Available online at www.sciencedirect.com
ScienceDirect
Nuclear Physics B Proceedings Supplement 00 (2014) 1–4
Nuclear Physics B Proceedings Supplement

Abstract

Data recorded at the CMS experiment are funnelled into streams, integrated in the HLT menu, and further organised in a hierarchical structure of primary datasets and secondary datasets/dedicated skims. Datasets are defined according to the final-state particles reconstructed by the high level trigger, the data format and the use case (physics analysis, alignment and calibration, performance studies). During the first LHC run, new workflows have been added to this canonical scheme, to exploit at best the flexibility of the CMS trigger and data acquisition systems. The concepts of data parking and data scouting have been introduced to extend the physics reach of CMS, offering the opportunity of defining physics triggers with extremely loose selections (e.g. dijet resonance trigger collecting data at a 1 kHz). In this presentation, we review the evolution of the dataset definition during the LHC run I, and we discuss the plans for the run II.

Keywords:

1. Introduction

The Compact Muon Solenoid (CMS) experiment [1] is an omni-purpose detector operating at the Large Hadron Collider [2] at CERN. The central feature of the CMS apparatus is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. Within the superconducting solenoid volume are a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass/scintillator hadron calorimeter. Muons are measured in gas-ionization detectors embedded in the steel flux return yoke outside the solenoid. In addition, the CMS detector has extensive forward calorimetry. The first level of the CMS trigger system, composed of custom hardware processors, uses information from the calorimeters and muon detectors to select the most interesting events in a fixed time interval of less than 4 μs. The High Level Trigger (HLT) computer farm further decreases the event rate from around 100 kHz to around 0.5 kHz, before data storage.

The data preparation at CMS at CERN is a complex set of inter-dependent workflows devised to assure the full physics exploitation of the CMS detector potential and of the collisions (proton-proton, ion-ion and ion-proton) delivered by the LHC. The data preparation workflows [3] supply the physics analyses with reconstructed collision events, be them from the experiment or from simulations, and are designed to use efficiently the distributed computing resources of CMS [4]. A high quality and timely production of physics results relies on the careful definition and efficient handling of primary datasets, skims and datasets for scouting, which will be described in this paper, see figure 1.

2. Primary Datasets

The stream of events acquired by CMS is organised in primary datasets, according to the results of the High Level Trigger selection [5]. For an event to be recorded by CMS, it has to be accepted by the first level hardware trigger and to satisfy at least one of the composite HLT selections which are commonly referred to as paths. The primary datasets are defined by a set HLT paths, and collect events which have passed at least one them; as such, primary datasets are non-exclusive subsets of the event stream acquired by the CMS experiment. During the run I data taking, the processing for event reconstruction of all the primary datasets started at the CERN tier-0 computing within two days of the event collection [3].

The design of the primary datasets is mainly centred around particle candidates reconstructed in the event final state by the HLT, and follows one basic principle:

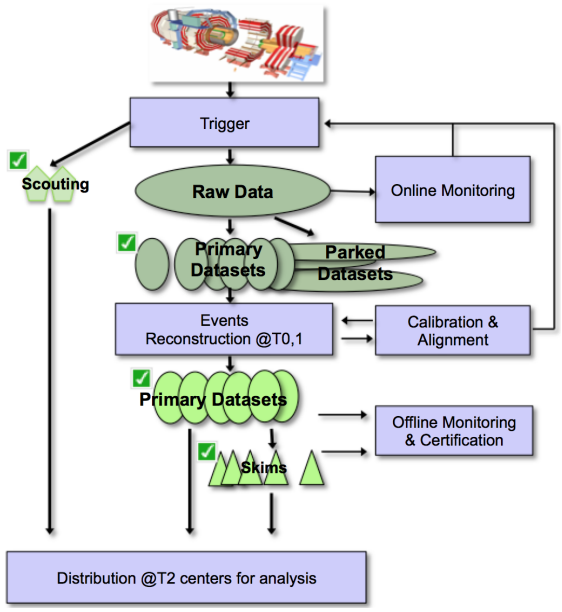

Figure 1: Workflows and Datasets devised to provide reconstructed data and simulated events for physics analysis at CMS. The green ticks indicate the datasets which will be described in this paper. Different shades of green indicate the event format: raw event (dark), reduced event (middle) and reconstructed event (bright).

