Chapter 11

Computing

11.1 Overview

The CMS offline computing system must support the storage, transfer and manipulation of the recorded data for the lifetime of the experiment. The system accepts real-time detector information from the data acquisition system at the experimental site; ensures safe curation of the raw data; performs pattern recognition, event filtering, and data reduction; supports the physics analysis activities of the collaboration. The system also supports production and distribution of simulated data, and access to conditions and calibration information and other non-event data.

The users of the system, and the physical computer centres it comprises, are distributed worldwide, interconnected by high-speed international networks. Unlike previous generations of experiments, the majority of data storage and processing resources available to CMS lie outside the host laboratory. A fully distributed computing model has therefore been designed from the outset. The system is based upon Grid middleware, with the common Grid services at centres defined and managed through the Worldwide LHC Computing Grid (WLCG) project [242], a collaboration between LHC experiments, computing centres, and middleware providers.

The nature of the CMS experimental programme poses several challenges for the offline computing system:

- The requirement to analyse very large statistics datasets in pursuit of rare signals, coupled with the fine granularity of the CMS detector, implies a volume of data unprecedented in scientific computing. This requires a system of *large scale*, supporting efficient approaches to data reduction and pattern recognition.

- The system is required to be *highly flexible*, allowing any user access to any data item recorded or calculated during the lifetime of the experiment. A software framework is required which supports a wide variety of data processing tasks in a consistent way, and which must evolve along with the goals of the experiment. Since the CMS programme centres on discovery of new phenomena, under new experimental conditions, analysis requirements cannot be wholly defined in advance.

- A complex distributed system of such large scale must be designed from the outset for *manageability*, both in the operation of computing resources for physics, and in terms of software

construction and maintenance. The *longevity* of the system, of 15 years or more, implies several generations of underlying hardware and software, and many changes of personnel, during the lifetime of the system.

Key components of the computing system include:

- An event data model and corresponding application framework;
- Distributed database systems allowing access to non-event data;
- A set of computing services, providing tools to transfer, locate, and process large collections of events;
- Underlying generic Grid services giving access to distributed computing resources;
- Computer centres, managing and providing access to storage and CPU at a local level.

At each level, the design challenges have been addressed through construction of a modular system of loosely coupled components with well-defined interfaces, and with emphasis on scalability to very large event samples [243].

11.2 Application framework

The CMS application software must perform a variety of event processing, selection and analysis tasks, and is used in both offline and online contexts. The software must be sufficiently modular that it can be developed and maintained by a large group of geographically dispersed collaborators. The chosen architecture consists of a common framework which is adaptable for each type of computing environment, physics modules which plug into the framework via a well-defined interface, and a service and utility toolkit which decouples the physics modules from details of event I/O, user interface, and other environmental constraints [212].

The central concept of the CMS data model is the *Event*. The Event provides access to the recorded data from a single triggered bunch crossing, and to new data derived from it. This may include raw digitised data, reconstructed products, or high-level analysis objects, for real or simulated crossings. The Event also contains information describing the origin of the raw data, and the provenance of all derived data products. The inclusion of provenance information allows users to unambiguously identify how each event contributing to a final analysis was produced; it includes a record of the software configuration and conditions / calibration setup used to produce each new data product. Events are physically stored as persistent ROOT files [244].

The Event is used by a variety of *physics modules*, which may read data from it, or add new data, with provenance information automatically included. Each module performs a well-defined function relating to the selection, reconstruction or analysis of the Event. Several module types exist, each with a specialised interface. These include: *event data producers*, which add new data products into the event; *filters* used in online triggering and selection; *analysers*, producing summary information from an event collection; and *input and output modules* for both disk storage and DAQ.

