The CMS Computing, Software and Analysis Challenge

David Futyan
Imperial College, University of London, London, United Kingdom

Rainer Mankel*
Deutsches Elektronen-Synchrotron, Hamburg, Germany

Christoph Paus
Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

for the CMS collaboration

*Corresponding author. Email: Rainer.Mankel@cern.ch

Abstract. The CMS experiment has performed a comprehensive challenge during May 2008 to test the full scope of offline data handling and analysis activities needed for data taking during the first few weeks of LHC collider operations. It constitutes the first full-scale challenge with large statistics under the conditions expected at the start-up of the LHC, including the expected initial mis-alignments and mis-calibrations for each sub-detector, and event signatures and rates typical for low instantaneous luminosity. Particular emphasis has been given to the prompt reconstruction workflows, and to the procedures for the alignment and calibration of each sub-detector. The latter were performed with restricted latency using the same computing infrastructure that will be used for real data, and the resulting calibration and alignment constants were used to re-reconstruct the data at Tier-1 centres. The paper addresses the goals and practical experience from the challenge, as well as the lessons learned in view of LHC data taking.

1. Introduction and scope
The goal of the CMS Computing Software and Analysis challenge (CSA08) has been to test the full scope of offline data handling and analysis activities which are required for the CMS data-taking. In contrast to earlier campaigns, it constituted the first full-scale challenge with large statistics under the conditions expected at LHC startup. This implied simulation of initial mis-alignments and mis-calibrations as expected before first collisions, and event signatures and rates typical for low instantaneous luminosity. A particular focus for this exercise was on the offline detector commissioning aspects, with emphasis on calibration and alignment. The full complexity of a large number of calibration and alignment workflows intended for startup was performed concurrently, which allowed in particular studying the effects of interdependencies between the individual processes, as well as the organizational challenges involved in producing such constants in quasi real-time. The obtained alignment and calibration constants were used in a full re-reconstruction of the data.
samples. In addition, realistic physics analyses were performed both on promptly reconstructed and re-reconstructed data samples.

A particular feature of the CSA08 challenge was that its time window had been immutably fixed in advance by its concurrency with the Common Computing Readiness Challenge (CCRC08). Contrary to CSA which was a CMS-specific exercise, CCRC was a multi-VO computing stress test, and its schedule had been defined several months in advance to ensure synchronization between the large experiments. The scope of the prompt alignment and calibration activities conducted during CSA08 has been to demonstrate the complete workflow in as realistic a manner as possible.

| Sample         | Scenario(s) | Number of events (requested) |
|----------------|-------------|------------------------------|
| MinBias        | S43+S156    | 25M                          |
| JetET20        | S43+S156    | 4M                           |
| JetET30        | S43+S156    | 4M                           |
| JetET50        | S43+S156    | 4M                           |
| JetET80        | S43+S156    | 4M                           |
| JetET110       | S43+S156    | 4M                           |
| JetET150       | S156        | 4M                           |
| GammaJets      | S156        | 2M                           |
| MuonPT5        | S43+S156    | 10M                          |
| MuonPT11       | S156        | 10M                          |
| Jpsi           | S156        | 750k                         |
| Upsilon        | S156        | 250k                         |
| Zmmumu         | S156        | 13k                          |
| Zee            | S156        | 13k                          |
| Wenu           | S156        | 119k                         |
| Wmmumu         | S156        | 119k                         |
| MuonCosmicBON  | S43+S156    | 10M                          |
| MuonCosmicBOFF | S43+S156    | 10M                          |
| TkCosmicBON    | S43+S156    | 10M                          |
| TkCosmicBOFF   | S43+S156    | 10M                          |
| MuonBeamHalo   | S43+S156    | 10M                          |
| TkBeamHalo     | S43+S156    | 200k                         |
| HCALNZS        | S43+S156    | 6M                           |
| HCALIsoTracks30| S43         | 5M                           |
| HCALIsoTracks50| S156        | 5M                           |

Table 1: Data simulations targeted for the CSA08 pre-production. The numbers in the sample names stand for transverse energy and momentum threshold parameters. Samples without noise suppression are denoted “NZS”.

