Dissertation ETH NO. 19429

CMS Grid Computing and Monitoring of CMS Data Quality Using Selected Data Samples

A dissertation submitted to
ETH ZURICH
for the degree of
DOCTOR OF SCIENCES

presented by

Zhiling Chen
M.Sc. Graduate School of the Chinese Academy of Sciences
born on July 31th, 1978
citizen of China

accepted on the recommendation of
Prof. Dr. Felicitas Pauss (examiner),
Prof. Dr. Günther Dissertori (co-examiner)

- 2011 -

Abstract

The Large Hadron Collider (LHC) at CERN, the European Center for Particle Physics in Geneva, is a proton-proton collider, designed to operate at a center-of-mass energy of 14 TeV at a nominal luminosity of 10^{34} cm^{-2}s^{-1}. The main physics goals of the LHC experiments are the search for the Standard Model Higgs boson and new physics phenomena beyond the Standard Model. The design and construction of the Compact Muon Solenoid (CMS) experiment at the LHC had to meet unprecedented challenges both for the detector operation as well as for the data handling. Due to the high event rate and large event size, the LHC experiments generate large amount of data of about 15 petabyte (10^{15} bytes) per year at the design luminosity, which thousands of scientists at hundreds of research institutes and universities around the world access and analyse. In addition, detailed Monte Carlo simulations of various physics processes also require large-scale computing power and huge amount of mass storage. To meet these requirements, a novel globally distributed model for data storage and CPU power was chosen: the Worldwide LHC Computing Grid (WLCG). The WLCG collaboration in Switzerland provides computing infrastructure and resources to physicists from Swiss institutions involved in the LHC experiments as well as to the experimental collaborations, by operating a high-performance Tier-2 center at CSCS in Manno and the Swiss CMS Tier-3 center at the Paul Scherrer Institute (PSI) in Villigen.

This thesis reports on my work for the Swiss Tier-2 and the CMS Tier-3 centers. The two facilities passed several benchmarks, were upgraded continuously over the past years and show excellent operation performance since the start-up of the LHC on 30 March 2010, providing proton-proton collisions at a center-of-mass energy of 7 TeV.

In the second part of this thesis studies of the detector performance and data quality monitoring are described, which are key issues for the physics output, especially at the start of data taking. A data sample of $Z \to e^+e^-$ recorded up to September 2010 is selected to study the performance of the CMS detector and to monitor the data quality. Electrons and positrons are reconstructed and identified in the electromagnetic calorimeter requiring a matching track in the tracking system. The measured invariant mass distribution obtained from the selected electron-positron pairs show a clear Z mass peak with very little background. This result is in good agreement with the Monte Carlo predictions and illustrates the good data quality at the start of the CMS operation.

Zusammenfassung

Der Large Hadron Collider (LHC) am Europäischen Zentrum für Teilchenphysik CERN in Genf ist ein Beschleuniger der für Proton-Proton-Kollisionen bei einer Schwerpunktsenergie von 14 TeV und einer Design Luminosität von 10^{34} cm^{-2}s^{-1} gebaut wurde. Die wichtigsten wissenschaftlichen Ziele der Experimente am LHC sind die Suche nach dem Standardmodell Higgs Boson und neuen physikalischen Phänomenen jenseits des Standardmodells. Design und Konstruktion des Compact Muon Solenoid (CMS) Experiments am LHC war eine grosse Herausforderung sowohl für den Betrieb des Detektors als auch für die Verarbeitung der aufgezeichneten Daten. Aufgrund der hohen Kollisionsrate und der grossen Menge der pro Kollision anfallenden Daten werden die LHC Experimente noch nie dagewesene Datenvolumen von etwa 15 Petabyte (10^{15} Bytes) pro Jahr erzeugen. Tausende von Wissenschaftlern an mehreren hundert Universitäten weltweit werden auf diese Daten zugreifen um sie zu analysieren. Zusätzlich benötigt auch die detaillierte Simulation der verschiedenen physikalischen Prozesse immense Rechen- und Speicherkapazitäten. Um diesen Anforderungen gerecht zu werden wurden die für die Analyse der Daten benötigten Rechen- und Speicherkapazitäten auf der ganzen Welt verteilt und als 'Worldwide LHC Computing Grid (WLCG)' vernetzt. Die WLCG-Gruppe der Schweiz stellt den Physikern an Schweizer Instituten sowie den LHC Kollaborationen Infrastruktur und Rechenkapazitäten für die Verarbeitung der LHC Daten zur Verfügung. Ein wichtiger Teil dieser Infrastruktur ist das der Hochleistungs-Tier-2-Cluster am Schweizer Supercomputing-Zentrum (CSCS) in Manno und das Schweizer CMS Tier-3 Zentrum am Paul Scherrer Institut (PSI) in Villigen.

Die vorliegende Doktorarbeit beschreibt meine Arbeit für den Schweizer Tier-2 und den Schweizer CMS Tier-3 Computer-Cluster. Diese zwei Tier Zentren bestanden mehrere Leistungstests und wurden kontinuierlich in den vergangenen Jahren erweitert. Sie bewährten sich bestens seit dem Start von Proton-Proton Kollisionen bei einer Schwerpunktsenergie von 7 TeV am 30. März 2010.

Ein weiterer Teil dieser Doktorarbeit beschreibt die Untersuchungen zur Leistungsfähigkeit des CMS Detektors und Qualität der aufgezeichneten Daten. Solche Untersuchungen sind besonders in den ersten Monaten der Datennahme von zentraler Wichtigkeit. Zu diesem Zweck wurden Ereignisse aus den Daten welche bis zum September 2010 aufgenommen wurden, analysiert, bei denen ein Z-Boson in ein Elektron und ein Positron zerfällt ($Z \to e^+e^-$). Elektronen und Positronen werden durch ein charakteristisches Signal im elektromagnetischen Kalorimeter identifiziert, wobei eine passende Spur in der zentralen Spurkammer gefordert wird. Die gemessene Verteilung der invarianten Masse dieser Ereignisse zeigt ein klares Z Signal mit sehr kleinem Untergrund. Dieses Ergebnis ist in guter Übereinstimmung mit der Vorhersage von Monte Carlo Simulationen und illustriert die sehr gute Qualität der aufgezeichneten Daten.

Contents

1 Introduction

The Standard Model (SM) of particle physics provides a remarkably accurate picture of the most fundamental structure of matter. The predictions of the SM are in good agreement with the experimental results. However, the origin of the masses of particles is still a major open question. In the SM, the electroweak symmetry breaking is introduced to solve the problem, predicting the Higgs particle, which is not yet observed. The search for the Higgs particle and the related symmetry breaking mechanism are one of the great challenges in particle physics.

The SM answered many questions about particle physics, but many fundamental questions are still open: the number of generations, the mass hierarchy, the C P violation, the mixing of quarks, the mixing of neutrinos, the unification of interaction and so on. The latest observations of the cosmic microwave background indicate that the matter described in the SM framework is only about 4% of the matter in the Universe and 23% are linked to dark matter and 73% to the dark energy [1]. Dark matter and dark energy cannot be explained within the framework of the SM. There are many theoretical models going beyond the SM, with Supersymmetry being one of the most promising theories beyond the SM [2].

The Large Hadron Collider (LHC) [3] is a proton-proton collider which has been designed to operate at a center-of-mass energy ($\sqrt{s}$) of 14 TeV and a nominal luminosity of 10^{34} cm^{-2}s^{-1}. The main physics goals of the LHC experiments are the search for the Higgs boson of the Standard Model and new physics phenomena beyond the Standard Model, such as the Supersymmetry, large extra dimensions, composite models, etc.

At the design luminosity of LHC a mean of about 20 inelastic collisions will be superimposed on the event of interest. This implies that about 1000 charged particles will emerge from the interaction region every 25 ns. The effect of this pile-up must be reduced by using high-granularity detectors with fast time resolution, which requires a huge number of detector channels. The LHC experiments have to meet great challenges both for the detector operation and data handing.

As one of the two general purpose detectors at the LHC, the Compact Muon Solenoid (CMS) experiment is designed to meet the requirements and is optimized for the precision measurements of muons, photons and electrons, as well as good measurements of hadron jets. The CMS detector is 24 m in length and 14.6 m in diameter with the total weight of 12500 tons. The CMS

detector consists of a 4 T superconducting solenoid and the four sub-detectors: Silicon tracker, electromagnetic calorimeter, hadron calorimeter and muon chambers. The commissioning of the CMS detector went smoothly. In November 2009 the CMS detector started data taking by recording the first *pp* collisions provided by LHC.

Because of the high event rate and large event size, the LHC experiments at the design luminosity of 10^{34} $\mathrm{cm}^{-2}\mathrm{s}^{-1}$, generate unprecedented about 15 PB of data annually, which thousands of scientists in hundreds of research institutes and universities around the world access and analyse. In addition, detailed Monte Carlo simulations of the physics processes and the detector responses also require large-scale computing power and huge amount of mass storage. However, it is impractical to build one computing center nearby LHC, which can fulfill such an enormous amount of the data processing and storage. Thus, a novel globally distributed model for data storage and analysis – a computing Grid: Worldwide LHC Computing Grid (WLCG) – was chosen because it provides several key benefits [4].

The WLCG collaboration in Switzerland provides computing infrastructures and resources to the LHC physicists from Swiss institutions as well as to the whole LHC Collaborations under the agreement of WLCG MoUs [5], including a high-performance Tier-2 center at the Swiss National Supercomputing Center (CSCS) in Manno and the Swiss CMS Tier-3 center at PSI. The CSCS Tier-2 center provides a part of its resources to the CMS, ATLAS and ALICE experiments for official production and analysis jobs. The CMS Tier-3 center is completely dedicated to the Swiss CMS groups. A large fraction of analysis jobs from Swiss CMS groups is carried out at the Tier-3 centers and the hardware schema and software configuration of the Tier-3 is optimized for the end-user analysis and completely dedicated to the Swiss CMS groups.

This thesis reports on my contributions to the Swiss CMS Tier-2 at CSCS and the Swiss CMS Tier-3 at PSI. I worked on the configuration and commissioning of the Tier-2 at CSCS, including the setup and configuration of the CMS Grid Computing environment and the CMS experimental software environment, the LHC and CMS data challenge and commissioning. I worked on (made important contributions to) the setup and commissioning of the Swiss CMS Tier-3 at PSI, including the design and setup of the Tier-3 Cluster, the configuration of the CMS Grid Computing environment and the CMS experimental software environment, as well as to the commissioning of the Swiss Tier-3.

The studies on the detector performance and the data quality monitoring are key issues for the success of the physics research of an experiment, especially also at the beginning of the data taking. The reaction $Z \to e^+e^-$ is one of the best processes to study the performance of the detector (electromagnetic calorimeters and the tracking system) and to monitor the data quality. The decay of Z bosons into electron pair provides a clean experimental signature for the event

selection. The properties of Z bosons are very well studied by the LEP experiments. At the LHC, the Z production cross-section is very large, i.e., ~ 1 nb at $\sqrt{s} = 7$ TeV. Thus one could obtain a large data sample of $Z \to e^+e^-$ to study the detector performance with the very early data at the LHC and to monitor the data quality of CMS. The reconstruction of $Z \to e^+e^-$ events relies only on the tracker and the electromagnetic calorimeter. An electron is reconstructed and identified from a 'super cluster' in the electromagnetic calorimeter and a matching track in the tracking system. The $Z \to e^+e^-$ events are selected from two oppositely charged isolated electrons.

The thesis is organized as follows. Chapter 1 provides an overall introduction. In Chapter 2, CERN and the Large Hadron Collider are briefly introduced. An overview of the SM and the physics beyond the SM, as well as the experiments at the LHC are discussed. The requirements for the detector design and the data analysis are summarized. Chapter 3 presents the design, construction and commissioning of the CMS detector. Chapter 4 introduces Grid computing, which is the only computing scenario to meet the great challenges for the data handing and processing for the LHC experiments. The WLCG project is discussed in detail. The configuration and commissioning of Swiss CMS Tier-2 at CSCS and the Swiss CMS Tier-3 at PSI are discussed in Chapter 5. The physics preparation and the data quality monitoring of $Z \to e^+e^-$, including the reconstruction, event selection and the backgrounds, are discussed in details in Chapter 6. The results from the first period of data taking are presented. Finally, Chapter 7 summarizes the results obtained.

2 Physics Goals at the Large Hadron Collider

The Standard Model (SM) of particle physics describes all known particles and their interactions, except the recent neutrino oscillation results. However, the origin of the masses of the particles is still the major open question. In the SM the electroweak symmetry breaking is introduced to solve this problem and a Higgs particle is predicted, which is not yet observed. The search for the Higgs particle and the related symmetry breaking mechanism is one of the great challenges for particle physics.

The Large Hadron Collider (LHC) at CERN is one of the largest scientific facilities ever built. It will help physicists to answer fundamental questions in particle physics, including the search for Higgs particles. The unprecedented energy at LHC may even reveal some new phenomena beyond the SM.

At the beginning of this chapter, CERN and the Large Hadron Collider are briefly introduced, including the design and the parameters of the LHC accelerator and experiments. This chapter mainly focuses on the physics goals of the LHC and the requirements for the experiments. An overview of the SM, as well as the experimental studies of the SM and physics beyond the SM at the LHC are given in details in Section 2.3. The requirements for the detector design and the data analysis are also summarized. The startup and commissioning of LHC are presented in Section 2.4.

2.1 CERN

CERN, the European Organization for Nuclear Research, is one of the world's largest and most respected centers for scientific research. Its mission is the fundamental research in particle physics, finding out what the Universe is made of and how it works. At CERN, the world's largest and most complex scientific instruments are used to study the basic constituents of matter. By studying what happens when very high-energy particles collide, physicists learn about the laws of Nature.

The instruments used at CERN are particle accelerators and detectors. Accelerators boost beams of particles to high energies before they are made to collide with each other or with stationary targets. Detectors observe and record the particles produced in these collisions.

Founded in 1954, the CERN Laboratory is located at the Franco-Swiss border near Geneva. It was one of Europe's first joint ventures and now has 20 member states.

Numerous experiments have been constructed at CERN by international collaborations for particle physics during its 56 years of history. CERN made important contributions to the development of particle physics, including:

1. The discovery of the neutral currents in 1973 by the Gargamelle bubble chamber experiment [6].
2. The discovery of the W and Z bosons in 1983 by the UA1 and UA2 experiments [7–10].

These discoveries are the historic milestones for the tests of the SM of particle physics. In the 1990s, precision measurements were performed using the Large Electron-Positron Collider (LEP), including the measurement of the number of the neutrino families. The physics results obtained with the LEP experiments tested the SM predictions with high precision.

Moreover, many remarkable technical accomplishments were also made at CERN during the construction of the Intersecting Storage Rings (ISR) commissioned in 1971 and the Super Proton Synchrotron (SPS), which came into operation in 1976 and was converted into the $p\bar{p}$ collider. The massive W and Z particles were produced at the $p\bar{p}$ collider for the first time, confirming the unified theory of the electromagnetic and weak forces – the electroweak theory.

To analyse an enormous amount of data from the experiments, CERN also has a large computer center with powerful data processing and mass storage facilities, primarily devoted to experimental data analysis. In the domain of computing and network, the revolutionary achievement at CERN was the invention of the World Wide Web ('WWW' or simply the 'Web') in the late 80s by Tim Berners-Lee and others [11]. On 30 April, 1993, CERN announced that the World Wide Web would be a free tool [12], available to everybody. WWW made a large impact on the information technology and changed the life of ordinary people. A recent study reported that there are worldwide at least 19.41 billion indexed Web pages on the Web as of January 2010 [13].

Most of the activities at CERN are currently directed towards the full exploitation of the physics potential offered by the LHC. At the full operation intensity, the LHC will produce roughly 15 PB[1] of data annually. To meet the great challenges for the LHC data analysis, a crucial active computing project, the Worldwide LHC Computing Grid (WLCG), was established [4]. It is a global collaboration linking grid infrastructures and computer centers worldwide, which

[1] A petabyte (PB) is 1000 terabytes, or 1×10^{15} *bytes*.

distribute, store and analyse the immense amount of data recorded by the LHC experiments. The mission of the WLCG is to build and to maintain a data storage and analysis infrastructure for the entire high energy physics community of LHC. At present, the WLCG combines the computing resources of more than 100,000 processors from over 170 sites in 34 countries, producing a massive distributed computing infrastructure that provides more than 8,000 physicists around the world with near real-time access to LHC data and the power to process it. The WLCG concept is discussed in depth in Chapter 4.

2.2 The Large Hadron Collider

The Large Hadron Collider (LHC) is the world's largest and highest-energy particle accelerator, colliding opposing particle beams of protons with energy up to 7 TeV, or lead nuclei with energy up to 2.76 TeV per nucleus. The LHC accelerator and its experiments provide a long-awaited and unprecedented tool for fundamental physics research for many years to come. The Large Hadron Collider was built to test various predictions of high-energy physics, including the existence of the hypothesized Higgs boson and new particles predicted by supersymmetry and other theories beyond the SM. The design luminosity for pp collision is 10^{34} cm^{-2}s^{-1} at the beam energy of 7 TeV, but at least until the end of 2011 LHC will run at lower luminosities and lower energies. The construction, commissioning and operation of LHC make CERN to be the leading research center of particle physics in the world.

2.2.1 Design of the LHC

The design of LHC was strongly influenced by the cost saving to be made by re-using the LEP tunnel and its injection chain. In 1989, CERN started the LEP operation, the world's highest energy electron-positron collider. In November 2000, LEP data taking was terminated to liberate the tunnel for the LHC construction. Some advantages were obtained as the tunnel and several infrastructures, including injectors, already exist. However, the radius of LHC limits the centre-of-mass energy to 14 TeV, since the beams must be bent by superconducting dipole magnets whose maximum reliable field is currently limited to about 8 T.

The choice of the proton beam provides the following advantages:

- Hadrons allow exploration of physics in a wide energy range with fixed-energy beams: they are the natural choice for a discovery machine. Protons are not elementary particles, and in hard collisions the interactions involve their constituents (quarks and gluons), which carry a non-fixed fraction of the proton energy.

- Protons allow the accelerator to reach higher luminosity with respect to anti-protons, as their production and storage are easier.
- In a circular collider of radius R, the energy loss per turn due to synchrotron radiation is proportional to $(E/m)^4/R$, where E and m are respectively the energy and mass of the particles accelerated. Therefore, due to their higher mass, using of protons implies a much smaller energy loss for synchrotron radiation with respect to electrons.

The LHC has to provide a very high luminosity to compensate for the low cross section of the interesting physics processes. The cross section of different processes as functions of the centre-of-mass energies for pp collisions is shown in Figure 2.1. The Higgs cross section increases steeply with the centre-of-mass energy, while the total cross section (i.e., the background) remains almost constant. The highest possible centre-of-mass energy of LHC should provide the best ratio of signal to noise for the Higgs search.

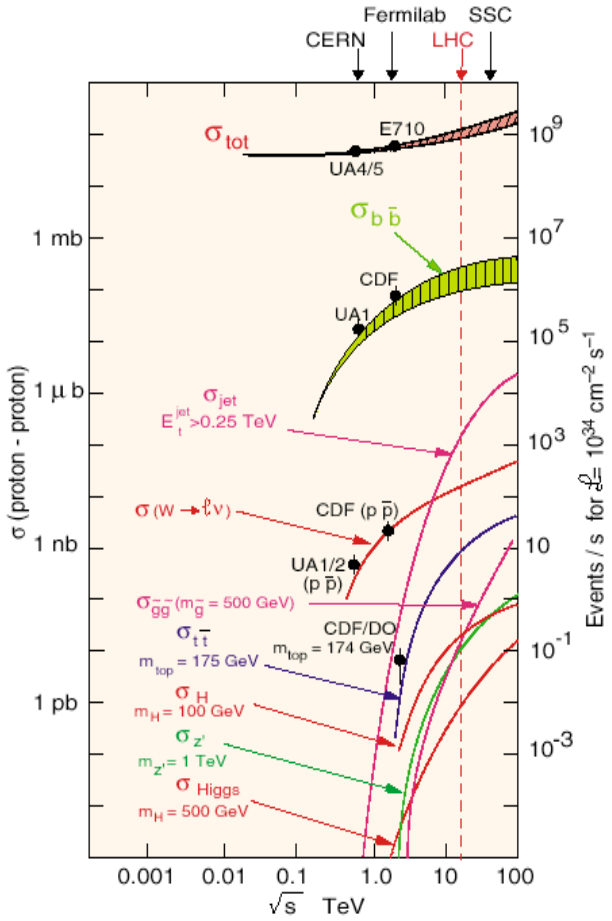

Figure 2.1: The cross sections and the event rates at the design luminosity for the hard scattering processes as a function of the centre-of-mass energy $\sqrt{s}$ [14].

2.2.2 LHC Accelerator

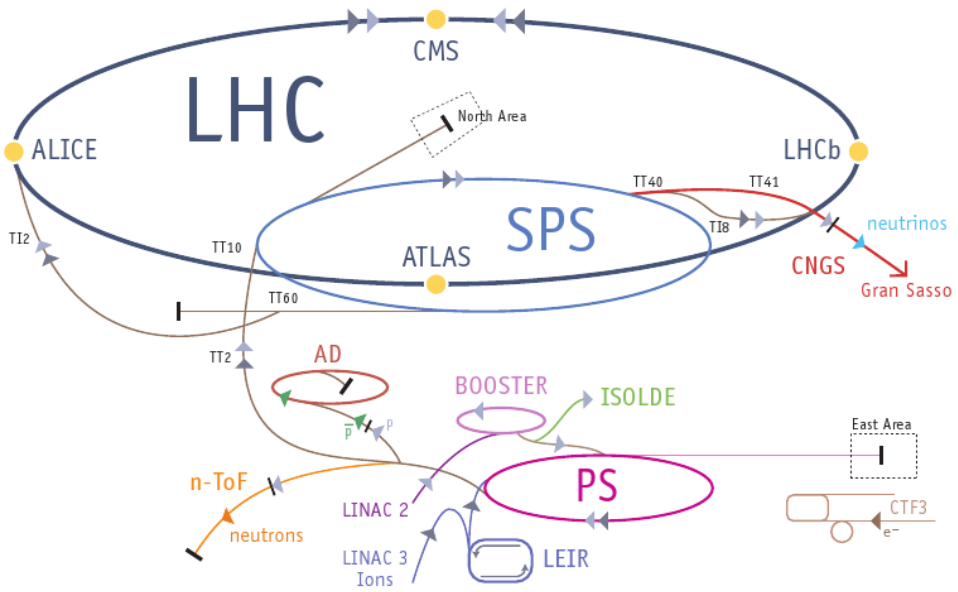

Figure 2.2: The Large Hadron Collider and its preceding accelerators. Protons are initially accelerated in Linac 2 and the Proton Synchrotron Booster before being injected into the Proton Synchotron (PS). The PS accelerates the protons to an energy of 25 GeV and subsequently injects these into the Super Proton Synchotron (SPS) which accelerates them to 450 GeV. The protons are then injected into the LHC and form the two counter-rotating beams. These are then accelerated to full energy before collisions are established by crossing the beams.

The Large Hadron Collider, as illustrated in Figure 2.2, consists of a ring with 26.7 km circumference with separate magnetic fields and vacuum pipes, but sharing the cryogenic structure (the so-called '2-in-1' magnet solution). The particle beam reaches its final energy via the four stages: Protons (heavy ions) are initially accelerated in Linac 2 and the Proton Synchrotron Booster (LEIR and Linac 3) before being injected into the Proton Synchrotron (PS); 25 GeV particles bunch are then formed in the Proton Synchrotron (PS); the Super Proton Synchrotron (SPS) pre-accelerates the beam to 450 GeV and injects it into the LHC counter clockwise, where they will reach their nominal energy of 7 TeV (for the proton beam).

The LHC itself comprises 1232 dipole magnets, with radio frequency cavities to increase the proton energy by ∼0.5 MeV/turn. Accelerating the beam from 450 GeV to 7 TeV takes about 20 minutes. The beams are injected in bunches separated in time by 25 ns (or multiple of 25 ns). To achieve collision conditions, each beam is focused by a complex array of magnets before they cross the interaction point. At collision the bunch length is $\sim 8\ cm$ with a diameter of $\sim 16\ \mu m$. The machine parameters are listed in Table 2.1. The luminosity is given by:

Parameter		pp	heavy-ion
Energy per nucleon (TeV)	E	7	2.76
Circumference (km)		26.7	26.7
Revolution frequency (Hz)	f_{rev}	11245	11245
Dipole field at 7 TeV (T)	B	8.33	8.33
Design Luminosity ($\mathrm{cm}^{-2}\mathrm{s}^{-1}$)	$\mathcal{L}$	10^{34}	10^{27}
Bunch time separation (ns)		25	100
No. of bunches	k_B	2808	592
No. particles per bunch	N_p	1.15×10^{11}	7.0×10^{7}
β -value at IP (m)	β^*	0.55	0.5
RMS beam radius at IP (μm)	σ^*	16.7	15.9
Luminosity lifetime (h)	τ_L	15	6

Table 2.1: The LHC machine design parameters [3]. For heavy-ion operation the design luminosity for Pb-Pb collisions is given.

$$\mathcal{L} = \frac{\gamma f_{rev} k_B N_p^2}{4\pi\varepsilon_n \beta^*} F, \tag{2.2.1}$$

where $\gamma = 1/\sqrt{1-\beta^2}$, and $\beta = v/c$ is the velocity in terms of the speed of light; f_{rev} is the revolution frequency; k_B is the number of bunches; N_p is the number of protons per bunch; ε_n is the normalized transverse emittance (with a design value of 3.75 μm); β^* is the betatron function at the interaction point, and F is the reduction factor due to the crossing angle. The event rate R of a given process with cross section σ is given by:

$$R = \mathcal{L}\sigma \tag{2.2.2}$$

The inelastic proton-proton cross-section is ∼70 mb [2] [15]. The collision of two proton bunches with nominal parameters consequently causes approximately 20 inelastic events at the design luminosity, as can be seen using equations 2.2.1 and 2.2.2:

$$\begin{aligned} N &= \frac{\mathcal{L}}{k_b f_{rev}}\sigma \\ &= \frac{1\times 10^{34}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}}{2808 \times 11245\ s^{-1}}\ 7\times 10^{-26}\ \mathrm{cm}^2 \\ &\approx 20 \end{aligned} \tag{2.2.3}$$

Where N is the number of pile-up events caused by the collision of two proton bunches. Most

[2] 1 mb = 10^{-27} cm^2

of these are minimum bias events, acting to obscure interesting interactions which have a much lower cross-section.

2.2.3 LHC Experiments

The proton-proton inelastic cross-section at $\sqrt{s} = 14$ TeV is roughly 70 mb. At design luminosity of $10^{34}\ cm^{-2}s^{-1}$, the general-purpose detectors, ATLAS and CMS, will therefore observe an event rate of 70×10^8 inelastic events/s. This leads to a number of formidable experimental challenges [15].

The event selection process 'trigger' must reduce the $\sim$ billion interactions/s to no more than $\sim 10^2$ events/s, for storage and subsequent analysis. The short time between bunch crossings, 25 ns, has major implications for the design of the readout and trigger systems. It is not feasible to make a trigger decision in 25 ns. Therefore, a chain of pipelined trigger processing and readout architectures are implemented.

At the design luminosity a mean of $\sim$20 minimum-bias events will be superimposed on the event of interest. This implies that around 1000 charged particles will emerge from the interaction region every 25 ns. The products of an interaction under study may be confused with those from other interactions in the same bunch crossing. This problem, known as pileup, clearly becomes more severe when the response time of a detector element and its electronic signal is longer than 25 ns. The effect of pileup can be reduced by using highly granular detectors with good time resolution, giving low occupancy at the expense of having large numbers of detector channels. The resulting millions of detector electronic channels require very good synchronization.

The particles coming from the interaction region lead to a high radiation level, requiring radiation-hard detectors and front-end electronics. Access for maintenance will be very difficult, time consuming and highly restricted. Hence, a high degree of long-term operational reliability, which is usually associated with spaceborne systems, has to be attained.

The online trigger system has to analyse information that is continuously generated at a rate of $40,000$GB/s and reduce it to hundreds of MB/s for storage. The many PB of data that is generated per year per experiment has to be distributed for offline analysis to scientists located across the globe. This data management and analysis requirements motivated the development of the LHC Computing Grid.

Four detectors have been installed at the LHC, seen in Figure 2.3. Two of them, **A** **T**oroidal **L**HC **A**pparatu**S** (**ATLAS**) experiment and the **C**ompact **M**uon **S**olenoid (**CMS**) experiment, are large general purpose detectors. The other two are devoted to specific topics: **A** **L**arge **I**on **C**ollider **E**xperiment (**ALICE**) to heavy ions and the **L**arge **H**adron **C**ollider beauty (**LHC**b)

experiment to b-physics.

Figure 2.3: The LHC Experiments.

2.3 Physics Goals at the LHC

The SM is the best theory so far to describe the interactions between the most fundamental building blocks of matter, quarks and leptons [16–18]. The SM of particle physics has been successfully tested with very high precisions at many experiments over a wide energy range. The latest tests were carried out by the experiments at the LEP and the Tevatron colliders. The LHC will allow to precisely measure the SM parameters in view of:

- Having a better understanding of the SM processes, performing the precision electroweak measurements and the Top physics studies. Especially, the precise measurements of the W boson mass, the Top mass and the Weinberg angle can constrain the prediction of the Higgs boson mass. If the Higgs was observed, this relation would be a very important test of the SM.
- Searching for direct and indirect deviations from the SM in experimental spectra to be observe signatures for physics beyond the SM.

However, one of the main missions of the LHC is the study of the origin of the electroweak symmetry breaking mechanism. Therefore the search for the Higgs particle is a major physics goal for the experiments.

2.3.1 Success of the Standard Model

Since the middle of the last century, enormous theoretical and experimental work have been carried out in order to find out the ultimate constituents of the elementary particles and to establish an accurate description of their interactions. The SM proposed by Glashow [16], Weinburg [17] and Salam [18] is based on the gauge group $SU(3) \times SU(2) \times U(1)$. The SM classifies all known particles properly, and describes the electromagnetic interaction, weak interaction and the strong interaction correctly. Particularly the SM unifies the electromagnetic interaction and the weak interaction into the so-called electroweak interaction. All particle physics experimental data, except for the recent experimental results of the neutrino mixing, are in excellent agreement with the prediction of the SM with very high accuracies, some of them as precise as $\sim 10^{-5}$. The SM asserts the smallest blocks of the matter in the Universe which are a limited set of fundamental spin $\frac{1}{2}$ particles, or *fermions*: six *quarks* and six *leptons* interacting through fields. The particles associated with the three interactions are spin-1 particles called gauge *bosons*. Gravity is the only interaction, which is not described by the SM.

Fundamental Fermions

The matter particles (fermions) consist of six types (or flavours) of leptons and six types (or flavours) of quarks.

The charged leptons ($e^{\pm}, \mu^{\pm}, \tau^{\pm}$) carry one unit of electric charge. The muon (μ) and the tau (τ) leptons are heavy 'versions' of the electron. The neutral leptons are called *neutrinos*, denoted by the generic symbol ν. The different 'flavours' of neutrinos are paired with the corresponding 'flavour' of the charged lepton, as indicated by the subscript, i.e., ν_e, ν_μ, ν_τ. The neutrinos are massless particles in the SM. The neutrino oscillation experiment results show that neutrinos do have a non-zero masses [19–22].

There are six types of quarks, known as flavours: up (u), down (d), charm (c), strange (s), bottom (b) and top (t). Quarks have various intrinsic properties, including electric charge, color charge, spin and mass. The quarks carry fractional electric charges, of $+\frac{2}{3}|e|$ or $-\frac{1}{3}|e|$. And, just as for the leptons, the quarks are grouped into pairs differing by one unit of electric charge. Since the electric charge of a hadron is the sum of the charges of the constituent quarks, all hadrons have integer charges: proton and neutron are the combination of three quarks, while

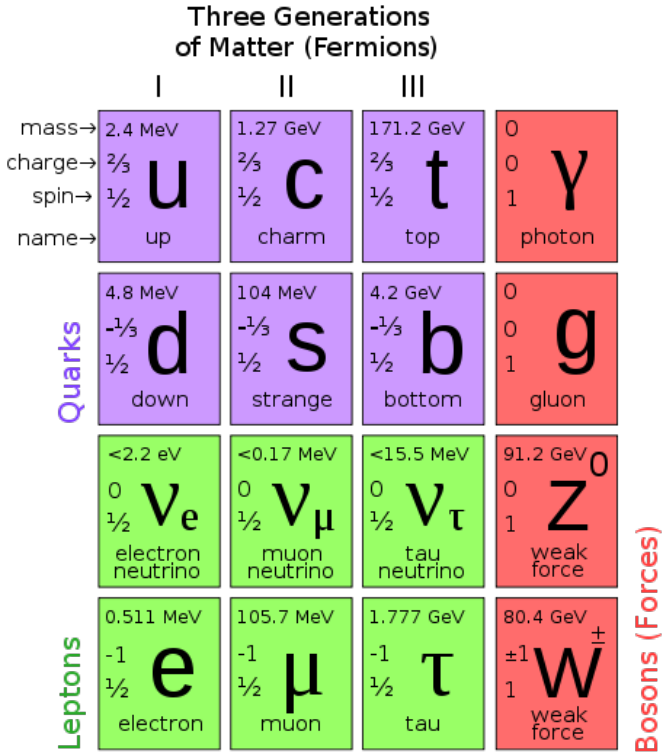

Figure 2.4: The SM of elementary particles, with the fermions in the first three columns [23].

their antiparticles are the combination of three antiquarks. Mesons are made of a quark-antiquark pair, always resulting in integer electric charges.

While leptons exist as free particles, quarks are confined in hadrons. Neutrons and protons are bound states of the lightest u and d quarks, three at a time: a neutron consists of ddu, a proton consists of uud. The heavier quarks s,c,b,t and their antiparticles also combine to form various hadrons but they are unstable and decay rapidly to protons, neutrons and other particles. They can only be produced in high energy collisions (such as those involving cosmic rays and particle accelerators) and the high energy astrophysics reactions.

All these elementary fermions can be grouped into three families or generations. Each family contains two leptons and two quarks as listed in Figure 2.4. They participate in the electroweak interactions, and quarks participate in the strong interactions. All stable matter is built from the first generation of fermions (u, d, e).

Besides the electric charge, quarks have another kind of charge named the colour charge. This is relevant for the strong interaction between quarks. The theory of the strong interaction, Quantum Chromodynamics (QCD), assumes each flavour of quark comes in three different colours: red, green and blue.

Interactions

The SM describes the interactions between particles. Each interaction is mediated by a particle with integer values of spin, known as gauge bosons. There are four types of fundamental interactions or fields:

- The *electromagnetic* interaction is the interaction between charged particles, in particular responsible for the bound states of electrons with nuclei, i.e., atoms and molecules, and for the intermolecular forces in liquids and solids. The theory of the electromagnetic interaction is the Quantum Electrodynamics (QED). This interaction is mediated by photon (γ) exchange. Since the photon is massless, the range of the electromagnetic interaction is infinite. QED describes the electromagnetic interaction with very high accuracy is consistent with all experimental results.
- The *strong* interaction is binding quarks together to form protons, neutrons and other particles. It is also the force that binds protons and neutrons together to form nuclei. The theory of the strong interaction is the Quantum Chromodynamics (QCD). Quarks with different colour charge attract one another as a result of the strong interactions, which is mediated by a massless particle called *gluon* (g). Because gluons are massless, they might be expected to have infinite range. However gluons, unlike photons, carry a colour-charge and interact among each other. This leads to a phenomenon called confinement which restricts the strong force to nuclear distances (R $\sim 10^{-13}$ cm). Quarks are confined within composite particles (hadrons). These composite particles also contain many gluons, which interact with the quarks and with each other. QCD introduces a new type of quantum number called colour. The colour is analogous to the electromagnetic charge but with three facets: red, green or blue and antiquarks are anti-red, anti-green or anti-blue. The gluons can be thought of as carrying these characteristics by being a mixture of colour and anticolour, which enables them to exchange colour between two quarks. This is noticeably different to electromagnetism where the photons do not themselves carry electric charge. Only colourless particles are allowed to exist freely in Nature.
- The *weak* interaction is typified by the slow process of nuclear β-decay, involving the emission by a radioactive nucleus of an electron and neutrino. It is the only interaction capable of changing flavour of particles and the only force affecting neutrinos (except for gravitation which is negligible on laboratory scales). The mediators of the weak interactions are the charged W^+ and W^- bosons and the neutral Z boson, which were discovered by UA1 and UA2 experiments at the $p\bar{p}$ collider of CERN earlier 1980s [7–10]. They are massive ($\sim$ 100 GeV) and therefore the weak interaction is short ranged (R $\sim 10^{-17}$ cm).

The electromagnetic and the weak force are treated theoretically as two aspects of one electroweak force in the SM.

- The *Gravitational* interaction acts between all types of particles with mass. On the scale of particle physics experiments, gravity is by far the weakest of all the fundamental interactions, although of course it is dominant on the macroscopic scale of the Universe. The gravitational force has been known about for longer than the other three forces. The general theory of relativity correctly describes the large scale gravitational effects. However, there is as yet no well developed quantum theory of gravity. The gravitational radiation is expected to exist although it has not yet been observed directly. The quantum of the radiation is called the *graviton* with spin of 2.

These four types of interactions have different coupling strengths. To indicate the relative magnitude of them, the comparative strengths of the interactions between two protons are roughly as given in Table 2.2.

strong	electromagnetic	weak	gravity
1	10^{-2}	10^{-7}	10^{-39}

Table 2.2: The strengths of the interactions [24].

All particles in the precise symmetry of $SU(3) \times SU(2) \times U(1)$ of the SM are massless. In the SM, the spontaneous symmetry-breaking mechanism is introduced to produce masses for particles. This mechanism predicts the *Higgs* (H) particle with spin zero to play a key role. The SM incorporates the idea that the massive fundamental particles acquire their mass via an interaction with the Higgs field.

The Higgs boson is the only unknown particle in the SM. The principal motivation for building the LHC is to search for the direct evidence for the existence of the Higgs field.

The Higgs boson is expected to decay with a very short lifetime, and will be detected by measuring its decay products. The decay modes depend on the mass of the Higgs, which is constrained by presently available data as shown in Figure 2.5. The CMS experiment is designed to detect the Higgs via different decay modes. An important one is $H \rightarrow \gamma\gamma$, which is the most promising channel for Higgs search if the Higgs is light which is favored by the current experiment results. The high quality photon detection capability is vital for CMS, contributing to the Higgs boson search and many of the other physics goals of the experiment.

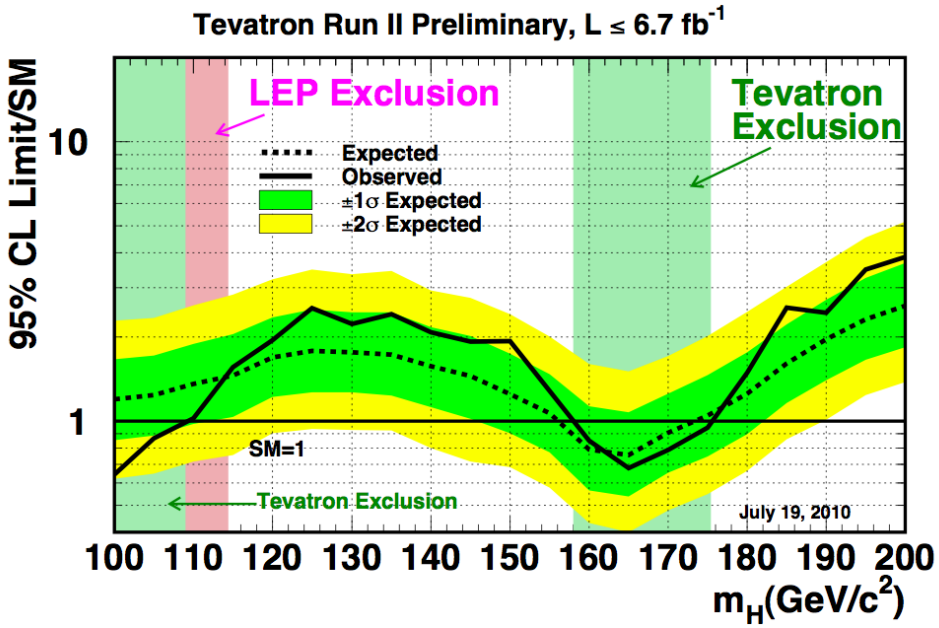

Figure 2.5: Observed and expected exclusion limits for a Standard Model Higgs boson at a 95 percent confidence level for the LEP (LEP Exclusion) and the combined CDF and DØ analysis (Tevatron exclusion). The yellow and green bands indicate a 68 and 95 percent probability regions, in the absence of a signal. The CDF and DØ data exclude a Higgs boson between 158 and 175 GeV/c^2 at a 95 percent confidence level [25].

2.3.2 Beyond the Standard Model

Even though the SM explains most of the basic properties of particles and forces, it is not the ultimate theory of particle physics. The SM answered many questions about particle physics, but many fundamental questions are still open:

- Why are there three and only three generations of matter particles?
- Why does there seem to be much more matter than antimatter in the Universe?
- Why are there huge mass differences between the three generations of quarks and leptons?
- Are quarks and leptons really fundamental?
- Are the fundamental force unified at the Planck scale?
- Many free parameters in the SM are required to be explained, such as the electroweak mixing angle, the masses of the particles and so on.

The latest observation of the cosmic microwave background indicates that the matter described in the SM framework is only about 4% of the matter in the Universe, and about 23% are linked to dark matter and about 73% to dark energy [1]. Dark matter and dark energy can not be

explained within the framework of the SM.

Many observations in astronomy confirmed the gravitation effects on dark matter. Dark matter does not emit or absorb electromagnetic radiation. According to the latest result, only a very small fraction of dark matter could be explained by neutrinos, so called the hot (or relativistic) dark matter. There is no room for cold (or non-relativistic) dark matter particles, which is the majority of dark matter, in the framework of the SM. The weakly interactive massive particle (WIMP) is a speculated particles for cold dark matter. The physics origin of dark matter and the search for dark matter particles are great challenges in particle physics.