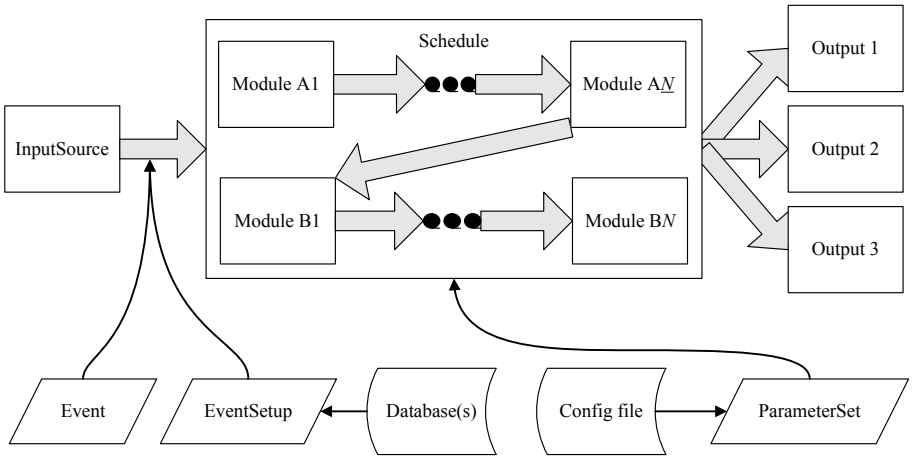

Figure 11.1: Modules within the CMS Application Framework.

Modules are insulated from the computing environment, execute independently from one another, and communicate only though the Event; this allows modules to be developed and verified independently. A complete CMS application is constructed by specifying to the Framework one or more ordered sequences of modules through which each Event must flow, along with the configuration for each. The Framework configures the modules, schedules their execution, and provides access to global services and utilities (figure 11.1).

11.3 Data formats and processing

In order to achieve the required level of data reduction, whilst maintaining flexibility, CMS makes use of several event formats with differing levels of detail and precision. Other specialised event formats are used for heavy-ion data. The process of data reduction and analysis takes place in several steps, typically carried out at different computer centres.

RAW format

RAW events contain the full recorded information from the detector, plus a record of the trigger decision and other metadata. RAW data is accepted into the offline system at the HLT output rate (nominally 300 Hz for pp collisions). An extension of the RAW data format is used to store the output of CMS Monte Carlo simulation tools. The RAW data is permanently archived in safe storage, and is designed to occupy around 1.5 MB/event (2 MB/event for simulated data, due to additional Monte Carlo truth information).

The RAW data will be classified by the online system into several distinct *primary datasets*, based upon the trigger signature. Event classification at the earliest possible stage has several advantages, including the possibility of assigning priorities for data reconstruction and transfer in the case of backlog, and balancing of data placement at centres outside CERN. CMS will also define one or more flexible "express streams" used for prompt calibration and rapid access to interesting or anomalous events.

2008 JINST 3 S08004

RECO format

Reconstructed (RECO) data is produced by applying several levels of pattern recognition and compression algorithms to the RAW data. These algorithms include: detector-specific filtering and correction of the the digitised data; cluster- and track-finding; primary and secondary vertex reconstruction; and particle ID, using a variety of algorithms operating on cross-detector information.

Reconstruction is the most CPU-intensive activity in the CMS data processing chain. The resulting RECO events contain high-level *physics objects*, plus a full record of the reconstructed hits and clusters used to produce them. Sufficient information is retained to allow subsequent application of new calibrations or algorithms without recourse to RAW data, though basic improvements in pattern recognition or event formats will probably require re-production of the RECO data at least once per year. RECO events are foreseen to occupy around 0.5 MB/event.

AOD format

AOD (Analysis Object Data) is the compact analysis format, designed to allow a wide range of physics analyses whilst occupying sufficiently small storage so that very large event samples may be held at many centres. AOD events contain the parameters of high-level physics objects, plus sufficient additional information to allow kinematic refitting. This format will require around 100 kB/event, small enough to allow a complete copy of the experimental data in AOD format to be held at computing centres outside CERN. AOD data is produced by filtering of RECO data, either in bulk production, or in a skimming process which may also filter a primary dataset into several analysis datasets.

Non-Event data

In addition to event data recorded from the detector, a variety of *non-event data* is required in order to interpret and reconstruct events. CMS makes use of four types of non-event data: construction data, generated during the construction of the detector; equipment management data; configuration data, comprising programmable parameters related to detector operation; and conditions data, including calibrations, alignments and detector status information. We concentrate here on the lattermost category.