grouping together events with similar physics content. Besides requirements on the physics content, the organisation of the primary dataset has to satisfy constraints related to the data processing and handling. The average event rate needs to be approximately uniform across the different primary datasets, in order to ease their distribution to the tier-2 computing centres which serve the analyses; in addition, the primary dataset event rate needs to be more than 10 Hz to avoid files of too small size, and less than 200 Hz, to assure that: (i) a dataset can be processed for event reconstruction by at least one of the tier-1 computing centres and that (ii) the number of events in a luminosity section (corresponding to 23 s of data taking) can be reliably processed by a single job.

Since a given event can pass more than one HLT path, it can be included in more than one primary dataset. We define as overlap of the primary datasets the relative increase in the rate of events when summing the primary datasets altogether, with respect to the rate of events acquired by CMS. The datasets overlap was required to be in the order of 20% or less, in order to fit within the processing and storage budget available to the tier-0 and tier-1 data centres.

The definition of the primary datasets has evolved during the LHC run I, following the unfolding of the

Table 1: Event Rate in Hz at the output of the High Level Trigger farm in the last semester of 2012 at CMS; two reference values of instantaneous luminosity are considered. Calibration triggers are not accounted for, only physics events are included. The average rate of the prompt reconstruction is also reported.

L_{inst} [$cm^{-2}s^{-1}$]	5×10^{33}	7×10^{33}
Start of the run [Hz]	460	560
Average over a run [Hz]	345	420
Prompt reconstruction [Hz]	430	525

LHC running conditions, notably the instantaneous luminosity and the number of multiple interaction per bunch crossing, and of the corresponding HLT menu. Most primary datasets have been split in 2 (4) time epochs, reflecting the major changes in conditions of 2011 (2012). A regular monitoring effort was deployed to keep the primary datasets aligned with the operations of the experiment, and to assign newly devised trigger paths to the existing datasets. Designed around particles candidates reconstructed by the HTL, the definition of the primary datasets could accommodate the evolution of the trigger menu naturally.

Table 2 presents all the active primary datasets in the second part of 2012 at the end of the run I, when the HLT deployed a menu commissioned for the instantaneous luminosity of $7 \times 10^{33} cm^{-2} s^{-1}$. The table is organised in topology sub-groups, with electron-gamma, muon, jet and missing energy, tau and b-decay candidates in the final state. The datasets in the last group are of technical nature and were implemented to monitor detector performance, physics backgrounds, and to provide an unbiased data sample for trigger studies. The measured rates have a statistical error of 15%.

In the last six months of the 2012 data taking, parked datasets have been introduced besides the core datasets already used by CMS in 2010/2011. The goal of such addition was the extending the physics program of CMS to allow precision Standard Model measurements and new physics searches which, were considered unaffordable at the beginning of run I due to the constraints posed by the processing infrastructure available for the prompt reconstruction. Table 3 provides the physics motivation for each parked dataset. 450 Hz of extra events were collected and funnelled in the parked datasets; while the prompt reconstruction continued to process the core dataset, the processing of the parked ones was deferred to the long shutdown of 2013-2014. The triggers paths used to define the parked datasets were either looser version of the core triggers (for instance relaxing candidates' transverse momentum

Table 2: Primary datasets the second part of 2012 at the end of run I. The measured rates have a statistical error of 15%.