2. The CSA08 scenarios

In a workable approximation of the realistic initial ramp-up of the machine, two benchmark scenarios were selected as they are expected to appear during the commissioning of the LHC:

- S43: for each beam, 43 bunches are filled, which is expected to result in a mean luminosity of about $2 \cdot 10^{30} \text{ cm}^{-2} \cdot \text{s}^{-1}$
for each beam 156 bunches are filled, which is expected to result in a mean luminosity of about $2 \cdot 10^{31}$ cm$^{-2}$s$^{-1}$.

The data-taking (=HLT output) rate was assumed to be a constant 300Hz, obtained by adjusting the trigger parameters. A preliminary trigger menu adjusted to start-up conditions was deployed for CSA08. Assuming further an effective duration of 6 days, and accounting for machine running efficiency, each scenario resulted in about 150 million events. The corresponding integrated luminosities were 1pb$^{-1}$ (S43) and 10pb$^{-1}$ (S156), respectively.

In consequence, the CSA08 data samples were governed by QCD-type interactions selected by minimum bias or jet triggers. In addition, small sets of high-$p_T$ muon data with $p_T>5$ GeV/c and $p_T>11$ GeV/c have been added, the latter only for the S156 scenario, which was also augmented with small samples of di-muon and di-electron resonance decays. This was vastly different from the scenarios at nominal luminosity, which earlier studies [1] in the past have been based on. For reasons of simulation economy, a large part of the S43 datasets was reused for the S156 exercise, under the assumption that the low-$p_T$ triggers would be pre-scaled for the S156 scenario, leading to approximately equivalent sized samples. The overall set of Monte Carlo datasets for the CSA08 challenge is listed in Table 1.

In retrospect some of these scenarios have turned out to be too pessimistic, as in the mean time the cosmic muon runs have resulted in much better results for several sub-detectors.

3. Offline workflow and time line

The CMS Offline Workflow used for CSA08 is illustrated in Figure 1. It represented a simplified version of the full design workflow and included the following steps:

- Pre-production of simulated data samples, resulting in RAW datasets, replacing the part that the detector and the trigger will perform in real life. This step was performed at various computing tiers. The resulting data samples were copied to the Tier-0 site at CERN
- Prompt reconstruction of the data at the Tier-0
- Production of alignment and calibration skims. These skims comprise a reduced form of the reconstructed data (named “AlCaReco” datasets), which is obtained by selecting precisely the minimal information needed by the alignment and calibration algorithms.
- Transfer of the AlCaReco datasets to the CERN Analysis Facility (CAF)
- Alignment and calibration at the CAF, upload of constants to the database
- Re-reconstruction at the Tier-1 centers
- Physics analysis at the Tier-2 centers and the CAF

With exception of the preceding pre-production, the CSA08 challenge was concentrated in four weeks from May to early June 2008. In the first week, transfer of the pre-produced datasets to CERN was still ongoing, and the prompt reconstruction of the S43 data sample extended for a few days into the second week of the challenge. Nevertheless, alignment and calibration based on S43 data started at the beginning of the second week and delivered a full set of constants well in time for the reprocessing of this sample, which started at the beginning of the third week. In parallel, alignment and calibration were performed on the S156 data sample, and again the constants were completed within the week. In the fourth and last week of the challenge, the S156 data sample was reconstructed with the updated constants.
Very importantly, all essential milestones for the CSA08 challenge that had been set months in advance were achieved with delays not exceeding two days, and the overall schedule did not slip. Some aspects had to be de-scoped during the challenge, including the production and use of cosmic muon and beam halo data for alignment of the muon system.