There are many theoretical models going beyond the SM. Supersymmetry theory is one of the most promising theories beyond the SM [2]. Supersymmetry assumes a new symmetry between fermions and bosons. The supersymmetric partners of fermions with spin half are the spin zero particles, e.g., selectron (scalar electron), smuon (scalar muon), squark (scalar quarks). All gauge bosons of spin 1 in the SM have the supersymmetric partners with spin half, such as photino, gluoino, wino. The most distinguishing feature of supersymmetry is that in the framework of the theory the strengths of the electroweak interaction and the strong interaction will converge i.e., unify, at the very high energy scale ($\sim 10^{15}$ GeV). Supersymmetry particles are candidates for dark matter particles. In Supersymmetry, there are more than one Higgs bosons (2 charged as well as 3 neutral ones). The search for the lightest Supersymmetry neutral Higgs is quite similar to the search for the SM Higgs. Some supersymmetric particles could be found in the energy range of the LHC.

The models of extra dimensions introduce extra dimensions of space, and could be another possible theory beyond the SM. The existence of the extra dimensions can lead to a characteristic energy scale of quantum gravity, M_D, which is the analogue of the Planck mass in a D-dimensional theory, and which could lie just beyond the electroweak scale. Some interesting phenomena from the decays of the particles predicted by possible scenarios of the extra dimension theories could be observed at the LHC [26, 27].

It is also possible that a scalar Higgs boson does not exist at all. However, if a scalar Higgs particle does not exist within the energy range of the LHC, some new physics phenomena or new particle must be observed around the energy scale of 1 TeV.

2.3.3 Study of the Standard Model at the LHC

The cross sections of the main physical processes of the SM are one or two orders of magnitude larger at the LHC compared to the Tevatron[3]. The LHC will be a factory of W, Z and Top quark particles, hence reducing the statistical error of the measurements significantly. In addition, the large amount of collected data of Z and W will also be used to understand the detector performance and to control the systematic errors.

Precision Measurements of the Electroweak Interaction

W mass measurement The precision measurements of the W boson properties, such as its mass and width, constitute an important consistency check of the SM. The most useful decay channels for the study are leptonic decays of $W \to l\nu$.

An improved precision on the measurement of the W mass (m_W), combined with other electroweak measurements (i.e., m_{top}, $\sin^2\theta_W$), will provide a strong indirect constraint on the mass of the SM Higgs boson (m_H). To ensure that the experimental errors on m_W and m_{top} equally contribute to the uncertainty on m_H, the precision on m_W and m_{top} has to satisfy the following relation [28]:

$$\Delta m_W \sim 0.7 \times 10^{-2} \Delta m_{top} \tag{2.3.1}$$

Since m_{top} is foreseen to be measured at the LHC with an accuracy better than 2 GeV [27], a global precision on the mass of W with about 15 MeV has to be achieved. At the LHC the total cross section for the production of W bosons decaying into a lepton (electron or muon) and the corresponding neutrino is about 30 nb, while the cross section for the process $pp \to Z \to l^+l^-$ (with $l = e, \mu$) is about 1/10 of the corresponding W cross section. Thus, combining results from ATLAS and CMS experiments with an integrated luminosity of 10 fb^{-1}, a precision measurement of the W mass within 15 MeV becomes reachable [28]. The huge data samples available at the LHC experiments will guarantee a nearly negligible statistical error and a good control of the systematic effects.

Top Quark Physics

The Top quark, discovered by the CDF and DØ experiments at the Tevatron in 1995 [29], completes the three family structure of fermions in the SM. The CDF and D0 experiments

[3]Tevatron is a circular proton-antiproton accelerator at the Fermi National Accelerator Laboratory in Batavia, Illinois and is the second highest energy particle collider in the world after LHC.

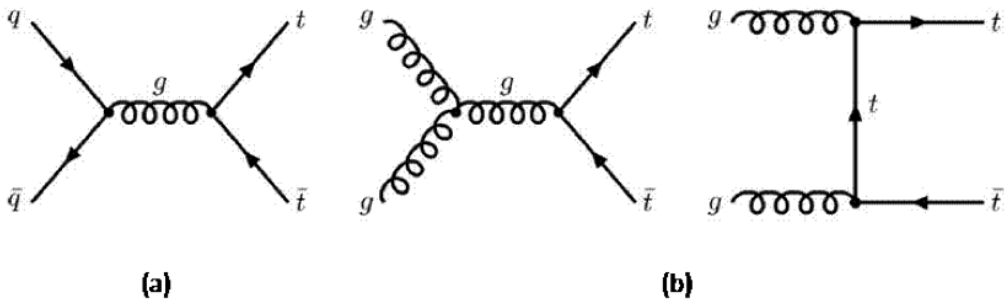

Figure 2.6: Feynman diagrams for $t\bar{t}$ production: (a) quark-antiquark annihilation and (b) gluon-gluon fusion.

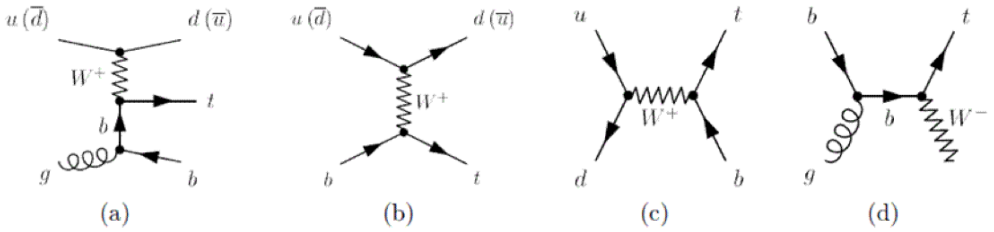

Figure 2.7: Feynman diagrams for three single top quark production channels: (a) and (b) the t-channel, (c) the s-channel, and (d) associated production.

performed many studies on Top physics. However, some properties of Top quarks are still unknown because of the limited Top data samples at the Tevatron. The large amount of Top events produced at LHC will allow to measure most of the Top properties with high precision.

Top quark can be produced by the strong interaction leading to a $t\bar{t}$ pair in the final state as shown in Figure 2.6, with a cross section of about (833 ± 100) pb at $\sqrt{s} = 14$ TeV [30], producing 8 million of Top pairs per per 10 fb^{-1}. Top quarks can also be produced in a single way from the weak interaction as illustrated in Figure 2.7 with a cross section of 330 pb.

The signature of $t\bar{t}$ events is dominated by the W decay leading to two quarks or a lepton ν pair as depicted in Figure 2.8. So, the three possible final states are:

- The fully hadronic channel where both W decay into quarks. Though this channel is dominant, it suffers from a huge background mainly from QCD processes.
- The di-lepton channel where both W decay into lepton ν pair. This channel benefits from very small backgrounds but the two neutrinos make the complete reconstruction of the Top quarks difficult.

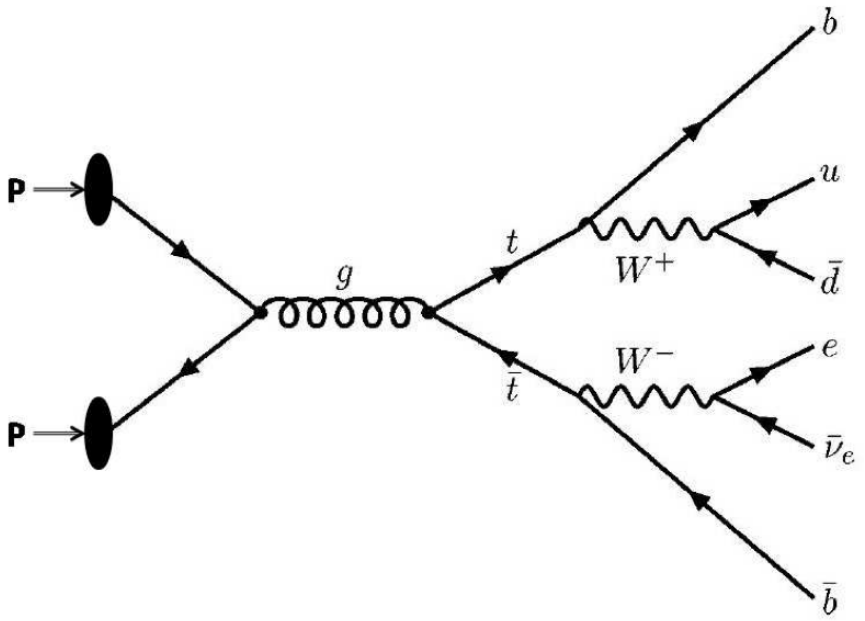

Figure 2.8: Feynman diagram for possible Top quark decays.

- The golden channels is that one where one Top decays into leptons, which is used to trigger the event, and another Top decays into hadrons. This channels benefits from a large event yield and a small background.

For the golden channel, events are selected requiring at least 4 jets with $p_T \geq 40$ GeV/c, one electron (or muon) with $p_T \geq 20(25)$ GeV/c according to the trigger menu and a missing transverse energy above 20 GeV.

2.3.4 Higgs Searches and Measurements at the LHC

The Higgs boson plays a key role in the SM. The detection of this particle was one of the primary design consideration for CMS. According to the SM, the main processes which contribute to the Higgs production at the LHC are expressed in the diagrams of Figure 2.9:

- Gluon-gluon fusion ($gg \rightarrow H$, via a top-quark loop).
- W and Z boson fusion ($qq \rightarrow Hq\bar{q}$).
- $t\bar{t}$ associated production ($gg \rightarrow Ht\bar{t}$).
- W and Z associated production ($q\bar{q} \rightarrow HW$).

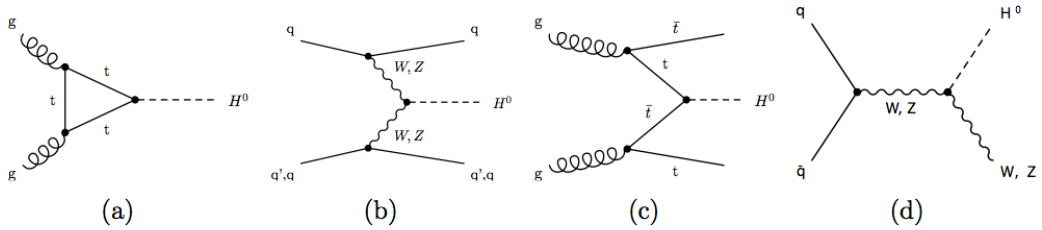

Figure 2.9: Higgs production mechanisms at tree level in proton-proton collisions: (a) Gluon-gluon fusion; (b) W and Z fusion; (c) $t\bar{t}$ associated production; (d) W and Z associated production.

Higgs production cross sections are shown in the left plot of Figure 2.10, illustrating that the gluon-gluon fusion is the dominant process over the whole mass spectrum. The vector boson fusion cross section is about one order of magnitude lower than the gluon-gluon fusion one for a large range of the Higgs mass. This process has a well known next-to-leading-order cross section, small QCD corrections and a very clear experimental signature, due to the presence of the two spectator jets with high invariant mass in the forward region.

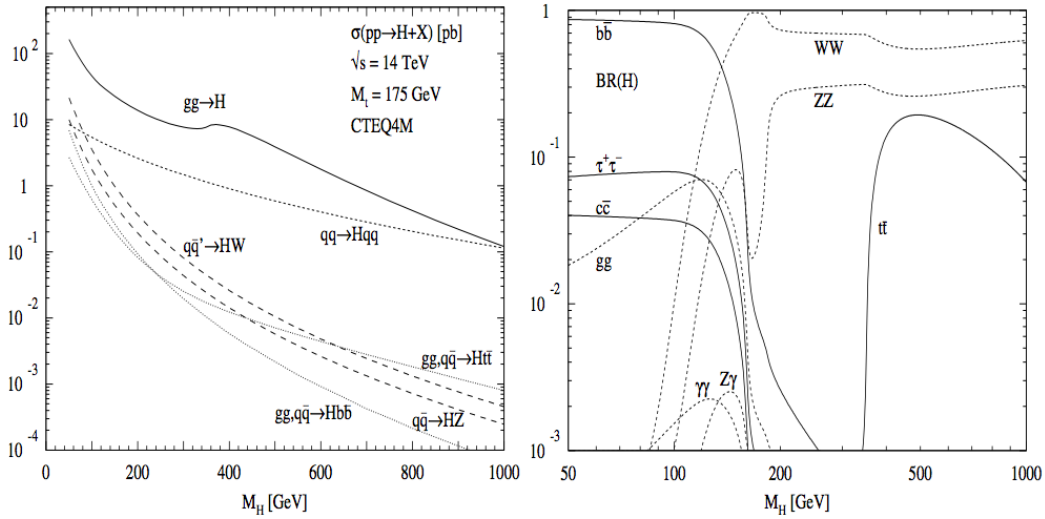

Figure 2.10: The production cross sections of the SM at LHC at $\sqrt{s} = 14$ TeV (left) and branching ratios (right) as a function of the SM Higgs mass [31].

The branching ratios for several Higgs decay channels as a function of the Higgs mass are shown in the right plot of Figure 2.10. They can be interpreted on the basis of the Higgs couplings to fermions and gauge bosons. The coupling is proportional to the fermion masses and to the square of the boson masses; basically three mass regions may be defined:

- Low mass Higgs ($m_H < 130$ GeV/c^2): the heaviest available fermion is the b quark, and $H \to b\bar{b}$ dominates. However, this decay channel could be difficult to be observed at the LHC because of the huge QCD background (except maybe by exploiting associated $t\bar{t}H$ or WH production). In this mass region the most promising channel is $H \to \gamma\gamma$, which despite the very low branching ratio ($\sim 10^{-3}$) has a very clean signature. The signal should appear as a narrow peak over the continuum $(q\bar{q}, gg) \to \gamma\gamma$ background, but excellent photon energy and angular resolution are required as well as good π^0 rejection.
- Intermediate mass Higgs (130GeV/c^2 < m$_H$ $\lesssim$ 500 GeV/c^2): the production of WW and ZZ pairs becomes possible; the branching ratio is high, but purely hadronic final states are again not accessible. $H \to ZZ^* \to 4\ell$ ($\ell = e, \mu$) is the most performant channel for Higgs search; The channel $H \to WW^* \to \ell\nu\ell\nu$ has the disadvantage that experimentally accessible final states have at least one neutrino that escapes detection; however it will be a good discovery channel, especially for $m_H \approx 2m_W$ where the WW production is at threshold and the ZZ branching ratio drops to 20%.
- High mass Higgs (above 500 GeV/c^2): the cross section becomes low and semi-leptonic final states ($2\ell 2j, \ell\nu 2j$) have to be used. The Higgs width becomes also very broad, as shown in Figure 2.11, so that the reconstruction of a mass peak becomes difficult.

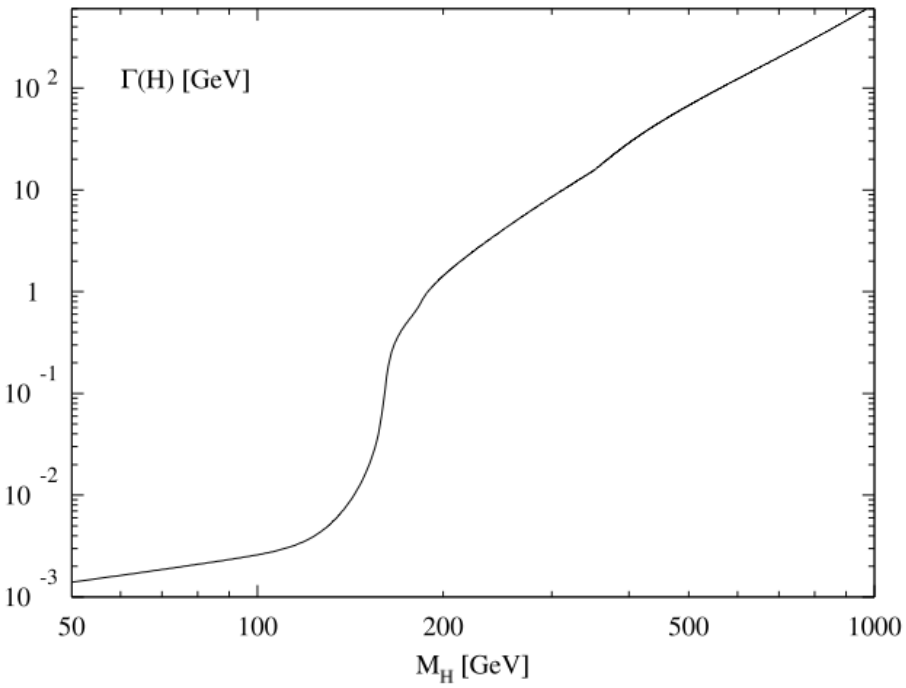

Figure 2.11: Higgs boson decay width as a function of the Higgs mass [31].

2.3.5 Searches for Phenomena Beyond the Standard Model at the LHC

There are many physics models beyond the SM. To search for the new particles predicted by various models were discussed extensively in the literature [27]. Only two scenarios are illustrated in this section:

Supersymmetric Particles

Some supersymmetric particles could be produced via different processes at the LHC. The search of supersymmetric particles is one of the major physics goals at LHC. The decays of supersymmetric particles, such as squarks and gluinos, involve cascades that, if R-parity [32] is conserved, always contain the Lightest SUSY Particle (LSP). The latter is expected to interact very weakly, thus leading to significant E^T_{miss} in the final state. The rest of the cascade results in an abundance of leptons and jets (particularly b-jets and/or τ -jets). In supersymmetric models with broken R-parity [33], the LSP can decay into a photon and gravitino, an increased number of hard isolated photons is expected.

New Massive Vector Bosons

Massive vector bosons are predicted by various models [32]. The golden channels to search for new massive vector bosons are the decays of high-mass objects, such as $Z' \rightarrow e^+e^-$ and $\mu^+\mu^-$ [34]. From the invariant mass distributions of e^+e^- and $\mu^+\mu^-$ one could observe a massive Z' boson. However, the discovery potential is limited by the statistical significance of the signal. Ways of distinguishing between different models involve the measurement of the natural width and the forward backward asymmetry, both of which require sufficiently good momentum resolution at high p_T ($\Delta p_T/p_T < 0.1$ at $p_T \sim 1$ TeV/c) to determine the sign of the leptons and a pseudorapidity coverage up to $\eta = 2.4$.

2.4 LHC Commissioning and First Operation

2.4.1 LHC Incident and Repairs

After an intense period of preparation, the LHC started beam commissioning on 10 September 2008. The first beam with the energy of 450 GeV was successfully steered around the full 27 kilometers tunnel of LHC. The initial progress was impressive. However, on 19 September 2008, during powering tests of the main dipole circuit in sector 3–4 for the beam energy of 5 TeV, an electrical fault occurred, producing an electrical arc and resulting in mechanical and electrical

damage, release of helium from the magnet cold mass and contamination of the insulation and beam vacuum enclosures. Proper safety procedures were in force and no one was put at risk, but material damage was important, eventually affecting some 700 m of the 3.3 km length of the sector 3-4 [35],[36]. The damage has required the removal of 53 main magnets (dipoles and quadruples) and the repair of the considerable collateral damage.

Details of the repair procedures are shown in Figure 2.12. It took more than a year to investigate and to repair the faults, to check the rest of the accelerator for similar problems and replace faulty magnets, and to install new safety devices to prevent a repeat [37, 38]. The quench protection system was upgraded to enhance the protection of all main quadruples and dipoles. Massive measurement campaigns were carried out to identify and to repair many bad splices. Additional release valves were installed (so far in half of the ring) to improve the pressure relief system to eventually cope with maximum Helium of 40 kg/s in the arcs and reinforced the quadruple supports of arc quadruples, semi-standalone magnets and so on.

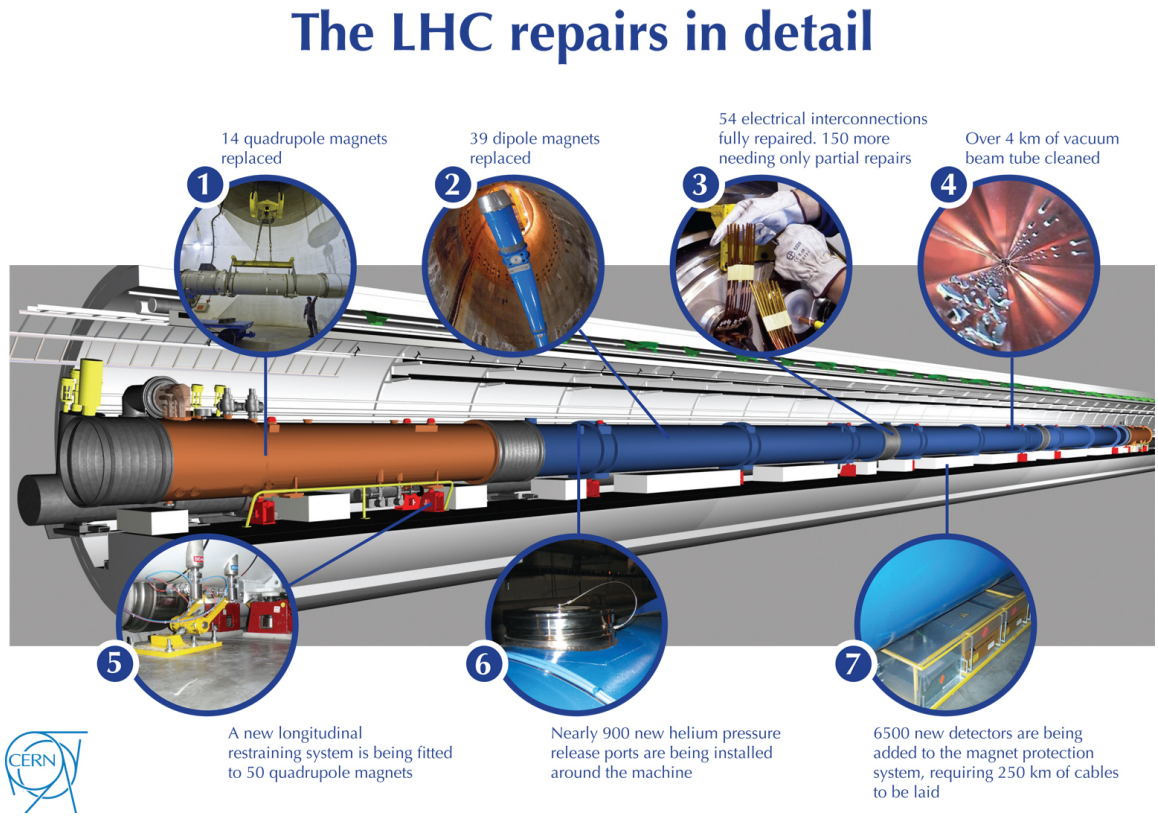

Figure 2.12: The LHC repairs in detail [http://cdsweb.cern.ch/record/1187560/].

With the LHC shutdown for repairs, physicists were busy upgrading both equipments and software, making minor fixes that originally had been scheduled for the LHC's first winter shutdown, and repairing nagging problems that cropped up during years of construction. They had also

Date	Day	Milestones Achieved
Nov 20	1	Each beam circulating. Key beam instrumentation working
Nov 23	4	First collisions at $E_{beam} = 450$ GeV. First ramp (reached 560 GeV)
Nov 26	7	Magnetic cycling established
Nov 27	8	Energy matching
Nov 29	10	Ramp to 1.18 TeV
Nov 30	11	Experiments solenoids on
Dec 04	15	Aperture measurement campaign finished. LHCb and ALICE dipoles on
Dec 05	16	Machine protection (injection, dump, collimators) ready for safe operation
Dec 06	17	First collisions with stable beams, 4 on 4 pilots at 450 GeV. Rates $\sim$ 1Hz
Dec 08	19	Ramp colliding bunches to 1.18 TeV
Dec 11	22	Collisions with stable beams, 4 on 4 pilots at 450 GeV; $> 10^{10}$/bunch; rates $\sim$ 10Hz
Dec 13	24	Ramp 2 bunches/beam to 1.18 TeV. Collisions for 90 minutes
Dec 14	25	Collisions with stable beams, 16 on 16 at 450 GeV; $> 10^{10}$/bunch; rates $\sim$ 50Hz
Dec 16	27	Ramp 4 on 4 to 1.18 TeV; squeeze to 7 m.

Table 2.3: Summary LHC operation in 2009 (S. Myers, 18. Dec 2009).

taken data using cosmic rays in preparation for recording actual collisions and process the data with the worldwide LHC Computing Grid.

2.4.2 Commissioning and First Operation in 2009 and 2010

Recommissioning the LHC began in the summer of 2009, and the successive milestones have regularly been passed (Table 2.3). The LHC reached its operating temperature of 1.9 Kelvin, or about -271° Celsius, on 8 October 2009. Particles were injected on 23 October. The LHC circulated its first beams on 20 November right after the rapid beam-commissioning phase. The first collisions were recorded on 23 November (a CMS event is shown in Figure 2.13), and a world-record beam energy of 2.36 TeV was established on 30 November. Following those milestones, a systematic phase of LHC commissioning led to an extended data-taking period to provide data for the experiments. Over the last two weeks before its planned Christmas shutdown, the LHC experiments have recorded over a million particle collisions, which have been distributed smoothly for analysis around the world on the LHC computing Grid.

After the 2009 winter shutdown, the LHC was restarted and the beam was ramped up to 3.5 TeV per beam. On 30 March 2010, the first planned collisions took place between two 3.5 TeV beams, which set a new world record for the highest-energy man-made particle collisions. The Monitoring page (OP Vistars) is shown in Figure 2.14.

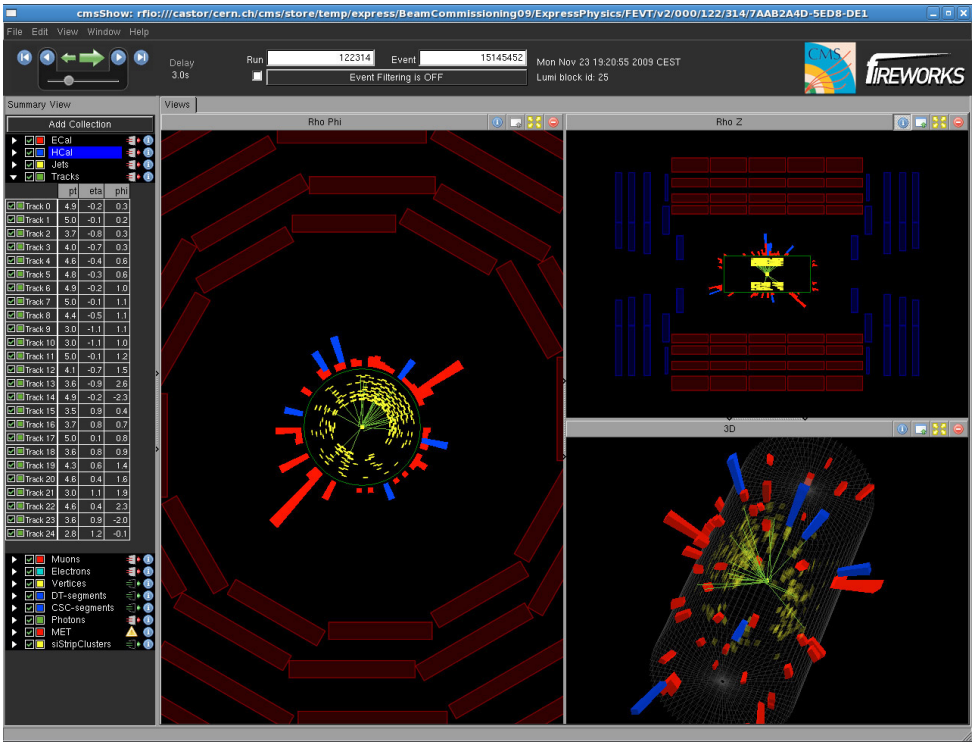

Figure 2.13: A CMS event display. This event occurred during the first LHC collisions at $E_{CM} = 450$ GeV, recorded by the CMS experiment on 23 November 2009.

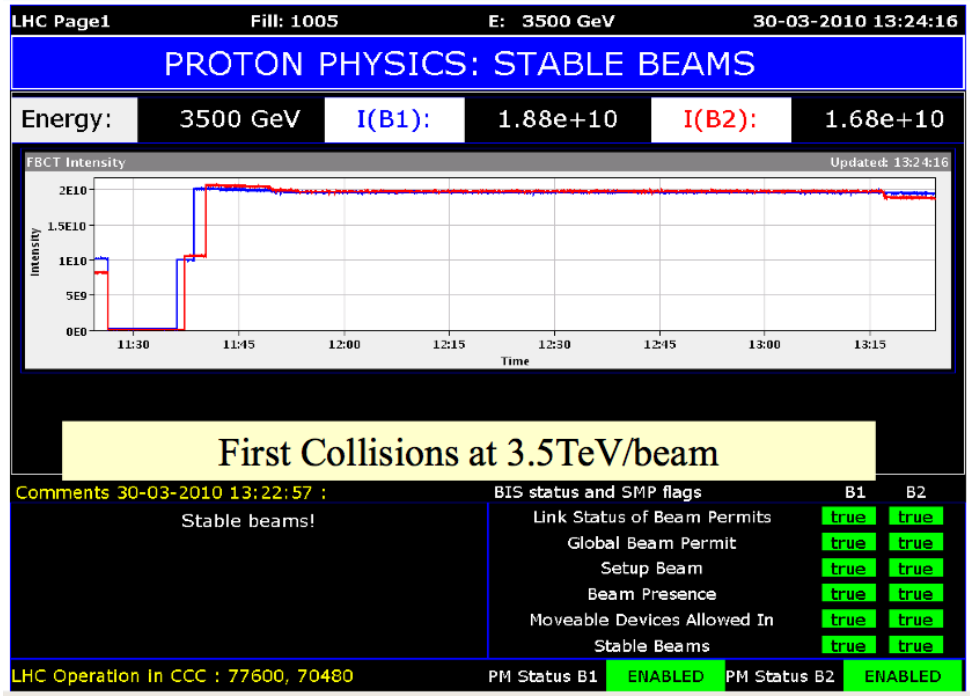

Figure 2.14: The Monitoring page (OP Vistars) during the first collisions at 3.5 TeV/Beam shows the status of stable beams [39].

Event	β*	Nb	Ib	Itot	MJ	MJ Factor	Nc	Peak luminosity	Date
1	10	2	1.00E+10	2.0E+10	0.0113	0.0000	1	8.9E+26	30 March 2010
2	10	2	2.00E+10	4.0E+10	0.0226	2.0000	1	3.6E+27	02 April 2010
3	2	2	2.00E+10	4.0E+10	0.0226	1.0000	1	1.8E+28	10 April 2010
4	2	4	2.00E+10	8.0E+10	0.0452	2.0000	2	3.6E+28	19 April 2010
5	2	6	2.00E+10	1.2E+11	0.0678	1.5000	4	7.1E+28	15 May 2010
6	2	13	2.60E+10	3.4E+11	0.1910	2.8167	8	2.4E+29	22 May 2010
7	3.5	3	1.10E+11	3.3E+11	0.1865	0.9763	2	6.1E+29	26 June 2010
8	3.5	6	1.00E+11	6.0E+11	0.3391	1.8182	4	1.0E+30	02 July 2010
9	3.5	8	9.00E+10	7.2E+11	0.4069	1.2000	6	1.2E+30	12 July 2010
10	3.5	13	9.00E+10	1.2E+12	0.6612	1.6250	8	1.6E+30	15 July 2010

Figure 2.15: The summary of luminosity evolution in 2010 [39].

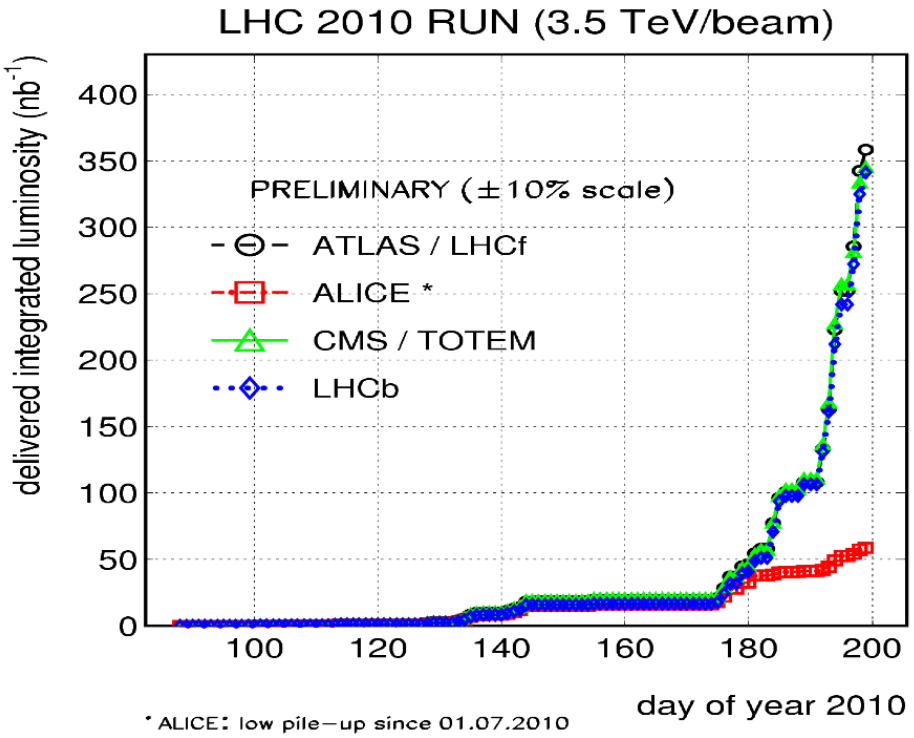

Figure 2.16: Integrated luminosity delivered to the LHC experiments till 14 July 2010. [39].

From 3 May until 9 June, most of the time was spent in understanding the machine protections, primarily collimators, the stability of the orbits and of the accelerating system while trying to inject higher and higher currents in each bunch and/or increasing the number of proton bunches. After each step in stored energy is checked out and declared under control, the accelerator aims at delivering physics collisions to the experiments: the typical pattern has been intense machine

development studies during the week days, with attempts to integrate the experience into short physics fills overnight and focus on steady delivery of physics fills during weekends. This has led to significant steps in the peak luminosity achieved, up to 1.6×10^{30} cm^{-2}s^{-1}. The summary of luminosity evolution till 15 July 2010 is shown in Figure 2.15. As shown in Figure 2.16, both ATLAS and CMS have recorded integrated luminosities of ~ 350 nb^{-1} till 14 July 2010.

2.5 Summary

The SM explains most of the fundamental behavior of particles and their interactions. However the SM is not the ultimate theory of particle physics. The SM answered many questions about particle physics, but more questions are still open. The most critical challenge within the SM is the symmetry breaking mechanism. In the SM, this mechanism of the symmetry breaking predicts the existence of the Higgs particle.

LHC is a hadron collider (proton-proton or heavy ion–heavy ion) built in the existing LEP tunnel of CERN with the proton beam energy up to 7 TeV. The design luminosity for pp collision is 10^{34} cm^{-2}s^{-1}. The major physics goals of LHC are the search for the Higgs particle and the new physics phenomena beyond the SM. The cross sections of the interesting physics processes are many orders of magnitudes smaller than the total cross section. Thus the LHC imposes an extremely challenging environment for the experiments due to the very high luminosity: very high total interaction rate with an average pile up of ~ 20 events in every beam crossing every 25 ns, the average particle multiplicity of 1000, and very high radiation dose. Another great challenge for the LHC experiments is to store and analyse the enormous data with an order of 10^7 GB per experiment per year. The LHC Grid Computing optimizes the global computing and storage resources to meet the challenge.

LHC started beam commissioning on 10 September 2008. However, on 19 September 2008, an incident occurred during powering tests of the main dipole circuit. It took more than a year to repair the faults and to prevent a repeat. Recommissioning the LHC began in the summer of 2009. The first collisions were recorded on 23 November, and a world-record beam energy of 2.36 TeV was established on 30 November. On 30 March 2010, the first planned collisions took place between two 3.5 TeV beams, which set a new world record for the highest-energy man-made particle collisions. Machine studies led to significant steps in the peak luminosity achieved, up to 1.6×10^{30} cm^{-2}s^{-1} by the middle of July 2010.

3 CMS Experiment

The main goals of the LHC are the search for new physics and precision measurements. One of the main tasks is to probe the existence of the Higgs boson or the origin of the electroweak symmetry breaking in general. Furthermore, many theoretical models with new physics wait to be investigated experimentally. To meet those physics goals, the LHC has the seven-fold increase in the beam energy and a hundred-fold increase in the design luminosity over the previous proton collider experiments. The total proton-proton cross-section at $\sqrt{s}$ = 14 TeV is expected to be roughly 70 mb. At the design luminosity of 10^{34} $cm^{-2}s^{-1}$, on average, about 20 inelastic collisions will be superimposed on each event of interest. The effect of this pile-up must be reduced by using high-granularity detectors with fast time resolution, which requires a huge number of detector channels. The resulting millions of detector electronic channels require very good synchronization. Moreover, the large flux of particles coming from the interaction region leads to high radiation levels, requiring detectors and front-end electronics to be radiation-hard. The Compact Muon Solenoid (CMS) [40] detector is designed to meet the requirements for the physics goals and the technology challenges at the LHC. The design and the performance of the CMS detector are described in details in this chapter. Section 3.1 discusses the general design concept of the CMS detector. The following sections discuss the design and performance of the sub-detectors: the superconducting magnet, the inner tracking system, the electromagnetic calorimeter, the hadron calorimeter, the muon system, the trigger and data acquisition. Section 3.8 reviews the CMS detector commissioning, including the cosmic run in 2008 and in 2009, the first collision in 2009 and the collision data at $\sqrt{s}$= 7 TeV in 2010.

3.1 General Design Concept

CMS is one of the two general purpose experiments at the LHC. In principle, an ideal general purpose collider detector should be designed with 4 π coverage homogeneously around the collision point in order to detect all particles produced in the collision. However, it's extremely difficult to design and to fabricate such a detector. Therefore a cylindrical shape together with end caps have been adopted, mainly limited by the solenoid magnet shape, still allowing for

an almost 4π coverage. The modern collider detectors are composed of several sub-detectors, positioned in concentric layers (onion structure) and each of them dedicated to different and complementary types of the measurements.

The design of the CMS experiment has been optimized for the physics goals of LHC:

- **The best possible electromagnetic calorimeter:** good electromagnetic energy resolution, good di-photon and di-electron invariant mass resolution ($\sim 1\%$ at 100 GeV), wide geometric coverage, π^0 rejection, and efficient photon and lepton isolation at the highest luminosities of the LHC.
- **A high quality inner tracking system:** good charged-particle momentum resolution and reconstruction efficiency at very high multiplicity in the inner tracker. Efficient triggering and offline tagging of τ's and b-jets, requiring the pixel detectors close to the interaction region.
- **A hadron calorimeter with almost 4π coverage:** good missing-transverse-energy and dijet-mass resolution, requiring the hadron calorimeters with large hermetic geometric coverage and with fine lateral segmentation.
- **An accurate and efficient muon system:** good muon identification and momentum resolution over a wide range of momenta and angles, good di-muon invariant mass resolutions ($\sim 1\%$ at 100 GeV), and the ability to determine the charge of muons up to $p \leq 1$ TeV.

The overall layout of the CMS detector is shown in Figure 3.1. CMS has a conventional cylindrical onion-like structure. It is composed of a barrel and two endcaps. The barrel is the cylindrical part coaxial to the beam pipe. The endcaps are installed perpendicular to the beam pipe at both sides of the barrel. The CMS detector is 21.5 m in length and 15 m in diameter. The total weight is 12500 tons. The CMS detector consists of several sub-detectors and a 4 T superconducting solenoid to optimize the precision measurements of muon, photon and electrons, as well as good measurements of hadron jets. Combining the measurements in the four sub-detectors, the CMS detector can provide particle identification with high efficiency. Figure 3.2 is a slice of the CMS barrel detector. The typical trajectories from muon, electron, hadron and photon are illustrated in the figure. From their different behaviors in the sub-detectors, one can distinguish these particles clearly. This is also one of the key issues for CMS to meet the physics goals.

The CMS coordinate system has the origin centered at the nominal collision point inside the vacuum pipe, the positive y-axis pointing vertically upward, and the positive x-axis pointing radically inward, toward the center of the LHC. Thus, the positive z-axis points along the beam direction toward the Jura mountains from the LHC Point 5. The azimuthal angle ϕ is measured from the x-axis in the x-y plane, and the radial coordinate in this plane is denoted by r. The

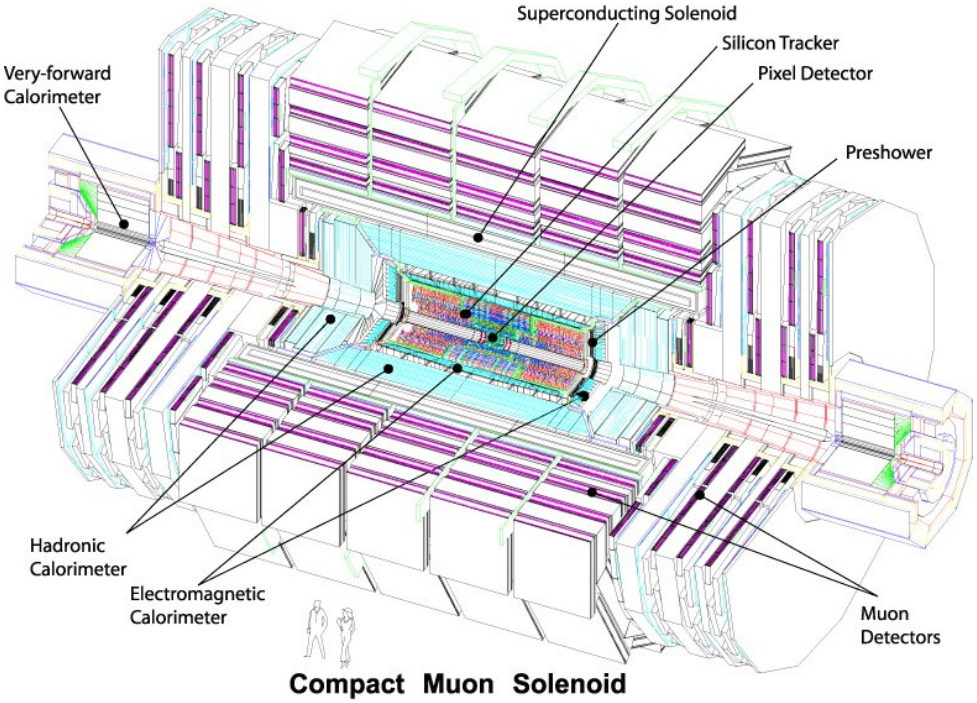

Figure 3.1: The three-dimensional view of the CMS detector [40].

