Validation of software releases for CMS

Oliver Gutsche\(^1\) on Behalf of the CMS Computing and Offline Projects

\(^1\) Fermi National Accelerator Laboratory, CMS/CD MS 205, P.O.Box 500, Batavia, IL, 60510, USA

E-mail: gutsche@fnal.gov

Abstract. The CMS software stack currently consists of more than 2 Million lines of code developed by over 250 authors with a new version being released every week. CMS has setup a release validation process for quality assurance which enables the developers to compare to previous releases and references. This process provides the developers with reconstructed datasets of real data and MC samples. The samples span the whole range of detector effects and important physics signatures to benchmark the performance of the software. They are used to investigate interdependency effects of software packages and to find and fix bugs. The samples have to be available in a very short time after a release is published to fit into the streamlined CMS development cycle. The standard CMS processing infrastructure and dedicated resources at CERN and FNAL are used to achieve a very short turnaround of 24 hours. The here described release validation process is an integral part of CMS software development and contributes significantly to ensure stable production and analysis. It’s success emphasizes the importance of a streamlined release validation process for projects with a large code basis and significant number of developers and can function as an example for future projects.

1. Introduction

The Large Hadron Collider (LHC) at CERN, Geneva, Switzerland \(^1\) is expected to record the first proton-proton collisions in September 2009. The Compact Muon Solenoid experiment (CMS)\(^2\), one of the two startup experiments of the LHC, is ready to start collecting data. A world-wide distributed computing infrastructure \(^3\) will be used to process the data, to simulate collisions and to analyze both to extract physics results. The binding element between all these activities is the CMS software stack \(^4\). It contains the persistency formats to store CMS’ data in files on disk and tape and the software to record, reconstruct and analyze it. The stack also contains all necessary components to perform Monte Carlo simulations (MC).

The CMS software stack consists of more than 2 Million lines of code apart from external third party packages and has a very active developer community of more than 250 individual developers \(^5\). CMS consolidates it’s code base regularly into releases. Different releases are grouped into release cycles dedicated to a specific purpose (e.g. data taking and reconstruction in 2009/2010 LHC collision period) and aggregating specific feature sets of the software stack. There are 3 main types of releases:

**Pre-Release:** consolidates the current state of the code base at a given point in time.
It is primarily a test release to present the developer with a consistent snapshot of the development status and allows to test interdependencies between different software components developed in parallel. Pre-Releases are not used for official tasks like data taking and reconstruction, MC production or analysis.

**Release:** closes a development cycle. It contains the final feature set of the respective cycle. New features are not allowed to be integrated anymore and have to go into the next release cycle. The release is used for official central activities like data taking and reconstruction, MC simulation and analysis. A release is distributed and installed world-wide on all levels of the CMS tiered computing infrastructure [3].

**Amendment Release:** only integrates bug fixes to an earlier release to fix specific problems. It supersedes the previous release and is used for official activities and distributed world-wide.

CMS is undergoing a very rapid development cycle in anticipation of first collision data. On average, a new pre-release is provided to the developers and the collaboration every week. In 2008, CMS released 78 different releases of the software stack in 4 different release cycles (see Fig. 1).

![Software Releases in 2008](image)

**Figure 1.** Distribution of release type of the software releases published by CMS in 2008.

The rapid development cycle of CMS and the resulting high number of releases requires a thorough quality assurance process. To guarantee stable and performant releases while supporting high turn around and diverse development, CMS implemented an advanced central release validation process.

### 2. Release Validation
CMS implemented early on a central release validation to accompany the rapid software development. It is now an integral part of the complete software development process and has an important purpose to fulfill:

- Release validation guarantees that all software components of a release work together without interference or failures during execution.
- It checks that a release is compatible with the global production and processing infrastructure of CMS [6].
• Release validation is used to validate the correctness of the produced physics output and performance of a release in terms of:
  - Algorithmic performance,
  - Stability at larger scales (e.g. number of events),
  - Memory and time consumption

In general, release validation produces reference MC simulation samples and reconstruction passes of detector data for every release. The samples are provided to the developers and the collaboration time near for the validation of the release. Dedicated resources are used for the production of these samples/datasets. They are chosen so that the release itself can be installed instantaneous after the announcement and consist of resources at the CERN T0 and the Fermilab T1 center:

**CERN:** 500 priority batch slots are provided in parallel to the T0 resources for a short turn around sample production.

**Fermilab:** Computing cycles of the 5000 batch slots of the Fermilab T1 center are used for release validation in parallel to central production activities like re-reconstruction and skimming. In this mode, Fermilab cannot be used for a short turn around production. It’s main usage is to supplement the short turn around production at CERN by providing with higher statistics samples for validation.

All samples/datasets are grouped into *validation sets*. The sets are optimized to use the available resources efficiently:

**Standard Set:** is produced *within 24 hours* at CERN to enable rapid feedback from validation before the next release is published (on average 1 week). The very low latency gives the developers enough time to validate the release and provide fixes for possible problems.

**High Statistics Set:** is produced *within 1 week* at Fermilab to provide higher statistics comparison samples/datasets to supplement the standard set. The high statistics set is not produced for every pre-release but at least twice per release cycle.

The samples and datasets of the sets span the whole range of detector effects and important physics signatures to benchmark the performance of the software. A summary of the sample composition can be seen in Tab. 1.

