Table of Contents

2.2 CMS Computing Model

Complete: 🟩

Contents

Goals of this workbook page

When you finish this page, you should understand:

- the tier structure of the CMS computing model.
- how detector data and MC data travel through the tiers.
- how data are distributed, stored, and accessed.

Introduction

CMS presents challenges not only in terms of the physics to discover and the detector to build and operate, but also in terms of the data volume and the necessary computing resources. Data sets and resource requirements are at least an order of magnitude larger than in previous experiments.

CMS computing and storage requirements would be difficult to fulfill at any one place, for both technical and funding reasons. Additionally, most CMS collaborators are not CERN-based, and have access to significant non-CERN resources, which it is advantageous to harness for CMS computing. Therefore, the CMS computing environment has been constructed as a distributed system of computing services and resources that interact with each other as Grid services. The set of services and their behaviour together comprise the computing, storage and connectivity resources that CMS uses to do data processing, data archiving, Monte Carlo event generation, and all kinds of computing-related activities.

The computational infrastructure is intended to be available to CMS collaborators, independently of their physical locations, and on a fair share basis.

Tier architecture of computing resources

The computing centres available to CMS around the world are distributed and configured in a tiered architecture that functions as a single coherent system. Each of the three tier levels provides different resources and services:

Tier-0 (T0)

The first tier in the CMS model, for which there is only one site, CERN, is known as Tier-0 (T0). The T0 performs several functions. The standard workflow is as follows:

1. accepts RAW data from the CMS Online Data Acquisition and Trigger System (TriDAS)
2. repacks the RAW data received from the DAQ into primary datasets based on trigger information (immutable bits). Roughly 10 datasets are expected in the 7TeV run when there is sufficient luminosity and eventually growing to 15.
3. archives the repacked RAW data to tape.
4. distributes RAW data sets among the next tier stage resources (Tier-1) so that two copies of every piece of RAW data is saved, one at CERN, another at a Tier-1.
5. performs PromptCalibration in order to get the calibration constants needed to run the reconstruction.
6. feeds the RAW datasets to reconstruction.
7. performs *prompt* first pass reconstruction which writes the RECO, Analysis Object Data (AOD) and mini-AOD extraction.
8. distributes the RECO datasets among Tier-1 centers, such that the RAW and RECO match up at each Tier-1.
9. distributes full AOD/mini-AOD to all Tier-1 centers.

The T0 does not provide analysis resources and only operates scheduled activities.

The T0 merges output files if they are too small. (This will affect RECO and AOD, and maybe AlcaReco; under certain repacker scenarios one could even imagine merging RAW data files but this will be avoided as much as possible.) The goal of CMS is to write appropriately sized data into the tape robots. Currently CMS typically imports 2-3GB files, though 5-10GB files are technically possible and are desirable for tape system performance.

At CERN, though logically separated from the T0 is the CMS-CAF (CERN Analysis Facility). The CAF offers services associated with T1 and T2 centers and performs latency critical, non-automated activities. The CAF is not needed for normal Tier0 operation; it is intended for short-term, high priority, human-operated calibration, physics validation and analysis. For example, the CAF would be used for very fast physics validation and analysis of the *Express Stream* (a subset of the data that is tagged by Online and then processed as quickly as possible).

Tier-1 (T1)

There is a set of seven Tier-1 (T1) sites, which are large centers in CMS collaborating countries (large national labs, e.g. FNAL, and RAL). Tier-1 sites will in general be used for large-scale, centrally organized activities and can provide data to and receive data from all Tier-2 sites. Each T1 center:

1. receives a subset of the data from the T0 related to the size of the pledged resources in the WLCG MOU
2. provides tape archive of part of the RAW data (secure second copy) which it receives as a subset of the datasets from the T0

3. provides substantial CPU power for scheduled:
    - re-reconstruction
    - skimming
    - calibration
    - AOD extraction
4. stores an entire copy of the AOD
5. distributes RECOs, skims and AOD to the other T1 centers and CERN as well as the associated group of T2 centers
6. provides secure storage and redistribution for MC events generated by the T2's (described below)

