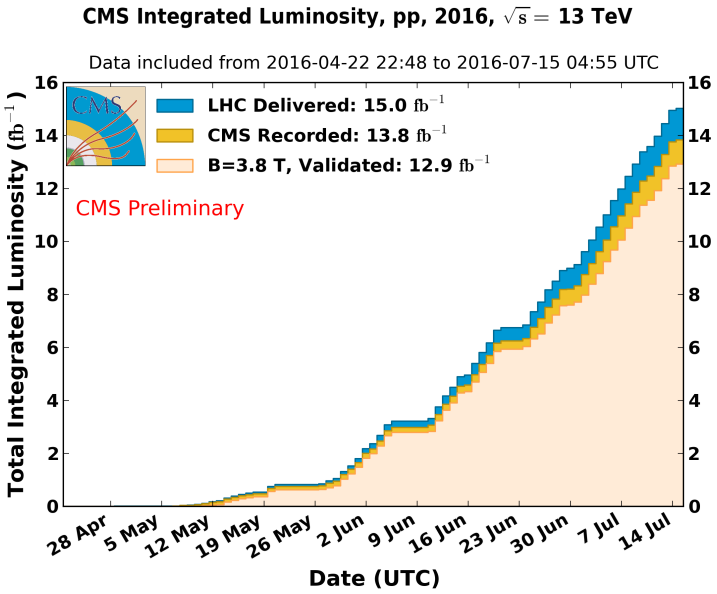

Figure 3: The plot shows the cumulative curves for the luminosity delivered by LHC (azure), recorded by CMS (orange) and certified as good for physics analysis during stable beams (light orange). The luminosity validated for physics analysis corresponds to data recorded with all detectors and reconstructed physics objects showing good performance. [10]

6. Conclusions

The performance of the CMS experiment during the ongoing LHC run rely also on a complex and powerful offline machinery which prepares the data for physics analysis with very short

turnaround and high efficiency in terms of utilization of the resources. The contribution presented a selection of these workflows focusing on their consolidation and operational experience in the 2015 and 2016 data-taking campaigns.

References

[1] R. Adolphi *et al.* [CMS Collaboration], ' *The CMS experiment at the CERN LHC*, *JINST* **3** (2008) S08004.

[2] [CMS Collaboration], *CMS: The electromagnetic calorimeter. Technical design report*, CERN-LHCC-97-33, CMS-TDR-4.

[3] E. Di Marco [CMS Collaboration], *CMS electromagnetic calorimeter calibration and timing performance during LHC Run I and future prospects*, CMS-CR-2014-410.

[4] G. Cerminara and B. van Besien, *Automated workflows for critical time-dependent calibrations at the CMS experiment*, *J. Phys. Conf. Ser.* **664** (2015) no.7, 072009.

[5] S. Chatrchyan *et al.* [CMS Collaboration], *Alignment of the CMS tracker with LHC and cosmic ray data*, *JINST* **9** (2014) P06009.

[6] G. Petrucciani and A. Rizzi and C.Vuosalo, *Mini-AOD: A New Analysis Data Format for CMS*, *J. Phys. Conf. Ser.* **664** (2015) no.7, 072052.

[7] [CMS Collaboration], *Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET*, CMS-PAS-PFT-09-001.

[8] [CMS Collaboration], *Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS detector*, CMS-PAS-PFT-10-001.

[9] F. De Guio [CMS Collaboration], *The CMS data quality monitoring software: experience and future prospects*, *J. Phys. Conf. Ser.* **513** (2014) 032024.

[10] Public CMS Data Quality Information , `https://twiki.cern.ch/twiki/bin/view/CMSPublic/DataQuality`, 16-09-2016.