4. Computing performance

The CSA08 challenge was relying on the new CMSSW software release 2, which contained a considerable number of new features. After appearance of the release on April 19th, ten more days were required to achieve stable running of the pre-production, which was less than 10 days before the scheduled start of the exercise. For this reason, pre-production was impossible to finish before the prompt reconstruction started. This led to unforeseen overlaps concerning the activities of the operations team, as well as the transfer infrastructure. While some congestion occurred on the CASTOR transfer pools at CERN, these problems could be solved rather quickly without causing major delays.

The computing performance during the pre-production is shown in Figure 2 over the first 12 days of May. This production is distributed over the Tier-0, Tier-1 and Tier-2 levels. On average about 8000 jobs were processing concurrently. Thus the bulk of the CSA08 pre-production could be finished in less than two weeks.
The main CSA08 production activity at the Tier-0 site were the prompt reconstruction and the prompt production of the AlCaReco skims. Execution of these workflows dominated the job bandwidth shown in Figure 3 during the prompt reconstruction of the S156 sample. At the same time, the production team also had to perform the reconstruction of the data from the cosmic run at zero magnetic field (CRUZET). Nevertheless, CSA08 reconstruction could be completed ahead of expectations within less than 5 days.
During the period of prompt production, the resulting data was copied to a pre-defined list of custodial sites. One of the main issues of the organization of tape-writing at the Tier-1 centers was the proper assignment of the LFN file names for the files which are coupled to the definition of tape families. This definition included differentiation at the level of processing versions, which turned out to be more fine-grained than necessary. As a result, new tape families had to be introduced whenever a new processing version arrived, which led to some additional latency and communication overhead with the Tier-1 centers. The conclusion from this lesson was that a reduced granularity of tape families, for example at the level of primary datasets, would be more efficient for this kind of production.

5. Calibration and alignment
The following alignment and calibration exercises have been performed during CSA08:

- tracker alignment with the MillePede [2], Kalman filter [3] and HIP [4] algorithms
- muon system alignment with MillePede and HIP algorithms
- ECAL calibration, exploiting azimuthal symmetry, and using the response to \( \pi^0 \rightarrow \gamma \gamma \) and \( Z^0 \rightarrow ee \) decays
- HCAL calibration, exploiting azimuthal symmetry, single-pion response and transverse energy balance of di-jet signatures
- Muon drift tube chamber calibration: time pedestals and drift velocity
- Pixel tracker calibration: Lorentz angle
- Strip tracker calibration: Lorentz angle and cluster charge
- Determination of beam spot, before and after alignment

In the following, we will show several examples of alignment and calibration results from the CSA08 challenge. The performance of ECAL and muon drift-tubes calibration in CSA08 are described elsewhere [5,6].

5.1. Alignment of the tracker
Three different algorithms have been used for tracker alignment: the HIP (hit and impact point), Kalman filter and MillePede-II algorithms. In the S43 exercise, only minimum bias data (6.6 million tracks) and muon data (p_T>5 GeV/c) have been used. In the S156 study, cosmic muon tracks, a sample of muons with higher transverse momentum (p_T>11 GeV/c) and di-muon samples have been added. Results from the different algorithms have been passed through an identical validation procedure for an unbiased assessment.

The alignment results from the MillePede-II algorithm were selected for the data reprocessing, and are shown in the following. The goodness of fit, a standard measure of track quality, is shown in Figure 4. While the values for unaligned startup geometry are large, barely visible in the figure and extend far beyond the right border of the plot, the results after the alignment exercises give histograms reaching their peak values near 1, and only a gradual improvement is visible going from the S43 to the S156 exercise. The S156 result is very close to the behavior with ideal (=true) geometry. A more detailed investigation within the individual tracker subsections gives mean alignment precisions in \( \phi \) direction of 6 \( \mu \)m (3 \( \mu \)m) for the pixel barrel, 24\( \mu \)m (10 \( \mu \)m) for the tracker inner barrel (TIB), and 48 \( \mu \)m (38 \( \mu \)m) for the tracker inner disks (TID), where the first number corresponds to the S43 and the second number in parentheses to the S156 sample result. The precision is measured after undoing global shifts and rotations.
Figure 4: Track quality (goodness of fit) shown for the misaligned (“Startup”) geometry, and after tracker alignment in the S43 and S156 exercises. The distribution for misaligned geometry is barely visible close to the abscissa in this linear plot.