Primary Dataset	rate [Hz]
SingleElectron	68
DoubleElectron	15
ElectronHad	14
SinglePhoton	16
SinglePhotonParked	69
DoublePhoton	33
DoublePhotonHighPt	9
PhotonHad	11
MuEG	19
SingleMu	63
MuHad	14
DoubleMu	23
DoubleMuParked	26
MuOnia	38
MuOniaParked	94
Multijet	23
Multijet1Parked	174
JetHT	17
HTMHT	16
JetMon	6
VBF1Parked	201
MET	16
METParked	43
Tau	21
TauParked	40
TauPLusX	36
BTag	3
BTagPlusX	3
Commissioning	15
HCALNZS	< 1
NoBPTX	7
MinimumBias	10

or isolation requirements), or completely new triggers with little overlap with those preexisting. Most parked datasets (eight out of ten) fully included one of the core datasets.

Figure 2 shows a matrix of event rates; one bin represents the rate in Hz of events common to a given pair of primary datasets, the overlap of the two. The bins in the rightmost column and along the main diagonal represent the single-dataset rates, while the bins in the top-most row are the overlap of a given dataset with all the others. The approximate block structure visible in the matrix stems from the design choice of funnelling in a primary datasets events from trigger paths with similar physics content, and from having ordered the datasets in the plot based on particle candidates defining the event final state. At the end of run I the average rate of events for the core dataset was 420 Hz, with 25% overall overlap. Considering also the parked datasets, the rate rose to 890 Hz, and the overlap to 35%.

Table 3: Primary parked datasets and the physics motivation for their introduction.

Primary Dataset	Motivation
MuOniaParked	Qiarkonium physics
DoubleMuParked	PDF @low m($\mu\mu$)
TauParked	3-prong τ decays, $h \to \tau\tau$
VBF1Parked	Vector Boson Fusion
SinglePhotonParked	Mono-γ, dark matter
METParked	Susy hadronic
HTMHTParked	Susy hadronic
MultiJet1Parked	Susy hadronic
HLTPhysicsParked	Unbiased HLT study
ZeroBiasParked	Unbiased HLT study

3. Skims

Skims were introduced for specific analysis or calibration tasks, with the goal of easing the data access and and facilitating the event processing; produced starting from primary dataset after their processing, the the definition of skims could rely not only on the HLT trigger results, but also on quantities made available by the offline full event reconstruction. The peculiarity of a given skim lied in: a reduced rate with respect its parent primary dataset, a customised event content, or both. By design the size on disk of a skim was required to be less than one third of the parent dataset. The skims size was typically a few per-cent of the parent primary dataset, which made multiple replicas at different computing centres possible and provided easier and faster processing workflows for specific analysis and calibration tasks. By the later half of 2012, 21 skim defini-

tions were devised and applied to 26 primary datasets, for a total of 80 skims produced: 28 for physics analysis or calibration purposes, and 52 for technical tasks such monitoring of hardware or processing errors.

4. Datasets For Scouting

Data scouting has been introduced by CMS in 2011 to recover sensitivity for physics measurements and searches in a phase space not accessible via the standard trigger selections. Novel trigger and data acquisition strategies were devised to afford 1kHz of events in addition to the core and parked datasets, and access events below the thresholds of the standard trigger paths. The seeding HLT selection was set to a very loose $HT > 250$ GeV, where HT is the scalar sum of the transverse momentum of reconstructed jets above 20 GeV. The event size was reduced, from approximately the 1 MB of the standard raw CMS readout, to 10 kB, by saving for particle candidates (jets, missing transverse energy, muons > 5 GeV, electrons > 8 GeV) only the quadri-momenta reconstructed by the HLT. The impact on the processing budget of the HLT was kept to a minimum in 2011, when scouting needed the computation of some reconstructed quantities based on the particle-flow algorithm [6], and to zero on 2012, when all the quantities saved for scouting were pre-computed by the HLT.