Conditions data are produced and required by both online and offline applications, and have a well-defined interval of validity (IOV). For instance, calibration constants for a given run may be derived from prompt reconstruction of a subset of recorded events, and then used both by the HLT system and for subsequent reconstruction and analysis at computing centres around the world. Non-event data are held in a number of central Oracle databases, for access by online and offline applications. New conditions data, including calibration and alignment constants produced offline, may be replicated between the databases as required. Conditions data access at remote sites takes place via the FroNTier system [245] which uses a distributed network of caching http proxy servers.

2008 JINST 3 S08004

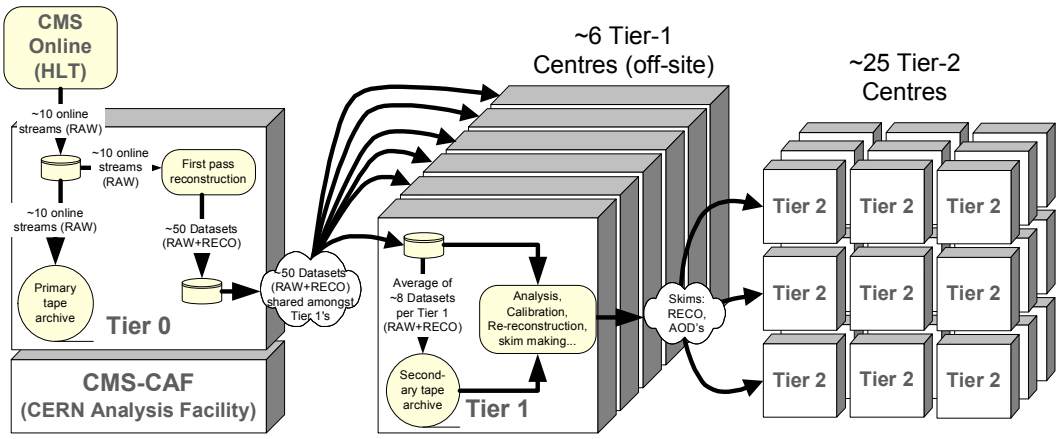

Figure 11.2: Dataflow between CMS Computing Centres.

11.4 Computing centres

The scale of the computing system is such that it could not, even in principle, be hosted entirely at one site. The system is built using computing resources at a range of scales, provided by collaborating institutes around the world. CMS proposes to use a hierarchical architecture of Tiered centres, similar to that originally devised in the MONARC working group [246], with a single Tier-0 centre at CERN, a few Tier-1 centres at national computing facilities, and several Tier-2 centres at institutes. A representation of the dataflow between centres is shown in figure 11.2.

The CMS computing model depends upon reliable and performant network links between sites. In the case of transfers between Tier-0 and Tier-1 centres, these network links are implemented as an optical private network (LHC-OPN) [247]. Data transfers between Tier-1 and Tier-2 centres typically takes place over general-purpose national and international research networks.

Tier-0 centre

A single Tier-0 centre is hosted at CERN. Its primary functions are to:

- Accept data from the online system with guaranteed integrity and latency, and copy it to permanent mass storage;
- Carry out prompt reconstruction of the RAW data to produce first-pass RECO datasets. The centre must keep pace with the average rate of data recording, and must provide sufficient input buffering to absorb fluctuations in data rate;
- Reliably export a copy of RAW and RECO data to Tier-1 centres. Data is not considered "safe" for deletion from Tier-0 buffers until it is held at at least two independent sites. (One of these is CERN computing centre, playing the role of a Tier-1.)

During the LHC low-luminosity phase, the Tier-0 is intended to be available outside data-taking periods for second-pass reconstruction and other scheduled processing activities. High-luminosity running will require the use of the Tier-0 for most of the year. The Tier-0 is a common CMS facility used only for well-controlled batch work; it is not accessible for analysis use.