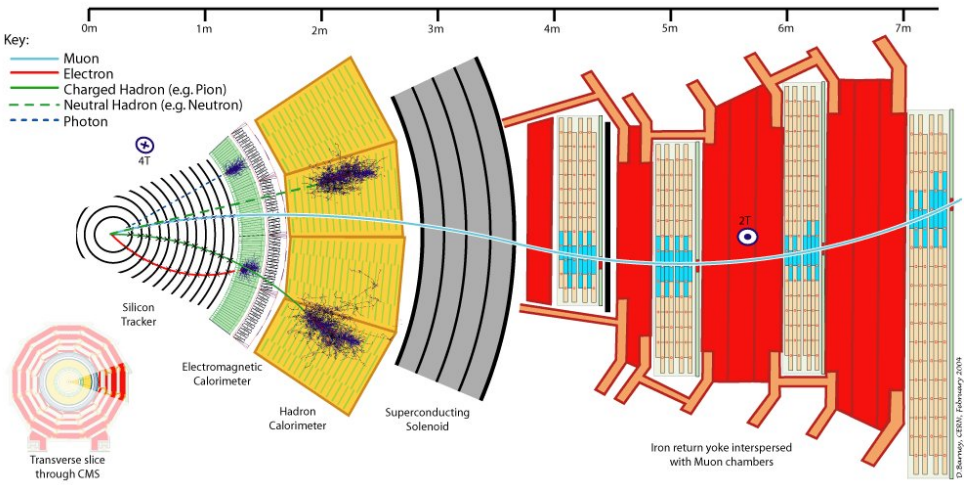

Figure 3.2: A slice of the CMS barrel in the x-y plane. The trajectories of a muon, electron, hadron and photon are illustrated.

polar angle θ is measured from the z-axis. The pseudorapidity is defined as $\eta = -\ln\tan(\theta/2)$. Typically, the barrel covers a pseudorapidity range of $\eta \leq 1.5$ while the endcaps cover the range of $1.1 \leq |\eta| \leq 3.0$. The precise limits depend on the sub-detector considered. Thus, the momentum and energy transverse to the beam direction, denoted by p_T and E_T, respectively, are computed from the x and y components. The imbalance of the energy measured in the transverse plane is

denoted by E_T^{miss} [1].

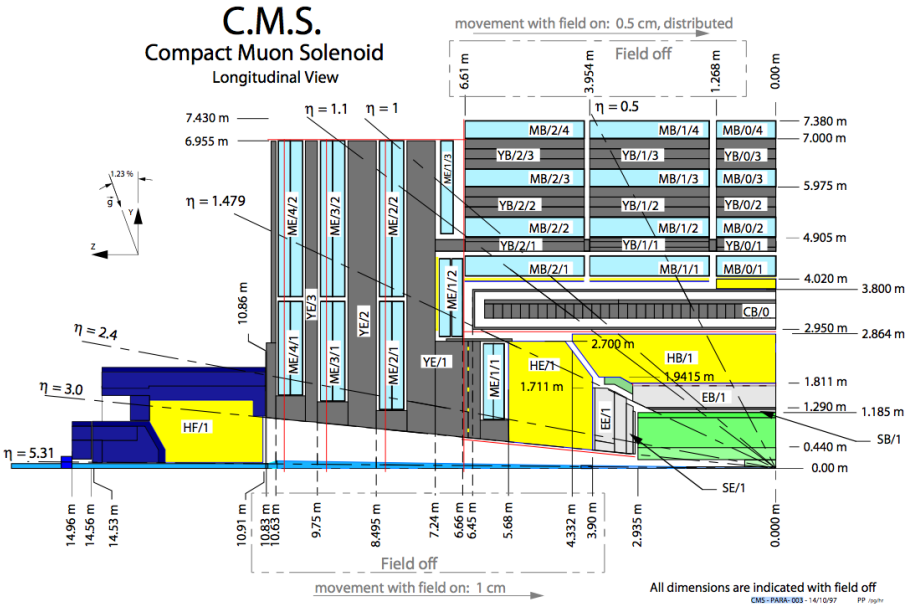

Figure 3.3: A quadrant of the CMS detector in the x-z plane. The sub-detectors are rendered in different colors respectively: the tracker (green), the electromagnetic calorimeter (light grey), the hadronic calorimeter (yellow) and the muon chambers (blue). The iron return yoke (YB,YE) and magnet (CB) are both rendered in dark grey [40].

One quadrant of the CMS detector is shown in Figure 3.3. At the heart of the CMS detector there is a 13-m-long, 6-m-inner-diameter, 4-T superconducting solenoid providing a large bending power (12 Tm) before the muon bending angle is measured by the muon system. The return field is powerful enough to saturate 1.5 m of iron, allowing 4 muon stations to be integrated to ensure robustness and full geometric coverage. Each muon station consists of several layers of aluminum drift tubes (DT) in the barrel region and the cathode strip chambers (CSC) in the endcap region, complemented by resistive plate chambers (RPC).

In order to cope with high track multiplicities, the CMS detector employs 10 layers of silicon microstrip detectors, which provide the required precision and granularity. In addition, 3 layers of the silicon pixel detectors are placed close to the interaction region to improve the measurement of the impact parameter of charged-particle tracks, as well as the position of secondary vertices. The ECAL uses lead tungstate ($PbWO_4$) crystals with coverage in pseudorapidity up to $|\eta| \leq 3.0$.

[1] The transverse missing energy is reconstructed with the sum of the electromagnetic calorimeter and hadron calorimeter tower raw energies, corrected for the energy contribution of each muon in the event. The events with large missing transverse energy are interesting. For example, the leptonic decay of the W boson has large missing transverse energy due to the neutrino. More important, many channels of discovery at the LHC present as a clear signature for new physics a large missing transverse energy (e.g., SUSY decays with a LSP escaping detection by CMS).

The ECAL is surrounded by a brass/scintillator sampling hadron calorimeter (HCAL) with coverage up to $|\eta| \leq 3.0$. This central calorimeter is complemented by a tail-catcher in the barrel region (HO) ensuring that hadronic showers are sampled with nearly 11 hadronic interaction lengths. Coverage up to a pseudorapidity of 5.0 is provided by an iron/quartz-fibre calorimeter. An even higher forward coverage is obtained with additional dedicated calorimeters (e.g., CASTOR and ZDC, not shown in Figure 3.1) and with the TOTEM [41] tracking detectors.

3.2 Superconducting Magnet

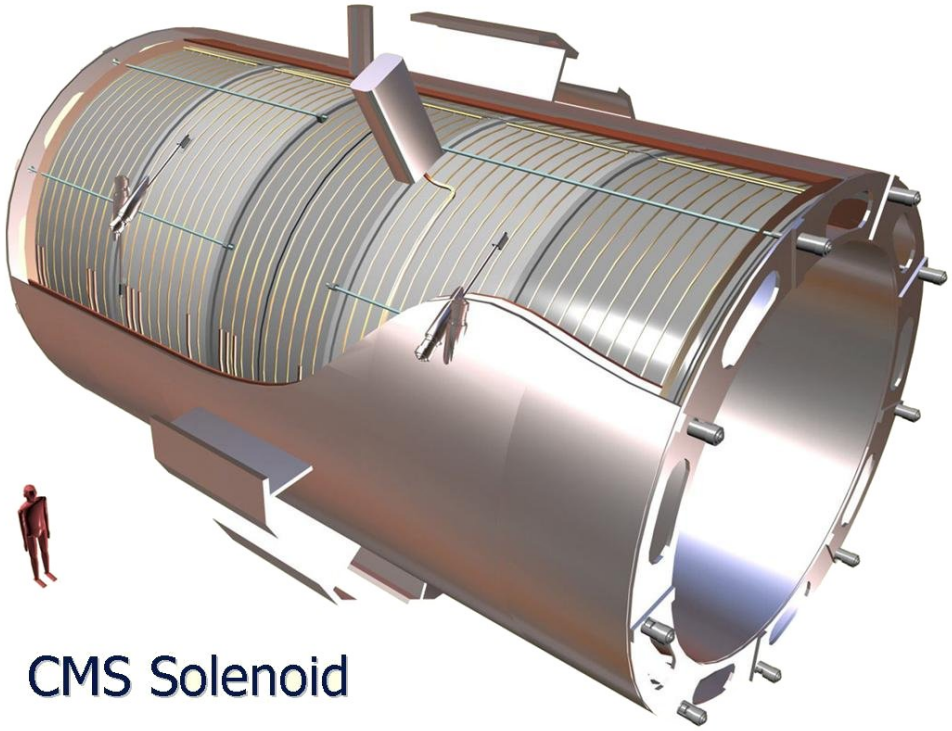

Figure 3.4: The general artistic view of the 5 modules composing the cold mass inside the cryostat, with details of the supporting system (vertical, radial and longitudinal tie rods) [40].

The physics motivation of the CMS experiment emphasizes on the precision muon measurement, since muons are very suitable for the experimental observation (efficient detection and precise reconstruction even at very high luminosities). Muons are also important signatures for many new physics phenomena. The good performance in the muon reconstruction is a challenging task that has driven the CMS design. To detect and measure the momentum of muons and other high-energy charged particles efficiently and precisely, large bending power is needed. CMS has chosen a very elegant solution with a large 4-T superconducting solenoid (an artistic view is shown in Figure 3.4) providing enormous bending power (12 Tm) with a stored energy of

parameter	value
Field	4 T
Inner bore	5.9 m
Length	12.9 m
Current	19.5 kA
Stored energy	2.7 GJ
Hoop stress	64 atm

Table 3.1: Parameters for the superconducting solenoid [42].

2.7 GJ at the full current with an instrumented iron return yoke. This solution provides both excellent momentum resolution using the tracker and adequate triggering capabilities outside the calorimeter with muon stations embedded in the iron return yoke. This design also leads to a more compact experiment.

The superconducting magnet [42] has been designed to reach a 4-T field in the free bore of 5.9 m diameter and 12.9 m length. The flux is returned through a 10 000 tons yoke comprising 5 wheels and 2 endcaps, composed of three disks each as shown in Figure 3.1. The distinctive feature of the 220 tons cold mass is the 4-layer winding made from a stabilized reinforced NbTi conductor with a larger cross-section that can withstand an outward pressure (hoop stress) of 64 atmospheres. The parameters of the CMS magnet are summarized in Table 3.1. The magnet was assembled and tested in the surface hall (SX5) above the CMS cavern, prior to being lowered 90 m below ground to its final position in the experimental cavern. After provisional connection to its ancillaries, the CMS magnet has been fully and successfully tested and commissioned in SX5 during autumn 2006.

3.3 Inner Tracking System

Encompassing the beam-pipe, the CMS tracker measures the trajectories and momenta of charged particles up to $|\eta| \simeq 2.4$ [44, 45]. The overall layout of the tracker is shown in Figure 3.5. It plays an important role to reconstruct tracks for the precision measurements of their momenta and vertex of origin. Furthermore, it's also important in jet flavour tagging, especially for the $b-$ and $\tau-$jets. The momentum measurement of tracks in the 1-5 GeV range is also crucial to define 'isolated' objects $(e, \nu, \gamma, \tau, \ldots)$. The measurements can also distinguish electrons and photons, as well as study muons and jets.

In order to deal well with such high track multiplicities, CMS employs many layers of silicon microstrip detectors, providing a relatively low number of precisely measured points, rather than

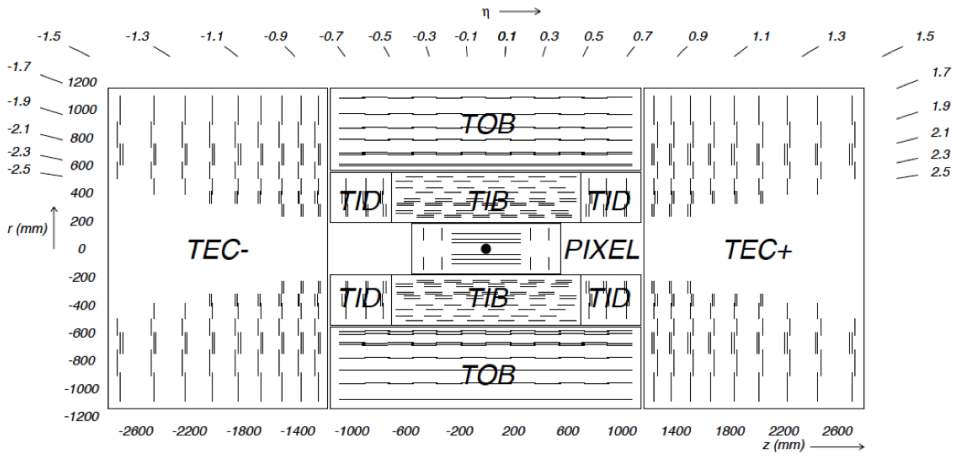

Figure 3.5: The rz-view of the CMS tracking detectors [43]. The single lines represent layers of modules equipped with one sensor, while the double lines indicate layers with back-to-back modules [40].

continuous tracks. In addition, fine granularity pixels are placed close to the interaction point, where the particle flux is highest to maintain a low channel occupancy and minimize the track ambiguities.

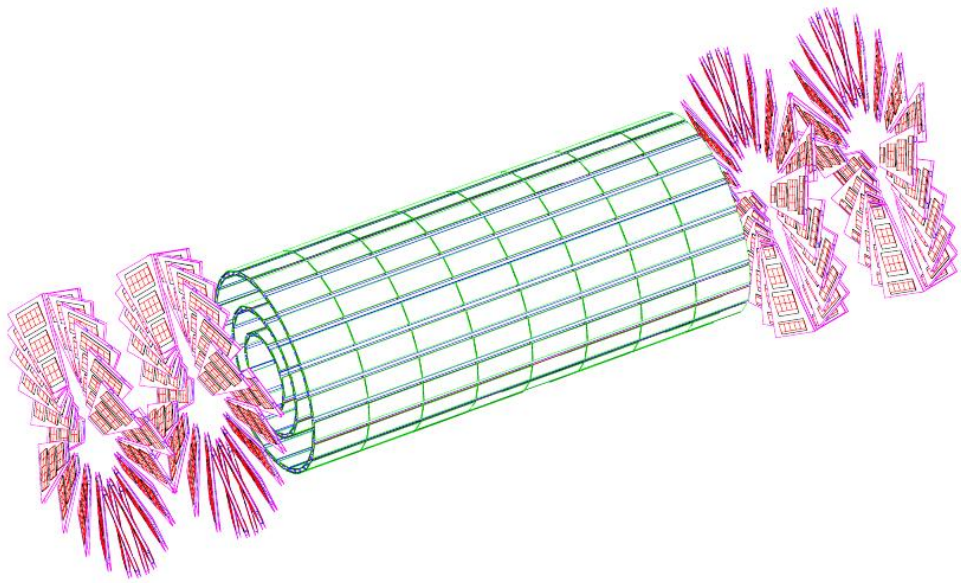

Figure 3.6: The schematic drawing of the pixel tracker. The barrel is coloured green, the endcaps red [40].

As illustrated in Figure 3.6, the pixel system consists of 3 barrel layers: at a radius of 4.4 cm, 7.3 cm and 10.2 cm from the beam-pipe with a length of 53 cm and 2 endcap discs extending from 6 cm to 15 cm in radius, at $|z| = 34$ cm and 46.5 cm. 66 million pixels of size $\sim 100 \times 150\ \mu m^2$

are arranged across 768 and 672 modules in the barrel and endcaps respectively. To maximize vertex resolution an almost square pixel shape has been adopted. A Lorentz angle of 23° in the barrel improves the $r - \phi$ resolution through charge sharing. The endcap discs are assembled with a turbine-like geometry with blades rotated by 20° to also benefit from the Lorentz effect. The resultant spatial resolution is 10 μm in $r - \phi$ and 20 μm in z, allowing a primary vertex resolution of ~ 40 μm in z.

The silicon strip tracker has a length of 5.8 m and a diameter of 2.4 m, and is composed of four subsystems: the four-layer tracker inner barrel (TIB), the six-layer tracker outer barrel (TOB) and on each side three-disk tracker inner disks (TID) and nine-disk tracker endcaps (TEC). An rz-view of the tracker geometry is shown in Figure 3.5.

The silicon strip tracker is built from 15148 single-sided modules that provide 9.3 million readout channels. The modules for the TIB, the TID and the first four rings of the TEC are single-sided while the TOB and the outer three rings of the TEC are equipped with double-sided modules. A double-sided module is constructed from two single-sided modules glued back-to-back at a stereo angle of 100 mrad.

The leakage current of the bulk of the detector increases exponentially with temperature and linearly with radiation dose. In turn, the sensor temperature increases with the power dissipated within it. This cyclic dependency can result in a thermal runaway and requires the sensor temperature to be maintained at $\sim -15°C$. This is achieved through a distributed cooling system.

Apart from the sensitive detector volumes, the CMS tracker contains lots of non-sensitive material, like mechanical supports, electrical supply cables and cooling services. For the $1.2 < |\eta| < 2.1$ region, this material can constitute more than one radiation length[2], as shown in Figure 3.7. The amount of material in the tracker must be kept as low as possible in order to avoid secondary interactions, excessive multiple scattering, bremsstrahlung and photon conversion which would compromise the performance of the electromagnetic calorimeter. Therefore, a compromise has to be sought between the number of hits per track (i.e., the number of active layers) for an efficient track reconstruction and amount of material in the tracker.

With the configuration described above, the expected performance is shown in Figure 3.8 for muons and pions. The track reconstruction efficiency for high energy muons is about 99% and drops at $\eta > 2.1$ due to the reduced coverage of the forward pixel detector. For pions the efficiency is in general lower because of interactions with the material in the tracker.

[2]The radiation length X_0 is a characteristic of a material, related to the energy loss of high energy electrons in the material.

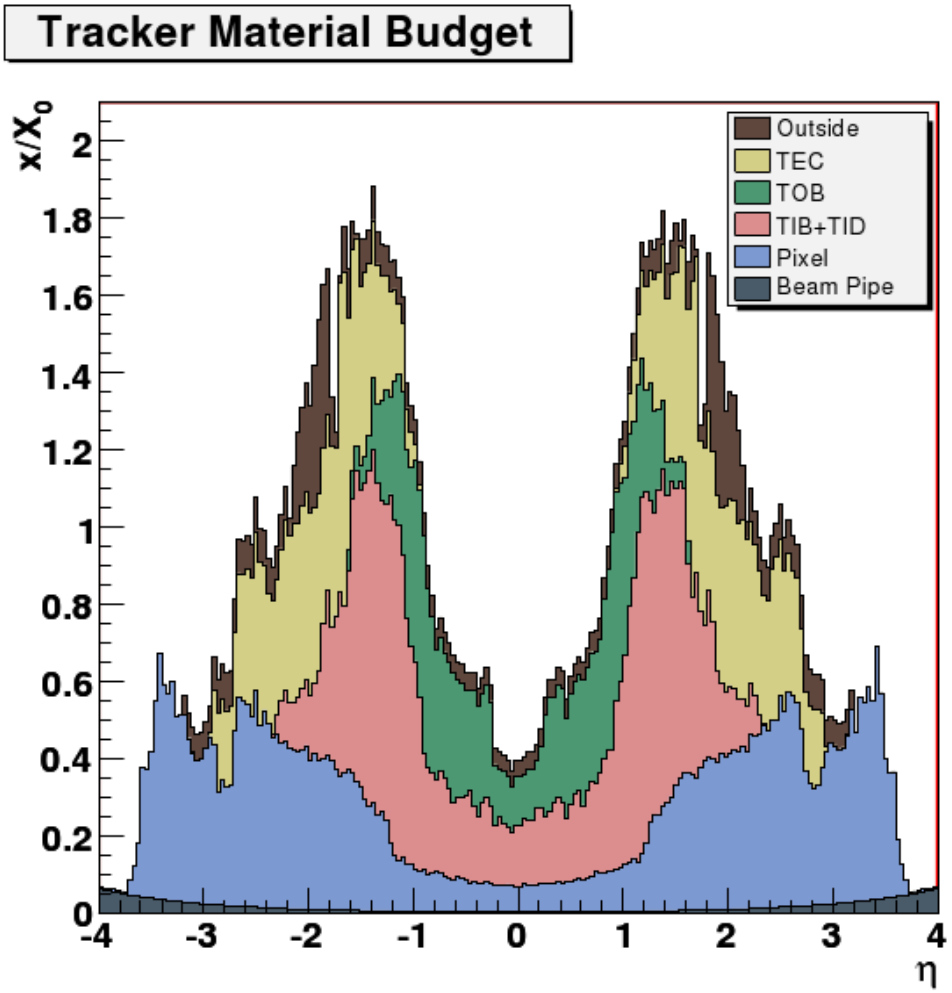

Figure 3.7: The material budget of the CMS tracker in terms of radiation lengths as a function of η for the different tracker subunits [40].

In Figure 3.9 the transverse momentum resolution for muon tracks with $p_T = 1$, 10 and 100 GeV is shown. At high momenta the resolution is around 1–2% for $|\eta| < 1.6$. The material of the tracker accounts for 20–30% of the transverse momentum resolution. At lower momenta, the resolution is dominated by multiple scattering. The resolution of the track impact parameters in the longitudinal and the transverse plane are also shown in Figure 3.9. At high momentum, the transverse impact parameter resolution is fairly constant and is dominated by the hit resolution in the first pixel layer. It is progressively degraded by multiple scattering at lower momenta. The same applies to the longitudinal impact parameter resolution. The improvement of the z_0 resolution up to $|\eta| = 0.5$ is due to the charge sharing effects among the neighboring pixels.

The calorimeters sit outside of the tracker: an electromagnetic calorimeter designed to measure the energies of electrons and photons and a hadron calorimeter designed to measure the energies of hadronic jets. These sub-detectors will be discussed in the following sections.

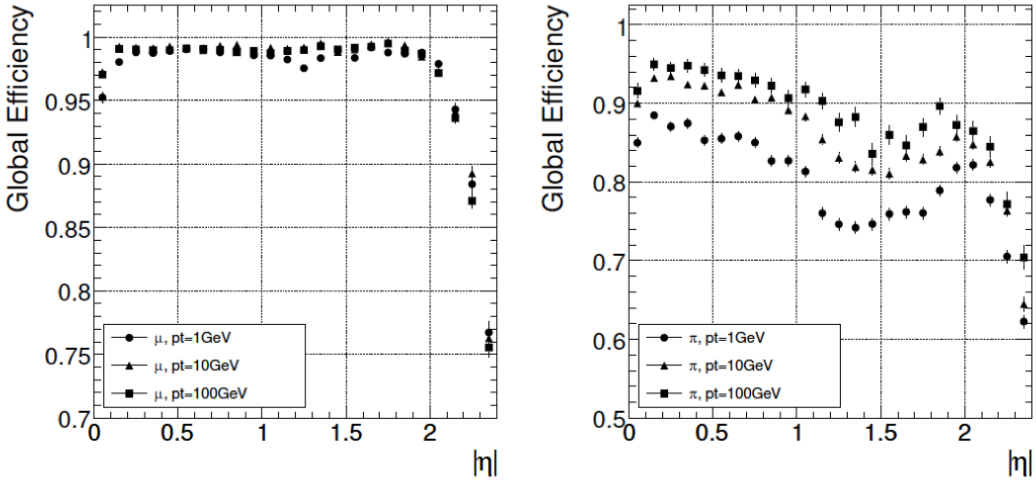

Figure 3.8: The track reconstruction efficiency for muons (left) and pions (right) with transverse momenta of 1, 10 and 100 GeV [43].

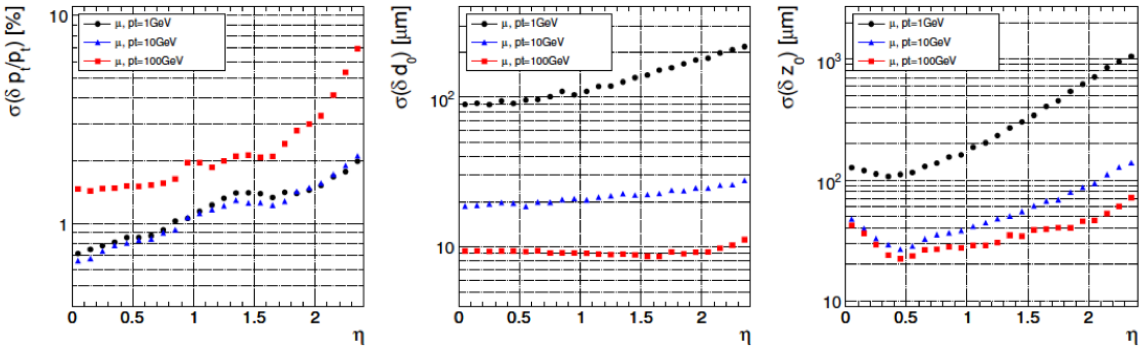

Figure 3.9: The resolution as a function of pseudorapidity of track transverse momentum (left), transverse impact parameter (middle) and longitudinal impact parameter (right). The resolution is shown for muons with transverse momenta of 1, 10 and 100 GeV [43].

3.4 Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) is placed around the tracker with an inner radius of 129 cm. The function of the electromagnetic calorimeter is to measure the energy of electrons and photons with high precision and together with the hadron calorimeter, to measure jets. The design of the CMS electromagnetic calorimeter is driven by the requirement to provide an excellent di-photon mass resolution for the crucial two photon decay mode of the Higgs boson $H \to \gamma\gamma$, which is the main Higgs discovery channel for $m_H \lesssim 130\text{GeV}$. This mass resolution depends on the resolution in energy of the two photons and the error on the measured angle

between them.

To achieve a good angular separation between photons, which is important to identify reducible background processes where two photons from energetic π^0s may be reconstructed as a single photon, a highly granular design of the electromagnetic calorimeter is crucial. In the endcaps the ECAL is complemented by a preshower detector in order to identify the neutral pions efficiently.

A design that fulfills all requirements consists of about 76000 lead tungstate ($PbWO_4$) crystals. This crystal was chosen because it has a high density (8.2 g/cm^3) leading to a short radiation length ($X_0 = 0.89$ cm) and a small Moliere radius ($R_M = 2.19$ cm), allowing for a compact ECAL design with narrow electromagnetic showers. The second advantage of $PbWO_4$ is that the scintillating process is fast: 85% of the light is emitted within 20 ns, matching the LHC bunch crossing time of 25 ns. The third reason for using lead tungstate is that the material is intrinsically radiation hard. However, the low light yield requires a read-out through photodetectors with high gain.

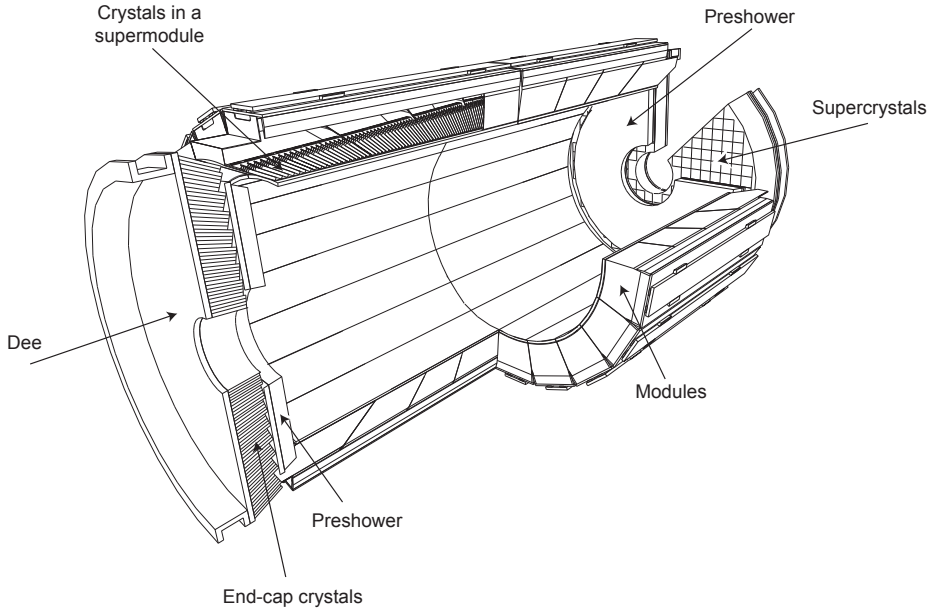

Figure 3.10: The layout of the CMS electromagnetic calorimeter, showing the barrel supermodules, the two endcaps and the preshower detectors [43].

The layout of ECAL is illustrated in Figure 3.10 and Figure 3.11. Each of the 36 supermodules in the ECAL barrel (EB) consists of 1700 tapered $PbWO_4$ crystals with a frontal area of approximately 2.2×2.2 cm^2 and a length of 23 cm (corresponding to 25.8 radiation lengths). The crystal axes are inclined at an angle of 3° relative to the direction of the nominal interaction point, in both the azimuthal (ϕ) and η projections. The two ECAL endcaps (EE) are constructed from four half-disk 'Dees', each consisting of 3662 crystals, with a frontal area of 2.68×2.68 cm^2

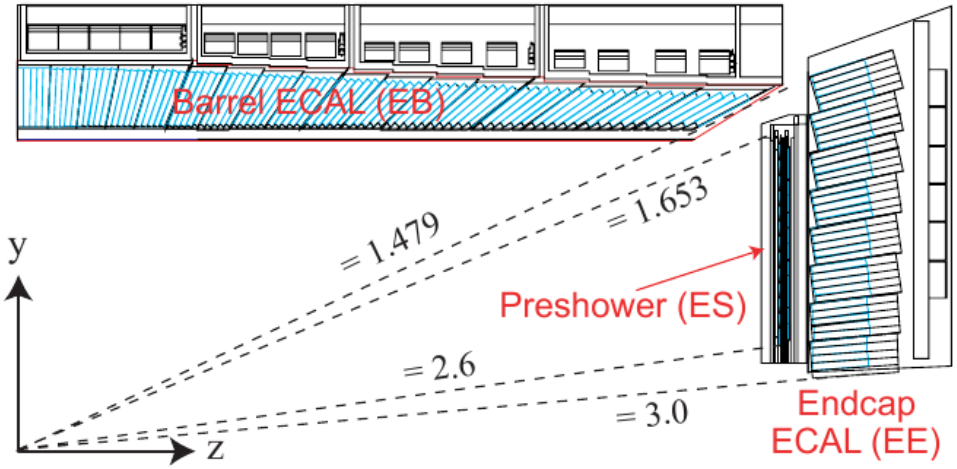

Figure 3.11: The transverse section of the CMS ECAL (one quarter) [43].

and a length of 22 *cm* (corresponding to 24.7 radiation lengths), arranged in a quasi-projective geometry. The crystals in each Dee are organized into 138 standard 5×5 super-crystal units, and 18 special shaped super-crystals that are located at the inner and outer radii. The crystals convert energy into light, and the scintillation light is detected by silicon avalanche photodiodes (APDs) in the barrel region and vacuum phototriodes (VPTs) in the endcap region.

The energy resolution of the ECAL can be parametrized as:

$$(\frac{\sigma}{E})^2 = (\frac{S}{\sqrt{E}})^2 + (\frac{N}{E})^2 + C^2 \tag{3.4.1}$$

where the first term is the stochastic term, due to the fluctuations of the shower containment and photo-statistics; the second term is the noise term, consisting of both electronics noise and pile-up energy; and C^2 is the constant term. The coefficients S and C are determined by the active detector material. The value of the three parameters were determined by an electron test beam measurement to be $S = 0.028\ GeV^{\frac{1}{2}}$, $N = 0.12\ GeV$ and $C = 0.003$. The ECAL energy resolution as a function of the electron energy is shown in Figure 3.12.

3.5 Hadron Calorimeter

The CMS hadron calorimeter (HCAL) surrounds the electromagnetic calorimeter and is used in conjunction with the latter to measure the energies and the directions of particle jets, as well as to provide hermetic coverage for measuring missing transverse energy. It also plays an important role in the identification of electrons, muons and hadrons. The pseudorapidity range of $|\eta| \leq 3.0$ is covered by the barrel and endcap hadron calorimeters which are located inside the 4 T field of the CMS solenoid. Because of the high magnetic field, the calorimeters are necessarily made out

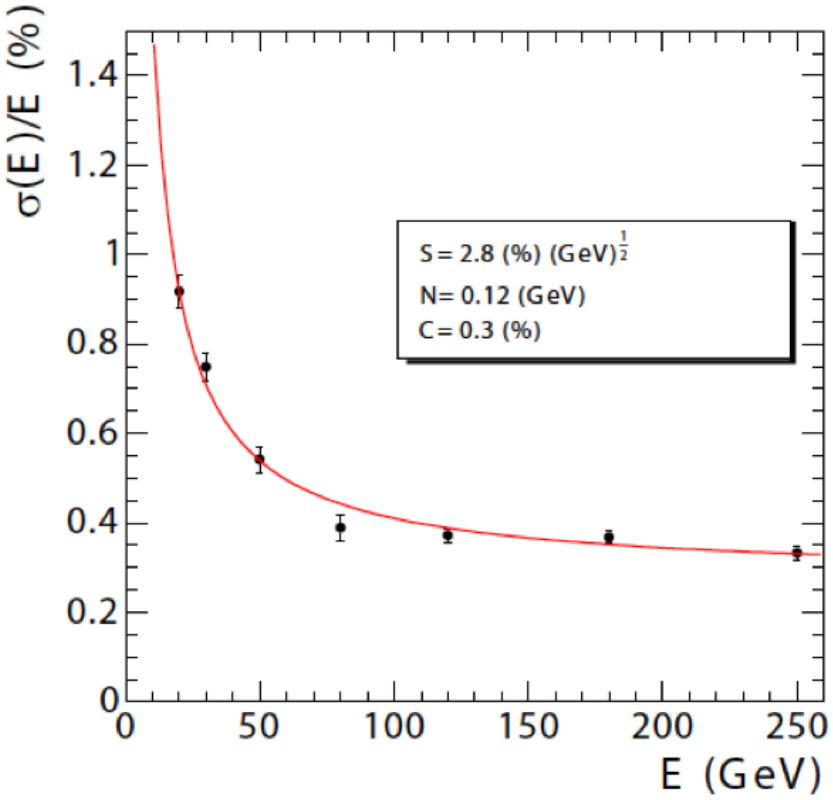

Figure 3.12: The ECAL energy resolution as a function of the energy measured in an electron test beam [43]. The measured values of the stochastic (S), noise (N) and constant (C) term are displayed in the legend.

of non-magnetic material. In order to minimize multiple scattering for traversing muons, low-atomic-number materials like copper alloy and stainless steel are chosen. The active elements of the barrel and endcap hadron calorimeter consist of plastic scintillator tiles with the wavelength shifting fiber readout. The layers of these tiles alternate with layers of brass absorber to form the sampling calorimeter structure. The tiles are arranged in projective towers with fine granularity to provide the good di-jet separation and mass resolution.

The HCAL detector is divided into four sub-detectors as shown in Figure 3.13, comprising a total of 9072 channels. The HCAL barrel (HB) and endcap (HE) detectors surround the electromagnetic calorimeter and are contained completely within the high magnetic field region of the solenoid. The HB provides coverage in the pseudorapidity range of $|\eta| < 1.4$, while the HE provides overlapping coverage in the range of $1.3 < |\eta| < 3.0$. The HCAL forward calorimeters (HF) provide the measurements of energetic forward jets and increase the hermeticity for the missing transverse energy measurement. The HF sub-detectors extend the HCAL pseudorapidity

coverage into the region of $|\eta|$ of 2.9–5.0. The effective HCAL thickness in the region of $|\eta| < 1.3$ is extended by the addition of an array of 'outer barrel' (HO) scintillators outside the magnet cryostat. Each sub-detector spans the full range of the azimuthal angle ϕ.

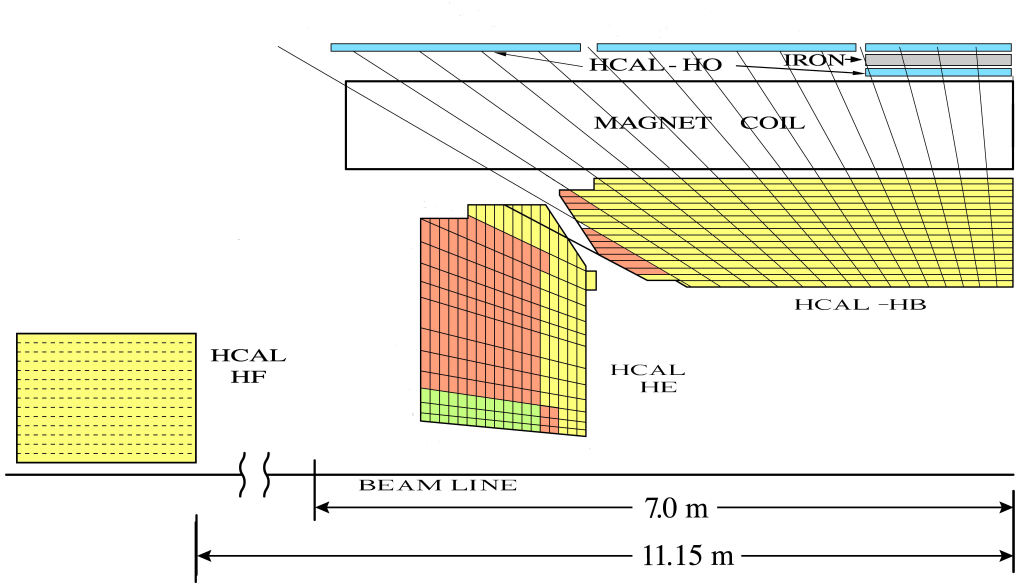

Figure 3.13: The quarter view of the CMS hadron calorimeter. The shading indicates the optical grouping of the scintillator layers into the different longitudinal readouts [43].

The HB and HE sub-detectors consist of layers of plastic scintillator within a brass/stainless steel absorber. These sub-detectors are segmented into readout channels that cover an area of 0.087×0.087 in $\eta - \phi$ space. In the regions where $|\eta|$ is greater than 1.74, the ϕ segmentation is more coarsely granulated. The scintillation light is detected by hybrid photodiodes (HPDs), with each HPD collecting signals from 18 different HCAL channels.

The HF sub-detector is a Cherenkov light detector made of quartz fibers embedded within a 165 cm long steel absorber. There are two types of fibers within HF: 'long' fibers that span the length of the sub-detector, and 'short' fibers that begin 22 cm into the detector. The differences between signals read out from the long and short fibers are used to distinguish between electromagnetic and hadronic showers. The long and short fibers are separately grouped to span 0.174 radians in ϕ, and intervals in η ranging between 0.111 and 0.178. Each group is read out separately as a single HF channel. The photomultiplier tubes (PMTs) connected to the fibers via light guides convert detected light to electrical signals. The seven-bit analog-to-digital converters (ADCs) digitize the signals from the calorimeter for readout. The signals from 4 HPDs or 72 PMTs are digitized within a single 72-channel readout box (RBX).

The results of the beam tests of the CMS calorimeter (including the ECAL) indicate the following

energy resolution between 30 GeV and 1 TeV [46]:

$$\frac{\sigma}{E} = (\frac{122}{\sqrt{E(\text{GeV})}} + 5)\% \tag{3.5.1}$$

For hadrons with transverse momenta below 20-30 GeV, the non-linearity of the response of the ECAL + HCAL system is more problematic.

3.6 Muon System

As mentioned in the previous sections, muons are the key signatures for most of the physics goals of CMS. In electroweak and top physics, Higgs physics, B-physics, as well as in most extensions of the Standard Model, such as supersymmetry and extra dimensions, muons are often present in the final state topology. The ability to trigger on and to reconstruct muons at the highest luminosities is central to the concept of CMS, as can be understood from the name of the experiment. Because of their high mass and long lifetime, muons are the cleanest experimentally measurable objects. Combining the information of the muon system with the tracker system, a percent level precision can be obtained on the transverse momentum of a 100 GeV muon. The muon system of CMS has therefore three purposes: to trigger on muons, to identify muons and to measure their momenta and charges.

The muon detector is the outermost sub-detector covering the region of $|\eta| \leq 2.4$. It is designed to record the trajectories of muons, which is the only particle (except neutrinos, which are very difficult to be detected due to the extremely small interaction cross section) to penetrate the calorimeter. Due to the low particle flux compared to the Silicon tracker and the large surface to be covered ($\sim 25000\ m^2$), the granularity is far coarser consisting of nearly 1 million readout channels. The schematic of one quadrant of the muon system is shown in Figure 3.14.

The muon detectors are distributed over four layers in the barrel and endcap sections of the return yoke. The latter is made of iron and designed to prevent the leakage of the strong magnetic field to large distance. Three gaseous muon tracking technologies are employed, reflecting the varying radiation and magnetic environments. In the barrel $|\eta| \leq 1.2$, 250 drift tube (DT) chambers are operated in up to 12 planes. Each consists of central anode wire surrounded by aluminum cathode and is filled with Ar and CO_2. The induced charge has a maximum drift time of 400 ns (cf. the 25 ns bunch spacing of LHC). Each chamber provides a muon vector of resolution 100 μm in $r - \phi$ and 1 mrad in direction.