<table>
<thead>
<tr>
<th></th>
<th>Full Simulation</th>
<th>Fast Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Particle</td>
<td>Physics</td>
</tr>
<tr>
<td>Generation</td>
<td>Gun Process</td>
<td>Process</td>
</tr>
<tr>
<td>Standard Set</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td>High Statistics Set</td>
<td>12</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 1. Composition and statistics of release validation sample sets. Detector data samples have only been added recently and are not mentioned here.
3. Validation Set Production
The production of each release validation sample consists of the following components:

(i) Generation of proton-proton collisions for various physics processes
(ii) Simulation of the detector response
(iii) Local reconstruction of individual detector component signals
(iv) Simulation of the trigger response
(v) Reconstruction of objects like tracks, jets and other global event quantities
(vi) Extraction of special event quantities for alignment and calibration of the detector.

For the full simulation samples, these components cannot be executed in one step. They have to be processed separately in the following steps:

(i) Generation, Simulation, local reconstruction and simulation of the trigger response
(ii) Reconstruction
(iii) Extraction of alignment and calibration event quantities

Each of the steps produces one or more output datasets. The result of step 1 is called the RAW dataset and is the input to step 2. The result of step 2 is called the RECO dataset and is the input to step 3. Step 3 has several different outputs which are called the AL&Ca datasets (although also step 2 writes out some AL&Ca datasets). The release validation sets contain a matrix of different configurations for the different steps combined according to the physics content of the generation.

Normally, one would process the different steps one after the other, finishing the requested statistics for step 1 before starting step 2 on the output of this step. This introduces significant bookkeeping overheads into the total processing time because of tail effects. It reduces the efficiency with which the special validation batch queue at CERN is used and the amount of samples which can be produced in a given time window (24 hours for the standard set). Because the demand for release validation samples increased significantly over time as more and more components of the software had to be thoroughly tested, CMS optimized the production infrastructure by introducing the concept of chained processing (see Fig. 2).

**Figure 2.** Chained processing is an optimization of the CMS production and processing infrastructure to minimize bookkeeping overhead. It runs the steps of the release validation sample production in one job using the output of previous steps directly on the workernode.
The production infrastructure was now able to process all steps of a release validation sample production in one job on one batch slot using the output of a previous step as the input of the next step. This requires that each processing job was able to stage out multiple output files corresponding to the outputs of the different steps. With chained processing, the number of samples could be increased significantly. Fig. 3 shows the special validation queue at CERN during production of several standard release validation sample sets. Each set fits nicely into the 24 hours time window.

![Special Validation Queue at CERN](image)

**Figure 3.** Special validation queue at CERN.

The very efficient usage of processing resources of the chained processing mode for multi-step production workflows was pioneered in release validation and is now also used in CMS’ central MC production.

Chained processing enables CMS to produce a significant number of release validation samples with enough statistics to fulfill most of the sample demands by development. Fig. 4 shows the total number of events and Fig. 5 the total number of different datasets for the standard release validation set.

![Events per Standard Set](image)

**Figure 4.** Total number of events for the standard release validation set.

![Datasets per Standard Set](image)

**Figure 5.** Total number of datasets for the standard release validation set.
The same is shown in Fig. 6 and Fig. 7 for the high statistics release validation set.

Because of the higher statistics and the competition with the official production at the Fermilab T1, the high statistics validation set is not produced for every pre-release or release. It is normally produced twice per release cycle.

4. Performance Overview
Using all the release validation samples produced for a certain pre-release or release, global performance numbers are extracted using information from the production jobs themselves. These global performance numbers track the maximum virtual memory consumption and the time per event for every step of the release validation sample production. Fig. 8 shows exemplary the maximum memory consumption for the simulation step and the time per events for the reconstruction step averaged over all processing jobs per sample.

These numbers are used to compare the performance within one release and between different releases to see trends in the memory consumption and processing time.

5. 2008 Release Validation
Release validation is an integral part of the CMS software development process. In 2008, CMS produced only for validation more than 5100 release validation datasets corresponding to over 114 million events and 110 terabyte of MC and data samples. Fig. 9 shows the total number of events produced for release validation in 2008 and Fig. 10 the total data size.

Developers rely on the timely provision of reference samples to validate their software components. The demand for release validation samples exceeds the production capabilities of the available resources many times. Therefore, compromises in the design of the release validation sample sets have to be found and synergies have to be exploited where possible. The samples are rapidly used after they have been made available to the collaboration. About 2/3 of the release validation samples are used for validation within 2 days of the announcement of the samples, 1/3 is used between 6 and 10 days after the availability of the samples.
Figure 8. Performance overview from the release validation sample production averaged over all processing jobs of individual samples. **Left:** maximum memory consumption for the simulation step, **Right:** time per events for the reconstruction step.

Figure 9. Total number of events produced in 2008 for release validation.

Figure 10. Total amount of terabyte produced in 2008 for release validation.

6. Conclusions
CMS release validation contributes significantly to ensure stable production and analysis. By providing reference MC and data samples for each pre-release and release, developers are able to validate their software components for performance and stability and check interference effects between different components developed in parallel. Without the release validation effort, CMS
would not have been able to guarantee the stability and performance of each release within the very rapid development cycle of almost one pre-release per week. In 2008, the release validation production was sizeable and resulted in over 114 million events and 110 terabyte of MC and data samples produced. In the future, the release validation sample sets will be adapted further to the needs of the developers and will use detector data samples as input rather than MC generations.

7. Acknowledgments

We would like to thank the funding agencies under which this work was conducted, and all CMS collaborators for their help, support and contributions.

References

[6] F. van Lingen, “CMS production and processing system - design and experiences”, Poster 82, CHEP '09