2008 JINST 3 S08004

Tier-1 centres

A few large Tier-1 centres are hosted at collaborating national labs and computing centres around the world. These centres are operated by a professional staff on a 24/365 basis, with the emphasis on extremely reliable delivery of data-intensive processing services. Each site provides large batch CPU facilities, a mass storage system including a robotic tape archive, and very high speed international network links including a dedicated link to the LHC-OPN. The primary functions of a Tier-1 are to:

- Provide long-term safe storage of RAW data from CMS, providing a second complete copy outside CERN distributed across the centres. Each Tier-1 takes long-term custodial responsibility for a fraction of the CMS dataset;
- Store and serve to Tier-2 centres simulated and derived data. Each Tier-1 holds a fraction of the CMS simulated and RECO data, and a complete copy of the AOD data. It can rapidly transfers these data to any Tier-2 centre which requires them for analysis;
- Carry out second-pass reconstruction: a Tier-1 provides access to its archive of RAW data to allow reproduction of RECO datasets using improved algorithms or calibrations;
- Provide rapid access to very large data samples for skimming and data-intensive analysis: a Tier-1 can support high-statistics analysis projects which would be infeasible at a Tier-2 centre.

Since each Tier-1 centre holds unique RAW and RECO datasets, it must be capable of serving data to any CMS Tier-2. However, for the purposes of Monte Carlo data receipt and AOD data serving, the Tier-1 serves a defined set of a few "associated" Tier-2 centres, usually defined by geographical proximity.

Tier-2 centres

Several Tier-2 centres of varying sizes are hosted at CMS institutes. A Tier-2 centre typically divides its resources between the local user community and CMS as a whole. Tier-2 centres are subject to less stringent requirements on availability and data security than a Tier-1 centre, making them feasible to manage with the resources available to a typical University group. The functions of a Tier-2 centre may include:

- Support of analysis activities, including local storage of data samples transferred from Tier-1 centres, and access to a flexible CPU farm; in particular, the Tier-2 centres are designed to support final-stage analysis requiring repeated passes over a reduced dataset;
- Support of specialised activities such as offline calibration and alignment tasks, and detector studies;
- Production of Monte Carlo data, and its transfer to an associated Tier-1 centre for long term storage.

2008 JINST 3 S08004

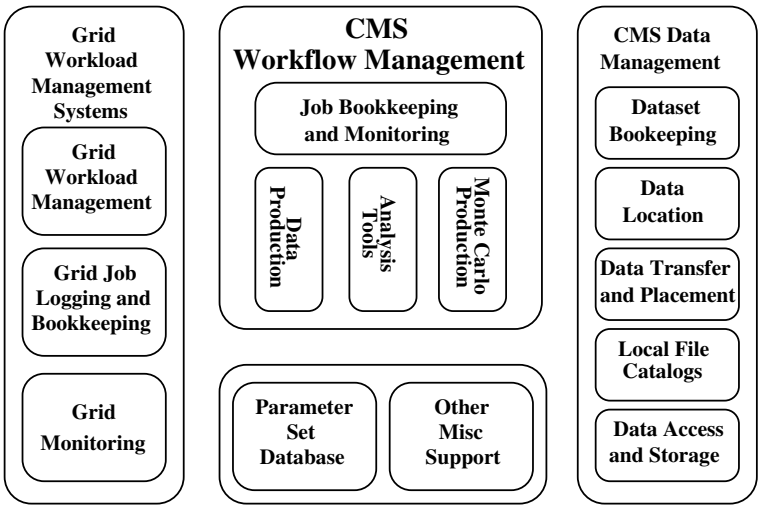

Figure 11.3: Overview of the CMS Computing Services.

CERN Analysis Facility

In addition to the Tier-0 centre, CERN also hosts an Analysis Facility which combines flexible CPU resources with rapid access to the entire CMS dataset. This centre supports fast turn-around analysis when required, and a variety of other specialised functions (calibration, performance monitoring) related to the operation of the CMS detector. The centre effectively combines the rapid data access capabilities of a Tier-1 with the flexibility of a very large Tier-2.

11.5 Computing services

Grid computing

The integration of the resources at CMS computing centres into a single coherent system relies upon Grid middleware which presents a standardised interface to storage and CPU facilities at each WLCG (Worldwide LHC Computing Grid) site. The Grid allows remote job submission and data access with robust security and accounting. The detailed architecture of the Grid is described in the WLCG Technical Design Report [242].