In the endcaps ($1.2 \leq |\eta| \leq 2.4$) with the high muon and neutron background environment 468 cathode strip chambers (CSCs) are used. CSCs provide precise space and time information in

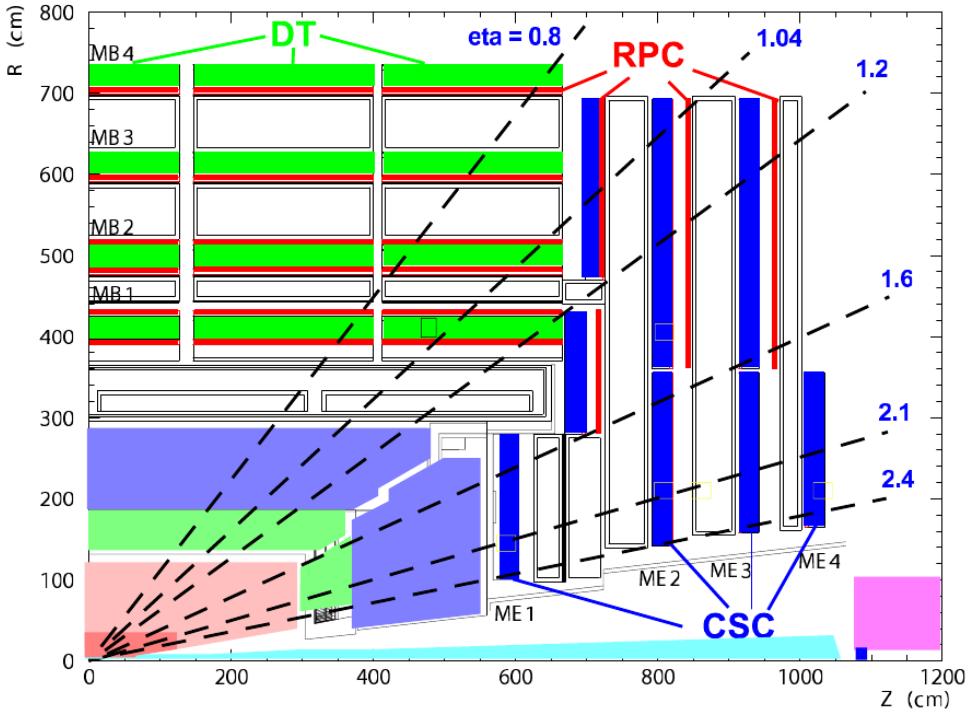

Figure 3.14: The schematic of one quadrant of the muon system. It shows three different detector technologies: the drift tubes (DT), the resistive plate chambers (RPC) and the cathode strip chambers (CSC) [43].

the presence of a high magnetic field and high particle rate. Each is trapezoidal and contains 6 gas gaps. Each gap has a plane of radial cathode strips with perpendicular anode wires. A single chamber provides a spatial resolution of 200 μm and an angular resolution of 10 mrad in ϕ.

To achieve a fast time response and hence accurate bunch-crossing identification, the resistive plate chambers (RPCs) are also used in both barrel and endcap regions. A RPC consists of a gas gap enclosed by two graphite-coated bakelite plates forming cathodes, operated in avalanche mode. The RPCs provide a time resolution of ~ 1 ns.

The described layout of the muon system ensures a reconstruction efficiency of muon tracks larger than 90% for 100 GeV muons in the entire pseudorapidity range. The precision of the momentum measurement in the muon system is essentially determined by the measurement of the bending angle in the transverse plane. The expected muon momentum resolution using only the muon system, using only the inner tracker, and using both sub-detectors is shown in Figure 3.15 for the barrel and the endcap region.

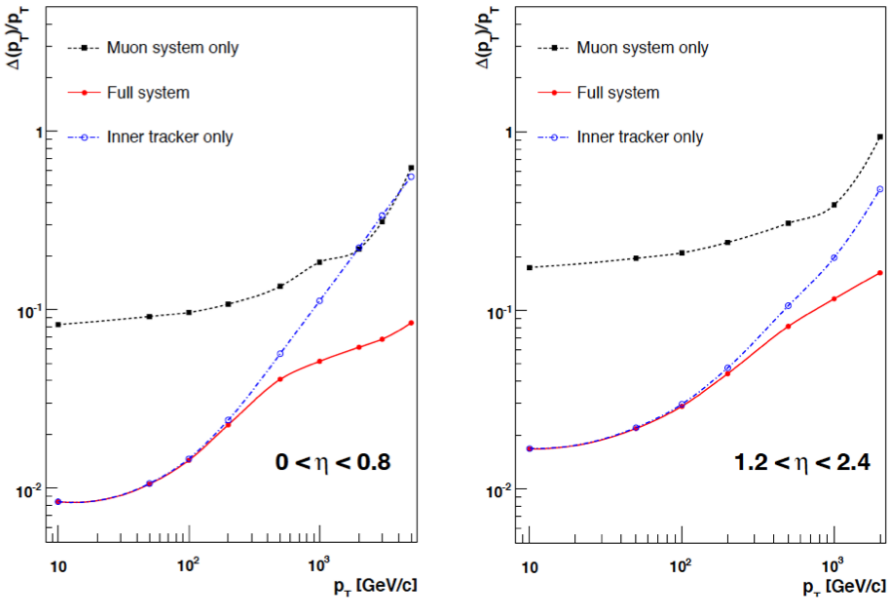

Figure 3.15: The muon transverse momentum resolution as a function of the transverse momentum for muons detected in the barrel (left) and the endcap (right) regions [43]. The resolution is given for the measurement using the muon system or the tracking system only and for a combined method.

3.7 Trigger and Data Acquisition

At the nominal LHC luminosity of 10^{34} cm^{-2}s^{-1}, an average number of $\sim$ 20 interactions is expected every bunch crossing of 25 ns. The CMS trigger system must ensure high data recording efficiency for a wide variety of physics objects and event topologies, while applying online selective requirements to reduce the 40 MHz event rate to an output rate of about 100 Hz, i.e., a data rate of $\sim$ 100 MB/s, allowing for the permanent storage of an event. This leads to a number of formidable experimental challenges. The reduction happens in the Level-1 trigger and High Level Trigger (HLT).

The Level-1 trigger is built of mostly custom-made hardware dedicated to analysing the detector information with a reduced granularity. The Level-1 triggers involve the calorimetry and muon systems, as well as some correlation of information between these systems. The Level-1 decision is based on the presence of trigger primitive objects such as photons, electrons, muons and jets above set E_T or p_T thresholds. It also employs global sums of E_T and E_T^{miss}. It operates at two levels. Firstly the calorimeter and muon triggers process information from their respective sub-detectors. The output is then fed into the global trigger which makes the final decision of the Level-1 trigger.

The calorimeter trigger uses information from individual trigger towers. A trigger tower consists of a single HCAL cell and various ECAL cells within the same η, ϕ region (5×5 crystal array in the barrel and larger in the endcaps). For the ECAL this involves the energy sum and the sum of the transverse energy, for the HCAL the energy sum and presence of minimum ionizing energy. The Regional Calorimeter Trigger (RCT) combines them to find candidate electrons and photons (isolated and non-isolated) along with taus and jets. The candidate information, along with their E_T, are forwarded to the Global Calorimeter Trigger (GCT). An η, ϕ grid of quiet regions is also forwarded to the global muon trigger for isolation cuts. The GCT sorts the RCT candidates by energy and calculates the total E_T and E_T^{miss} from them. The top 4 candidates, along with the global E_T information are forwarded to the global trigger.

The global muon trigger sorts the RPC, DT and CSC muon tracks, normalizes them to the same p_T , $|\eta|$ and $|\phi|$ scale and validates the muon sign. It then correlates the CSC and DT tracks with those of the RPC. Tracks are also deemed isolated if they fall within an area of quiet calorimeter towers. The final set of tracks are sorted by their quality, correlation and p_T. The top 4 energetic muons are sent to the global trigger.

The HLT processes all events accepted by the Level-1 in a single processor farm. The fully programmable nature of the processors on the filter farm enables the implementation of very complex algorithms in the reconstructed event. The strategy follows that of traditional multi-level trigger systems, where the selection process is optimized by rejecting uninteresting events as quickly as possible. With this in mind, each trigger path consists of a sequence of software modules with increasing complexity and physics sophistication. Each module fulfills a well defined task such as reconstruction, intermediate trigger decisions or the final trigger decision for that path. If an intermediate decision on a trigger path is negative, the remainder will not be executed.

All HLT algorithms have been implemented using the CMSSW software suite, which is also be used for reconstruction and offline analysis.

3.8 CMS Detector Commissioning

The CMS detector was assembled in the surface hall prior to the experimental cavern completion. In 2006, the wheels and the disks of the return yoke were closed and a magnet test took place successfully. Afterwards the detector components were completed and lowered in the experimental cavern. Right after the part of the detectors were available in 2007, the data acquisition system was integrated in the central one and several cosmic ray runs without the magnetic field were performed, for a total of more than 300 million events collected in 2007.

3.8.1 Data Collected in 2008

Between May and August 2008, a total of 350M cosmic ray events were collected with the magnet off in common (with all the available sub-detectors) data taking configuration during the four cosmic data taking campaigns called cosmic run at zero tesla (CRUZETs).

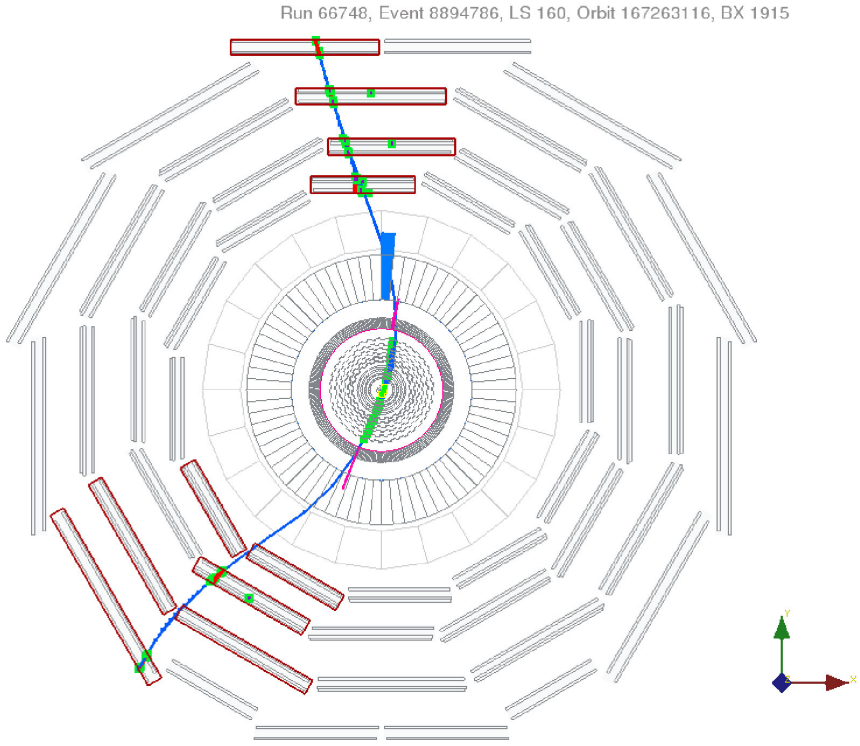

Figure 3.16: A cosmic muon that traversed the barrel muon systems, the barrel calorimeters, the inner strip and pixel trackers.

The CMS detector was ready and protons circulated in the LHC ring in September 2008. Several event types were collected during the first beam days: the beam halo events and the beam splash events (when a single beam of 2×10^9 protons was dumped on the closed collimators 150 meters upstream). These events contain horizontal particles, useful for the forward detectors commissioning and in particular the splash events delivered large energy deposits in the calorimeters. After the accident on 19 September 2008, the CMS experiment was kept closed and a long cosmic run at the nominal magnetic field was taken during cosmic run at four Tesla 08 (CRAFT'08). CRAFT'08 collected 290 million events at the magnetic field of 3.8 T, 87% of the events have a muon track in the muon chambers, 3% of the events have a muon track with tracker hits and about 30000 events have pixel hits. These runs represented a very useful data sample to complete the CMS commissioning with the real physics events. In particular for each sub-detector it was important to evaluate the efficiency and eventually to repair faulty channels, and to check the resolution and to improve the detector performance measuring the calibrations

and the alignment constants. The other essential aim of the global runs was to integrate the data acquisition, trigger systems and other online tools such as the detector control system and data quality monitoring. Figure 3.16 shows a typical event with a cosmic muon traversing the whole detector.

3.8.2 CRAFT'09 and First Collisions in 2009

After the final commissioning steps of the magnet, CMS started the continuous operation of the CMS data-taking with cosmic ray muons, called CRAFT'09 (Cosmic Run At Four Tesla 09). The runs successfully collected at least 160 Million events at the magnet field of 0 T and 300 Million events at the magnet field of 3.8 T in order to prepare the detector for the LHC data taking.

Delivered by the LHC at a centre-of-mass energy of 900 GeV, the first collisions recorded by the CMS detector in November 2009 were used to commission the particle-flow event reconstruction algorithm. The event displays for the first collisions are shown in Figure 3.17, 3.18 and 3.19. The first results on particle-based jets, missing transverse energy, isolation and tau identification were presented and confirm the performance predicted by the simulation [47].

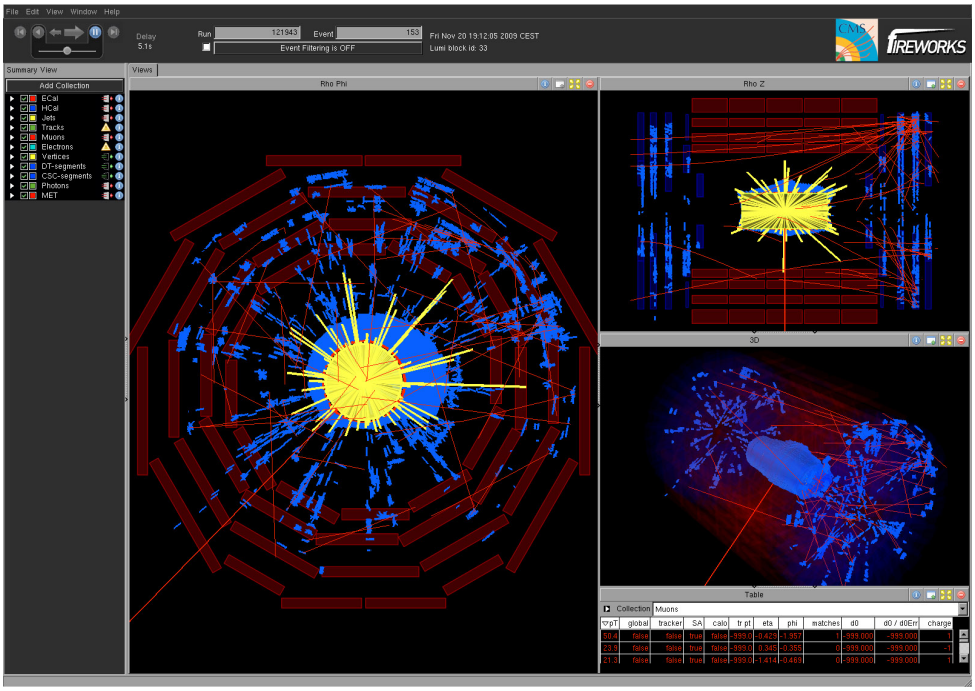

Figure 3.17: The first CMS event displays from LHC running on 20 November 2009: a splash event [48].

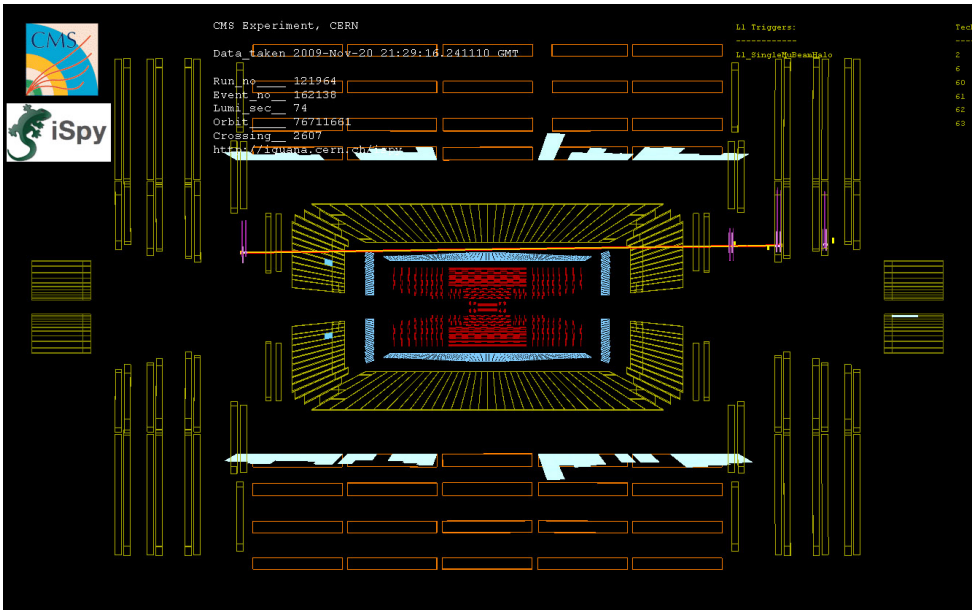

Figure 3.18: The first CMS event displays from LHC running on 20 November 2009: a halo muon [48].

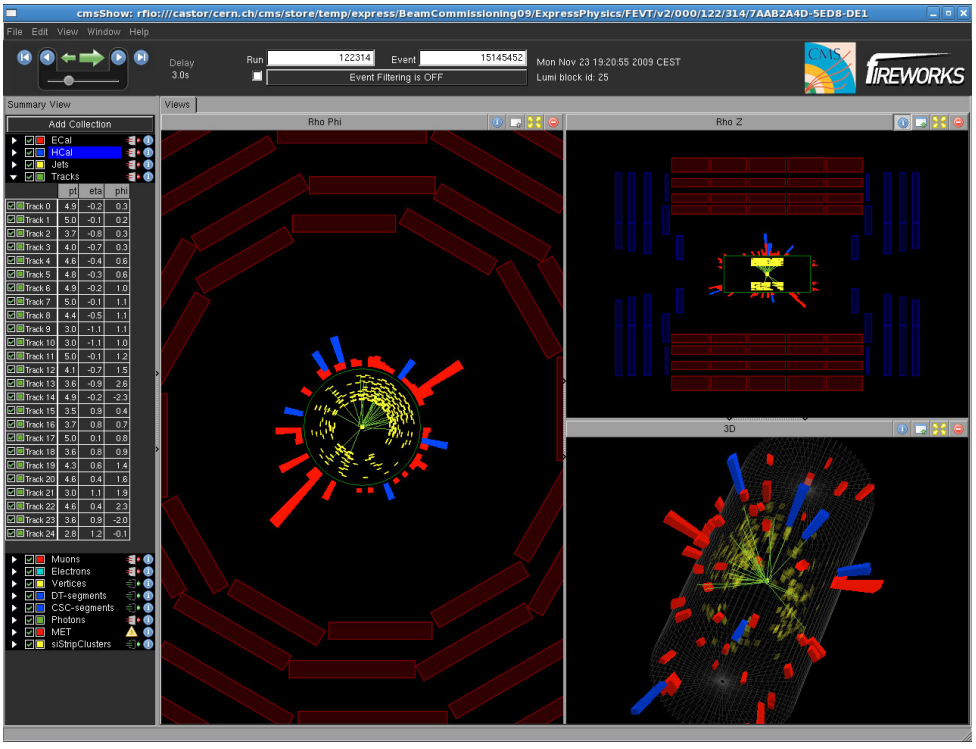

Figure 3.19: CMS 900 GeV collision candidates from 23 November 2009 [48].

3.8.3 7 TeV Collisions at CMS in 2010

After the shutdown in the winter of 2009, the LHC was restarted and the beam was ramped up to 3.5 TeV per beam. On 30 March 2010, the first collisions took place between two 3.5 TeV

beams, which set a new world record for the highest-energy man-made particle collisions. The sub-detectors reached the expected performance and showed very high availability, as shown in Figure 3.20. By the middle of July 2010, commissioning is essentially finished. The Integrated luminosity delivered by LHC and recorded by CMS was increasing very fast, as shown in Figure 3.21. Excellent detector performance has been archived since the very beginning of data taking. Calibration is advancing very fast with the LHC luminosity ramp-up.

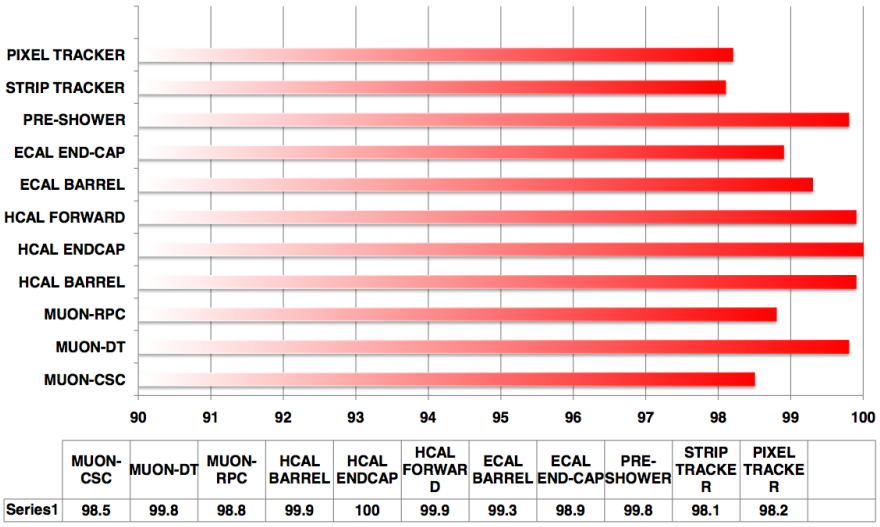

Figure 3.20: High availability of the channels in the CMS sub-detector systems in July 2010 [49].

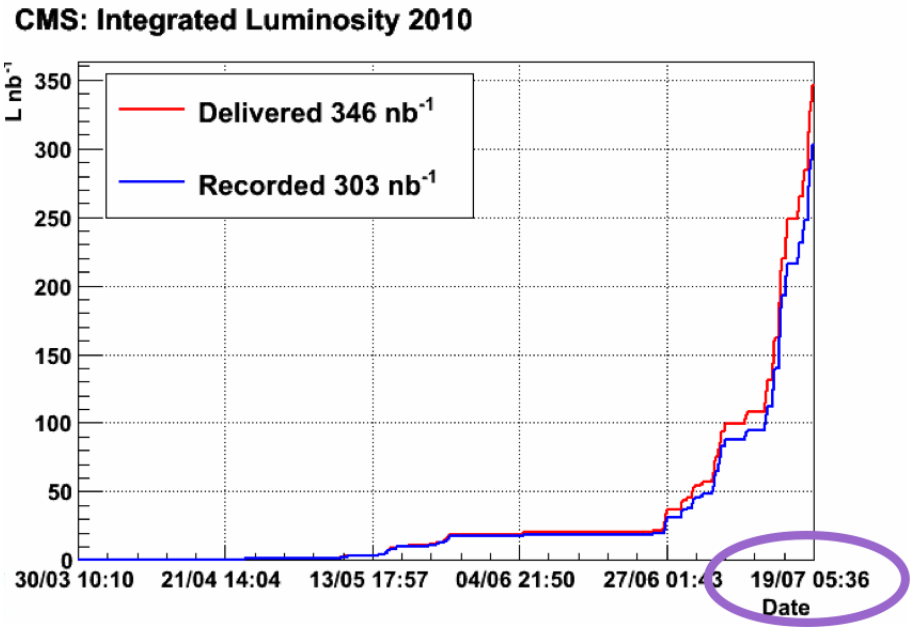

Figure 3.21: Integrated luminosity delivered by LHC and recorded by CMS till July 2010 [49].

3.9 Summary

The CMS design has been optimized to fulfill the LHC environment and to reach the physics goals, with emphasis on precision measurements of photons, electrons and muons to search for Higgs and new physics phenomena. The CMS detector was constructed successfully and the commissioning of the CMS detector started in 2007. The large quantity of the cosmic ray data were collected to commission and to validate the sub-detector systems, as well as to test the off-line data analysis system during 2008 and 2009. The sub-detectors reached the expected performance and showed very high availability. The beam collisions at 7 TeV at the LHC started on 30 March 2010, providing the first data sample for the LHC physics studies. Till the middle of July 2010, commissioning has been essentially finished. Excellent detector performance was archived since the very beginning of data taking. Calibration is advancing very fast with the LHC luminosity ramp-up.

4 Worldwide LHC Computing Grid and CMS Computing

The experiments at the LHC will produce roughly 15 PB of data annually at the design luminosity of 10^{34} $cm^{-2}s^{-1}$. Thousands of scientists around the world access and analyse those data. It is impractical to build one computing center in one location, which can fulfill such an enormous amount of data processing and storage. Therefore, the data processing and storage has to take place in hundreds of computing sites distributed around the world. To meet this challenge, the Worldwide LHC Computing Grid (WLCG) project, was approved by the CERN Council on 20 September 2001 [4]. The goal of the project is to develop, build and maintain a distributed computing infrastructure for storage and analysis of data from the LHC experiments.

In Section 4.1, the challenging demands of the LHC experiments for computing power, data storage and management capability are introduced. The basic concept of Grid and the reason why it is chosen as the solution for this challenge will be discussed in Section 4.2. The WLCG architecture, services and CMS computing model are described in detail from Section 4.3 to 4.5.

4.1 LHC Experiments' Requirements

The LHC experiments will generate roughly 15 PB of data annually at the design luminosity of 10^{34} $cm^{-2}s^{-1}$ [50], which will be accessed and analysed by thousands of scientists in hundreds of research institutes and universities around the world. All data needs to be available over the estimated 15-year lifetime of the LHC. Before the data gets available to physicists for physics studies, the collision events filtered through online triggers need to be reprocessed and skimmed at large-scale computing centers. In addition, the detailed Monte Carlo simulation of the physics processes and the detector responses also requires large-scale computing power and huge amount of storage. It's estimated that the LHC computing requires a CPU capacity of 140 million

SPECint2000 [1]. A traditional approach would be to centralize all the capacity at one location near the experiments. However, it is impractical to build one computing center nearby LHC, which can fulfill such an enormous amount of the data processing and storage. Accordingly, a novel globally distributed model for data storage and analysis – Grid computing – was chosen because it is able to provide several key benefits [4]. In particular:

- The cost to maintain and upgrade necessary resources for such a computing challenge is more easily operated in a Grid computing environment, where individual institutes and participating national organizations can fund local computing resources while still contributing to the LHC experiments;
- a Grid computing system can significantly reduce single-failure-points. Multiple copies of data and automatic management for computational tasks ensures load balancing of resources and facilitates, independent of geographical location.

Of course, a distributed system also presents a number of major challenges, including:

- Ensuring adequate network bandwidth between distributed resources;
- managing and protecting the data over the lifetime of the LHC;
- maintaining coherence of software versions deployed on distributed heterogeneous hardwares;
- providing policy-based accounting mechanisms;
- providing a transparent way for data access and process as efficient as possible;
- providing geographical distributed physicists a uniform way for data analysis.

4.2 Grid Computing

A Grid can be considered to be a collaborative group of the computers and the storage systems, communicating via the Internet. Whereas the Internet provides the seamless access to the information hosted on computers all over the world, the Grid aims to provide the seamless access to the computing power and storage systems distributed across the world.

Many factors introduce the complexity to share computing power and storage systems, including:

- Heterogeneous hardwares as well as operating systems;

[1] SPECint2000 is an integer benchmark suite maintained by the Standard Performance Evaluation Corporation (SPEC). It provides a comparative measure of compute intensive performance across the widest practical range of hardware. The measure has been found to scale well with typical HEP applications. As an indication, a powerful Pentium 4 processor delivers 1700 SPECint2000.

- resource discovery and fair share policies of resources for all Grid users;
- the security and traceability for owners of a Grid infrastructure;
- strategy for collaboration on a global scale, i.e., each contributing site could have different policies;
- high availability of Grid resources.

Although these issues bring lots of technology challenges, there is great potential for Grid computing to cause a revolution on the same scale as the Internet has.

4.2.1 Definition of Grid Computing

Ian Foster presents a definitive three point checklist defining the Grid as a system [51]:

- Coordinates resources that are not subject to centralized control;
- uses standard, open, general-purpose protocols and interfaces;
- delivers non-trivial qualities of service.

Systems of the Grid need standard protocols and interfaces to provide services to each other. Therefore, the second point on the checklist implies that a common infrastructure should be defined to provide functions such as: the authentication, the authorization, the resources discovery and the resources access. Some of the emerging standards in Grid computing will be presented in Section 4.2.2.

4.2.2 Grid Standards

Grid standards are essential to ensure connectivity between components of a Grid and connectivity between different Grids. Web Service and the other Internet standards, which most Grid standards are based on, will be briefly outlined. The Open Grid Forum (OGF) and the Open Grid Service will also be introduced, followed by a description of the Web Service Resource Framework and the Globus Toolkit.

Web Service

The Internet is the base of Grid computing. Several main Grid standards had employed Web Service to ensure the interoperability over Internet. Web Service, defined by the main international standards organization World Wide Web Consortium (W3C), allows communication between applications running on different platforms and developed in different programming languages over

Internet. This is accomplished via a standard mechanism for data exchanges, e.g., using eXtensible Markup Language (XML). Another standard, WSDL (Web Service Description Language), is used for providing the means to access Web Service.

Open Grid Service

Grid Service is designed to integrate Web Service and Grid technology. Open Grid Services Architecture (OGSA) [52] and Open Grid Services Infrastructure (OGSI) [53] were created by the Global Grid Forum (GGF). In OGSA, each entity in a Grid environment becomes a set of Web Services, allowing access components via a common framework.

OGSI is a companion standard that formally specifies Grid services in more technical detail. For example, OGSI defines interfaces and protocols for the interaction of Grid services. The use of OGSI ensures interoperability between different Grid platforms based on OGSA.

The Globus Toolkit

The Globus Toolkit (GT) is an open source project developed by the Globus Alliance [54], aiming to provide the essential software infrastructure for building Grid and Grid applications. The key components of GT include: security, information services, data management and resource management. The object-oriented approach of GT and its open source license allows developers to use GT free and develop Grid middleware easily. The WLCG middleware originated from GT. Some elements of GT will be discussed in the overview of components of a typical Grid in Section 4.2.3.

4.2.3 Components of a Typical Grid

This section presents an overview of the general components in a 'typical' Grid.

Security

Security is one of the foremost considerations to develop a Grid system. A security infrastructure must provide a robust system to deter illegal access to resources. Moreover, Grid systems are obligated to trace any possible misuse of resources. The main requirements of a Grid security infrastructure are mechanisms for authentication, authorization and encryption.

The Globus Grid Security Infrastructure (GSI) is based on the use of Grid certificates. A Grid

certificate is a 'digital identity' which uses public-key cryptography[2] to identify genuine users. Regional Certification Authorities (CAs) issue certificates to users after registration. Certificates use X.509 standard[3]. Each certificate has an accessible public part for user information and a password-protected private portion which is used to confirm their identity. A Grid user typically creates a proxy-certificate, which is valid for a finite time period, e.g., 48 hours. GT provides credential management services such as MyProxyServer to minimize unnecessary human involvement in automated operations.

Information Service

Any Grid system requires information of connected resources. The information is used for many essential tasks of the Gird, e.g., testing overall configuration of systems, collecting resource usage statistics and site administration.

In GT, the Grid Index Information Service (GIIS), also referred as the Monitoring and Discovery Service (MDS), collects Grid resources information for resource broker and Grid job scheduler.

The implementation of WLCG information system will be discussed in Section 4.3.

Job Scheduling

Job Scheduling is the process of finding suitable resources for the submitted job when a job is submitted to a Grid. GT does not provide concrete mechanisms for job scheduling, although several elements can be used to implement it.

The job scheduling of WLCG will be discussed in Section 4.3.

Data Management

Data management in a Grid system that covers many aspects including data storage and access. A set of tools must be provided by a Grid to facilitate data movement and replication between sites.

In GT, the middleware for data replication operations is known as the Data Replication Service (DRS). To archive robust data management, most Grids implement file catalogues with DRS,

[2]Public-key cryptography is a cryptographic approach which involves the use of asymmetric key algorithms instead of or in addition to symmetric key algorithms. Unlike symmetric key algorithms, it does not require a secure initial exchange of secret keys to both sender and receiver.

[3]In cryptography, X.509 is a standard for the public key infrastructure (PKI) for single sign-on and Privilege Management Infrastructure.

which will be discussed in Section 4.5.1. While the DRS offers tools for discovery and replication of files, file catalogues maintain records of data, which make data management independent of other Grid components such as the information system.

GT also provides tools for data transfer, such as GridFTP (Grid File Transfer Protocol). GridFTP integrates FTP (File Transfer Protocol) for data transfer and GSI for user authentication and authorization, which provides a secure, fast and reliable mechanism for data transfer on the Grid.

Job Management

Once a job has been scheduled to a particular resource, a set of services are necessary to enable job execution, job monitoring and output retrieval. The GT middleware Grid Resource Allocation Manager (GRAM) provides those functions.

4.3 Worldwide LHC Computing Grid (WLCG)

As a Grid, the WLCG consists of an set of Grid services and applications running on the Grid infrastructures developed by the WLCG partners [50]. These infrastructures are provided by the Enabling Grids for E-scienceE (EGEE) project in Europe, the Open Science Grid (OSG) project in the USA and the Nordic Data Grid Facility in the Nordic countries. The EGEE infrastructure brings together many of the national and regional Grid programs into a single unified infrastructure. In addition, many of WLCG sites in the Asia-Pacific region run the EGEE middleware stack as integral parts of the EGEE infrastructure.

The essential Grid services provided to the LHC experiments are based on the demand of the experiments and the agreements between WLCG and the sites.

To meet the challenge of large-scale data processing and management, the WLCG developed innovative use of its distributed computing resources and mass storage management systems to organize a hierarchical distribution of the data. The data transfer requires high-speed point-to-point replication facilities and a system checks and maintains the data consistency. In addition to backup, duplicated data is used to balance the search and network loads according to the response of sites. Consequently, the computing model adopted by the LHC collaborations is designed as a tiered distributed hierarchy of 'Regional Centers'.

4.3.1 Hierarchical Architecture

The WLCG organizes computing centers in a 'tiered' hierarchy, as illustrated in Figure 4.1. Data coming from the experiment data acquisition systems is written to the tape in the CERN Tier-0 facility, and a second copy of the raw data is simultaneously transfered to Tier-1 sites, with each site accepting an agreed share of the raw data. The sharing policy depends on the computing models of the experiments.

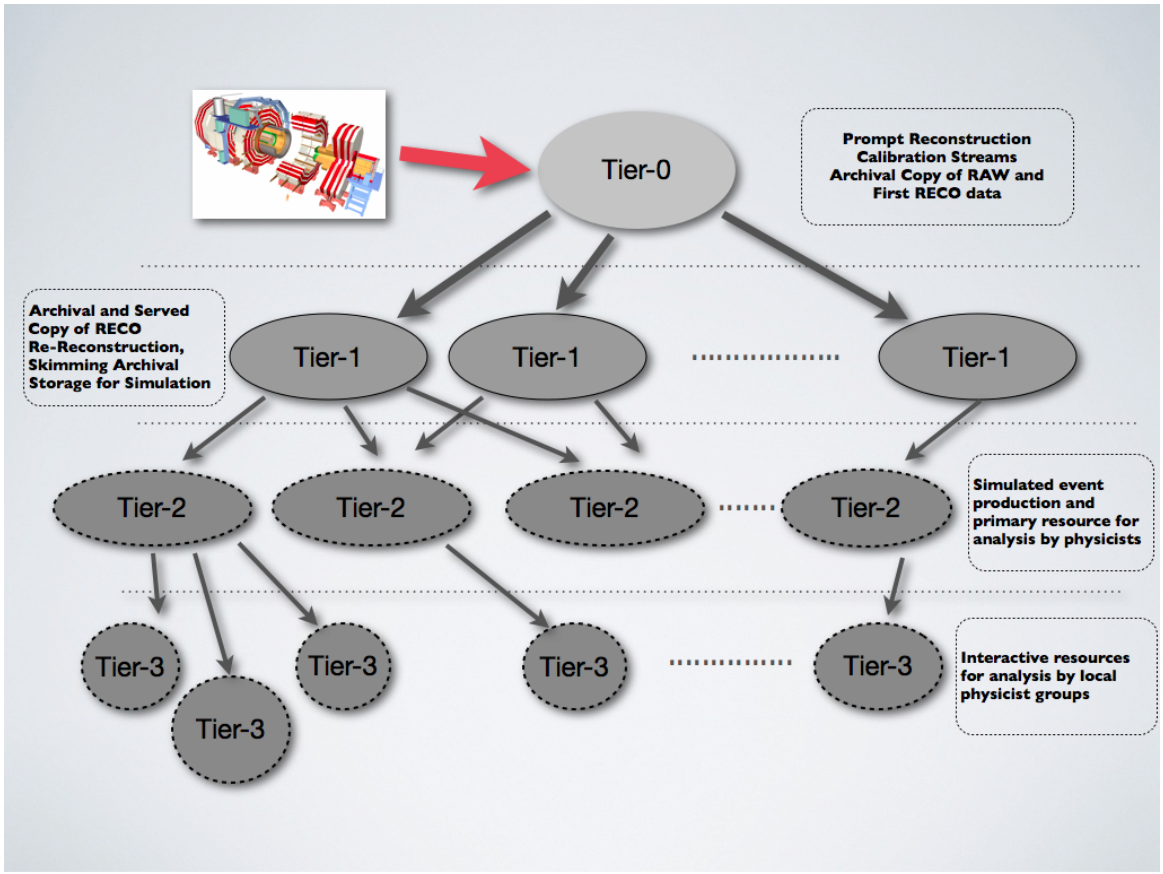

Figure 4.1: Tiered hierarchy of the WLCG.

The functionalities of four tiers of WLCG sites can be summarized as follows:

- **Tier-0:** The original raw data emerging from the data acquisition systems of the experiments is recorded by the Tier-0 center at CERN. The first-pass reconstruction takes place at the Tier-0 center, where a copy of the reconstructed data is stored. The Tier-0 distributes a second copy of the raw data to the Tier-1 centers associated with the experiment. Additional copies of the reconstructed data are also distributed to Tier-1s. As this data arrives at the Tier-1, it must ensure that it is written to tape and archived in a timely manner.
- **Tier-1:** There are 11 Tier-1 centers[4] (including one at CERN). These centers keep another

[4]Till June 2010

archive copy of the data, and are responsible for performing re-reconstruction of older data with improved calibration and algorithms, and making skims of primary datasets that are enriched in particular physics signals. They also provide archival storage of simulated samples produced at Tier-2 centers.

- **Tier-2:** There are 162 tier-2 centers[4] (including one at CSCS, Manno in Switzerland). The role of Tier-2 centers is to provide computational capacity and appropriate storage services for Monte Carlo (MC) event simulation and for end-user analysis. Because Tier-2 centers have no tape systems, they have to obtain data as required from Tier-1 centers and transfer the data generated at Tier-2 centers to Tier-1s for permanent storage.
- **Tier-3:** Other computing facilities in universities and laboratories take part in the processing and analysis of LHC data. They are called Tier-3 centers. Comparing with the setup of Tier-2 centers, that of Tier-3 centers are optimized for data analysis of the local physicist community.

4.3.2 Fundamental Resource Services

Like other Grids, LHC and its Tier centers provide their computational resource by means of providing a set of services. The services that various WLCG sites must provide, resource requirements and availability targets, are defined in a Memorandum of Understanding (MoU) [5] that is signed by each site. The key services include:

Storage Element Services

A Storage Element (SE) is a logical entity that provides the following services and interfaces [50]:

- **Mass Storage System (MSS)**, either disk cache or disk cache front-end backed by a tape system. Mass storage management systems currently in use include CASTOR, Enstore-dCache, HPSS and Tivoli for tape+disk systems, dCache, LCG-dpm and DRM for disk-only systems.
- **Storage Resource Manager (SRM)** interface provides a universal way to access the MSS no matter what the implementation of the MSS. It defines a set of functions and services that a storage system provides in a MSS-implementation independent way. Existing SRM implementations currently deployed include CASTOR-SRM, dCache-SRM, DRM/HRM and LCG-dpm.
- **GridFTP** is the basic-level data transfer service in WLCG. The implementation of GridFTP

service is scaled to the bandwidth required. Normally the GridFTP transfer is invoked indirectly via the FTS or through SRM commands.

- **Local POSIX-like input/output facilities** enable applications access to the data on SE in the local site. Currently this is available through rfio, dCap, AIOD and rootd, according to the implementation of the Mass Storage System. Various mechanisms for hiding this complexity also exist, including the Grid File Access Library in LCG-2 and the gLiteIO service in gLite. Both the mechanisms include connections to the Grid file catalogues to enable an application to open a file based on LFN (Logical File Name) or GUID (Grid Unique Identifier).
- **Authentication, authorization and audit/accounting facilities:** SE provides and respects ACL (Access Control List) for files and datasets. The access control bases on the use of extended X509 proxy certificates with a user Distinguished Name (DN) and attributes based on VOMS (Virtual Organization Membership service) roles and groups. It is fundamental that a SE provide necessary information to allow tracing of all activities at regular intervals, permitting audit on activities. It also provides information and statistics on the use of the storage resources, according to the schema and policies.

A site may provide multiple SEs for different qualities of storage. For example, it may be considered convenient to provide a SE for data intended to keep for extended periods and a separate SE for data that is transient. Large sites with MSS-based SEs may also deploy disk-only SEs for such a purpose or for general use.

File Transfer Services

The basic-level data transfer service is GridFTP. This may be invoked directly via the globus-url-copy command or through the srmcopy command which provides 3rd-party copy between SRM systems. However, for reliable and robust data transfer, gLite File Transfer Services (FTS) above srmcopy or GridFTP was implemented. The service is installed at the Tier-0 (for Tier-0 $\leftrightarrow$ Tier-1 transfers) and at the Tier-1s (for Tier-1 $\leftrightarrow$ Tier-2 transfers). It can also be used for 3rd-party transfers between sites that provide an SE. No service needs to be installed at the remote site apart from the basic SE services described above. However, tools are available to allow the remote site to manage the transfer service.

Computing Element Services

The Computing Element (CE) services provide an interface between computing resource managed by a local batch system running on a computer cluster and WLCG. Typically a CE provides

access to a set of job queues of the local batch system.