Figure 5: Reconstructed transverse momentum for particles with a simulated momentum of $p_T=100$ GeV/c

Another important criterion is the reconstruction of high momentum particles, which relies entirely on the coordinate resolution. The resulting resolution is visible in Figure 5 for a particle with $p_T=100$ GeV/c. With the geometry misaligned according to the expected startup situation, the uncertainty is huge and even charge reconstruction will be uncertain. Alignment with 1 pb$^{-1}$ (S43 sample) brings the resolution to 3%, and the 10 pb$^{-1}$ exercise improves the resolution further to 2.2%. This is already close to the ideal geometry value of 1.7%. It is important to note that the last step of improvement is largely due to the addition of cosmic muon tracks, which had not been available for the S43 exercise due to problems with the production workflows. There is no indication of momentum bias visible in
these distributions, which might otherwise indicate the presence of certain global systematic errors which are referred to as weak modes.

A main conclusion from this exercise is that a very good alignment of the CMS tracker should be obtainable already within the first few weeks of low-luminosity collision data.

5.2. Alignment of the muon system

Due to the large amount of material in the CMS magnet yoke and the calorimeter, track-based alignment of the muon system with collisions relies on significant statistics of high $p_T$ muons. The rate of such muons is limited by the luminosity and generally cannot be increased much by lowering trigger thresholds. The general expectation has therefore been that at least 50-100 pb$^{-1}$ of data are needed for a significant track-based alignment of the muon system. The integrated luminosity of the CSA08 exercises were clearly below this threshold; nevertheless, they still allowed to generally commission both the Millepede- and HIP-based muon alignment workflows.

![Figure 6: Comparison of simulated misalignment parameters and the alignment parameters determined by the MillePede procedure](image)

While generally in the S43 sample the statistics was too low for significant alignment, the 10 pb$^{-1}$ exercise demonstrated the proper operation of the track-based alignment algorithms. An example from the intrinsic alignment of the barrel muon chambers is displayed in Figure 6. Three alignment parameters, namely two displacement parameters in the chamber plane, and the rotation angle around the axis normal to it have been determined. The comparison with the true parameters used in the simulated misalignment, indicating a typical accuracy of 700-800 $\mu$m, shows clearly the onset of alignment even at this low integrated luminosity.

5.3. Tracker calibration

The cluster charge calibration of the strip tracker has been studied after applying a purely artificial Gaussian mis-calibration at the level of 10% in the S43 sample, and 5% in the S156 sample.
Calibration proceeded using 23 million minimum bias events by fitting a Landau function to the cluster charge spectrum of each sensor, and obtaining the calibration factor from the peak position (denoted by MPV=most probable value).

The resulting MPV distributions before and after cluster charge calibration are shown in Figure 7. Sharp peaks indicate that calibration has been successful, with an accuracy of better than 1%.

![MPV before calibration](image1)

![MPV after calibration](image2)

Figure 7: Distribution of the most probable value of the fitted Landau distribution for the strip tracker sensors before (top) and after cluster charge calibration (bottom). The result is also shown separately for the sensors of 320 μm and 500 μm thickness.