A number of CMS-specific distributed computing services operate above the generic Grid layer, facilitating higher-level data and workload management functions. These services require CMS-specific software agents to run at some sites, in addition to generic Grid services. CMS also provides specialised user-intelligible interfaces to the Grid for analysis job submission and monitoring, and tools for automated steering and monitoring of large-scale data production and processing. An overview of the CMS Computing Services components is shown in figure 11.3.

2008 JINST 3 S08004

Data management

CMS requires tools to catalogue the data which exist, to track the location of the corresponding physical data files on site storage systems, and to manage and monitor the flow of data between sites. In order to simplify the data management problem, the data management system therefore defines higher-level concepts including: *dataset*, a logical collection of data grouped by physical-meaningful criteria; *event collection*, roughly corresponding to an experiment "run" for a given dataset definition; and *file block*, an aggregation of a few TB of data files, representing the smallest unit of operation of the data transfer system.

To provide the connection between abstract datasets and physical files, a multi-tiered catalogue system is used. The *Dataset Bookkeeping System* provides a standardised and queryable means of cataloguing and describing event data [249]. It is the principle means of data discovery for the end user, answering the question "which data of this type exists in the system?" A second catalogue system, the *Data Location Service* provides the mapping between file blocks to the particular sites at which they are located, taking into account the possibility of replicas at multiple sites. *Local File Catalogues* at each site map logical files onto physical files in local storage.

The *data transfer and placement system* is responsible for the physical movement of file-blocks between sites on demand; it is currently implemented by the PhEDEx system [248]. This system must schedule, monitor and verify the movement of data in conjunction with the storage interfaces at CMS sites, ensuring optimal use of the available bandwidth. The baseline mode of operation for the data management system is that the collaboration will explicitly place datasets at defined sites, where they will remain for access by CMS applications until removed.

Workload management

Processing and analysis of data at sites is typically performed by submission of batch jobs to a remote site via the Grid workload management system. A standard job wrapper performs the necessary setup, executes a CMSSW application upon data present on local storage at the site, arranges for any produced data to be made accessible via Grid data management tools, and provides logging information. This process is supported by several CMS-specific services.

A *parameter set management system*, implemented with either global or local scope according to the application, allows the storage and tracking of the configuration of CMSSW applications submitted to the Grid. A lightweight *job bookkeeping and monitoring system* allows users to track, monitor, and retrieve output from jobs currently submitted to and executing at remote sites [250]. The system also provides a uniform interface to a variety of Grid-based and local batch-system based submission tools. In addition, a suite of software distribution tools provide facilities for automated installation of standard CMS applications and libraries at remote sites.

Bulk workflow management

For very large-scale data processing (including Monte Carlo production, skimming and event reconstruction), a specialised bulk workflow management tool has been developed. The ProdAgent system comprises a collaborative distributed network of automated job managers, operating at Tier-0, Tier-1 and Tier-2 sites [250]. The system provides facilities for large-scale Grid job submission,

2008 JINST 3 S08004

interface to the CMS data catalogues and data management system, and handling of large flows of logging and status information. A highly automated system such as ProdAgent is essential in order to allow the CMS data processing system to be controlled and monitored by a moderately-sized data operations team.

User workflow management

For a generic CMS physicist, a dedicated tool (CRAB) for workflow management is available [250]. It allows to submit user-specific jobs to a remote computing element which can access data previously transferred to a close storage element. CRAB takes care of interfacing with the user environment, it provides data-discovery and data-location services, and Grid infrastructure. It also manages status reporting, monitoring, and user job output which can be put on a user-selected storage element. Via a simple configuration file, a physicist can thus access data available on remote sites as easily as he can access local data: all infrastructure complexities are hidden to him as much as possible. There is also a client-server architecture available, so the job is not directly submitted to the Grid but to a dedicated CRAB server, which, in turn, handles the job on behalf of the user, interacting with the Grid services.