A CE provides the following general functions and interfaces:

- A mechanism by which work may be submitted to the local batch system. This is implemented typically at present by the Globus gatekeeper in LCG-2 and Grid/Open Science Grid. NorduGrid (the ARC middleware) uses a different mechanism.
- Publication of accounting information in an agreed schema and at agreed intervals. Presently the schema used in both LCG-2 and OSG follows the GGF accounting schema.
- Publication of information through the LHC information system and associated information providers.
- A mechanism by which users or Grid operators can query the status of jobs submitted to that site.
- A CE and the associated local batch systems provide authentication and authorization mechanisms based on the VOMS model. It is implemented in terms of mapping DN of a Grid user to a local user and group. The basic requirement is: the user presents an extended X509 proxy certificate, which includes a set of roles, groups and subgroups for who is authorized, and the CE/batch system respects those through appropriate mappings locally.

4.3.3 Middleware

The logical layer of software, which connects all Grid elements, is the so-called middleware. The description of middleware in this thesis is restricted to the gLite middleware. The middleware implements the Grid services and client software, while trying to hide the complexity of the Grid environment from users and applications, giving the impression that all of these resources are available in a virtual computer center. The following sections describe the middleware components, as well as the relation between middleware components as sketched in Figure 4.2.

Virtual Organizations

A virtual organization (VO) is a dynamic collection of individuals, institutions, and resources that is defined by certain sharing rules. In that sense a VO might represent an experiment collaboration as in the case of the WLCG. A single user asks for a Grid certificate through a Certification Authority (CA), which issues the user a personal Grid certificate (X.509 certificate). With this certificate a user can request the membership to a certain virtual organization like CMS. This certificate is then the key (authentication and authorization) to all resources belonging to

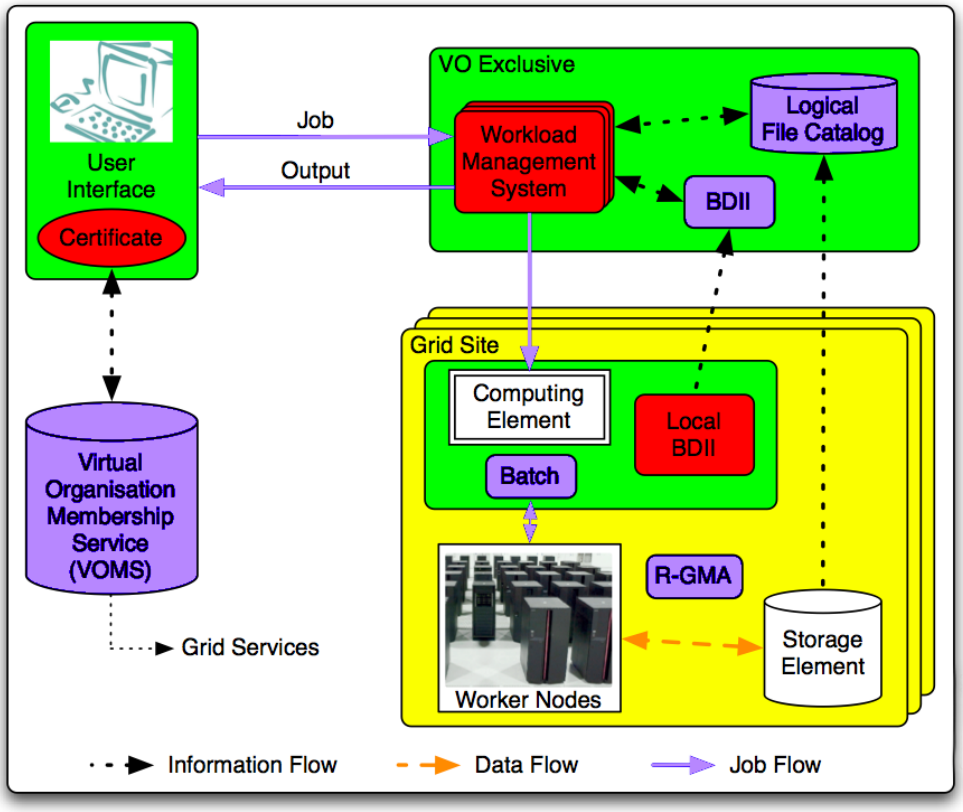

Figure 4.2: WLCG middleware infrastructure and the components.

the virtual organization. For security reasons, proxy certificate, which is a temporary copy of the certificate, but with a limited lifetime of typically some hours or days, are delegated across the Grid. For example, they can be attached to the Grid job for authorization and authentication. Following the Grid principles all users within a certain virtual organization are equal and share the resources on a fair basis. However, authorized users may equip themselves with different roles within a VO such as software manager or Monte Carlo production operator. Also VO sub-groups are supported, which allow users affiliated with a Swiss university or laboratory to obtain higher priorities for processing at the Swiss Grid sites.

The User Interface

The access point to the WLCG Grid is the User Interface (UI). This can be any machine which has the gLite UI software installed. It can be compared to the web browser as an interface to the World Wide Web, although the UI for the WLCG is performed via a set of command line tools instead of a graphical user interface. The UI provides access to the functionalities offered by the information, workload and data management systems, such as:

- discovery of all resources suitable for the execution of a given job;
- job submission and cancelation;
- status checks for submitted jobs;
- output retrieval for finished jobs;
- access to logging and bookkeeping information of jobs for debugging purposes;
- copy, replication, and deletion of files from/to the Grid storage elements;
- retrieval of the status of different resources from the information systems.

The Information System

The information system is a critical part of the Grid infrastructure, which allows users and services to discover which resources and services are available within the Grid or at a certain site. The precision and up-to-date of the information determine the quality of the service of the whole Grid.

At a Grid site the CEs and SEs are equipped with so-called information provider software, which generate data about the resource (e.g., general availability/status, free/used storage space/batch slots). The data of the different information providers are aggregated by a local/site-level BDII (Berkeley Database Information Index). The BDII middleware stores and publishes the data. Finally a top-level BDII polls the data from all available sites within the specific Grid. Effectively the top-level BDII defines a view of the overall Grid resources and serves, e.g., as an input source for the workload management systems. A different source of information is the R-GMA (Relation-Grid Monitoring Architecture). While BDII is based on Lightweight Directory Access Protocol (LDAP) information system, R-GMA provides data as a global distributed relational database. R-GMA is used for accounting and both system- and user-level monitoring.

The Workload Management System

The Workload Management System (WMS) acts as job distributor and load balancer in WLCG. Its task is to accept jobs and to assign them to the most appropriate computing element. The WMS regularly checks the status of the jobs and retrieves the output upon the end of each job. By calls to the WMS via UI, the user can get information about the jobs.

The user can specify certain requirements within the jobs, such as the operating system, the SE, input files, or time requirements. Upon the submission of a job into WLCG it is handed over to one of the independent WMSes of the VO. Among all available CE, which fulfill the requirements

expressed by the user, the WMS passes the job to the CE with the best ranking. The ranking is based on quantities derived from the CE status information expressing the quality of the CE (typically a function of the numbers of running and queued jobs). In addition to the submission of single jobs, WMS allows to submit a collection of jobs in bulk. This allows for a much more efficient job submission and improves the limit of jobs/day hit.

Monitoring and User Support

As an evolving system, a key component of WLCG is a reliable and up-to-date monitoring. Apart from the site and experiment specific monitoring, the central WLCG/EGEE monitors the basic functionality of all Grid sites by submitting test jobs regularly. Only sites which pass these so-called Site Availability Monitoring (SAM) tests, are visible in the top-level BDII and thus are available for user jobs. These tests do not only spot problems, but equip the Grid with a robustness against failures: unstable sites are flagged and the jobs are routed to more reliable clusters.

The Global Grid User Support (GGUS) provides centralized support for WLCG sites and users. The service consists of a ticket system for an efficient solution of problems by the direct involvement of Grid site administrators and Grid experts. In addition, known bugs are tracked, lists of frequently asked questions and documentation are maintained. The GGUS portal is supposed to be the key entry point for Grid users looking for help.

4.4 The Overview of CMS Computing Model

CMS has been developing its distribution computing model based on the LHC Computing Grid from the very early days of the experiment [55]. The goal of the CMS computing system that is tightly integrated with WLCG is to implement a dedicated Grid environment to support the storage, transfer and manipulation of the recorded data for the lifetime of the CMS experiment. The system accepts real-time detector information from the data acquisition system at the experimental site; ensures the safety 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, e.g., equipment management data, configuration data and calibration data.

The key components of the CMS Computing Model 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, to locate, and to process large collections of the CMS events;
- underlying generic Grid services giving access to distributed computing resources; the computing centers, managing and providing access to storage and CPU at a local level.

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

4.4.1 Event Data Model and Data Flow

CMS uses a number of the event data formats with varying degrees of details, sizes and refinement to meet the different requirements of different data processing tasks. Starting from the raw data produced from the online system successive degrees of processing refine this data, apply calibrations and create higher-level physics objects.

The CMS DAQ system writes the DAQ-RAW (Detector data, L1 + HLT information) events (1.5MB) to the HLT computer cluster input buffer. The HLT computer cluster writes RAW events (1.5 MB) at a rate of 150 Hz. RAW events are classified in $\mathcal{O}(50)$ primary datasets[5] depending on their trigger history (with a predicted overlap of less than 10%). Primary datasets are grouped into $\mathcal{O}(10)$ online streams in order to optimize their transfer to the offline farm and the following reconstruction process.

The first event reconstruction is performed immediately on the Tier-0 cluster at CERN which writes RECO (reconstructed physics objects and hits/cluster) events. The RAW and RECO versions of each primary dataset are archived on the Tier-0 mass storage system, a copy is transferred to a CMS Tier-1 which takes backup responsibility for this. The transfer to other Tier-1 centers is subject to additional bandwidth being available. Thus RAW and RECO versions of dataset are available either in the Tier-0 archive or in at least one Tier-1 center. The AODs (Analysis Object Data, which contains reconstructed physics objects, some hits information for physics analysis) which are derived from RECO events and contain a copy of all the high-level physics objects plus a summary of other RECO information sufficient to support typical analysis

[5]In order to take advantage of the highly parallelized CMS data processing model and make the data access easier for the Physics groups, the data is split into Primary Datasets, where events are grouped according to similar analysis use-cases. The original full set of triggers (about 150) has been reduced to a core set (about 30-50), still capable of retaining the same coverage. This makes for a simpler, more robust, and easier to maintain startup trigger table.

actions (e.g., the re-evaluation of calorimeter cluster positions or track refitting, but not pattern recognition) are produced in the Tier-0 reconstruction step and distributed to the Tier-1 centers (one full copy at each Tier-1).

The CMS Tier-1 centers produce subsequent AOD versions, and distribute these new versions between themselves. The additional processing (skimming) of RAW, RECO and AOD data at the Tier-1 centers is triggered by the requests from physics groups and produce second and third (etc.) generation versions of the AOD which contain high level physics objects and pointers to events (e.g., the run and event number), which allow their rapid identification for further study.

Selected skimmed data, all AOD of selected primary streams, and a fraction of RECO and RAW events are transferred to Tier-2 centers. Those Tier-2 centers, each consisting of one or several collaborating computing facilities, provide capacity for analysis, calibration activities, and Monte Carlo simulation. CMS Tier-2s support iterative analysis of authorized groups of users. Grouping is expected to be done not only on a geographical but also on a logical basis, e.g., supporting physicists performing the same analysis or the same detector studies. Individual scientists access these facilities through Tier-2/3 computing resources. Corresponding to their tasks the different Tiers have to meet certain resource requirements for CMS.

So, the CMS computing system is geographically distributed. Data are spread over a number of centers following the physical criteria given by their classification into primary datasets. Replication of data is given more by the need of optimizing the access of most commonly accessed data than by the need to have data 'close to home'. Figure 4.3 shows the Tier centers in the CMS computing model and the schematic flow of the real event data.

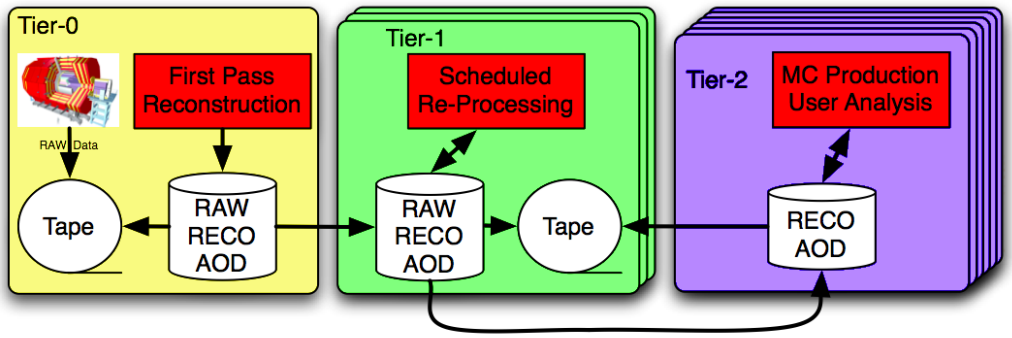

Figure 4.3: Schematic flow of bulk (real) event data in the CMS Computing Model. Not all connections are shown, e.g., flow of MC data from Tier-2's to Tier-1's or peer-to-peer connections between Tier-1's.

4.4.2 Application Framework

The CMS application software is designed to be capable of a variety of event processing, selection and analysis tasks, and is used in both offline and online contexts. The software is modular that it can be developed and maintained by different groups of geographically distributed collaborators. The chosen architecture consists of a common framework which is adaptable for all of the computing environment, physics modules which plug into the framework via a pre-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 [56].

The Event is the central concept of the CMS data model. It provides access to the recorded data from a single triggered bunch crossing, and to new data derived from it including raw digitized data, reconstructed products, or high-level analysis objects. 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 [57, 58].

Event is operated by a variety of physics modules, which may retrieve data from it, or append new data, with provenance information automatically included. Each module performs a pre-defined function relating to the selection, reconstruction or analysis of events. Several module types exist, each with a specialized 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.

Modules are isolated 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, as shown in Figure 4.4.

4.5 CMS Computing Services and Operations

The CMS distributed computing system is based on services implemented by the Worldwide LHC Computing Grid (WLCG). The overall architecture of the CMS computing system along

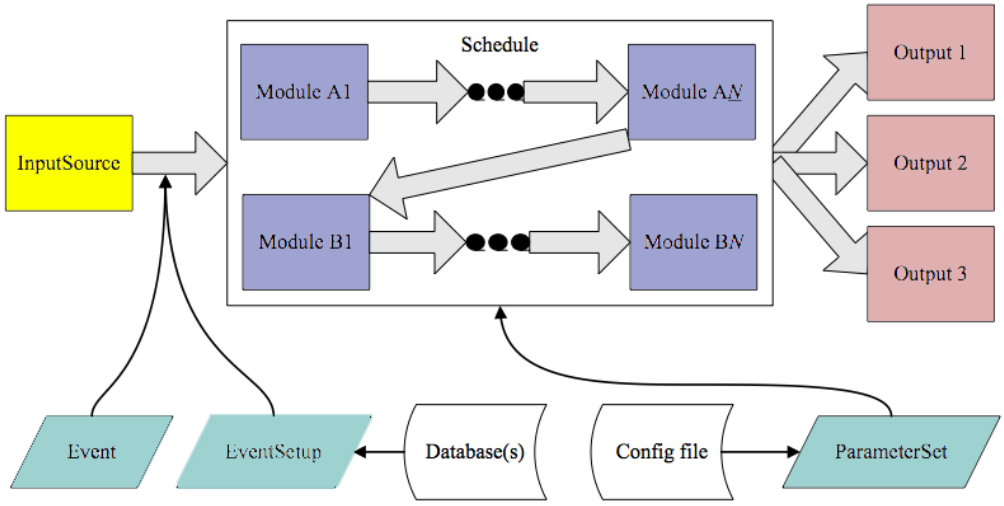

Figure 4.4: Modules within the CMS Application Framework [55].

with the most important systems and services can be divided into a Grid Workload Management System, a CMS Data Management system, and other CMS-specific services, needed to support specific needs of experiment applications and software, as shown in Figure 4.5. A CMS Workflow Management system holds all of these pieces together into a coherent system supporting all CMS necessary workflows (data re-reconstruction, calibration activities, Monte Carlo production, AOD production, skimming and general user analysis) and shields users/operators of these systems from the full complexity of the underlying architecture.

4.5.1 The CMS Data Management System

The CMS data management is based on a set of loosely coupled components which allow physicists to discover, access and transfer event data.

Data Organisation

It's not feasible to manage and transfer data on physics event level. The computing system needs to support both physicist abstractions, such as 'dataset' and 'event collection', as well as physical 'packaging' concepts native to the underlying computing and Grid systems, such as files.

CMS defines an 'event collection' as the smallest unit that a user is able to select through the dataset bookkeeping system described below, i.e., without using an analysis application which reads individual events. An event collection may correspond to the event data from a particular

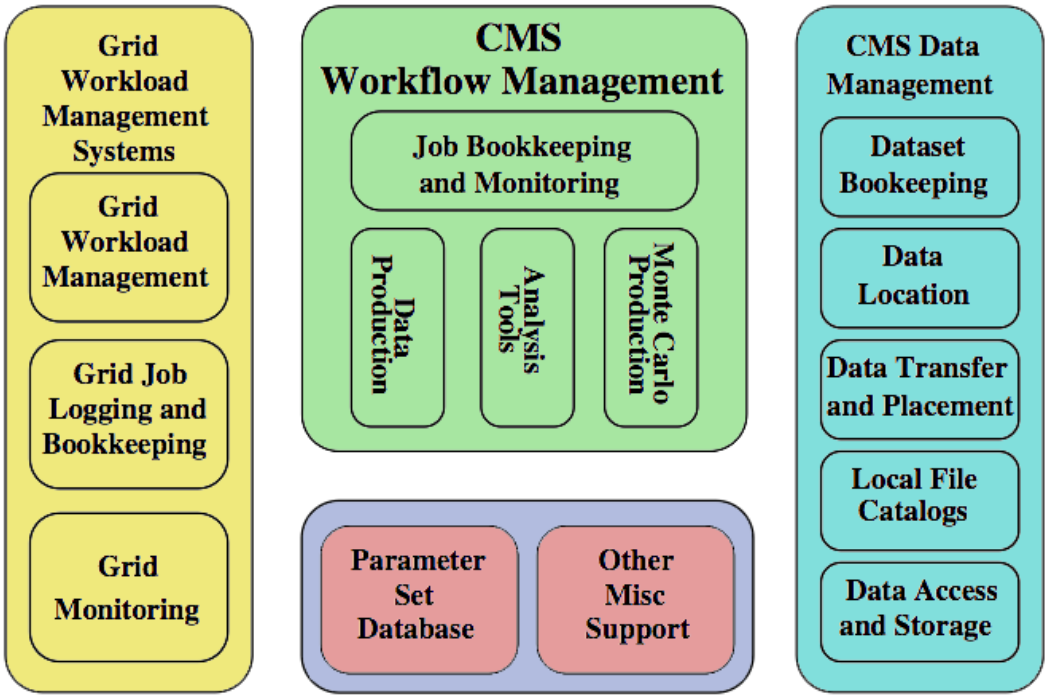

Figure 4.5: Overview of systems and services supporting the CMS workflow management system [59].

trigger selection from one given 'run'.

CMS generically defines a dataset as any set of 'event collections' that would naturally be grouped and analysed together as determined by physics attributes, like their trigger path or Monte Carlo physics generator, or by the fact that they represent a particular object model representation of those events (such as the RAW, FEVT, RECO and AOD data formats).

In the CMS computing model, event collections are the basic job configuration concept used at application run-time, while dataset as a concept is only used by physicists prior to job submission.

Behind the physicist view of datasets and event collections, the event data is organized into files, which can be handled easily by the storage and transport systems. Event collections are in general mapped to one or more files and that there be some easy means for the framework application to know which files to open by the name of an event collection.

The packaging of events into files are done in such a way that the average file size is kept reasonably large (e.g., at least 1GB) in order to avoid a large number of practical scaling issues that arise with storage systems, catalogues, etc. Small files can also be created and handled in a transient way, e.g., as the output of individual jobs, but that the data production systems do 'merge' steps in their workflows in order that the files tracked by the data management system

in the long term are of adequate size.

In addition to 'files' as a unit of packaging, the 'file block' was introduced. This is just a set of files which are likely to be accessed together. It is convenient to group data in 'blocks' of 1-10TB for bulk data management reasons. Any given file is assigned immutably to a single unique file block and global replication within the system is likely be done by file block rather than the single file.

Data Location

The CMS Data Location used to define and discover the data and Monte Carlo simulated samples is the central Dataset Bookkeeping System (DBS). The DBS maintains the semantic information associated to the datasets such as which files belong to which dataset, their grouping into blocks, but also stores detailed meta-information about the files itself (type, size, checksums, content). It keeps track of the data parentage through their processing history and allows to discover which data exist. In addition it maps the file-blocks to sites holding a replica of them and allows to find the location of desired data. It is synchronized with the CMS data replacement and transfer tool PhEDEx [60]. Figure 4.6 shows the web site of DBS. Users can use the web site to search datasets by the meta-information and locate the files of datasets.

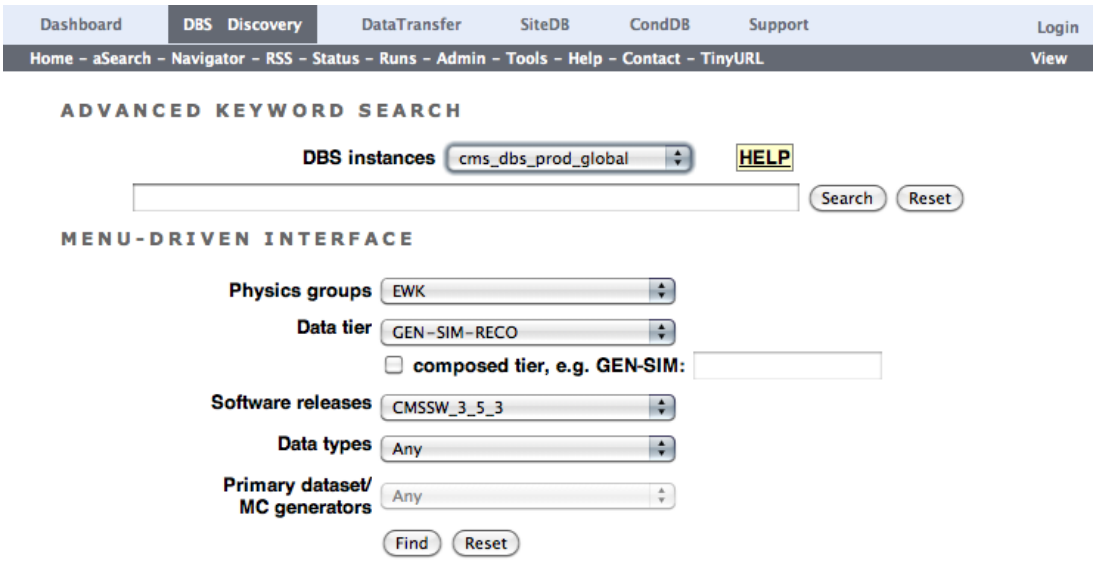

Figure 4.6: Discovery page of Dataset Bookkeeping System (DBS).

Local Data Access

For the simplified handling of files, the central databases store and deal only with logical file names. In order to access the files at the sites, e.g., through an analysis Grid job, the logical file name has to be resolved into a physical file name (a path to a local disk or a mass storage system like CASTOR or dCache). For this purpose each site maintains an XML-based file, containing simple, generalized rules to build physical paths from logical names and vice versa. The rules may depend on the desired access protocol and provide a fine-grained handle for the data organization to the site administrator. A common Tier-1 use case which is covered in that way, is the separation of data: files that go to tape and data which stay on disk only.

Data Transfer and Placement System

The Physics Experiment Data Export (PhEDEx) system manages the transfers of data among sites, dealing with Grid FTS and different storage systems. PhEDEx interacts with the CMS catalogues, cross-checks the file-level information in DBS for datasets mentioned in transfer requests, and updates the storage location when the data transfers are complete. Technically it is based on software agents that run autonomously at each site and exchange information via a central database. PhEDEx has been exercised in progressively increasing complexity and scale during several years of use in daily production and computing challenges. The PhEDEx website, as shown in Figure 4.7, provides monitoring web-pages with complex drill-down operations, suitable for debugging or presentation from many aspects. PhEDEx also providing access to PhEDEx information and certificate-authenticated services for other CMS data-flow and workflow management tools such as CRAB, WMCore, DBS and the dashboard. A PhEDEx command-line client tool provides one-stop access to all the functions of the PhEDEx Data Service interactively, for use in simple scripts that do not access the service directly.

Handling of Calibration and Alignment Data

For the delivery of condition data to a worldwide community of distributed processing and analysis clients, CMS uses a multi-tiered web approach well-suited to the Grid environment. Condition data include calibration, alignment, and configuration information used for online and offline event data processing. The conditions, which are stored in a central Oracle database, are keyed by time and have a limited validity. Since these data is used by many thousand jobs in parallel all around the world, the caching of such information close to the processing activity results in a significant performance gain.

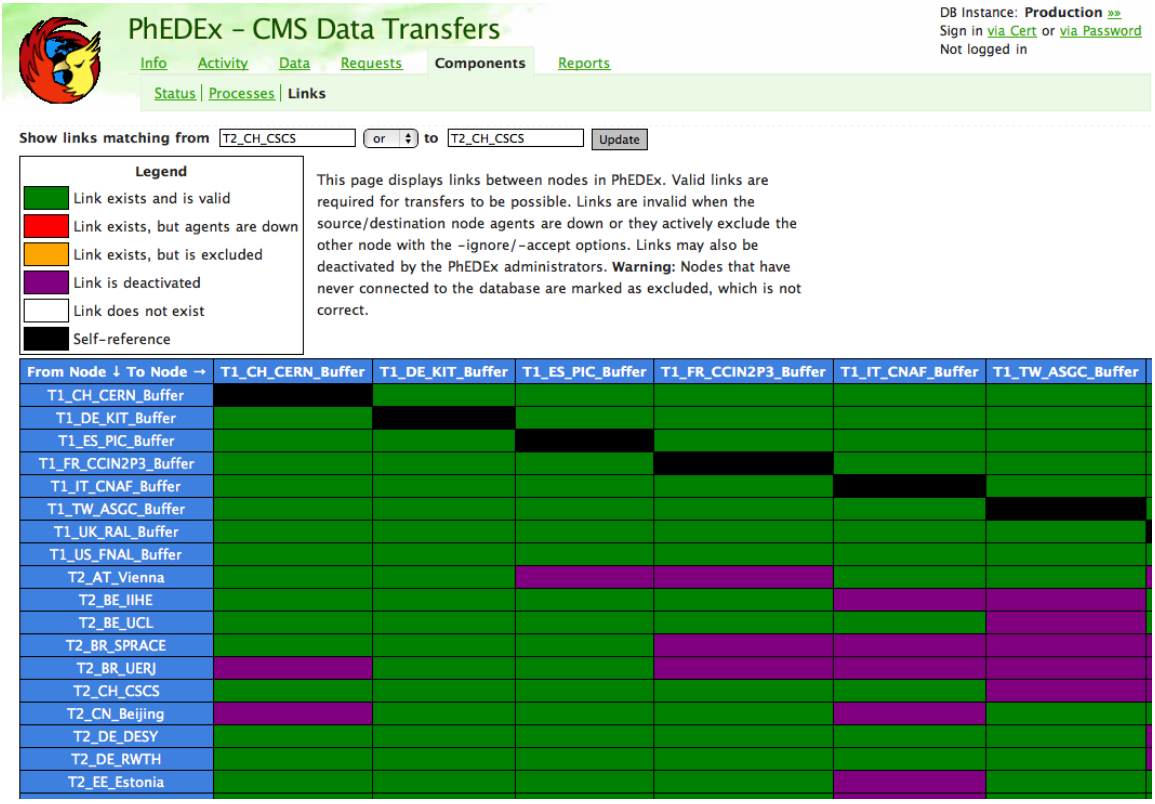

Figure 4.7: This web page displays links to/from the Swiss CMS Tier-2 at CSCS (Manno, Switzerland) in PhEDEx as of October 2010. [http://cmsweb.cern.ch/phedex/]

Therefore, each site deploys one or more squid caches which provide high performance access to the condition data requested by the jobs through the CMS framework and its interface FroNTier [61]. FroNTier is a simple Web Service approach providing client HTTP[6] access to a central database service. The cache is loaded on demand and manages itself automatically.

4.5.2 CMS Workload Management System

The CMS Grid Workload Management System (Grid WMS) relies on the core WLCG services to allow CMS to access distributed computing resources. CMS expects the WLCG and sites to provide a Grid WMS that has certain characteristics [50]. Basic functionalities are: to schedule jobs onto resources according to the policy and priorities of the CMS Virtual Organization; to collect information 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.

[6]The Hypertext Transfer Protocol (HTTP) is a networking protocol for distributed, collaborative, hypermedia information systems.[1] HTTP is the foundation of data communication for the World Wide Web.

A typical distributed processing workflow that illustrates the interactions with data management components and the Grid middleware is shown in Figure 4.8. The basic steps are:

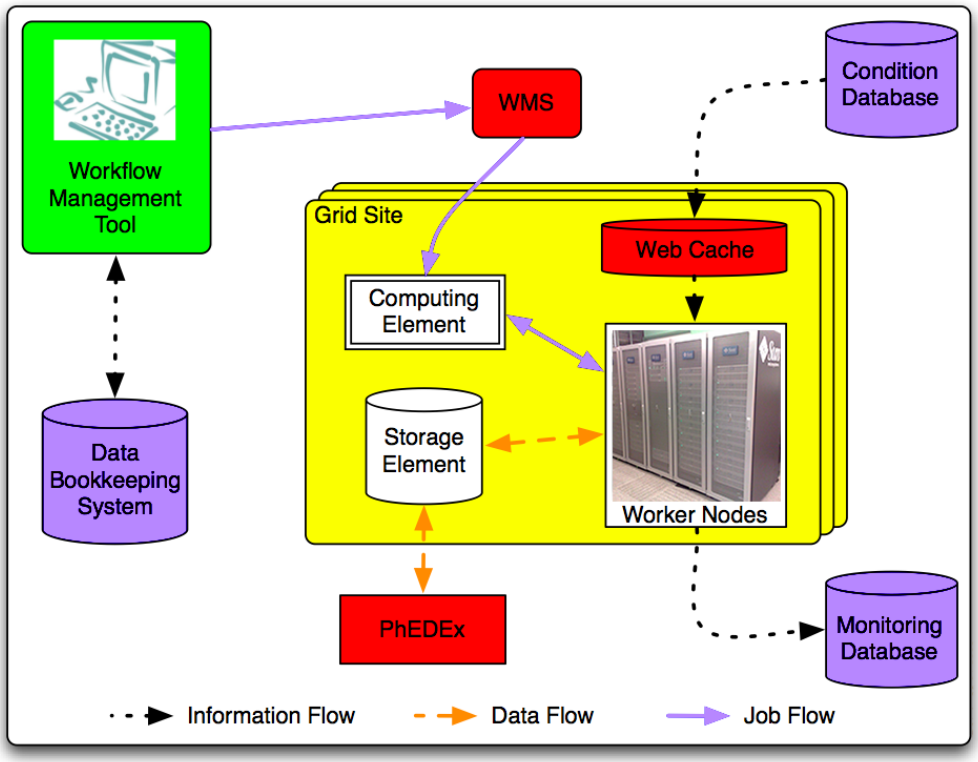

Figure 4.8: Overview of the CMS Grid workflow: The user interface provides access to the Grid world wrapped by the CMS workflow management tools such as CRAB for user analysis and ProdAgent for official Monte Carlo production. The CMS tools explore the available datasets and its location within the Dataset Bookkeeping System (DBS) and send the jobs through the Grid Workload Management System (WMS) to the site holding the data. At the site the job is handed from the Computing Element (CE) to the next free worker node, which accesses data stored on the associated SE. Condition data are retrieved from a central condition database which is cached through a web cache at the site. To allow constant monitoring, the CMS jobs report their state on a regular basis to a central monitoring database. Once the job has finished, the output might be stored on a local or remote storage element or can be retrieved together with the log files at the User Interface. In addition the processed files might be registered in DBS and CMS transfer tool (PhEDEx) for further processing or distribution.

- Data discovery and location via DBS;
- job submission to the site where the data are located;
- handling of the output data stored on local storage or passed to the transfer system

(PhEDEx);

- publication of the produced data with the relevant provenance information in DBS.

Monte Carlo Production

CMS has a long-term demand to perform large-scale Monte Carlo simulations. In addition it provides a way for testing the tools and infrastructure needed to process large amounts of events that have to be available at detector startup. The MC production system consists of three components: ProdRequest, ProdManagers and ProdAgents. The request system (ProdRequest) acts as a front-end application for production request submissions into the production system. The production manager (ProdManager) manages these requests, performing accounting and allocating work to a collection of production agents (ProdAgents). The ProdAgent consists of a set of loosely coupled components executing production workflows in the Grid environment. ProdAgents are responsible for job submission, job tracking, error handling, and automatic resubmissions, as well as data merging, and publication into the CMS cataloguing and data transfer system.

Analysis Tools

CRAB (CMS Remote Analysis Builder) [62] is developed to provide a user friendly interface for CMS physicists' interactions with data management and Grid submissions. CRAB supports the direct submission to the Grid, but also the submission with a CRAB server that aims at improving further automation and scalability of the whole system. Furthermore, CRAB can submit jobs to the local batch system with the help of the local scheduler. Therefore, CRAB provides users a unified approach of the CMS analysis job submission regardless of the kind of computing resource schedulers.

CRAB server and ProdAgent host BossLite tables in a MySQL database while the stand-alone CRAB relies on an SQLite database. Both ProdAgent and the pure client version of CRAB access the gLite functionality through BOSSLite miming basic UI functions. CRAB Server reuses part of the ProdAgent architecture, but implements multi-threaded submission to allow many users to submit tasks concurrently. CRAB Server has a robust handling of delegated proxies to avoid clashes and security flaws. With respect to the other tools, CRAB Server uses a different mechanism for sandbox handling. Being coupled with a GridFTP server, using gLite features, the sandboxes are directly transferred to/from the worker node. This allows implementing specific CMS policies on sandbox sizes, bypassing the WMS limits.

4.5.3 Monitoring

A key component of the Grid is the monitoring. It allows the system to react on failures and to alert site managers to check the health of the site. And it also provides valuable input for the users about the reliability of the resources to use.

The Experiment Dashboard

The CMS Dashboard aims to provide a single entry point to the monitoring data collected from the CMS Grid environment and the jobs executed within this distributed system. By the inclusion of experiment-specific information (via MonALISA [63]) in addition to R-GMA data, Dashboard is able to display quantitative and qualitative characteristics of the experiment and is thus able to indicate problems of any nature. General monitored quantities are: how many jobs are running, pending, accomplished successfully or failed on a per user, per site, per input data collection basis. Also the distributions evolving with time are available. Further resource usage (CPU, memory consumption, input/output rates) are aggregated. A detailed analysis of the job behavior (success rate, reasons of failures as a function of time, execution center, data collection) is possible and provides valuable feedback to the user to detect and identify the problem. Figure 4.9 shows the page summary types of jobs running on the Swiss CMS Tier-2 at CSCS.

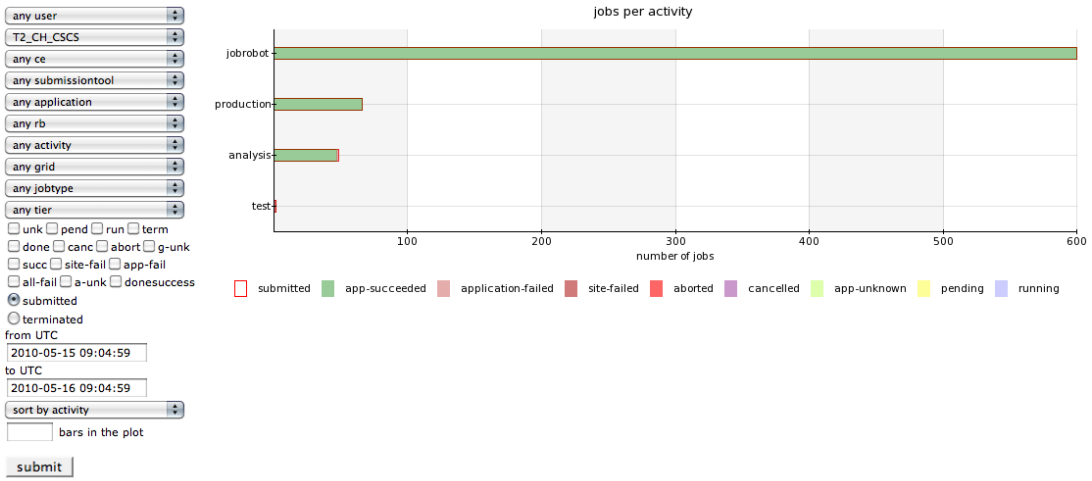

Figure 4.9: The Dashboard Monitoring provides a real-time monitoring for users and production jobs within the CMS. With its detailed output it could be used to debug cause of failures concerning CMS software or site problem. [http://dashb-cms-job.cern.ch/dashboard]

Site Availability Monitoring (SAM)

The Dashboard includes the collection of Site Availability Monitoring (SAM) plots, as shown in Figure 4.10. SAM subsumes a collection of tests which check the basic functionality in terms of the CMS needs. These dedicated jobs, which run roughly every hour, imitate analysis, production, or software installation jobs accessing computing and storage resources as well as CMS specific services such as FroNTier or the local CMS catalogues. Only sites which pass these tests on a regular basis are available for the usage within CMS.

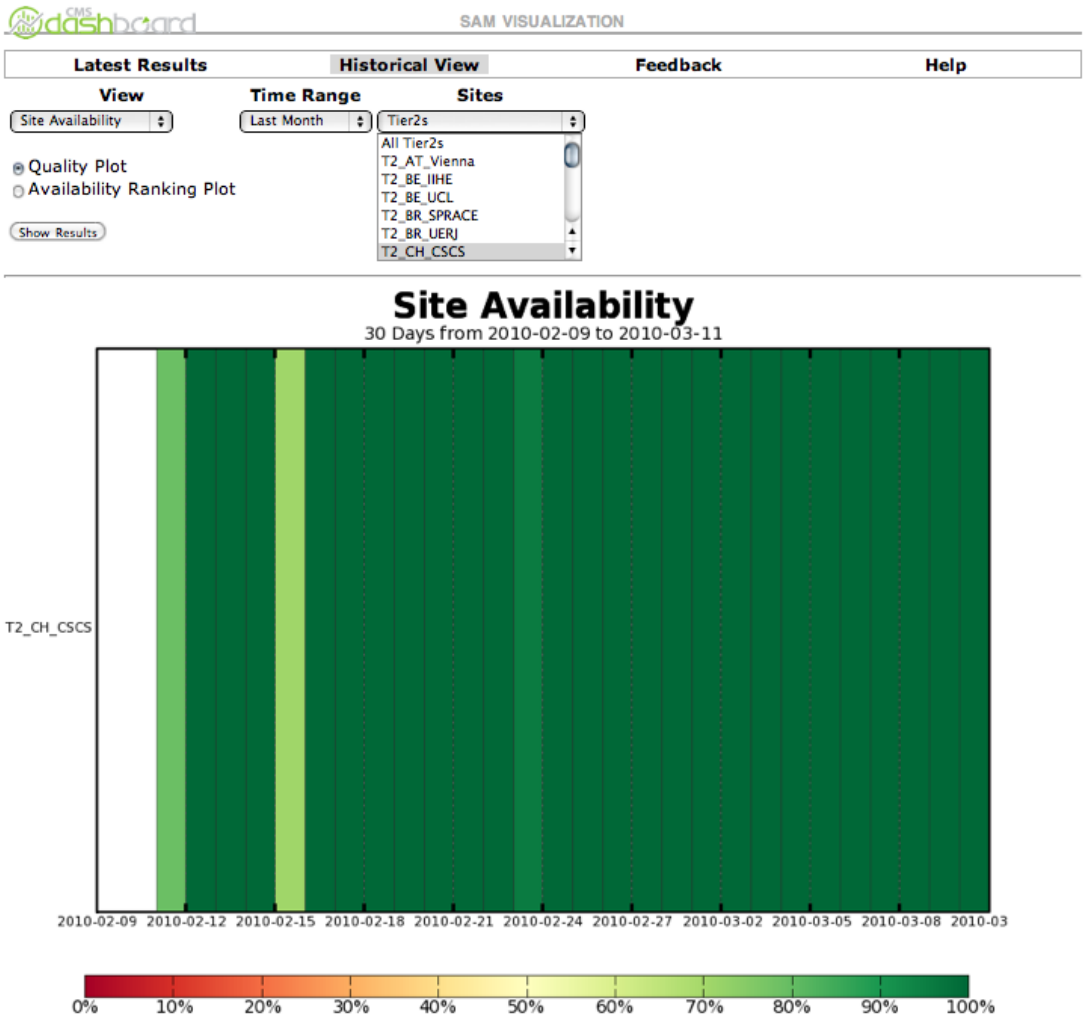

Figure 4.10: Site Availability of the CMS Tier-2 at CSCS, Manno during 9 February to 11 March 2010. The site availability monitoring enables the sites to follow their status. Only if all tests succeed a site is available for user jobs and MC production.

Other Monitoring tools

Job Robot

The task of Job Robot is regularly submitting jobs similar to real analysis jobs. The difference with respect to the SAM tests is the fact that the statistics are ~ 100 times higher, the fact that the accessed data can be spread on several disks and a higher load on the site storage system. A tool called Job Robot was developed to implement such an automatic job submission system using CRAB, the CMS analysis job submission tool [62].

At regular time intervals, a new analysis task is created for each site, to be run on a specific dataset. The task is then split into several jobs, which are submitted as a collection to the gLite WMS [64]. Each job performs a trivial data analysis on a fraction of the dataset. All submitted jobs are classified as successful, as failed at the application level or as aborted at the Grid level. It is used as a commissioning tool to test if a site is capable to run certain CMS workflow at the required scale.

Site Status Board

The site status board [65] is a meta monitoring system which conflates the information from the various specific CMS monitoring tools. Within one view all relevant monitoring information are available including their involvement with time.