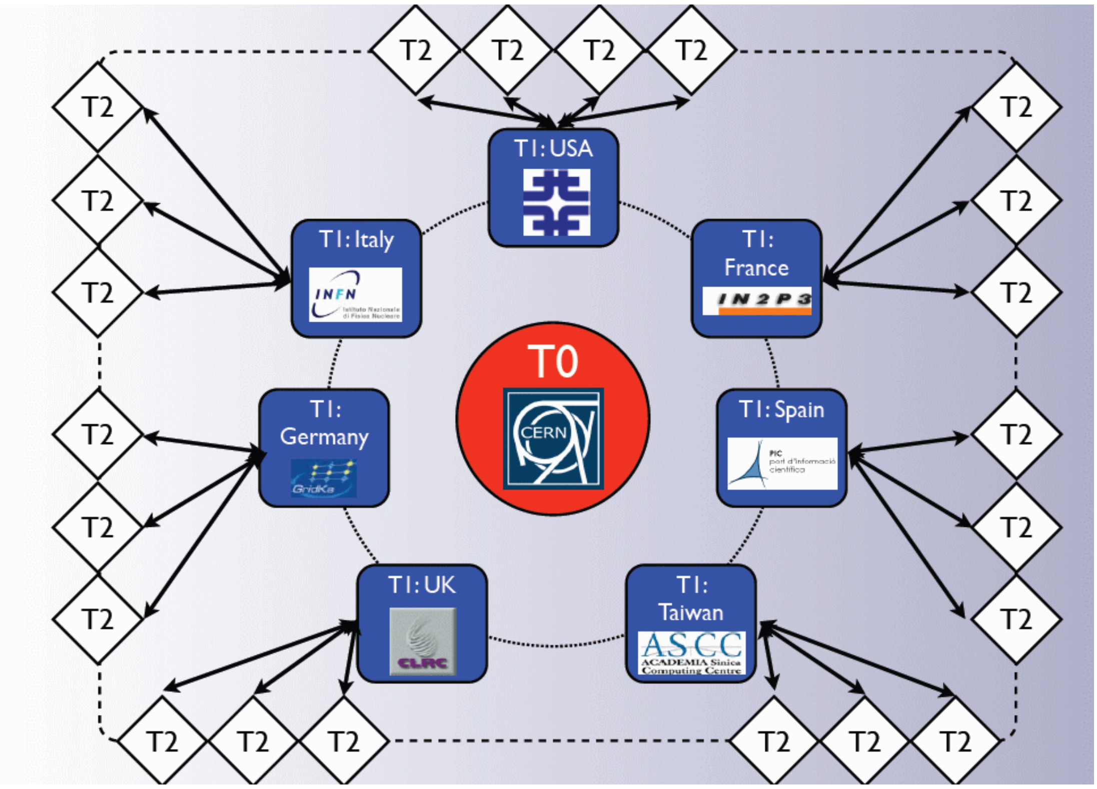
Tier-2 (T2)

A more numerous set of smaller Tier-2 (T2) centres ("small" centres at universities), but with substantial CPU resources, provide capacity for user analysis, calibration studies, and Monte Carlo production. T2 centers provide limited disk space, and no tape archiving. T2 centers rely upon T1s for access to large datasets and for secure storage of the new data (generally Monte Carlo) produced at the T2. The MC production in Tier-2's will in general be centrally organized, with generated MC samples being sent to an associated Tier-1 site for distribution among the CMS community. All other Tier-2 activities will be user driven, with data placed to match resources and needs: tape, disk, manpower, and the needs of local communities. The Tier-2 activities will be organized by the Tier-2 responsibles in collaboration with physics groups, regional associations and local communities.