11.6 System commissioning and tests

It has been recognised since the very start of preparations for LHC that the construction and organisation of the experiment computing systems would be a key challenge. Each component of the system must be designed with attention to both scalability and flexibility, and rigorously tested at realistic scale. The reliance on distributed computing, using the relatively new Grid approach, has many advantages, but adds further complexity in controlled deployment and testing compared to a system located primarily at a single site.

The relatively large cost of the computing system dictates that centres must build up their resources in a carefully controlled way; the rapidly falling price of hardware dictates that full-scale resources will only become available shortly before they are required, and that efficient use of resources is a strong requirement. The emphasis in CMS has been on a series of increasing scale full-system tests ("data challenges") over the last three years, exercising all available components in a realistic way.

In 2006 and 2007, CMS carried out large-scale Computing, Software and Analysis challenges (CSA06, CSA07). The scale of the two tests was set at 25% and 50% of the nominal 2008 performance, respectively, with the computing system operated continuously at this level for more than four weeks. The challenges were carried out using realistic application software and computing tools. Typical targets for the tests were:

- Preparation of large Monte Carlo datasets ($\approx$ 100 million events) at around twenty CMS Tier-1 and Tier-2 centres in the weeks preceding the challenge, and upload to CERN;
- Playback of the MC dataset for prompt Tier-0 reconstruction at around 100 Hz, including the application of calibration constants from the offline database, and splitting the event sample into around ten datasets;

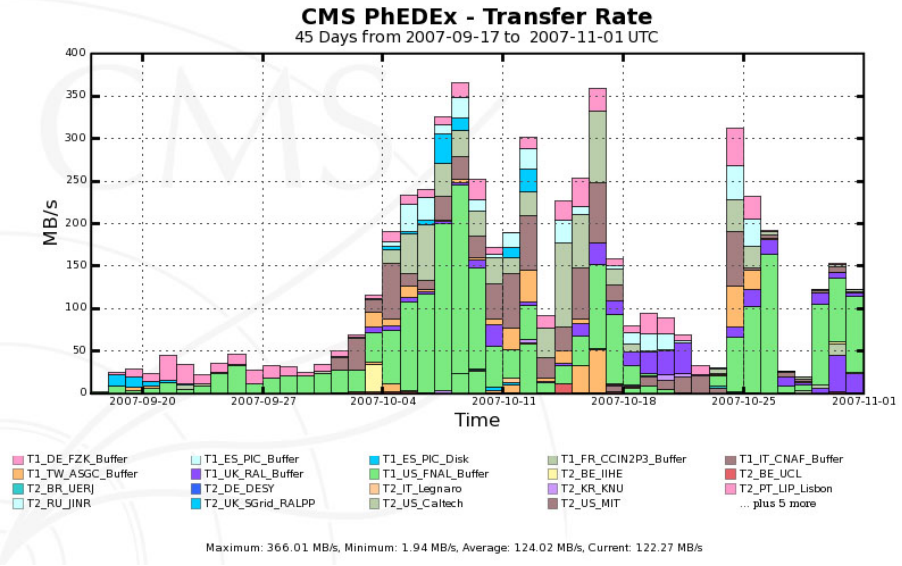

Figure 11.4: Dataflow from CERN during the CSA07 Data Challenge.

- Distribution of AOD and RAW to all Tier-1 centres, and subsequent alignment / calibration, reconstruction and skimming operations at several sites;
- Transfer of skimmed data to Tier-2 centres and running of physics analysis jobs.

Overall, many of the key metrics for success in the challenges were met: the reconstruction rate at the Tier-0 exceeded 100 Hz for periods of time; an export rate of over 350 MB/s was achieved from CERN (figure 11.4). CMS will finalise its data challenge programme with additional scale tests during 2008, which are in the final stages of preparation at the time of writing. In parallel with data challenges, continuous programmes are under way to deploy, commission and test the increasing hardware resources at the computing centres, and to debug and demonstrate reliable and high-speed data network links between them. The CMS computing model itself is also under ungoing review, with many new lessons expected to be learnt as detector data begins to flow.

2008 JINST 3 S08004