The first implementation of data scouting in 2011 made possible a physics search for narrow resonances in inclusive dijet mass spectra, see [7], for which the sensitivity was significantly extended below 1 TeV.

5. Outlook and Conclusions

The definition of datasets is a crucial ingredient for the performance and full physics exploitation at CMS. During the LHC run I, 30 core primary datasets were defined for physics analysis, for a total rate of 420 Hz and 25% overlap; in 2012 10 parked primary datasets were added, for an extra 470 Hz leading to a 35% overlap. Event scouting was used to provide an additional 1 kHz of events with reduced event content.

The structure of the datasets for the LHC run II will be put in place once the high level trigger menu will be ready. The logic of the datasets definition will reflect the one used in run I, with the HLT output rate set to 1 kHz. No parked datasets are foreseen, at least at the startup. An improved implementation of the scouting strategy is also being prepared.

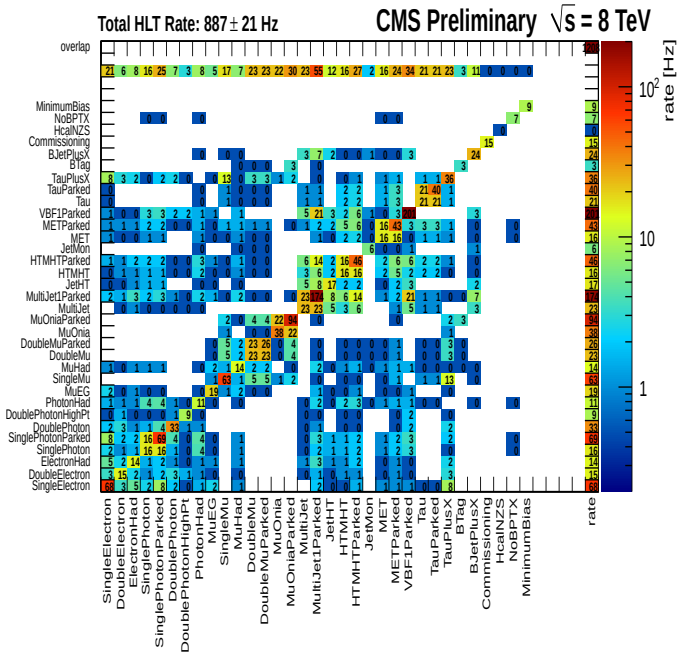

Figure 2: Event rates common to all the possible pairs of primary dataset, Hz. For a detailed description, see the text.

References

[1] CMS Collaboration, *The CMS experiment at the CERN LHC*, JINST 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[2] Lyndon Evans and Philip Bryant, *LHC Machine*, 2008 JINST 3 S08001, doi:10.1088/1748-0221/3/08/S08001.

[3] G. Franzoni, for CMS Collaboration, *Data preparation for the Compact Muon Solenoid experiment*, Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2013 IEEE, doi:10.1109/NSSMIC.2013.6829574.

[4] D. Bonacorsi, for the CMS Collaboration *Experience with the CMS Computing Model from commissioning to collisions*, 2011 J. Phys.: Conf. Ser. 331 072005, doi:10.1088/1742-6596/331/7/072005.

[5] V. Gori, for CMS Collaboration, *The CMS High Level Trigger*, PIC2013 conference, arXiv:1403.1500v1.

[6] F. Beaudette, for the CMS Collaboration *The CMS Particle Flow Algorithm*, Proceedings of the CHEF2013 Conference - Eds. J.C. Brient, R. Salerno, and Y. Sirois - p295 (2013), ISBN 978-2-7302-1624-1, arXiv:1401.8155v1.

[7] The CMS collaboration, *Search for narrow resonances and quantum black holes in inclusive and b-tagged dijet mass spectra from pp collisions at* $\sqrt{(s)} = 7$ *TeV*, JHEP 1204 (2012) 06, doi:10.1007/JHEP01(2013)013.
doi:10.1007/JHEP01(2013)013