6. Physics analyses

Physics analyses were carried out in four areas:

- measurement of charged particle spectra, and analysis of the underlying event
- early observation of muons, measurement of the di-muon mass spectrum, observation of J/Ψ, Υ and Z resonances
- early observation of electrons, observation of the Z resonance in the electron channel
- early observation of jets, determination of their corrections and the extraction of early jet physics

These analyses were successfully performed during the CSA08 challenge itself using the results of the prompt reconstruction of the S43 and S156 samples and the re-reconstructed data. They were continued during the two weeks following CSA08 using the re-reconstructed S156 data samples. Besides testing the overall workflow, these analyses represent also an important validation of the alignment and calibration constants derived during CSA08. While a detailed description of the physics analyses is beyond the scope of this paper, one can conclude that the analysis model was found to work well both at the CAF and the Tier-2 centres, the assigned computing resources were generally adequate, and on the whole the CSA08 analyses achieved a major milestone in establishing the end-to-end analysis process for these early physics topics.
7. Lessons from CSA08

On the side of data operations, although pre-production and prompt reconstruction were partially concurrent, the overall traffic was still manageable, and the system coped relatively well with the additional load. The fact that the data processing from global cosmic muon runs was ongoing in parallel, demanding for additional production represented an additional challenge, but both the organizational and computational demands were managed well. Some overhead in the datasets merging and registration procedures was observed, which have been traced back to shortcomings in the underlying software and subsequently corrected.

For the alignment and calibration workflows, important interdependences between the individual workflows have been observed. For example, muon system alignment methods based on tracks reconstructed in the tracker are markedly sensitive to the quality of the tracker alignment, and should therefore be executed after completion of the latter. Tracker alignment in turn is in general influenced by the results of the Lorentz angle calibration. The beam spot position depends on the tracker alignment and thus must be recalculated after alignment has changed. For alignment and calibration of the S156 sample, these dependencies were accounted for by executing the workflows in the appropriate order.

Technically, all the alignment and calibration workflows executed in CSA08 at the CAF were found to fit within a 24 hour window. This is important for the future implementation of a prompt calibration workflow, where small latency of these workflows is essential.

It is important to note that, due to the extended cosmic muon samples recorded since CSA08, in several aspects the CMS initial alignment and calibration will in reality be better than assumed in the CSA08 scenarios.

8. Summary

The CSA08 challenge has successfully demonstrated significant components of the CMS computing workflow in a realistic stress-test. A major pre-production effort of more than 150 million simulated events has been performed in a time scale of about two weeks. Prompt reconstruction, alignment and calibration skims, alignment and calibration exercises, re-reconstruction and the distribution of the data have been performed such that the overall time scale of the challenge could be kept.

In particular also the functionality of CMS alignment and calibration framework has been proven. Both the 1pb\(^{-1}\) and 10pb\(^{-1}\) exercises have been completed well in time by all sub-detector groups, such that the re-reconstruction could proceed on schedule. Significant organizational challenges were mastered, in particular with regard to handling the complexity of a large number of individual workflows, and the interdependencies between them. Validation and sign-off of alignment and calibration constants as well as the management of the corresponding database conditions have been successfully exercised.

Also realistic physics analyses under startup conditions have been performed with low latency, representing an important preparatory step for early observations with the first collision data at the LHC.

References

[1] CMS collaboration (G.L. Bayatian et al.), CMS technical design report, volume II: Physics performance, *J.Phys.*G34:995-1579,2007
[2] V. Blobel and C. Kleinwort, A new method for the high-precision alignment of track detectors, *arXiv*: hep-ex/0208021
[3] R. Fruhwirth, T. Todorov and M. Winkler, Estimation of detector alignment parameters using
the Kalman filter with annealing, J. Phys. G29 (2003) 561.

[4] V. Karimaki, A. Heikkinen, T. Lampen and T. Linden, Sensor alignment by tracks, arXiv: physics/0306034 (2003)

[5] D. Futyan, Commissioning the CMS alignment and calibration framework, Proceedings of the 17th International Conference on Computing in High Energy and Nuclear Physics (CHEP’09), 21 - 27 March 2009 Prague, Czech Republic.

[6] S. Maselli, Calibration of the barrel muon drift tubes system in CMS, Proceedings of the 17th International Conference on Computing in High Energy and Nuclear Physics (CHEP’09), 21 - 27 March 2009 Prague, Czech Republic.