4.6 CMS Computing Commissioning

The commissioning of the several hundred end-to-end links between the WLCG sites is a challenge for CMS computing. Actually, the individual sites in WLCG vary both in computing power (a few hundreds CPU cores to a few thousands CPU cores) and storage sizes (10 TB to a few PB). In addition, the expertise of operation teams at different sites are quite different. Before the LHC beam running in 2009, the amount of data collected by LHC experiments during the cosmic ray tests were not sufficient to exercise the WLCG to its capacity at the LHC rates. Every year since 2006, CMS undertook a dedicated stress test (CSA – Computing, Software and Analysis challenge) of the computing model using generated data. CMS created a 'Debugging Data Transfers (DDT) Task Force to coordinate the debugging of data transfer links in the preparation period and during the CSA07 data transfer test. The CSA07 service challenge was a data challenge in 2007 designed to test the transfer system at 50% of the design goal for 2008.

4.6.1 CCRC'08

Since the gLite architecture is a system scaling linearly with the number of gLite WMSs used, the CMS Monte Carlo production and analysis jobs are balanced over many WMS. During CCRC'08, 7 WMSs were deployed for the analysis, 4 for the Monte Carlo production. The typical instantaneous load of a single WMS, when no particular operations are scheduled, may reach 15k jobs per day, with peaks of 5k active jobs, running or idle, simultaneously handled. The dedicated stress tests on wms218 during CCRC'08, as shown in Figure 4.11, showed that a single WMS can handle 30k jobs per day without problems. Figure 4.12 covers the period from May 2008 to March 2009. In this period, about 23 million jobs were submitted through the gLite WMS, with an average of about 75k jobs per day.

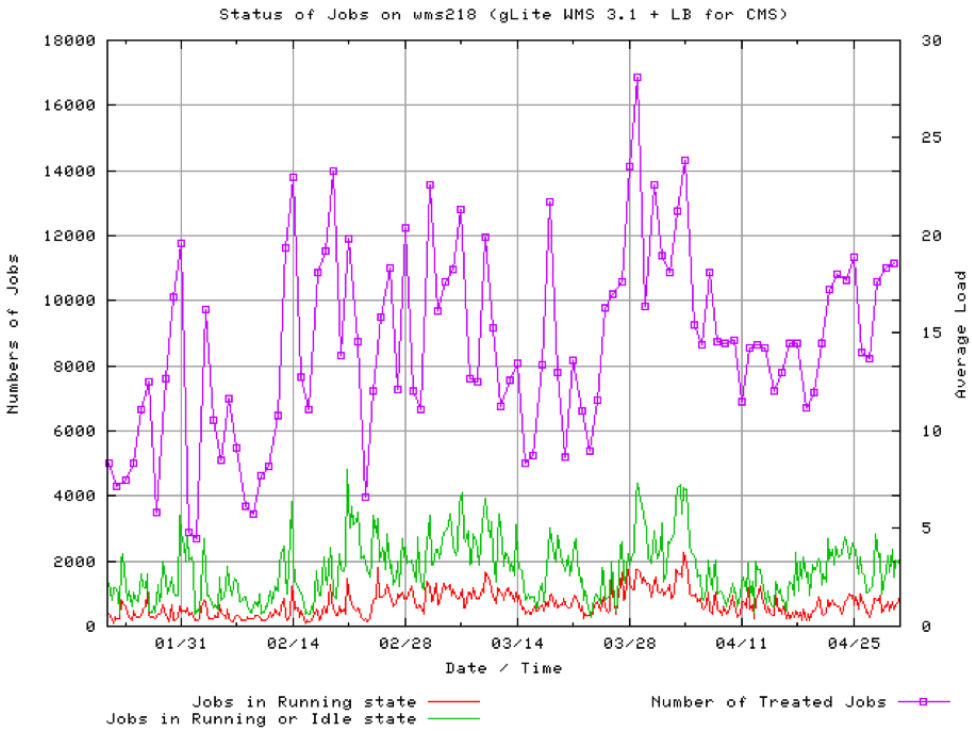

Figure 4.11: The plot shows the status of jobs on wms218 from 31 January till 25 April 2008. It presents typical instantaneous load of a gLite WMS used for CMS operations [66].

There are 44 'active' Tier-2 centers, meaning that a Tier-2 site successfully tested at least one data transfer link according to the procedures described in the following. There are also additional CMS Tier-2 centers that have not yet succeeded in testing at least one link.

The CMS computing model commissioned all links between:

- CERN to Tier-1 sites, and Tier-1 sites to CERN (14 links);

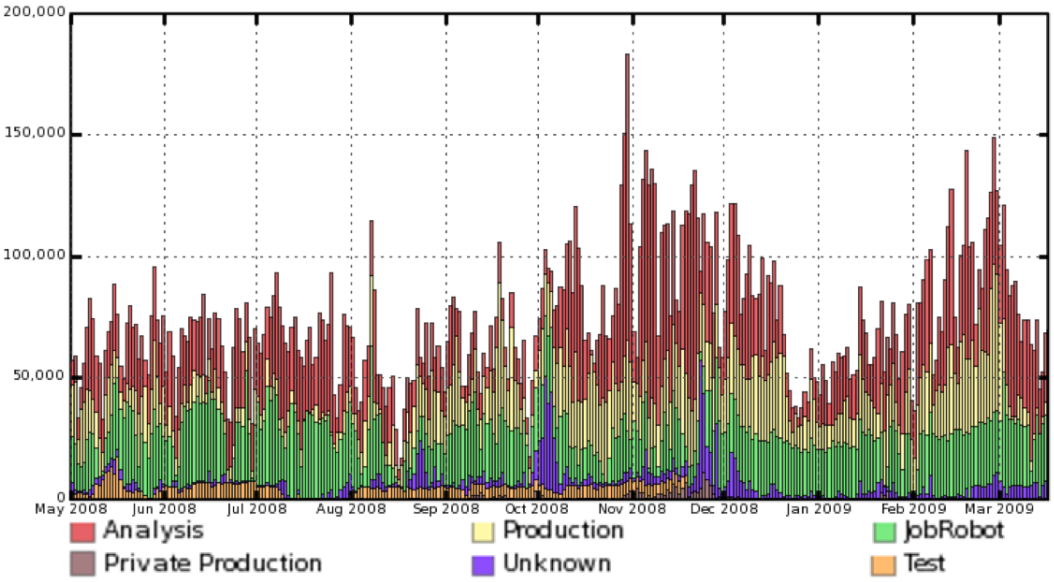

Figure 4.12: Distribution of the CMS jobs submitted to the gLite WMSs divided by activity during CCRC'08 [66].

- All other Tier-1-Tier-1 cross-links (42 links);
- All Tier-1 to Tier-2 downlinks (352 links);
- All Tier-2 to 'regional' Tier-1 uplinks (44 links).

Therefore, the total number of links to be commissioned in the computing model was 452 during CCRC'08.

Figure 4.13 illustrates the success rates of jobs submitted to the gLite WMS in the main CMS activities during CCRC'08. Only 58% of the analysis jobs terminated successfully. The pie-chart shows that the main reason for the failures is the application failures, not the Grid problems. The application failures are expected, since the analysis jobs run user codes which may not have been tested thoroughly, but more detailed analysis showed that the main reason for the analysis job failures is the stage out of the data output files to remote SRM servers (often those files are too large to be retrieved by WMS inside the OutputSandBox). Therefore, CMS developed a system to asynchronously copy user data to the remote final destination using temporary buffering at the site where the jobs run, in order to mitigate the problem. The same detailed investigation also showed that a good fraction of the Grid failures are not due to the middleware layer, but are simply jobs that for one reason or another spend too much time on the worker node and are killed by the local batch system, appearing as aborted by the Grid.

In the Monte Carlo production the application failure rate is lower, because of the usage of the

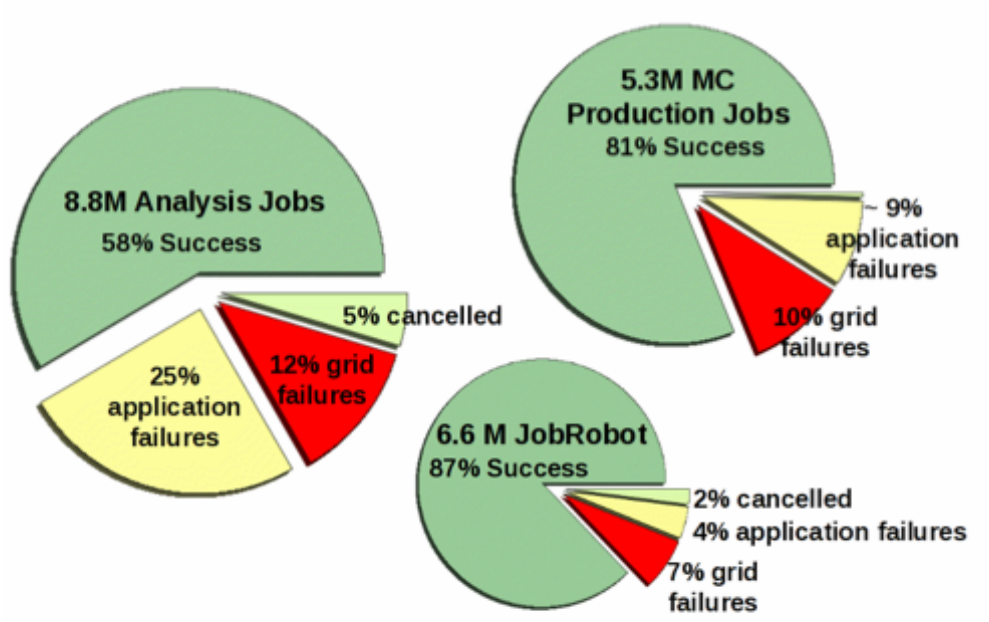

Figure 4.13: Success rate of the CMS jobs submitted to gLite WMSs divided per activity during CCRC'08 [66].

validated codes and the local stage out. The Grid failures are also fewer, but still jobs can be killed by the local batch systems for various reasons. For the JobRobot, the Grid failure rate is reduced to 7% of the submitted jobs. It was observed that many of the remaining Grid failures, mostly due to CE overload or mis-configured Worker Nodes, are cured by simple resubmission of the jobs.

4.6.2 CRAFT'08

The CMS Collaboration conducted a month-long data-taking exercise known as the 'Cosmic Run At Four Tesla' in late 2008 (CRAFT'08) in order to assure the commissioning of the experiment for extended operation. The month-long data taking exercise performed a major test for CMS computing workflows.

Data Handling

Table 4.1 gives an overview of the volumes of data produced from the central data-handling perspective during CRAFT'08. CMS collected over 2 billion events including technical events for monitoring and calibrations purposes.

Number of primary datasets produced	11
Number of events recorded	2×10^9
Number of events in Cosmic primary dataset	370×10^6
Number of runs recorded	239
Total data volume recorded and produced	396 TB
Total data volume recorded and produced in Cosmics primary dataset	133 TB

Table 4.1: Overview of data produced during the CRAFT'08 run, from the central data-handling perspective [67].

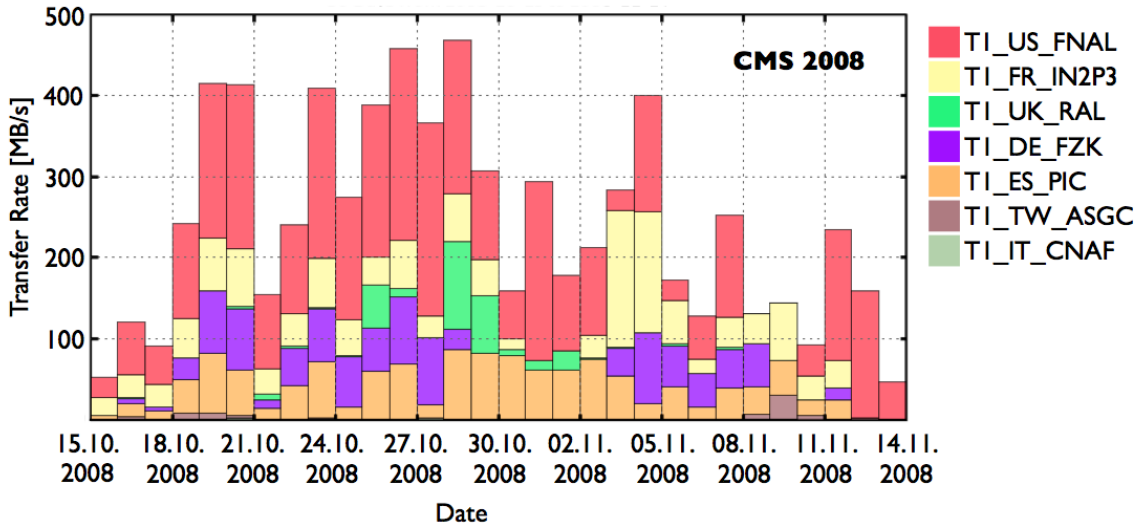

Figure 4.14: Transfer rates from Tier-0 to Tier-1 centers over the duration of CRAFT'08. The average was about 240 MB/s (taken from monitoring sources).

During CRAFT'08, the recorded and processed primary datasets were distributed amongst the Tier-1 sites according to available free tape space, taking into account processing capacity and reliability of the Tier-1 sites. For the Cosmics primary dataset, the average size per event for the RAW data tier was 105 kB/event and for the RECO data tier 125 kB/event.

Figure 4.14 shows the transfer rate during CRAFT'08 from the Tier-0 to the Tier-1 sites. The transfers averaged 240 MB/s with rates exceeding 400 MB/s on several occasions.

During CRAFT'08, a total of 600 TB was transferred out of CERN to the Tier-1 sites. Figure 4.15 shows the cumulative transfer volume per Tier-1 site. Overall, the transfer system performed very well and all CRAFT'08 data were transferred reliably to the Tier-1 sites.

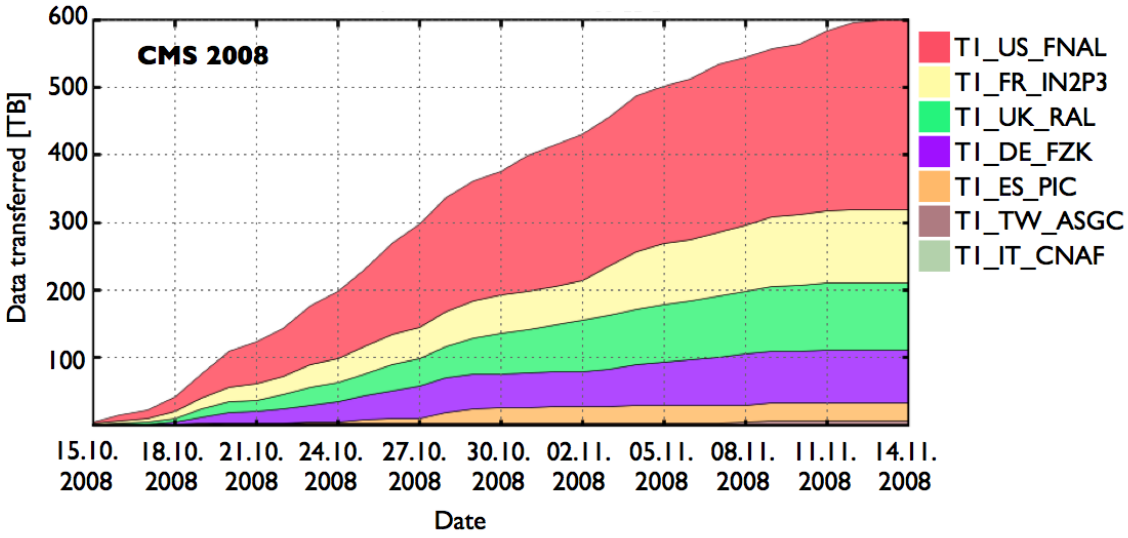

Figure 4.15: Cumulative transfer volume from Tier-0 to Tier-1 centers over the duration of CRAFT'08 (taken from monitoring sources).

CRAFT'08 Analysis Activity

The CRAFT'08 data were analysed both on the CERN Analysis Facility (CAF) and on the Grid making use of distributed resources (Tier-2). While access to data on Tier-2 sites were performed exclusively by CRAB, the CAF queues were used to run both CRAB and non-CRAB jobs.

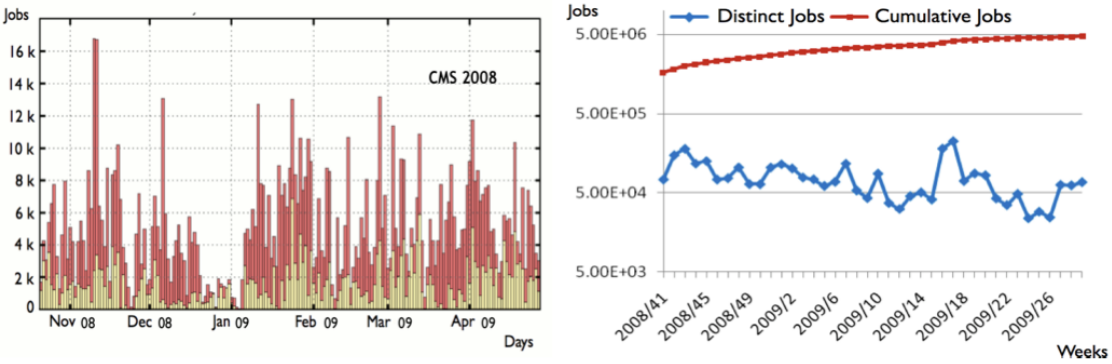

Figure 4.16: CRAFT'08 job distributions as a function of time. Left: Daily distribution of analysis jobs submitted using CRAB and accessing CRAFT'08 data. Grid (dark shading, red) and CAF (light shading, yellow) activities are shown (taken from monitoring sources). Right: CRAFT'08 jobs submitted only at CAF (with and without CRAB). The upper line shows the cumulative number of jobs, the lower line shows the number of jobs submitted each week. The time window extends well beyond the end of CRAFT'08 data taking to cover the extensive period of analysis [67].

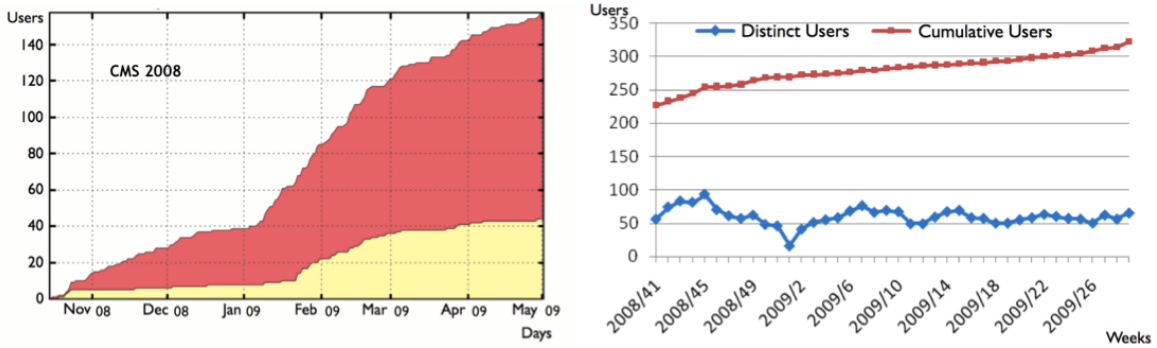

Figure 4.17: Cumulative plot of number of different users accessing CRAFT'08 data as a function of time. Left: users using CRAB to submit Grid (dark shading, red) and CAF (light shading, yellow) jobs (taken from monitoring sources). Right: number of users submitting jobs only at CAF (with and without CRAB). The lower line shows the number of users per week, the upper line the integrated number over a long period. The time window extends well beyond the end of CRAFT'08 data taking to cover the extensive period of analysis [67]

From October 2008 to the beginning of May 2009 more than 2 million analysis jobs accessed CRAFT'08 data, including both CRAB and non-CRAB jobs. The quoted value takes into account both CAF and Grid activity (Figure 4.16). Figure 4.17 shows the cumulative numbers of distinct users which performed CRAFT'08 data analysis in the considered time window. The shapes, combined with daily jobs distribution, give a clear indication of how the user community increased continuously. Referring to the same time interval it is estimated that more than 200 distinct users in total performed CRAFT analysis activities. The overall efficiency of CRAFT analysis jobs was approximately 60%. Local submissions on the CAF were 85% efficient. The main source of failures of Grid CRAFT jobs were remote stage-out problems, which was addressed by a new workload management infrastructure in 2009. In general, there is a 10% failure rate due to problems within the user code. No relevant bottlenecks were experienced by the system during CRAFT.

4.6.3 Collision Data Collected with CMS in 2009 and 2010

After the 2009 winter shutdown, the LHC was restarted and the beam was ramped up to 3.5 TeV per beam. Right before the beginning of May, the analysis of the about 50 million 7 TeV pp collisions in CMS during the first 30 days of 3.5 TeV running went very well under way. The consistency of the CMS computing model was confirmed during these first weeks of data taking after restart of LHC in 2010. Computing activities at the CMS Analysis Facility (CAF)

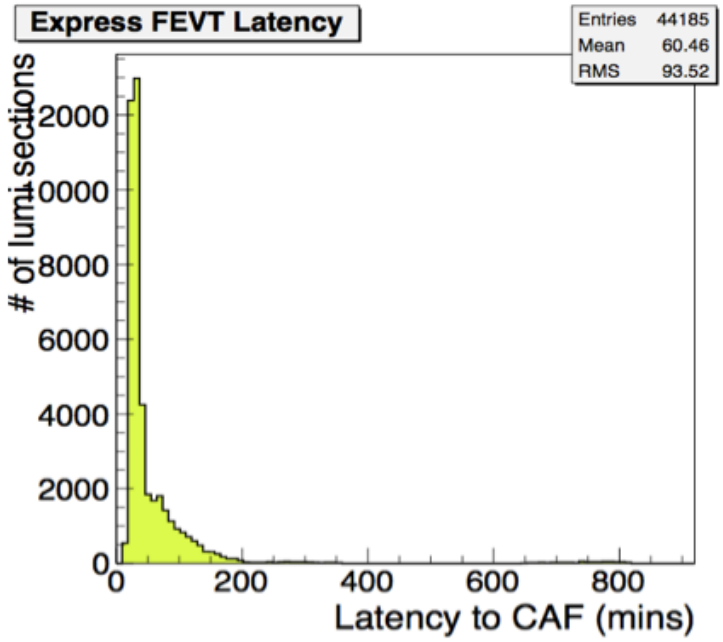

Figure 4.18: Data processing latencies at CAF were well within design goals [68].

at CERN were marked by a good response time for a load almost evenly shared between ALCA (Alignment and Calibration tasks with highest priority), commissioning and physics analysis. Latencies, in particular at T0 and CAF, were well within design goals as shown in Figure 4.18, allowing prompt reconstruction to be performed and calibration constants to be produced in a timely fashion.

Since 30 March 2010, data was continuously exported from CERN, with high peaks during the first LHC 'squeeze fills' (increasing the density of the protons in the bunches) at the end of April, initial transfer rate did not show any difficulties. Aggregated transfer rates of processed data from CERN to all Tier-1s and Tier-2s were well in the range of several hundreds MB/sec as shown in Figure 4.19. The system showed flexibility in dealing with occasional backlogs. The observed quality of service at Tier-1s for prompt skimming (selecting samples of data for particular analysis) and reprocessing is satisfactory.

CMS Tier-2s was also running very well. 400 active users had submitted 120000 jobs in April 2010: half for MC production and half for analysis (Figure 4.20). The very high proportion of successful jobs can be directly linked to the readiness of the Tier-2s.

In conclusion, the whole CMS computing system, including hardware and software, is stable and reliable. By the beginning of May, the data volume amounts to 14 TB, which is modest compared to what is expected for the whole period of the CMS operations with higher luminosities (around 3 orders of magnitude).

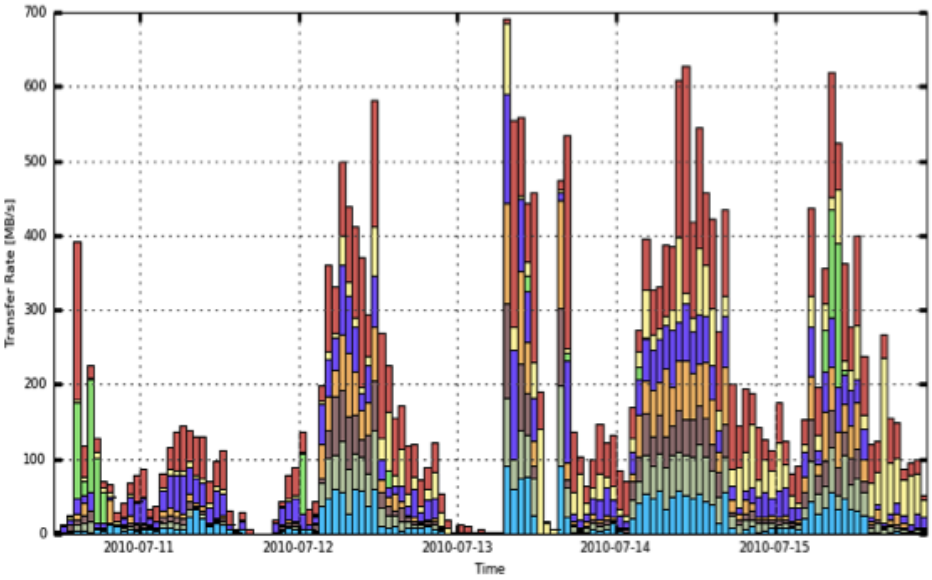

Figure 4.19: Hourly Peaks to Tier-1s of 600MB/s [68].

4.7 Summary

To meet the data analysis challenge on the unprecedented scale, WLCG develops innovative use of its enormous distributed computing resources and mass storage management systems to organize a hierarchical distribution of the data. WLCG uses gLite middleware to provide a complete set of services to a production Grid infrastructure. The CMS computing system is tightly integrated with WLCG to implement a dedicated Grid environment to support the storage, transfer and manipulation of the recorded data for the lifetime of the CMS experiment. The CMS computing system along with the most important systems and services are organized in a Grid Workload Management System, a CMS Data Management system and other CMS-specific services to support specific experiment applications and software.

The commissioning of the WLCG sites and several hundred end-to-end links between them is a huge challenge for CMS computing. Since 2006, CMS undertook every year a series of stress test of the computing model and data challenges (e.g., CCRC'08 and CRAFT) using generated data or cosmic ray events collected with the CMS detector. Since the restart of LHC in the winter of 2009–2010, the analyses of 7-TeV collision data have proceeded very well. The consistency and flexibility of the CMS computing model was confirmed during these tests and the first months of data taking in 2010.

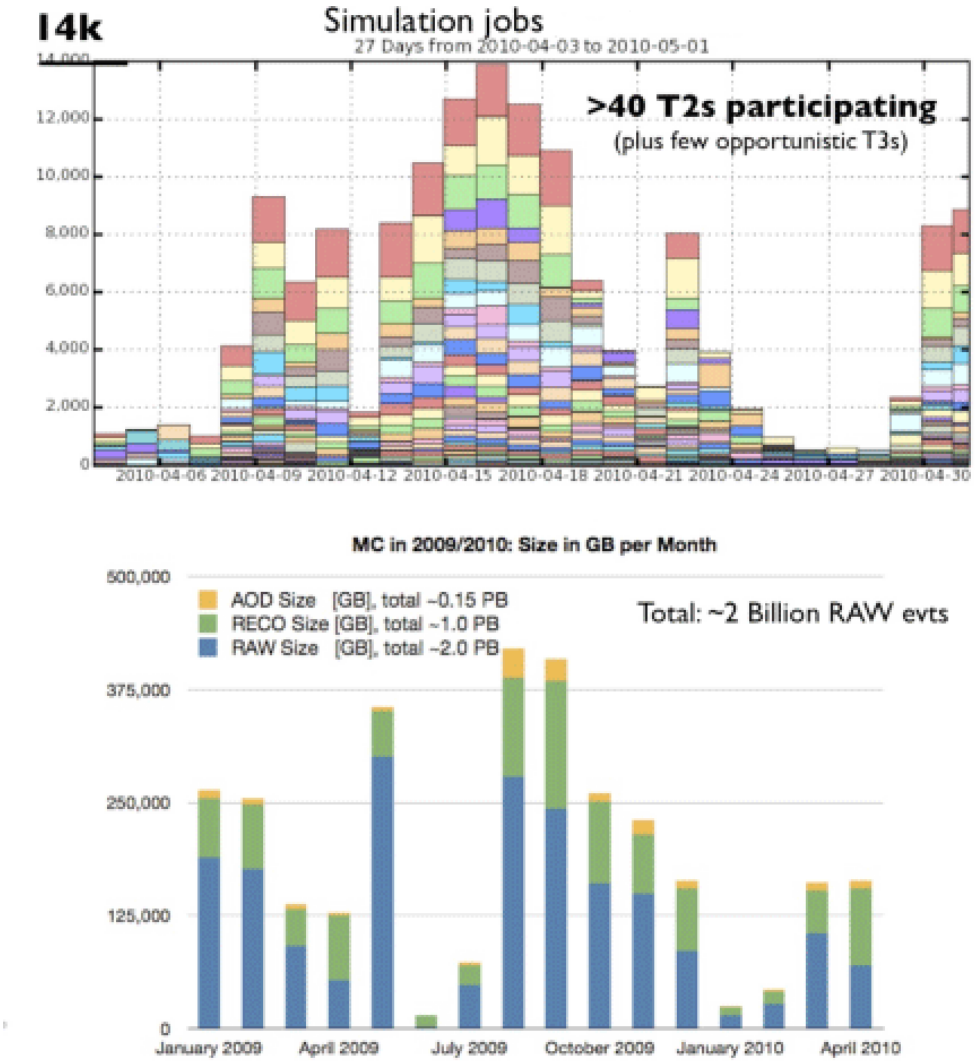

Figure 4.20: Number of jobs in April 2010 (top plot) and monthly data volume resulting from the massive MC production campaigns from January 2009 to April 2010 (bottom plot) [68].

5 CMS Computing in Switzerland

As a founding member of CERN and one of the two host states, the major Swiss universities, the two Swiss Federal Institutes of Technology in Zurich (ETHZ) and Lausanne (EPFL), and the Paul Scherrer Institute (PSI) engage actively in the experiments at the LHC, with strong participations in CMS, ATLAS and LHCb, as well as WLCG, as shown in Figure 5.1.

The WLCG collaboration in Switzerland provides computing infrastructures and resources to the LHC physicists from Swiss institution as well as the whole LHC Collaborations under the agreement of WLCG MoUs [5]. As a major part of collaboration, a high-performance Tier-2 center was set up at the Swiss National Supercomputing Center (CSCS) in Manno, near Lugano, as a part of the Worldwide LHC Computing Grid. The Tier-3 centers are operated by PSI for CMS, the University of Bern and the University of Geneva for ATLAS. I participated in the configuration and commissioning of Swiss Tier-2 and the setup, configuration and commissioning of Swiss CMS Tier-3 under the supervision of Dr. Derek Feichtinger.

Section 5.1 presents the evolvement of the Swiss Tier-2 at CSCS and its current configuration and commissioning. Section 5.2 reports our work on the commissioning of the Swiss CMS Tier-2. Section 5.3 provides our work on the setup, configuration and commissioning of Swiss CMS Tier-3 at PSI.

5.1 Swiss Tier-2 Center at CSCS

The Tier-2 center located at CSCS is a part of the Germany-Switzerland region in WLCG. It is the only Tier-2 in Switzerland. As there is no Tier-1 in Switzerland, the Tier-2 is associated with the FZK at Karlsruhe in Germany. CSCS operates the Tier-2 center on behalf of the Swiss Institute of Particle Physics (CHIPP), as laid down in the ETH Zurich/CSCS Memorandum of Understanding. In order to maximize the resource usage, it is a multi-VO Tier-2 supporting the ATLAS, CMS and LHCb experiments in Switzerland. The Tier-2 center is not only devoted to the Swiss LHC physicists for data analysis but also represents the Swiss global contribution to the enormous distributed computing effort of the LHC experiments. As a member of the EGEE-II Project, CSCS participates in the Grid Operation Center and the Regional Operation Center

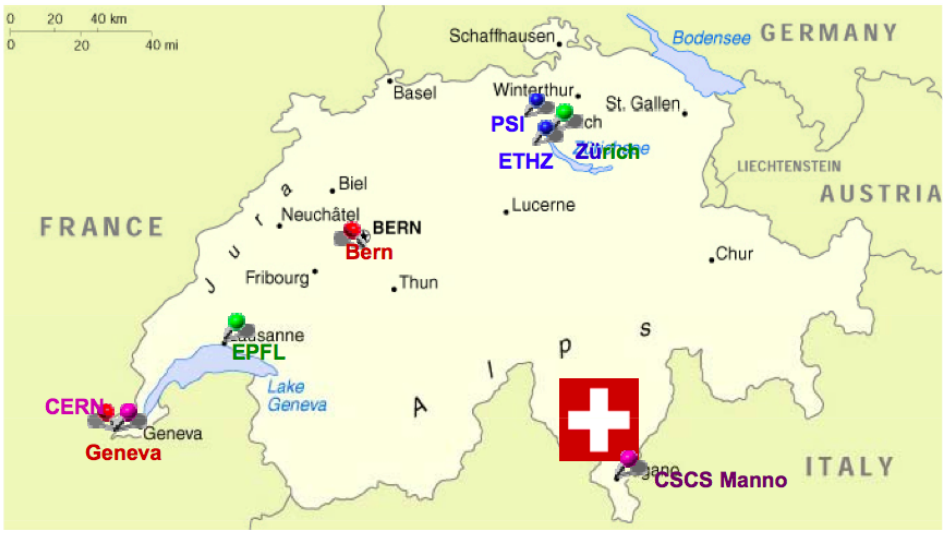

Figure 5.1: Swiss Participation in the LHC experimental programme: ATLAS (red): Geneva University, Bern University; CMS (blue): ETH Zurich, PSI, Zurich University; LHCb (green): EPFL, Zurich University.

Germany-Switzerland (DECH) for EGEE. The GridMap in Figure 5.2 shows a visualization of the status of WLCG. The CPU numbers of sites and their status are represented by rectangles of different sizes and colors respectively.

5.1.1 From Prototype to Production

A prototype cluster, named 'Phoenix' (Phase 0, shown in Figure 5.3) was built and operated at CSCS from December 2006 to January 2008. This prototype system featured a CPU processing power of 220 kSI2K[1] [70] and offered $\sim$ 55 TB of storage space. These were made available on the Grid by running services for an LCG Computing Element (CE) and a LCG Storage Element (SE). Instead of the scheme of one cluster for one experiment, the resources were shared between the ATLAS, CMS and LHCb experiments. The hardware and basic middleware maintenance was provided by CSCS, whereas the higher level experiment specific services were setup and maintained by the members of the participating institutes. During the building, setup and running the prototype system, lots of experiences on the LHC middleware and the specific experiment software were gained.

In the second half of 2006 the Tier-2 successfully participated in the CMS challenges of SC4 (Service Challenge 4) and CSA06 (Computing, Software, and Analysis challenge 2006).

During the CSA06 exercise the goal was to test the workflows and the data-flows associated

[1] 1 kSI2K = 1000 SPECint2000

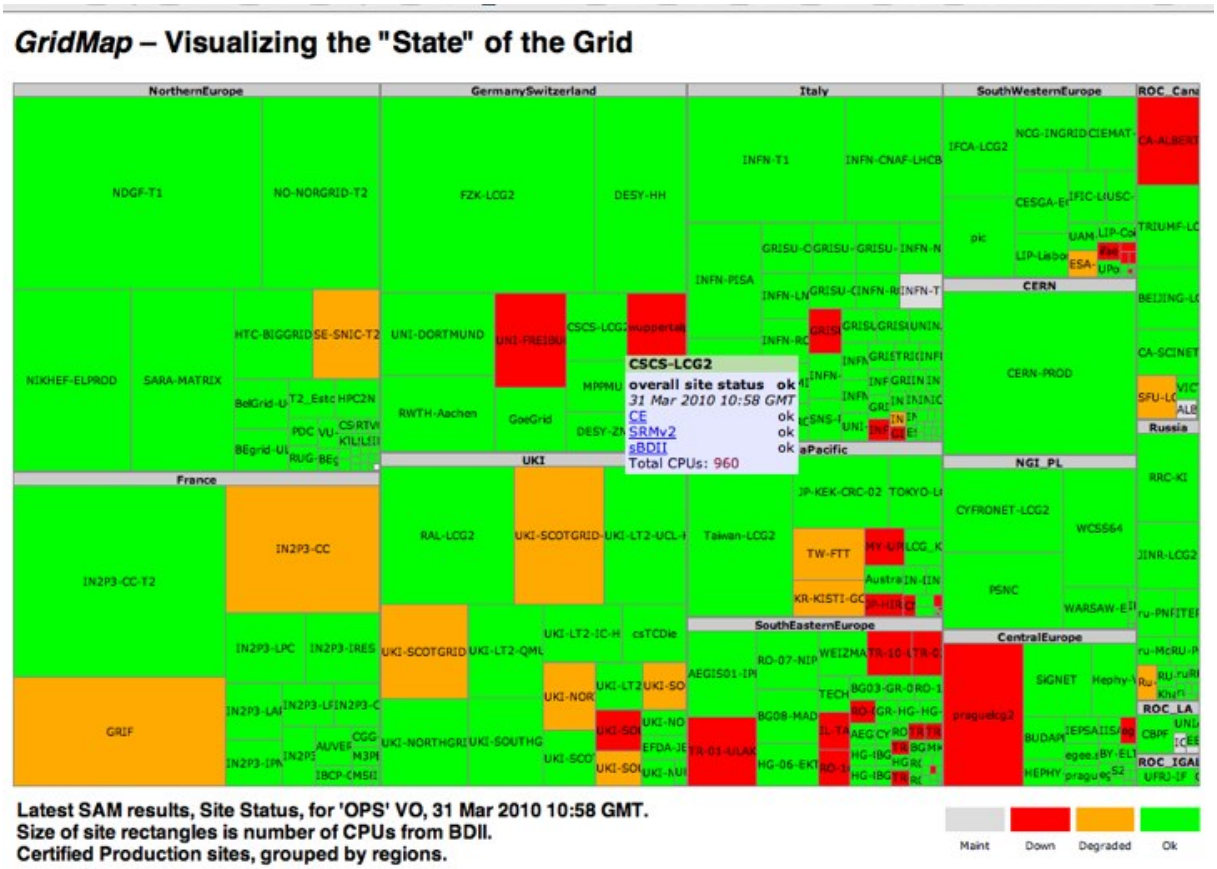

Figure 5.2: GridMap shows a visualization of the status of WLCG. The number of CPUs and status of sites are represented by rectangles of different sizes and colors respectively. The Swiss Tier-2 (CSCS-LCG2) is shown in the middle of the graph. The pop-up dialog box shows the overall status of the Swiss Tier-2 site. [http://gridmap.cern.ch]

with the data-handling model of the CMS experiment. First, 50 million events were generated and their detector response was simulated. Then, the events were reconstructed at the Tier-0 (CERN) at a rate of 40 Hz using the CMSSW framework and calibration constants from the offline database. Full Event Data (FEVT) and Analysis Object Data (AOD), as well as some (fake) High Level Trigger decision tags were produced. These data were then distributed over all participating Tier-1 centers. 'Skim jobs', that select the relevant events for a specific physics channel, were run at the Tier-1 centers and the resulting data were propagated to the Tier-2 centers, including the Swiss Tier-2 center in Manno.

In early 2007 CSCS began to evaluate a new cost effective and reliable storage solution composed of massive network attached storage and the dCache storage management system.

A new computing cluster 'Phoenix' (Phase A) was installed in December 2007. The system features 400 CPU cores which provide processing power of $\sim$ 800 kSI2k and offers $\sim$ 225 TB of

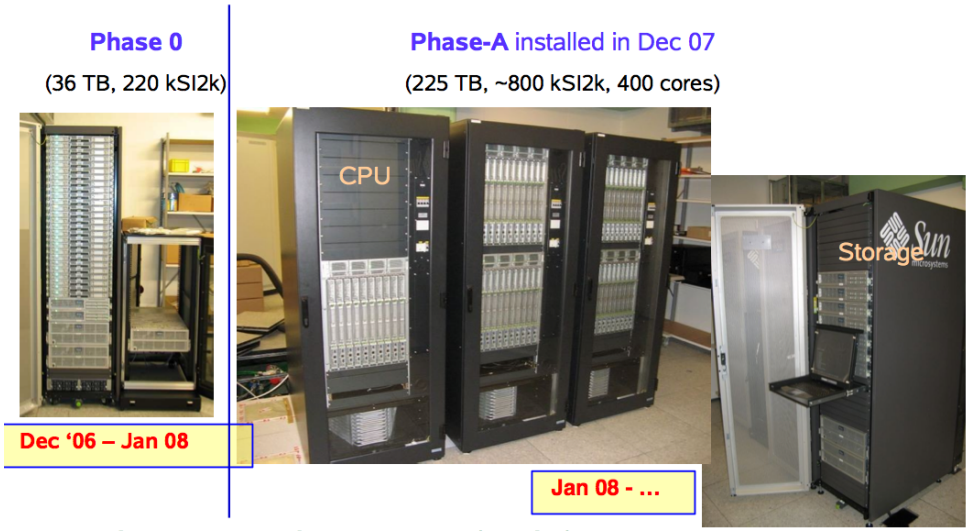

Figure 5.3: Photos of Phase 0 (left) and Phase A (right) of the Swiss Tier-2 cluster 'Phoenix' [69].