In summary, the Tier-2 sites provide:

1. services for local communities
2. grid-based analysis for the whole experiment (Tier-2 resources available to whole experiment through the grid)
3. Monte Carlo simulation for the whole experiment

As of July '18 there are about 55 T2 sites, each associated with one of the seven T1 sites or directly to CERN (the following image does not represent the actual T2 groupings under the T1s):



One can refer to the Dashboard site status monitoring page for the most up-to-date information about available sites along with their status.

Data Organization

Data a physicist wants to see

To extract a physics message for a high energy physics analysis, a physicist has to combine a variety of information:

- reconstructed information from the recorded detector data, specified by a combination of trigger paths and possibly further selected by cuts on reconstructed quantities (e.g., two jets),
- MC samples which simulate the physics signal under investigation, and
- background samples (specified by the simulated physics process).

The physics abstractions physicists use to request these items are *datasets* and *event collections*. The datasets are split off at the T0 and distributed to the T1s, as described above. An event collection is the smallest unit within a dataset that a user can select. Typically, the reconstructed information needed for the analysis, as in the first bullet above, would all be contained in one or a few event collection(s). The expectation is that the majority of analyses should be able to be performed on a single primary dataset.

Data are stored as ROOT files. The smallest unit in computing space is the file block which corresponds to a group of ROOT files likely to be accessed together. This requires a mapping from the physics abstraction (event collection) to the file location. CMS has a global data catalog called the Dataset Aggregation System (DAS) which provides mapping between the physics abstraction (dataset or event collection) and the list of fileblocks corresponding to this abstraction. It also gives the user an overview of what is available for analysis, as it has the complete catalog. The locations of these fileblocks within the CMS grid (several centers can provide access to the same fileblock) are resolved by the PhEDEx, the Physics Experiment Data EXport service. PhEDEx is responsible for transporting data around the CMS sites, and keeps track of which data exists at which site. The mapping thus occurs in two steps, at the DAS and PhEDEx. See WorkBookAnalysisWorkFlow for an illustration (note that, in that illustration, the role of the data-location service is represented by 'DLS', which was eliminated as being functionally redundant with the information contained in PhEDEx).

The CMS Data Hierarchy

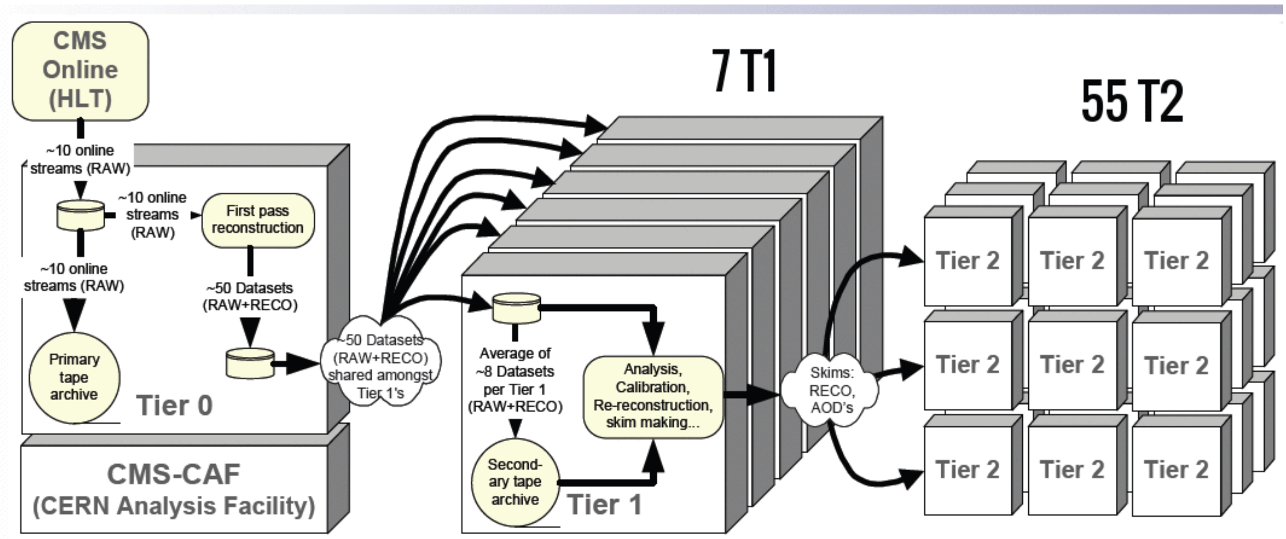
CMS Data is arranged into a hierarchy of data tiers. Each physics event is written into each data tier, where the tiers each contain different levels of information about the event. The different tiers each have different uses. The three main data tiers written in CMS are:

1. RAW: full event information from the Tier-0 (i.e. from CERN), containing 'raw' detector information (detector element hits, etc)
    - RAW is not used directly for analysis
2. RECO ("RECOnstructed data"): the output from first-pass processing by the Tier-0. This layer contains reconstructed physics objects, but it's still very detailed
    - RECO can be used for analysis, but is too big for frequent or heavy use when CMS has collected a substantial data sample.
3. AOD ("Analysis Object Data"): this is a "distilled" version of the RECO event information, and is expected to be used for most analyses
    - AOD provides a trade-off between event size and complexity of the available information to optimize flexibility and speed for analyses

The data tiers are described in more detail in a dedicated WorkBook chapter on Data Formats and Tiers It is the desire of CMS that the data tiers are written into separate files, though applications will be able to access more than one file simultaneously (an application will be able to access Reco and the corresponding RAW events from separate files.)

Detector data flow through Hardware Tiers

The following diagram shows the flow of CMS detector data through the tiers.



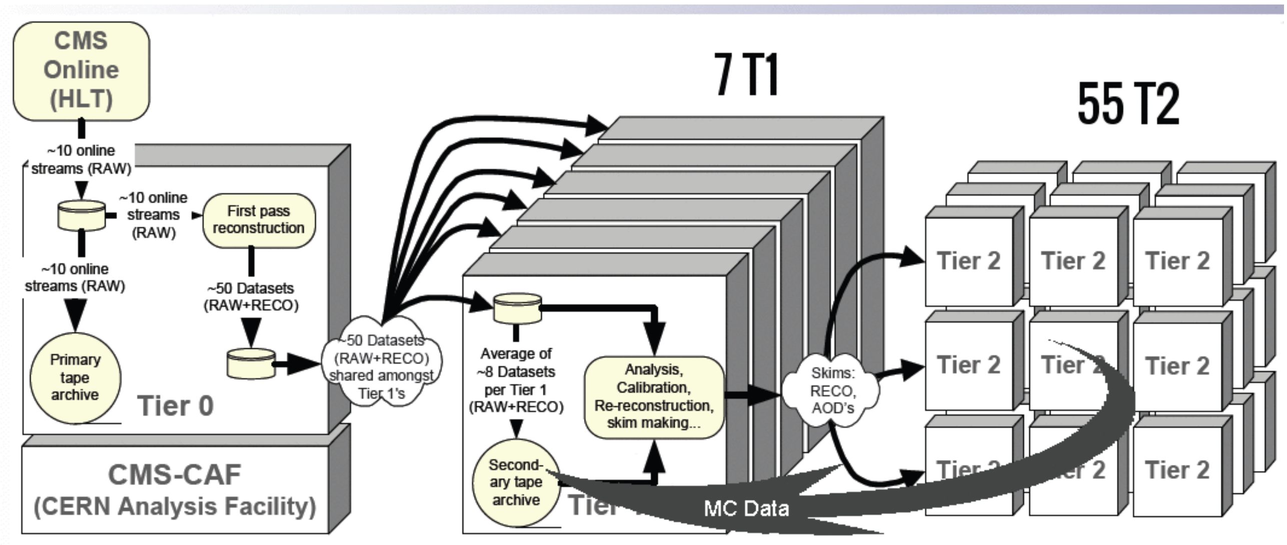
The essential elements of the flow of real physics data through the hardware tiers are:

- T0 to T1:
    - Scheduled, time-critical, will be continuous during data-taking periods

    - reliable transfer needed for fast access to new data, and to ensure that data is stored safely
- T1 to T1:
    - redistributing data, generally after reprocessing (e.g. processing with improved algorithms)
- T1 to T2:
    - Data for analysis at Tier-2s

Monte Carlo data flow through Hardware Tiers

Monte Carlo generated data is typically produced at a T2 center, and archived at its associated T1 and made available to the whole CMS collaboration.



Workflows in CMS Computing

A workflow can be described simply as "what we do to the data". There are three principle areas of workflow in CMS:

1. *At Tier2 Centres*: Monte Carlo events are generated, detector interactions simulated, events reconstructed in the same manner as will be applied to data, and the events are then moved to tape storage for later use
2. *At The Tier0 Center*: Data is received from the CMS detector experiment, it is "repacked" - i.e. events from the unsorted online streams are sorted into physics streams of events with similar characteristics. Reconstruction algorithms are run, AOD is produced, and RAW, RECO and AOD are exported to Tier1 sites
3. *The user - i.e. YOU!*: prepare analysis code, send code to site where there is appropriate data, then run your code on the data and collect the results
    - the process of finding the sites with data and CPU, running the jobs, and collecting the results is all managed for you (via the grid) by CRAB

Managing Grid Jobs

The management of grid jobs is handled by a series of systems, described in WorkBookAnalysisWorkFlow. The goal is to schedule jobs onto resources according to the policy and priorities of CMS, to assist in monitoring the status of those jobs, and to guarantee that site-local services can be accurately discovered by the application once it starts executing in a batch slot at the site. As a user, these issues should be invisible to you.

The datasets are tracked as they are distributed around the globe by the CMS Dataset Aggregation Service (DAS), while the Physics Experiment Data Export service (PhEDEx) moves data around CMS.

A major bottleneck in the data analysis process can be retrieval of data from tape stores, so storage and retrieval are major factors in optimising analysis speed.

Information Sources

- **The CMS Computing Model**, C. Grandi, D. Stickland, L. Taylor, CMS NOTE 2004-031 (2004) and CERN LHCC 2004-035/G-083.
- **CMS Computing Technical Design Report**, CERN-LHCC-2005-023 and CMS TDR 7, 20 June 2005.
- **CMS Computing Project Technical Design Report** at http://cms.cern.ch/iCMS/ (select *CPT* on left menu, find *Technical Design Reports* underneath the table in main section of page).
- Material on this page taken also from Tony Wildish's Computing Model lecture on the 2007 CERN Summer Studentship Programme.

Review status

Reviewer/Editor and Date (copy from screen)	Comments
JennyWilliams - 27 Sep 2007	updates based on Tony Wildish's summer students lecture
PeterElmer - 19 Feb 2007	updates and improvement in description of roles of Tier1 and Tier2 centres
TonyWildish - 28 Jan 2008	correct description of T0 workflow
IanFisk - 21 Feb 2008	Clarifications to the number of Tier-1s and a few minor corrections and additions. Added concept of 2 file reading

Detailed comments 23-Sep-2012 ▸ Hide ▾
I went through chapter 2 section 2. The information is relevant and clear. This section had a few links out of date, I updated them.

Open a new link at "DBS"

Install at several places link " PhEDEx "

Update link at "AOD"

Update link at " CERN LHCC 2004-035/G-083 "

Responsible: TonyWildish
Last reviewed by: PatriciaMcBride - 22 Feb 2008

This topic: CMSPublic > WorkBookComputingModel
Topic revision: r60 - 2018-07-22 - NitishDhingra

Copyright &© 2008-2019 by the contributing authors. All material on this collaboration platform is the property of the contributing authors.
Ideas, requests, problems regarding TWiki? Send feedback