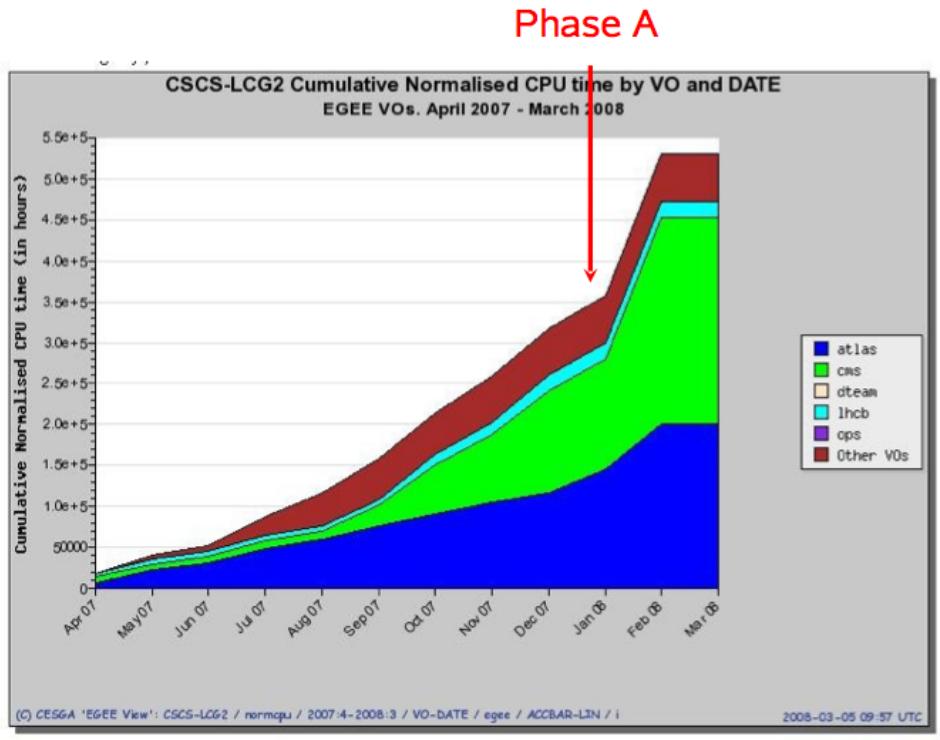

Figure 5.4: Cumulative normalized CPU time by VO and date of the Tier-2 center at CSCS.(`www4.egee.cesga.es/gridsite/accounting/CESGA/egee_view.php`)

storage space. The cluster showed high efficiency for user analysis jobs. Figure 5.4 shows that the cumulative normalized CPU time of the 'Phoenix' cluster was increasing fast, especially after

upgrading to Phase A.

Figure 5.5: Photo of the Swiss Tier-2 cluster 'Phoenix' (Phase B).

In January of 2009 the cluster 'Phoenix' was upgraded to the planned phase B, and the new hardware was successfully commissioned and immediately entered into the production level activities of WLCG. The new cluster, as shown in Figure 5.5, features $\sim$ 1440 CPU cores and $\sim$ 490 TB storages. The new storage service infrastructure used dCache to manager massive file servers with 48 disks per server. It had been proven to be an efficient and cost effective solution for our use cases.

A big progress was achieved in the automation of administrative tasks and to put the operations on a professional level exceeding the previous prototype operations. A central configuration management system was implemented, which allows to add new cluster components with minimal increasing in human effort. This is complemented by a centralized monitoring infrastructure which is able to trigger alarms in case of local system failures. Part of this information is also made available to the users, so it is easier for the users to check the state of the local resources (e.g., storage space, free nodes) or their jobs. The standard procedures for a number of routine tasks in the Tier-2 (downtimes, upgrades) have been formalized and tested. A web portal was built up to serve as single point of entry to the more static information for the members of the Swiss user community and also to help in the information exchange with other WLCG centers. The more dynamic information is managed through a number of community specific

archived mailing lists. All efforts significant improved the service quality of the Swiss Tier-2 center. Furthermore, the valuable knowledge and experience we learned from the Tier-2 greatly benefited the design and operation of the Swiss CMS Tier-3.

The Tier-2 successfully met the various data and service challenges during the past years, most prominently in CCRC'08 (Combined Computing Readiness Challenge in 2008) and STEP09 (Scale Testing for the Experimental Program in 2009) where the center was continually exercised by multiple experiments for several weeks each. It also was one of the centers which ran advanced user analysis tests for CMS, targeted at creating an approximation of the real load patterns to be expected during the operation of LHC experiments.

During the first half of 2010, the Phoenix cluster was upgraded to Phase C by CSCS. The upgrade scheme of the cluster is illustrated in Figure 5.6. The Phoenix cluster of Phase C roughly doubled the computing and storage capacity of the cluster of Phase B. The new cluster consists of: 96 Sun X6275 worker nodes with 768 cores (Intel E5540) with $\sim$ 3000 kSI2K; 115 TB Lustre[2] storage with a 7.6 GB/s write speed used as scratch; 10 Sun Thors X4540 with 480 TB total space for experimental data storage; high-performance Infiniband QDR used as interconnect.

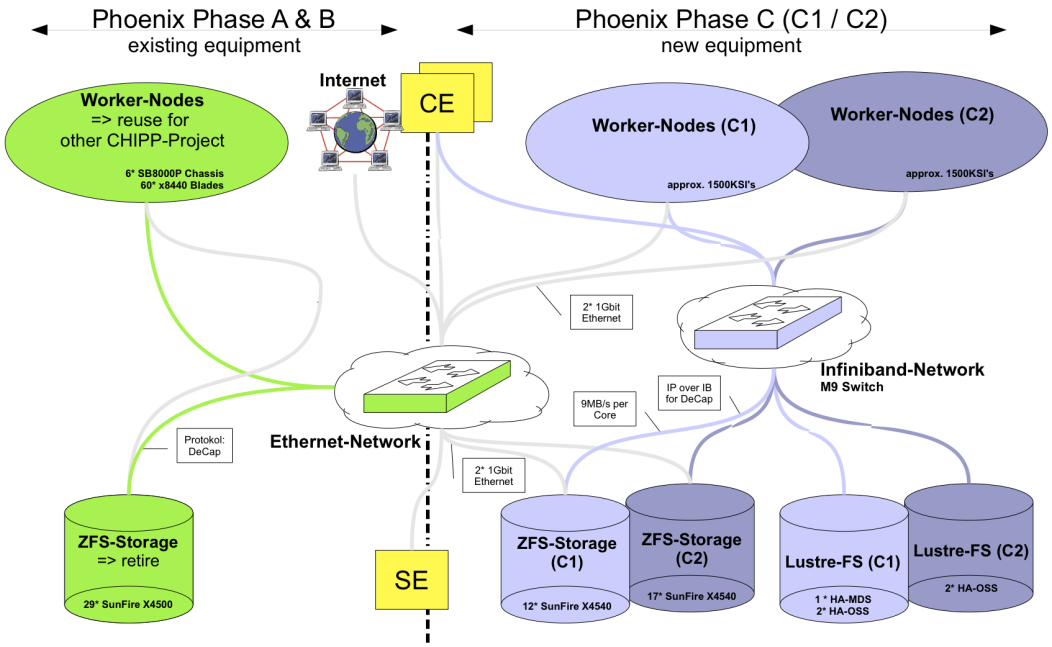

Figure 5.6: The scheme of the Phoenix Cluster upgrading from Phase B to C. Both clusters (phase A&B and phase C) were in operation from April to June 2010 for the smooth upgrade.

[2]Lustre is a massively parallel distributed file system, generally used for large scale cluster computing.

The Phoenix Phase C upgrade passed the acceptance test on 31 March. After that, system administrators at CSCS migrated the middleware from phase B to C during April. On 8 June, the Phoenix cluster (Phase C) was put fully operational. At the same time the cluster Phase B had been switched off and was being decommissioned. The old worker nodes were transferred to Swiss universities and used in ATLAS Tier-3 environments.

5.1.2 Infrastructures and WLCG Services

The Swiss Tier-2 at CSCS consists of the large computer cluster Phoenix managed by the local resource manager TORQUE, the storage systems managed by dCache, and network resources. A set of WLCG middleware running over the cluster perform WLCG defined services.

The Swiss Tier-2 is associated with the FZK Tier-1 center hosted by the Karlsruhe Institute of Technology (KIT) in Germany, which provides principal data access and storage services for the Swiss Tier-2. The Swiss Tier-2 though is able to transfer data from any other Tier-1 center. Since the Germany-Switzerland regional support system is also based around the FZK Tier-1 center, the Swiss Tier-2 can receive direct support from the FZK Tier-1. Data transfer is a prime example of the support and coordination between the Swiss Tier-2 and the FZK Tier-1 center.

Phoenix Cluster

After the Phase C upgrade finished in June 2010, the server nodes of the cluster are listed as follows in Table 5.1:

Host name	Server	Service
arc01	X4500	ARC Grid Computing Element
ce01	X4200	gLite CE,TORQUE
storage01	X4200	dCache storage manager main node
storage02	X4200	dCache pNFS database node, PostgreSQL
se01-10	X4540	dCache Pool (File servers) for experiment data
ui	Xen	UI
mon	Xen	Ganglia Monitoring Node
bdii	Xen	Local BDII node
cmsvobox	Xen	PhEDEx and Frontier

Table 5.1: Configuration of the Phoenix Cluster Server Nodes.

Currently the Tier-2 has 1156 'job slots' available for running computational tasks. A 'job slot' conceptually represents a CPU core capable of running a program with its associated memory

(e.g., a typical CMS (ATLAS) reconstruction job needs 1 (2) GB memory) and scratch disk. Typically, input files are staged to local storage on the worker node which runs the job, and a series of processes (e.g., filter, transform) is performed on the data. The worker nodes of the cluster are managed by TORQUE Resource Manager [71] running on CE (ce01.lcg.cscs.ch). TORQUE is an open-source distributed resource manager providing control over batch jobs and worker nodes. TORQUE is based on the code of popular batch system PBS with significant improvements in the areas of scalability, fault tolerance, and feature extensions. On the Swiss Tier-2, TORQUE is configured to integrate with the open source MAUI Cluster Scheduler [72] to improve overall utilization, scheduling and administration on a cluster.

The CSCS Tier-2 site is connected by the shared SWITCHlan dark fibre network with 4Gb/s, i.e., the bandwidth can be adjusted by illuminating the optical fibers with multiple frequencies [73]. This network is currently operated with one 10 Gb/s channel. The network map is shown in Figure 5.7. Corresponding to the latest version of the LHC experiment computing models, globally the current conclusion is that 1Gb/s links between a given Tier-2 and 'its' Tier-1 should be sufficient. Thus the Swiss capacity of the network meets the estimated requirements for connectivity.

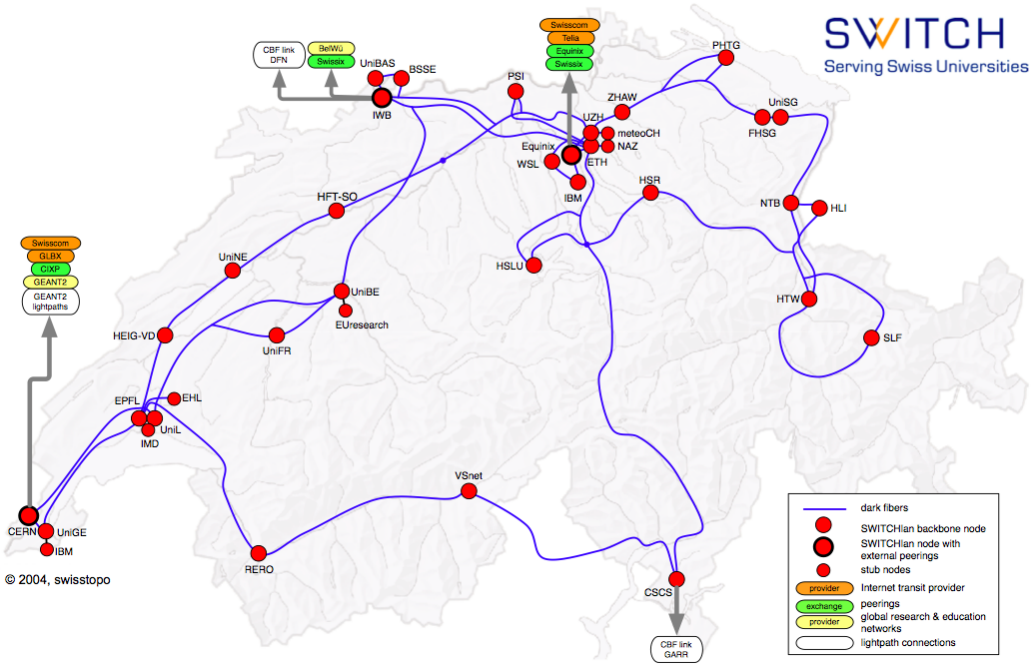

Figure 5.7: The SWITCHlan backbone. The Swiss CMS Tier-2 is located at CSCS, while the Swiss CMS Tier-3 at PSI.

Computing Element

The Computing Element (CE) deployed at CSCS Tier-2 implements a set of services representing the computing resource. Its main functionality is the Grid job management (Grid job submission, Grid job control, etc.). The CE interacts with the Workload Manager of WLCG, which submits a given job to an appropriate CE found by a matchmaking process. The CE provides the Local Credential MAPping Service (LCMAPS). For example, when Grid jobs were pushed from the Workload Manager of WLCG, based on the X509 certificate of the Grid user and the job specification (JDL), the LCMAPS plug-ins will acquire temporary local credentials, e.g., User ID and Group ID. Then the local credentials will be delegated to jobs.

Another component of the CE, the Batch Local ASCII Helper (BLAH), acts as an interface to a Local Resource Management System (LRMS). The BLAH service provides the interface for job submission, job hold, job resume, job status, job cancel and proxy renewal between WLCG and TORQUE, the current LRMS of the Phoenix Cluster.

Besides job management capabilities, the CE of the Tier-2 also provides information describing the status of the cluster. This information is used by the match making engine which matches available resources to queued jobs. The Computing Element Monitoring (CEMon) service is responsible for monitoring and collecting the information relevant to the Computing Element.

The following CEMon sensors are implemented on the CE at CSCS:

- The sensor for managing information relevant to the CE itself, according to the Glue Schema;
- the sensor which publishes job status information of the CE.

Storage Element

Data storage system is one of the most important and challenging system in WLCG. The administrators from CSCS and LHC experiments put large amounts of effort towards deploying a set of integrated data management services with the WLCG middleware to enable movement and replication of data at a high speed, reliable management of distributed replicated data and associated metadata and optimized access to data. The Storage Element at CSCS is composed of a group of Sun Fire X4540 file servers, named from se01 to se10, and two Sun Fire X4200 servers running the dCache disk pool management software providing SRM interfaces.

dCache is a system jointly developed by Deutsches Elektronen-Synchrotron (DESY) and Fermilab. The dCache project provides a mechanism for storing and retrieving huge amounts of data among a large number of heterogeneous server nodes, which can be of varying architec-

tures (x86, ia32, ia64). It provides a single namespace view of all of the files that it manages and allows access to these files using a variety of protocols and interfaces of SRM, GridFTP, RFIO for storage resource access remotely or locally. A dCache system is composed of a number of domains, each running in its own Java virtual machine. Each domain (and its constituent cells) has a specific role to play, such as dealing with file access requests from clients, updating the file system namespace or facilitating communication between domains. The disk storage is partitioned into a set of disk pools, each of which can be assigned properties which control how files within the dCache behave, depending upon particular client requests. It is possible for all dCache services to exist on a single node. However, particularly at the Tier-2 level at CSCS, dCache deployed the SRM and namespace services separated from the data transport and disk pool services.

Configuration

The Swiss Tier-2 includes dozens of servers and hundreds of worker nodes. It is an extremely heavy daily work to manage, configure and upgrade such systems. To meet this challenge, the cluster machines configuration is managed through Cfengine. Cfengine is a software to high level policy language and autonomous agent for building expert systems to administrate and to configure large computer clusters. It is ideal for large-scale cluster management and is highly portable across varying computer platforms. CSCS developed a set of configuration to simplify the following tasks:

- To duplicate files from a central repository to all managed hosts, enforcing permissions and ownership;
- to create or to modify users and groups on all managed hosts;
- to start/restart system daemons on managed hosts;
- to mount Network File System (NFS) for sharing files on all managed hosts.

The Cfengine server process (cfservd) runs on host Cfengine.lcg.cscs.ch; once per hour, client hosts connect to the server (via the cfagent command, executed by the cfexecd daemon) and enforce the configuration as described by the server config files. With the present configuration, Cfengine will copy files from a central repository to the hosts in the cluster; that is, any local changes to a Cfengine-controlled file will be undone on the next Cfengine run. It means to actually edit a Cfengine-managed file, manager must edit the copy in the repository. Also, Cfengine will restore files ownership and UNIX permissions to what is recorded in the repository or the configuration file. The entire Cfengine configuration tree is saved in a SubVersion repository. Integration into a SubVersion repository provides several advantages:

- The configuration of the whole cluster is backed up safely;
- any change can be reverted: history of the configuration is preserved;
- subversion[3] logs provide a (minimal) documentation about what configuration changes were effected and by whom;
- edits can be done on another check-out and only committed to the main Cfengine repository when ready.

Scheduling Policies

Since there are three LHC experiments sharing the computing resources of the Swiss Tier-2, it is important to establish and to realize policies to share the resources among the VOs according to the MoU and to maximize the usage of the Swiss Tier-2. Currently, MAUI scheduling configuration follows the following goals:

- Each VO/ group gets a pre-set share of the cluster CPU time, specifically, ATLAS, CMS and LHCb jobs can run for about 2/5, 2/5, 1/5 of the total cluster CPU time; additional VOs can only run when the cluster has free CPUs;
- it should be possible to tune the parameters so that, among all jobs belonging to the same VO, those submitted by specified users or groups (e.g., production jobs) have a higher priority;
- some resources are reserved to each VO, so that the VO users are able to start jobs quickly even if they have not submitted jobs for a certain amount of time. (i.e., the non-dedicated resources have all been taken by jobs from other VOs.)

To achieve resource reservation, MAUI provides the Quality of Service (QoS) mechanism for implementing resource partitioning and reservation. Basically, a QoS is a set of MAUI configuration directives, which will be applied only to jobs tagged with that QoS; in particular, a QoS can alter the prioritization policies or enable access to a class of resources. Jobs can be assigned a QOS on the basis of submitting batch queue, UNIX user or UNIX group, etc.

The following classes of jobs deserve special treatment in the current configuration:

- WLCG Operations jobs and 'software manager' jobs: jobs coming with the 'Role=lcgadmin' VOMS attribute should be given the maximum priority, as they are either availability tests, or jobs for the cluster software maintenance;
- physics jobs from LHC VOs (ATLAS, CMS, LHCb): each VO is assigned its own QoS in

[3]Subversion is an open-source revision control system

order to grant access to reserved resources.

The scheduling policies and related setting were reviewed regularly to assure that all of VOs obtain reasonable resource according to the MoU regardless of the dynamic changes of the requirements of experiments.

Site Status and Grid monitoring

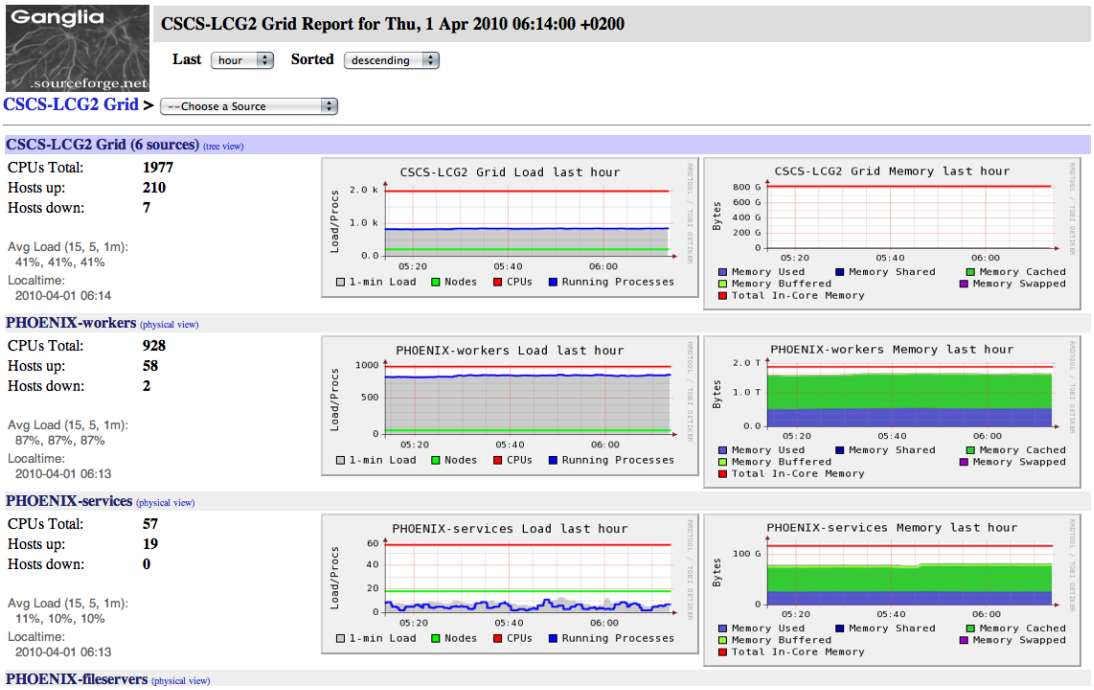

Figure 5.8: The ganglia monitoring page of the Swiss CMS Tier-2 at CSCS.

In order to monitor the status of the Tier-2 cluster, the Ganglia system is setup on all servers and worker nodes of the Swiss Tier-2. Ganglia is a scalable distributed system monitor tool for high-performance computing systems such as clusters and Grids. It allows users to check live or historical statistics (such as CPU load averages or network utilization, shown in Figure 5.8) for all machines that are being monitored.

By means of checking the web pages of Ganglia regularly, the administrators of the Tier-2 can monitor the status of individual machines or networks. It is also the essential means for administrators to investigate the misbehavior of services of the Tier-2. Moreover, to monitor the status or to examine the information of Grid services provided by the Tier centers, WLCG provides a series monitoring services for administrators of Tier centers:

- Central site functional tests (SFT): SAM test for the CE and SE;
- GSTAT monitoring page;

- GOC data base;
- GOC Accounting Graphs;
- LCG2 real time monitoring (VO summary reports);
- FZK monitoring, includes the FTS channel STAR-CSCS;
- Grid View Monitor;
- Grid Map Monitor.

The Swiss Tier-2 organizes some most useful plots and information sources in a web page 'Phoenix Monitoring Overview' as shown in Figure 5.9. This page contains:

- Plots of load of servers and worker nodes from Ganglia;
- status of batch jobs on the CE;
- status of CE;
- networking and file transfers of CMS and ATLAS respectively.

5.1.3 Configurations of CMS Specific Services

CMS runs two main workflows on the CMS Tier-2s: MC simulation and user analysis. In addition to the minimum WLCG infrastructure to be set at the Tier-2 sites, both activities require at least two services: PhEDEx and FroNTier. Those two services were implemented on the CMS VObox running on a virtual machine of the Swiss CMS Tier-2. I took part in the deployment of those services and their management and troubleshooting task.

PhEDEx

PhEDEx is the key service for CMS data management on WLCG. To monitor the status of PhEDEx, we developed our own PhEDEx log analyser which runs as a Cron job and gives statistics about the downloads to our site. The log analyser outputs a text summary on instances status, error analysis, error messages and site statistics. The overall volumes and average transfer speed are also summarized for administrators. However, the central based transfer details page and quality plots page are also very useful for the monitoring. To investigate download errors, every running transfer and all failed transfers write their full logs to files at the status directory.

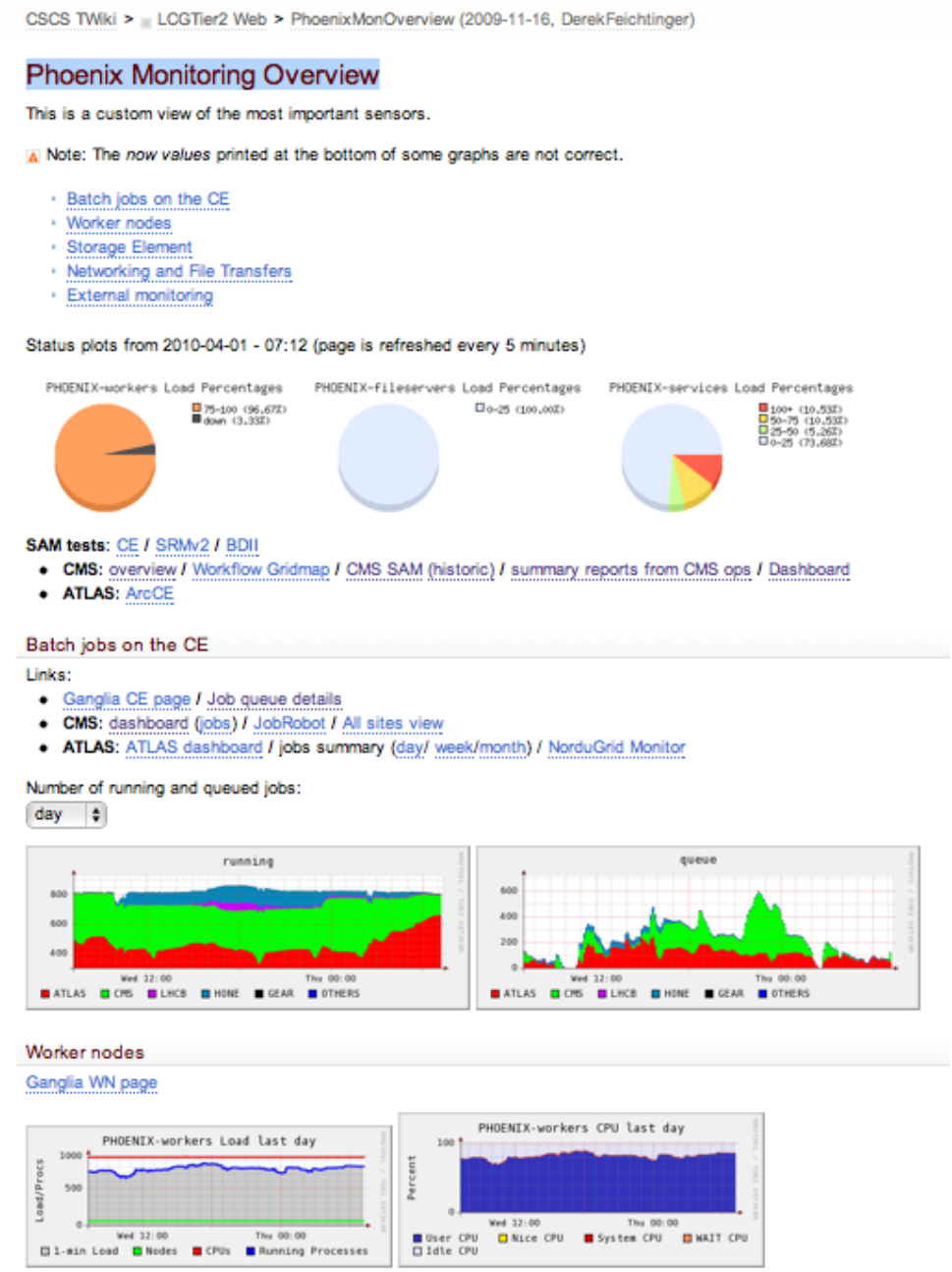

Figure 5.9: Phoenix Monitoring Overview web page collects most useful plots and links on a single page.

FroNTier

FroNTier is a http-proxy-server that is used to cache queries to the central databases, and thereby to reduce their load. Like PhEDEx, the service is running under a dedicated user dbfrontier on the CMS VObox. We found access logs can fill up the space rather fast, we usually run squid

with the access logs turned off. There is a central monitoring page[4] for the administrators.

5.2 Commissioning of Swiss CMS Tier-2

The successful commissioning of the Swiss CMS Tier-2 relied on a decent deployment of the CMS services on CMS VObox, experience gained from the maintenance works, reliable WLCG middleware (e.g., SE, CE) setup and high-performance infrastructure (e.g., servers, networks) provided by the Swiss Tier-2 center at CSCS. During the commissioning of the Swiss CMS Tier-2, we successfully met the CMS Data Challenges. The performance and reliability of the CMS Tier-2 were satisfied. Furthermore, we apply those experience obtained from the CMS Tier-2 at CSCS on the design and deployment of the Swiss CMS Tier-3 at PSI.

This section presents the experience we gained at Swiss CMS Tier-2 from its deployment and commissioning.

5.2.1 Administration of CMS Specific Services

The administration work could be divided into three types of tasks: to maintain the configuration the CMS software and upgrades of the software were necessary; to monitor the status of the services and status of running/terminated jobs; to identify and to resolve problems found from monitoring.

The monitoring tasks for the administrators for CMS Tier-2 are:

- To monitor CMS Specific Service: PhEDEx and FroNTier;
- to monitor and to manage datasets transfer and SE for CMS;
- to monitor CMS jobs, trace and debug problems.

One of the major tasks of administration is to ensure that all the CMS specific services are under the right status. By means of examining the plots of the worker nodes, service nodes and file servers on the web page 'Phoenix Monitoring Overview':

- To check the workload plots of the worker nodes, service nodes, and file servers. The service and fileserver pie charts must show no black parts (e.g., nodes down). A few worker nodes that are down are not so critical, but all service and fileserver should be in running status;
- to check all Site Availability Monitoring (SAM) tests, especially the CMS SAM tests;
- to check the graphs for running and queued jobs. For example, one should only see a

[4] http://frontier.cern.ch/squidstats/indexcms.html

number of queued CMS jobs, if the cluster is filled with running jobs. If jobs stay in the queue despite free slots on the cluster, something must be wrong with the scheduling;

- to check the free storage space graph for CMS, and to take note of the trend shown over the last week. The administrator can check how much space is taken up by users and datasets;
- to check the graphs for the dCache movers. If a large number of queued movers (especially if it is still growing) were observed, one should notify the CSCS administrators;
- to check PhEDEx activities by looking at the log analyser output on the PhEDEx download and export pages: if there is zero activity, the PhEDEx process status should be checked. If there are lots of transfer errors, it is necessary to analyse them based on the log analyser and post a support request on savannah;
- to check whether there are any pending data set requests. The decision whether to allow the request must be based on the available space and policy.

However, in most cases, it is not difficult to discover problems, but to settle problems usually needs a serious investigation and a thorough understanding of the system. In general, there are three kinds of issues we usually faced. First kind of issues could be addressed by the development and upgrade of software. Second kind of issues could be address by the miss-configuration on the CMS Tier-2 sites. The last kind of issues resulted from bugs or misuse of resources by the WLCG jobs.

We developed our own PhEDEx logs analyser and many other customized tools to speed up the problem identification. The PhEDEx logs analyser shown in Figure 5.10 summarizes the errors and other transfer information, e.g., average speed for every destination and source. It is very useful to check the transfer performance and identify source of problems.

5.2.2 Commissioning of Swiss CMS Tier-2

STEP'09 (Scale Testing for the Experiment Program in 2009) was a multi-VO exercise in the context of WLCG. The scope of the tests for CMS Tier-2 was the following:

- to demonstrate that CMS can have analysis at a scale that uses all pledged resources at Tier-2;
- the aim was to have most of the submissions by physicists doing real analysis while MC production went on in parallel. The goal was to completely use the portion of the Tier-2 processing dedicated to analysis (50%). Monitoring the totality of jobs should also be used to understand the available monitoring capabilities, and investigate the fair-share situation at Tier-2 sites.

```
 *** ERRORS from T1_TW_ASGC_Buffer:***
      9   TRANSFER error during TRANSFER phase: [GRIDFTP_ERROR] an I/O operation was cancelled globus_xio: Unable to connect to
se38.lcg.cscs.ch:2811 globus_xio: Operation was canceled globus_xio: Operation timed out
      7   TRANSFER error during TRANSFER phase: [GRIDFTP_ERROR] an I/O operation was cancelled globus_xio: Unable to connect to
se36.lcg.cscs.ch:2811 globus_xio: Operation was canceled globus_xio: Operation timed out
      3   TRANSFER error during TRANSFER phase: [GRIDFTP_ERROR] an I/O operation was cancelled globus_xio: Unable to connect to
se37.lcg.cscs.ch:2811 globus_xio: Operation was canceled globus_xio: Operation timed out
      1   TRANSFER error during TRANSFER phase: [GRIDFTP_ERROR] an I/O operation was cancelled globus_xio: Unable to connect to
se33.lcg.cscs.ch:2811 globus_xio: Operation was canceled globus_xio: Operation timed out

 *** ERRORS from T2_US_Caltech:***
      1   SOURCE error during TRANSFER phase: [GRIDFTP_ERROR] an end-of-file was reached globus_xio: An end of file occurred (possibly the
destination disk is full)

 *** ERRORS from T1_US_FNAL_Buffer:***
     38   TRANSFER error during TRANSFER phase: [GRIDFTP_ERROR] an I/O operation was cancelled globus_xio: Unable to connect to
se38.lcg.cscs.ch:2811 globus_xio: Operation was canceled globus_xio: Operation timed out
     27   TRANSFER error during TRANSFER phase: [GRIDFTP_ERROR] an I/O operation was cancelled globus_xio: Unable to connect to
se37.lcg.cscs.ch:2811 globus_xio: Operation was canceled globus_xio: Operation timed out
     26   TRANSFER error during TRANSFER phase: [GRIDFTP_ERROR] an I/O operation was cancelled globus_xio: Unable to connect to
se36.lcg.cscs.ch:2811 globus_xio: Operation was canceled globus_xio: Operation timed out
     14   SOURCE error during TRANSFER_PREPARATION phase: [GENERAL_FAILURE] AsyncWait
      7   TRANSFER error during TRANSFER phase: [GRIDFTP_ERROR] an I/O operation was cancelled globus_xio: Unable to connect to
se33.lcg.cscs.ch:2811 globus_xio: Operation was canceled globus_xio: Operation timed out
      4   TRANSFER error during TRANSFER phase: [TRANSFER_TIMEOUT] gridftp_copy_wait: Connection timed out

SITE STATISTICS:
==================
                        first entry: 2010-05-21 21:00:02      last entry: 2010-05-21 08:59:55
    T1_ES_PIC_Buffer (OK:  83  Err:  10  Exp: 107  Canc:   0  Lost:   0)   succ.: 89.2 %   total:   88.1 GB  ( 2.0 MB/s)
   T1_IT_CNAF_Buffer (OK:   1  Err:   0  Exp:   0  Canc:   0  Lost:   0)   succ.: 100.0 %   total:    1.8 GB  ( 0.0 MB/s)
   T1_TW_ASGC_Buffer (OK: 924  Err:  20  Exp:   0  Canc:   0  Lost:   0)   succ.: 97.9 %   total: 1300.0 GB  (30.1 MB/s)
    T1_UK_RAL_Buffer (OK: 764  Err:   0  Exp:   0  Canc:   0  Lost:   0)   succ.: 100.0 %   total: 1443.0 GB  (33.4 MB/s)
   T1_US_FNAL_Buffer (OK: 910  Err: 116  Exp:   0  Canc:   0  Lost:   0)   succ.: 88.7 %   total: 1392.6 GB  (32.2 MB/s)
       T2_CN_Beijing (OK:   6  Err:   0  Exp:   0  Canc:   0  Lost:   0)   succ.: 100.0 %   total:    7.6 GB  ( 0.2 MB/s)
       T2_US_Caltech (OK:  54  Err:   1  Exp:   0  Canc:   0  Lost:   0)   succ.: 98.2 %   total:   57.4 GB  ( 1.3 MB/s)
       T2_US_Florida (OK:   1  Err:   0  Exp:   0  Canc:   0  Lost:   0)   succ.: 100.0 %   total:    0.0 GB  ( 0.0 MB/s)
          T2_US_MIT (OK:  25  Err:   0  Exp:   0  Canc:   0  Lost:   0)   succ.: 100.0 %   total:   27.0 GB  ( 0.6 MB/s)
       T2_US_Purdue (OK:   2  Err:   0  Exp:   0  Canc:   0  Lost:   0)   succ.: 100.0 %   total:    1.5 GB  ( 0.0 MB/s)

TOTAL SUMMARY:
==================
                        first entry: 2010-05-21 21:00:02      last entry: 2010-05-21 08:59:55
total transferred: 4022.4 GB  in 12.0 hours
avg. total rate: 95.4 MB/s = 762.9 Mb/s  = 8046.1 GB/day
```

Figure 5.10: The output of the PhEDEx logs analyser. It summarizes the errors and other transfer information, e.g., average speed for every destination and source.

The Swiss CMS Tier-2 performed well during the STEP'09. In week 24 (8 - 14 June 2009), the Tier-2 got full load, and Tier-2 delivered 98% of the pledged slots, while also being under considerable load from ATLAS (though not from LHCb). The job failures due to the site fail category are small, except for the night where 3 WNs died because of disk failures and acted as black holes. In week 25 we delivered 185% of the pledged resources (shown in Figure 5.11, since at this time CMS was basically alone on the cluster. There were almost no ATLAS jobs. We did not encounter many site fail entries.

Job Slot Usage

Week	Days Site Not Ready	Jobs Terminated	Avg Slots Used	Analysis Pledge	Percentage of Analysis Pledge Used (%)	CPU (sec)	WC (sec)	CPU/WC (%)	CPU Success (sec)	WC Success (sec)	CPU/WC Success (%)	CPU Success Rate (%)	WC Success Rate (%)	Avg Job (min)
19	2	2133	35	170	20	11926698	21609030	55	10471281	15066706	69	87	69	168
20	0	5240	34	170	20	19932893	21143856	94	15768291	16587457	95	79	78	67
21	0	4587	28	170	16	15348115	17160032	89	14563797	16222269	89	94	94	62
22	0	5661	27	170	15	10913560	16797031	64	10485177	16149701	64	96	96	49
23	0	8642	111	170	65	55876532	67285766	83	54465292	65531487	83	97	97	129
24	1	13092	168	170	98	67627192	102032006	66	62003189	95844803	64	91	93	129
25	3	20014	316	170	185	163142897	191311244	85	158617941	184031306	86	97	96	159

Figure 5.11: In week 25 the CMS Tier-2 delivered 185% of the pledged resources. The low values for Percentage of Analysis Pledge Used have the primary reason of not enough jobs being sent.

The CPU efficiency of the CMS jobs was good, according to the local monitoring (shown in Figure 5.12). The access to files by CMS over dCap[5] was good, with one exception when too many jobs wanted access in parallel to the JobRobot data sets. The ATLAS jobs did have an efficiency problem. There were some ATLAS jobs which did heavy presaging of files from the SE to the WN using dccp command. Many nodes ended up with massive I/O wait due to the many dccp processes fighting for the bandwidth and the access to the local scratch space. This can be detected and alleviated by replication of the files over the cluster. But without a HSM behind dCache this replication must get triggered by hand.

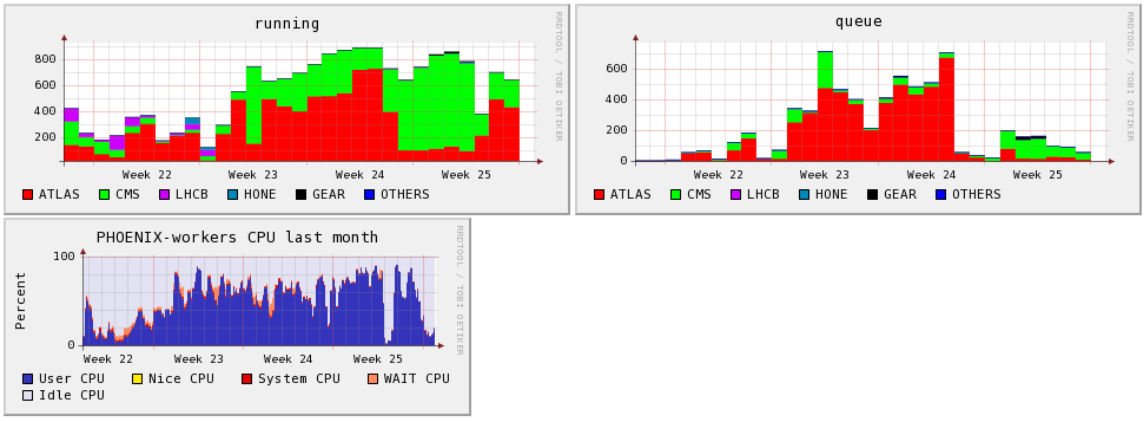

Figure 5.12: The numbers of running jobs (upper left), queued jobs (upper right) and CPU loads (lower left) in the Swiss Tier-2 during weeks 22–25, 2009. When all CPU slots were occupied by running jobs, the total CPU loads topped more than 90%.

With respect to CE configuration, we realized that we still need to implement correct fair share between the user 'CMS prod' for official CMS production jobs and other (analysis) jobs. This has been done in the downtime after STEP'09.

5.2.3 Performance of Swiss CMS Tier-2

The most important performance benchmark for the CMS Tier-2 is the overall transfer speed of datasets to or from other Tier sites. On September 2009, with our monitoring, steady PhEDEx throughput of >130 MB/s over last 12 hours was archived, which was much higher than the 20MB/s of requirements laid down in the MoU. The rate plot from the PhEDEx central web site is shown in Figure 5.13. Even though we were doing multiple pool migrations over the file servers, the infrastructure was managed to keep up a high and quite reliable throughput over a period of 12 hours from 18 to 19 September 2010.

The activity table for a period of 24 hours (Figure 5.14) shows the average transfer rate for

[5]DCache Access Protocol (dCap) is the native random access I/O protocol for files within dCache

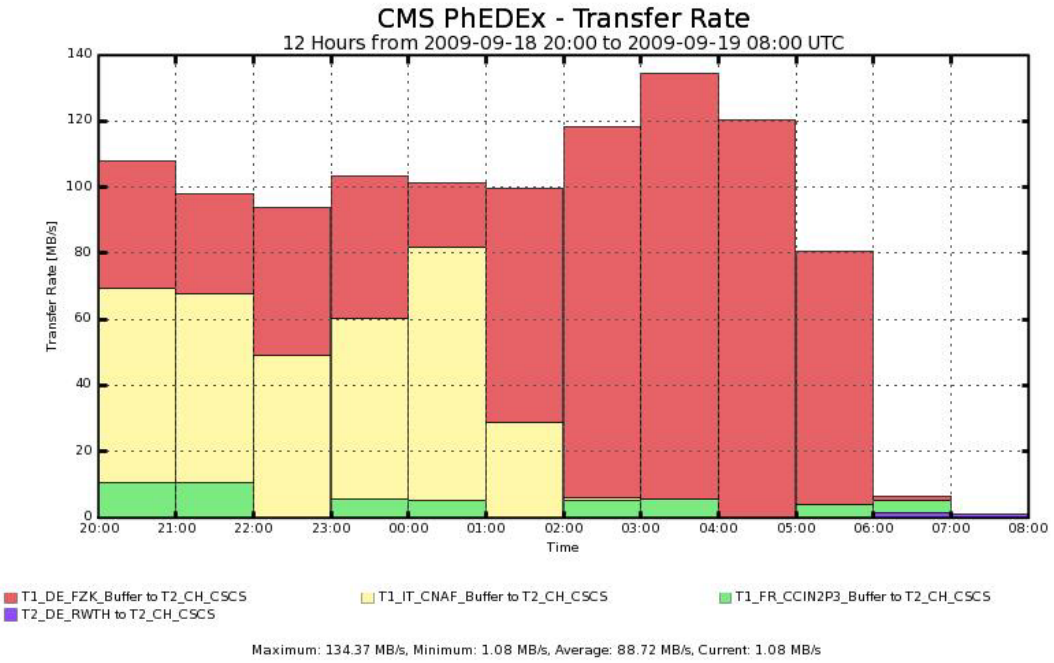

Figure 5.13: The rate plot for a period of 12 hours from 18 to 19 September 2010 from the PhEDEx central web site shows that the maximum transfer rate is slightly higher than 130MB/s. [http://cmsweb.cern.ch/phedex/prod/Activity::RatePlots]

Time span Last day | Include links with nothing but errors | Update

Last day

To		From	Files	Total Size	Rate	Errors	Expired	Avg. Est. Rate	Avg. Est. Latency
.*CSCS.*	or								
T2_CH_CSCS		T1_DE_FZK_Buffer	2208	3.1 TB	37.3 MB/s	238	22	40.0 MB/s	1d8h06
T2_CH_CSCS		T1_IT_CNAF_Buffer	1714	2.4 TB	29.2 MB/s	7	-	44.2 MB/s	11h56
T2_CH_CSCS		T1_FR_CCIN2P3_Buffer	387	363.1 GB	4.3 MB/s	13	8	9.0 MB/s	23h12
T2_CH_CSCS		T1_UK_RAL_Buffer	342	335.8 GB	4.0 MB/s	16	-	10.8 MB/s	2h36
T2_CH_CSCS		T3_US_FNALLPC	36	33.8 GB	410.0 kB/s	4	-	4.3 MB/s	0h55
T2_CH_CSCS		T2_DE_RWTH	22	27.2 GB	329.7 kB/s	24	-	7.4 MB/s	0h25
T2_CH_CSCS		T1_US_FNAL_Buffer	6	8.8 GB	106.6 kB/s	2	-	3.1 MB/s	0h20
Total			4715	6.2 TB	75.6 MB/s	304	30	-/s	0h00

Figure 5.14: The table from the PhEDEx central web site shows the summary of all of activity transfers to the Swiss Tier-2 during a period of 24 hours from 18 to 19 September 2010. [http://cmsweb.cern.ch/phedex/prod/Activity::Routing]

different sources. The average rate of the transfers from the FZK and CNAF Tier-1 sites were higher than 40MB/s.

Figure 5.15 shows the plot of the requested volume for the past 72 hours till 19 September 2010. The sharp rise corresponds to a number of the B-physics datasets ordered for our users. The rapid fall shown the high throughput of the transfer links.

From the monitoring plots (Figures 5.16, 5.17 and 5.18), it is clear that this feat was accomplished

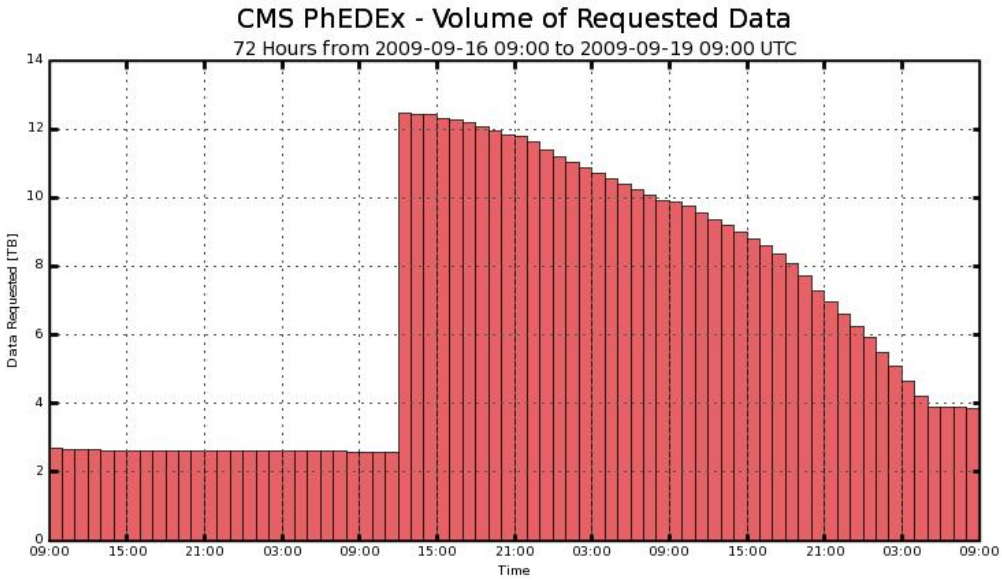

Figure 5.15: The plot of the requested volume for a period of 72 hours. The sharp rise corresponds to a number of B-physics datasets ordered for our users. [http://cmsweb.cern.ch/phedex/prod/Activity::RatePlots]

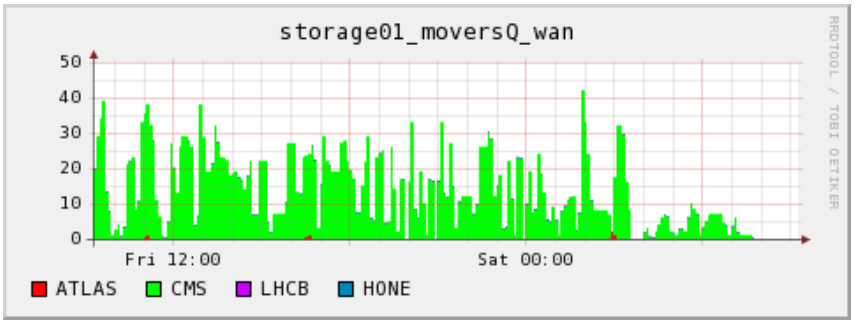

Figure 5.16: The rate plot from the Swiss Tier-2 local monitoring. [https://twiki.cscs.ch/bin/view/LCGTier2/PhoenixMonOverview]

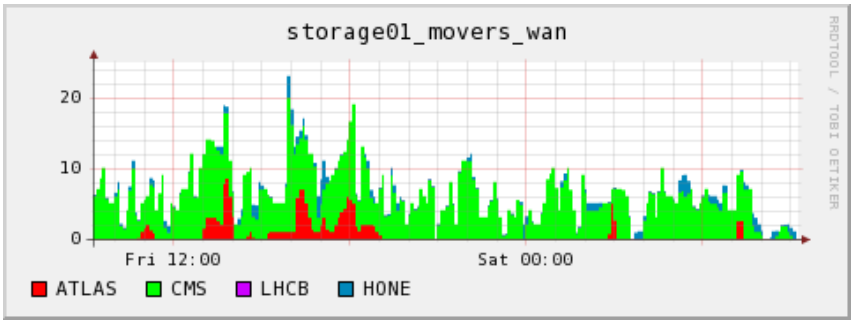

Figure 5.17: The rate plot from the Swiss Tier-2 local monitoring. [https://twiki.cscs.ch/bin/view/LCGTier2/PhoenixMonOverview]

by a limited amount of WAN[6] movers (always < 10). A lot of the active pools have set their

[6] A wide area network (WAN) is a computer network that covers a broad area (i.e., any network whose communications links across sites).

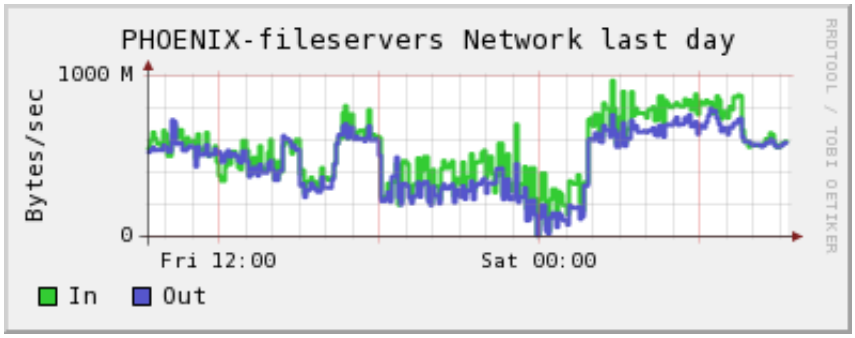

Figure 5.18: The rate plot for networking from the Swiss Tier-2 local monitoring. [https://twiki.cscs.ch/bin/view/LCGTier2/PhoenixMonOverview]

maximal movers to 2 only, while the others got queued. Therefore, the performance of transfer could be improved by increasing the WAN movers of dCache. However, the performance could be also be limited by the links of remote source sites.

5.3 Swiss Tier-3 Center at PSI

Whereas the CSCS Tier-2 center provides a part of its resources to the LHC experiments as a whole, the CMS Tier-3 center is completely dedicated to the Swiss CMS groups. A large fraction of the end-user analysis is carried out on the Tier-3 resources according to the Tier architecture of the CMS computing model. In July 2008, a CMS Tier-3 center has been established at PSI in Villigen. The Tier-3 was running smoothly in the testing mode in September 2008, and in the production mode since November 2008. After the smooth and efficient running of the Tier-3 during 2009, the computing cluster and storage system was upgraded to the phase B at the beginning of 2010. Currently, the cluster specification is listed in Table 5.2.

No. of WNs	Processors	Cores/node	kCINT2000/core	No. of Cores	kSI2k
8	2*Xeon E5410	8	3.34	64	213.76
20	2*Xeon X5560	8	6.2	160	992

Table 5.2: Cluster Nodes of the CMS Tier-3 center at PSI.

5.3.1 Scheme of Swiss CMS Tier-3 center

The main difference between the Tier-2 and the Tier-3 center of CMS is not the scale of the computational and storage resources of sites, but the services they implemented and their major tasks in WLCG. Since the CMS Tier-3 at PSI is dedicated to meet the analysis demands from the Swiss CMS groups, the setup of the Swiss CMS Tier-3 at PSI is optimized for developing and

running end-user physics analysis jobs. Furthermore, to be consistent with the WLCG running environments and to provide users and their jobs ability to access storage resources integrated by WLCG, all necessary WLCG services were implemented in the Tier-3, e.g., SE and UI. However, to reduce the complexity of the system, there is no full-function WLCG CE. The layout of the Tier-3 is shown in Figure 5.19. I have been involved in the setup and maintenance the Tier-3 center.

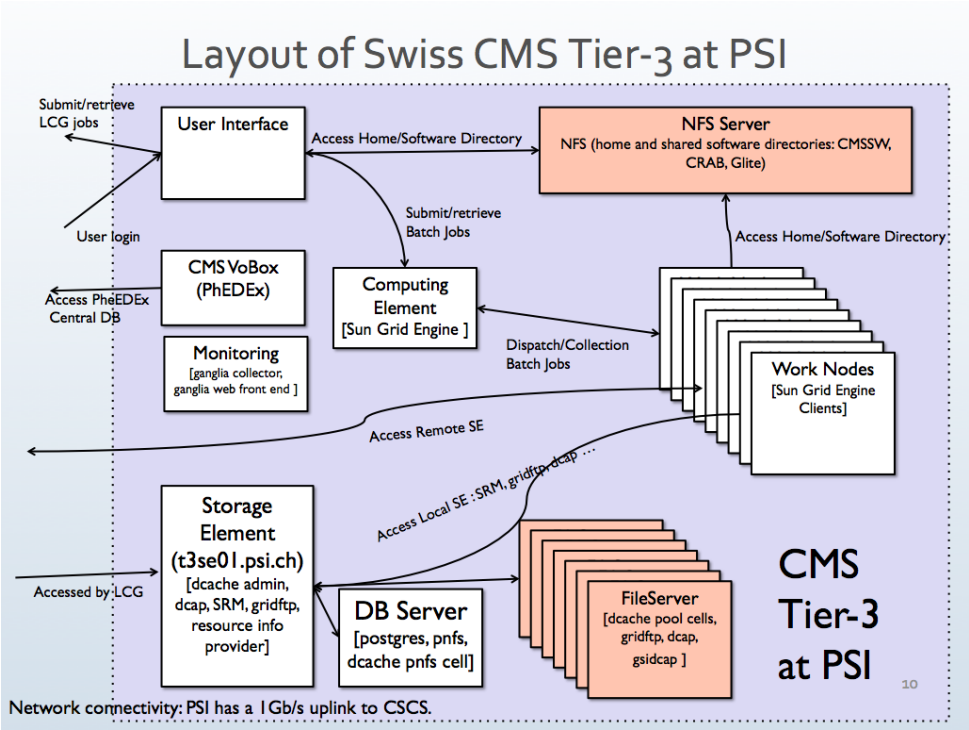

Figure 5.19: The scheme of the Tier-3 center at PSI.

The User Interface (UI) is the unique access point for our users to use the computational and storage resource of the Swiss Tier-3. However, users can also submit jobs from the UI to Grid with the gLite middleware installed on the UI. From UI, users can prepare their analysis programs, test and submit jobs to the worker nodes of the cluster.

The Sun Grid Engine SGE local batch system is used on the Tier-3. Many other sites reports SGE is more stable and has much less scalability faults than other popular batch systems, e.g., PBS and PBS Pro. The master node of SGE is the CE, which takse responsibility of job scheduler and management.

After the study of the performance of NFS, we proved that the NFS has good performance on file read. We did the test on the cluster for multiple remote accesses in parallel. The result is shown in Figure 5.20. Even though it is not the latest technology for the files sharing on the cluster, it is very stable and shows no serious read performance problem on the cluster with the

size of the Tier-3.

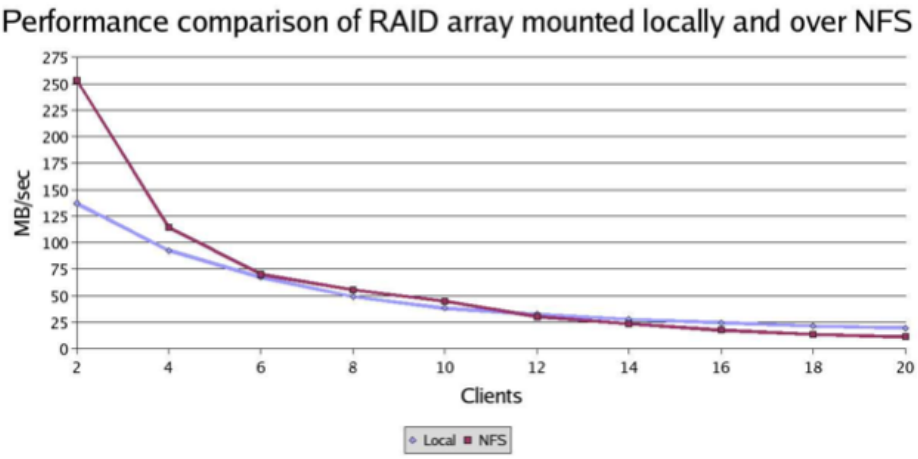

Figure 5.20: The read performance of files remote accesses in parallel on NFS (red line) against on local file system (light purple line).

Storage

The Tier-3 deployed ZFS as the file system and dCache as the storage management software. Since this is also the file system as well as the software used in the Swiss Tier-2 at CSCS. Therefore, our experience from the Swiss Tier-2 at CSCS benefited the deployment and maintenance of storage systems at the Tier-3. The ZFS file system has many advantages. Unlike traditional file systems, which reside on single devices and thus requiring a volume manager to use more than one device, the ZFS file system is built on top of the virtual storage pools called zpools. The zpool consists of one or more groups of disk drives. Those disks may be configured in different ways, depending on needs and space available. In addition, pools can have hot spares[7] to compensate for failing disks. ZFS also supports both read and write caching, for which special devices can be used. The zpool composition is not limited to similar devices but can consist of ad-hoc, heterogeneous collections of devices, which ZFS seamlessly pools together, subsequently doling out space to diverse file systems as needed. Therefore, arbitrary storage device types can be added to existing pools to expand the size for the storage system of the Tier-3 in the future.

ZFS uses a copy-on-write transactional object model. Besides ensuring data integrity, another advantage of copy-on-write is that when ZFS writes new data, the blocks containing the old data can be retained, allowing a snapshot or clone version of the file system to be maintained. ZFS

[7] A hot spare is used as a fail-over mechanism to provide reliability in system configurations.

snapshots and clone are created very quickly, since all the data composing the snapshot and clone is already stored; they are also space efficient, since any unchanged data is shared among the file system and its snapshots. These function makes data backup and recovery much easier than any traditional file system.

However, we cannot assign per user quotas within the same ZFS file system. Every user is entitled to a share of 100 GB home area. This should be plenty space for all normal use cases. All larger amounts of data should be placed on the storage element in the user's personal directory. The Tier-3 monitoring page contains a link to the web page that summarizes detailed space consumption of users. So we rely on a certain self control by the users. A user may temporarily use more than his/her share as long as he/she does not fill up the user home system and create inconveniences for other users.

PhEDEx

PhEDEx was deployed on the Tier-3. However, we have only established several links to the Swiss Tier-2 at CSCS and the Tier-1 at FZK. Currently PhEDEx will only download data sets via direct links. So, if a data set is not located at one of these directly linked centers, it is necessary to order the data set to both the Swiss Tier-2 at CSCS and the Swiss Tier-3 at PSI. Even though PhEDEx has the capability to transfer data through multiple hops, it is deactivated, since it causes problems with the management of the storage space of the centers that are used as caches.

To monitor the status of PhEDEx, we use the similar log analyser which runs as a Cron job and gives statistics about the downloads to our site. The log analyser output a text summary about instances status, error analysis, error messages and site statistics. Administrators can check the result of the analyser on the monitoring page of the Tier-3.

CMSSW Deployment

The CMSSW is the core software collection of the CMS data analysis. It was used in most of CMS production and data analysis programs. We deployed the CMSSW on the NFS. UI and all worker nodes access CMSSW via NFS. Since the NFS has good read performance and CMSSW is read-only for jobs, it should be not the bottleneck for the worker nodes even there were many parallel reading. The local disk of the worker nodes is used for the runtime disk cache for the jobs.

5.4 CRAB and Adaptation for SGE

5.4.1 CRAB

CMS Remote Analysis Builder (CRAB) is the official CMS analysis software to provide users easier interfaces with the Grid environment by hiding the system complexities. It allows an easy access to the data distributed over Grid in a transparent way and the users do not require any deep knowledge about the Grid.

The development of CRAB focuses on improving the interface to the CMS users and increasing the automation ability for CMS job submission. Since the CRAB did not support SGE batch system when we were setting up the Swiss Tier-3, I programmed a new SGE scheduler module to adapt CRAB into the Tier-3 local analysis environment. This section introduces the overall design of CRAB and my development work of the SGE scheduler module for CRAB.

CRAB is installed on the UI which is the user access point to the WLCG. It supports any CMSSW (the CMS software framework) based program, with any modules/libraries. It also deals with the output produced by the program. From a user point of view, the basic steps of the workflow of CRAB are:

- Job Creation: interact with data discovery services (DBS and DLS); split the task into smaller jobs; prepare input sandbox for jobs;
- Job Submission: interact with Resource Broker, Workload Management System and proxy services to submit jobs to sites matching the user requirements;
- Job Status: check the status of the jobs using the CMS Bookkeeping System (BOSS);
- Job Management: retrieve information of aborted jobs; kill jobs and resubmit failed jobs if necessary;
- Job Output: retrieve the job output from WLCG (output sandbox) or the local cluster; transfer the output to SE specified by the user.

The programming language to develop CRAB is Python [74]. It reduces development time and simplifies maintenance. The architecture of CRAB is applied with a modular software approach: independent components are implemented as agents communicating through an asynchronous and persistent message service.

CRAB uses the module of BOSSLite to submit jobs. BOSSLite is a Python library developed to interface the CMS Workload Management tools with the Grid middleware or local batch systems. It relies on a database to track and to log information into an entity-relation schema. Information is logically remapped into Python objects that can be transparently used by the

CMS framework and tools. It's important to understand well the structure of BOSSLite before the implementation of new local scheduler.

The BOSSLite architecture is showed in Figure 5.21. BOSSLite has to be highly efficient to avoid introducing bottlenecks in the Grid operations and to provide safe operation in a multi-processing/multi threaded environment. For that reason, the interaction with the database is performed through reliable sessions and connection pools. A database session provides access to database operations through a generic abstract interface, enabling standard operations such as open/close connection, query, insert/update rows and so on.

In the similar way, the Scheduler part of BOSSLite provides a generic abstract interface, enabling standard operations such as job submission, tracking, cancel and output retrieval. The specific interfaces for scheduler are implemented in specific plug-ins, loaded at Runtime. At that time, plug-ins were implemented including gLite (EGEE), OSG (Open Science Grid) and ARC (NorduGrid) middleware stacks and the LSF batch system.

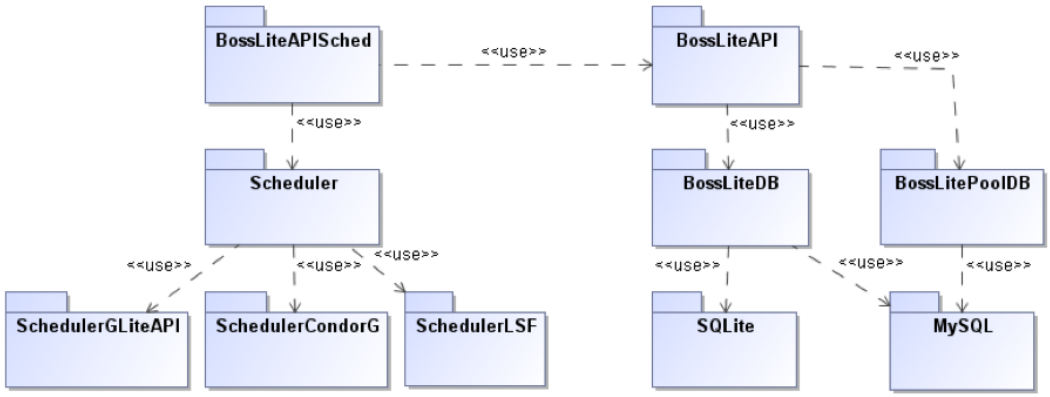

Figure 5.21: Schematic view of the BOSSLite Architecture.

A set of high level API connects the schedulers and the database part, implementing the default behavior as well as standard methods of general utility. The high level scheduler API depends on the database high level API, allowing coherent update and handling of the information to be stored in the database.

The typical CMS analysis requires access to a huge amount of data, usually split in many files of datasets. Because in general the same processes have to run over the whole sample, the analysis task can be split in many jobs that may run in different CPU cores in parallel. The only difference among them is given by the parameters identifying the portion of data to be accessed. The same applies for Monte Carlo jobs, instead of being analysed, data are produced by parallel jobs in small files to be merged later.

The design of the BOSSLite database structure reflects the typical CMS analysis task. A top entity called task groups similar jobs. It stores common information such as common jobs requirements, dataset to be accessed/produced and so on. Jobs are then provided with a static part, characterizing the job itself by recording for instance arguments and file names, and a dynamic part. The latter stores information, such as scheduling timestamps, job status and destination. Since a job may fail for many reasons, there should be as many resubmissions as needed: in order to record the execution information of every submission for logging purposes, they are stored in the so-called running instances. A job may have a running instance for each resubmission.

Since different CMS workload management tools (e.g., for production or for analysis, as well as for Grid or for local systems) may have a different setup and workflow, the scheduler interface of BOSSLite has to be flexible enough to ensure the coherent usage of the features optimizing the tool workflow. Since the CMS computing model uses its own data location system, the WMS match-making has the main role to perform a load balancing among sites that are hosting the data to be accessed or Monte Carlo simulated data to be produced. The usage of the bulk match-making further reduces the actual load of the WMS.

The access to the gLite features is made through the Workload Management Proxy (WMProxy) Python API. This allows the association between BOSSLite jobs and their Grid identifiers. On the other hand, the access to the features of local batch systems is through the command line. The scheduler interface of BOSSLite operates the command line interface of local batch systems and then parse their complicate standard output and error output. Among other things, it has to re-implement UI features such as configuration files interpretation and sandbox transfer.

Since the CRAB did not support the SGE local batch system when we started building the Swiss CMS Tier-3 at PSI, I developed the scheduler for CRAB to submit jobs to the SGE batch system. The development was divided into two parts: (i) the development of the CRAB scheduler interface for SGE; (ii) the development of the BOSSLite scheduler interface for SGE.

5.4.2 CRAB Scheduler Interface for SGE

The UML diagram for Scheduler Interface Class of Scheduler module of CRAB is shown in Figure 5.22. During the Runtime, the main module of CRAB calls the scheduler module of CRAB. Scheduler module is a catch-all class for all kinds of scheduler regardless Grid or local scheduler. The Scheduler module implements all functions and holds the status of the jobs. The functions include submission of jobs, query of status and so on. The Scheduler module will call the SchedulerLocal module if a user sets the type of Scheduler as a local scheduler, e.g., PBS

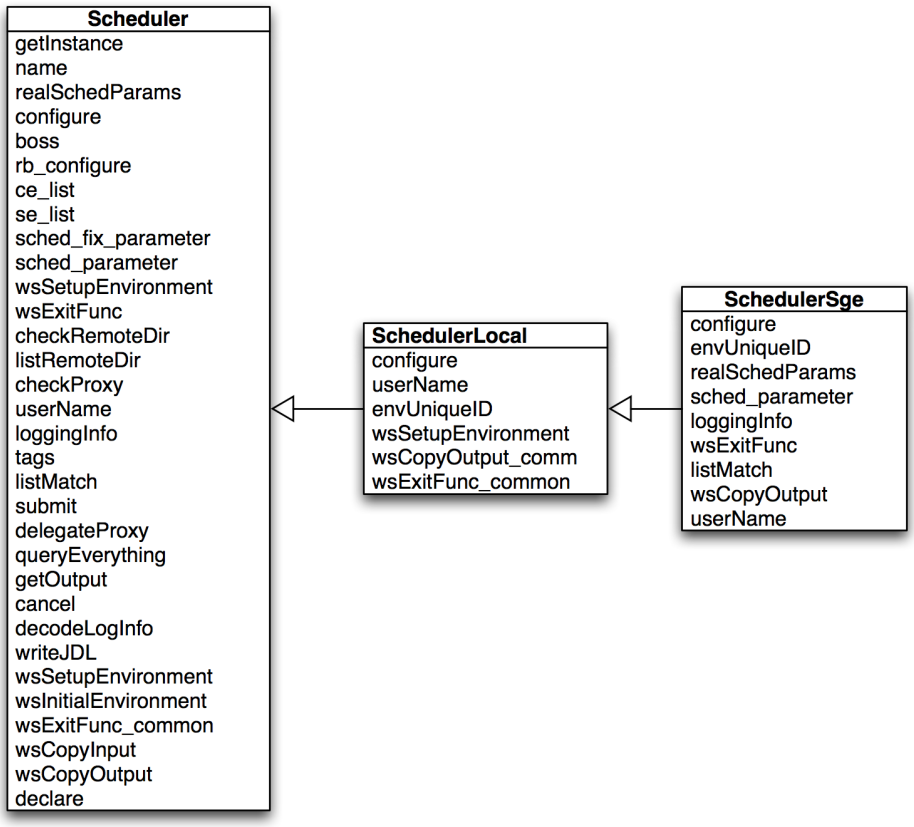

Figure 5.22: The UML diagram for Scheduler Interface Class of Scheduler module of CRAB.

or SGE in the configuration file of jobs. The SchedulerLocal module implements all specific functions for the configuration of local scheduler, e.g., retrieval of the output of jobs, return the status of exit code of jobs and so on.

Obviously the different local batch systems have different command line user interfaces and command sets. The mission of the SchedulerSge module of CRAB which contains the class SchedulerSge is to prepare the configuration and parameters for the local batch system. The class inherits from SchedulerLocal, conceals those function and store the status of jobs for CRAB.

To adapt CRAB for SGE local batch system, the following member functions of SchedulerSge class were implemented:

configure

The mail function of configure is to set the path for the CMSSW data storage.

envUniqueID

The function of envUniqueID is to create a unique ID for the job during the session.

realSchedParams

The function of realSchedParams is to create the parameters to be used by the SGE scheduler.

sched_parameter

The function of the sched_parameter is to create the parameter to be used by the BOSSLite Scheduler Interface Module.

loggingInfo

The function of loggingInfo is to return the log information for CRAB.

wsExitFunc

The function of wsExitFunc is to execute some operations before exit. These operations include the collection of the standard output and error output of jobs.

listMatch

The function of listMatch is to set the blacklist or whitelist of the destination sites.

wsCopyOutput

The function of wsCopyOutput is to configure the blacklist or whitelist of the destination sites for local scheduler.

5.4.3 BOSSLite Scheduler Interface for SGE

The UML diagram for Scheduler Interface Class of Scheduler module of BOSSLite is shown in Figure 5.23. During the Runtime, the Scheduler Interface module of CRAB calls the scheduler module of BOSSLite for the actual operation of the local batch system. The member functions of Scheduler module for SGE of BOSSLite are:

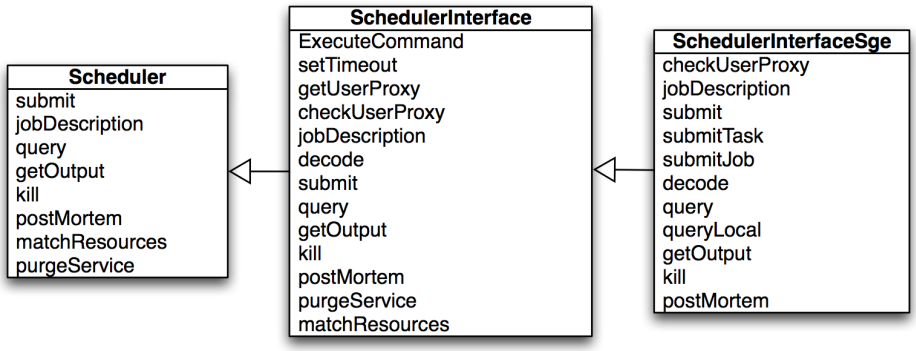

Figure 5.23: The UML diagram for Scheduler Interface Class of Scheduler module of BOSSLite.

jobDescription

Retrieve scheduler a specific job description from the output of the SGE command.

submit

This function sets up the submission parameters and submit jobs to SGE via function submitJob or submitTask.

submitTask and submitJob

To submit all jobs (submitTask) or specific job (submitJob) to the SGE local batch system by SGE command line interface.

decode

To prepare files for submission and retrieval of output jobs. These files includes input sandbox and output sandbox from worker nodes.

query,queryLocal

Query status and eventually other scheduler related information. These information will return to BOSSLite and be stored for future use.

getOutput

Retrieve output or just put it in the destination directory.

kill

Kill the submitted jobs.

The Scheduler Interface for SGE was used in the Swiss Tier-3 since the end of the 2008. I keep improving the codes according to the requests from our users. CRAB is our essential tools to submit CMSSW jobs from UI to cluster of the Swiss Tier-3. In addition to the advantages of hiding the complexity of the Grid and the local batch system, it also provides a unified way to submit jobs regardless of the sites. The Scheduler for SGE was accepted by the official release of CRAB since the version 2.5 in the middle of 2009.

5.5 Summary

As an important part of the contributions from Switzerland within the CMS collaboration, a high-performance Tier-2 center for CMS was set up at the Swiss National Supercomputing Center (CSCS) and a CMS Tier-3 center was set up at PSI. I participated in the construction, software deployment and configuration, commissioning and operation of the Swiss Tier-2 and the Tier-3 respectively at different levels:

CMS Computing Environment deployment, configuration and commissioning at the Swiss Tier-2:

The CMS computing environment has been deployed:

- CMS VoBox (CMS VO Specific Services) was deployed on a virtual machine for the Tier-2. The services included:
  - PhEDEx data management service and associated Authentication/Authorization, FTS for dataset transfer service;
  - FroNtier data base caching service for reducing load of servers;
  - an Apache installation used for displaying PhEDEx and PNFS monitoring information.
- User Interface for CMS users, including CMSSW experiment software and tools (CRAB, etc.);
- a set of tools were developed for monitoring, accounting, summarizing the dataset transfer errors for a better understanding of the systems as well as troubleshooting.

During the commissioning, we participated and met all of the CMS specific and globe WLCG data challenges and the CMS Tier-2 site showed high availability and good performance. Many

issues and problems were addressed and solved. There main issues were:

- Problematic links for dataset transfers were investigated. The configuration of the Tier-2 was improving during the commissioning and good communication between administrators of other Tier-sites was established;
- dCache was not fully mature during the commissioning. Many problems and potential issues were discovered and reported to the developers;
- one accidental issue happened from time to time are lost data on the SEs on the Tier-2. Files loss occurred because of massive hardware failure or dCache failure. We tried to minimize the contingency by now running everywhere on RAID6 / RAIDZ2, but it still could happen. The major part of the CMS files at the Tier-2 are replicated files from central data sets, thus these files are not a terrible loss. But some files belong to user's areas and they are most damaging. We developed a set of tools to prepare a pnfs name list of the lost files re-establish dCache/pnfs consistency by deleting leftover pnfs names with no on-site replicas.

Scheme design, construction software configuration at the Swiss CMS Tier-3 center:

With the experience gained from the Swiss Tier-2, the scheme of the Swiss CMS Tier-3 was optimized for the end-user analysis. Similar hardware to the Swiss CMS Tier-2 were chosen. But the scheme of deployment was different from the CMS Tier-2. With the advanced local batch system SGE and NFS share-based file system, the program development, submitting and debugging of the CMS analysis jobs on the UI of Tier-3 is much easier than that on the Grid. The software deployment and configuration was carried out smoothly in September 2008. A large fraction of analysis jobs from Swiss CMS groups was performed on the Swiss Tier-3. The Tier-3 has shown excellent performance and usability since the production phase started in November 2008.

CRAB is the official utility to create and submit CMSSW jobs to WLCG. Since the CRAB did not support our SGE batch system when we were setting up the Swiss Tier-3, I programmed a new SGE scheduler module to adapt CRAB into the Tier-3 local analysis environment. The CRAB on the Tier-3 provides a friendly and uniform way for Tier-3 users to submit CMSSW job to the Tier-3 SGE batch system or WLCG. The code of scheduler module was committed to the official code repository of CRAB.

The tools developed for monitoring, accounting, summarizing the dataset transfers errors at the Tier-2 were also developed on the Tier-3.

During the commissioning and operation of the Tier-3, due to the experience gained from the Tier-2, much less failures and problems have been encountered. Most of them were hardware

failure, especially the hard disk failure happened from time to time. With the RAID6 / RAIDZ2, most of them did not affect the running of the Tier-3.

6 Physics Preparation and Data Analysis of $Z \to e^+e^-$

With the fast increase of the delivered luminosity, the LHC collider will soon become a unique factory for the production of Z bosons. The reaction $pp \to Z + X$ with subsequent leptonic decays of the massive electroweak vector boson $Z \to \ell^+\ell^-$, has a large cross section and is theoretically well understood. Detailed measurements of the Z boson during the commissioning phase of LHC are very important for the understanding of the CMS detector as well as for testing the predictions of the Standard Model at $\sqrt{s} = 7$ TeV. The leptonic decay channels of the Z boson, in particular, provide one of the cleanest standard candles for a comprehensive understanding of both the CMS detector and the various Standard Model processes which are also important backgrounds for searches of new physics [75]. The decay of $Z \to e^+e^-$ provides a distinct signature in the silicon tracker and the crystal calorimeter of CMS [76].

At $\sqrt{s} = 7$ TeV a cross section of ~ 1 nb is predicted at the LHC for the $Z \to e^+e^-$ channel. Thus the event rate of $Z \to e^+e^-$ is very high and the Z boson event reconstruction based on the measurements in the silicon tracker and the crystal calorimeter can be used to monitor the data quality of the two sub-detectors. Experiments at LEP and the SLC have determined the properties of the Z boson with a precision of 0.1% or better. Currently, the Z boson mass is known to ± 2.2 MeV. Therefore, the process of $Z \to e^+e^-$ can be used to study the performance of the CMS detector and to monitor the data quality.

This chapter will discuss the physics preparation and data analysis of $Z \to e^+e^-$, as well as the data quality monitoring with electrons and positions. Since the reconstruction and identification must be done for both electrons and positions, the word 'electron' in this chapter means both electron and positron unless otherwise stated, to simplify the description.

6.1 $Z \to e^+e^-$ Production at the LHC

The dominant production mechanism for the electroweak gauge boson Z in proton-proton collisions is the weak Drell-Yan production process, where a quark and an antiquark annihilate to

form a vector boson. The corresponding Feynman diagram is shown in Figure 6.1. The $pp \to Z$ production, as shown in Figure 6.2, is dominated by the annihilation of the $u\bar{u}, d\bar{d} \to Z$ [77].

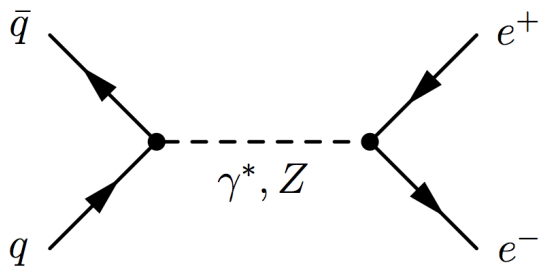

Figure 6.1: Leading order Feynman diagram for $q\bar{q} \to \gamma^*/Z \to e^+e^-$ process.

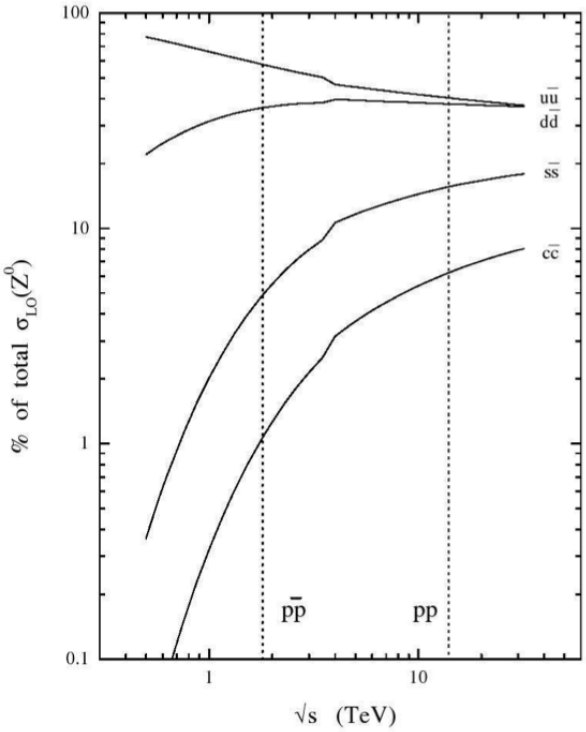

Figure 6.2: The parton decomposition of the total cross section of the Z production in pp and $p\bar{p}$ collisions. Individual contributions are shown as a percentage of the total cross section in each case [77].

The calculation of the total production cross sections of W and Z bosons, as shown in Figure 6.3, incorporate parton cross sections, parton distribution functions, higher-order QCD effects, and factors for the couplings of the different quarks and antiquarks to the W and Z bosons. The accuracy of the current calculations are limited by uncertainties in the parton distribution functions, as well as the higher-order QCD and electroweak radiative corrections.

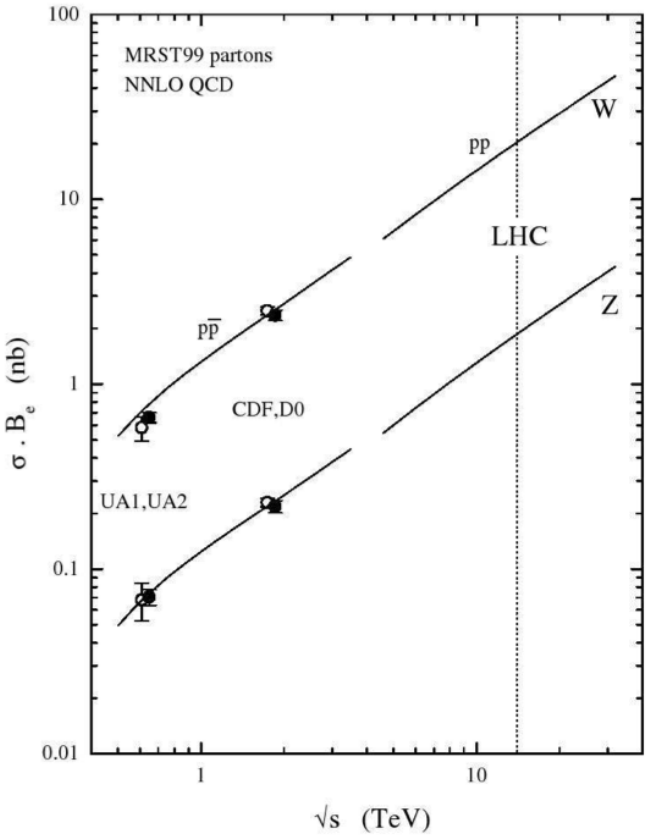

Figure 6.3: The prediction for the total cross section of W and Z production times the branching ratio to electrons in pp and $p\bar{p}$ collisions, as a function of the collider energy $\sqrt{s}$ [78].

Figure 6.4 shows the rapidity distribution of an on-shell Z boson at the LHC. Figure 6.5 illustrates the geometrical acceptance of $\gamma^*/Z \to e^+e^-$ events as a function of the rapidity. The acceptance is calculated as the fraction of the generated events in which both electrons fall within the CMS electromagnetic calorimeter fiducial region ($|\eta_{electron}| < 2.5$ with $1.4442 < |\eta_{electron}| < 1.560$ excluded). For rapidity close to zero, the acceptance is maximized but without reaching 1.0, meaning that there are some electrons expected outside the geometrical acceptance of ECAL. The geometrical acceptance drops to zero for rapidity close to 2.5.

6.2 Data and Monte Carlo Samples

The data and Monte Carlo samples were processed with CMSSW version 3_6_x.

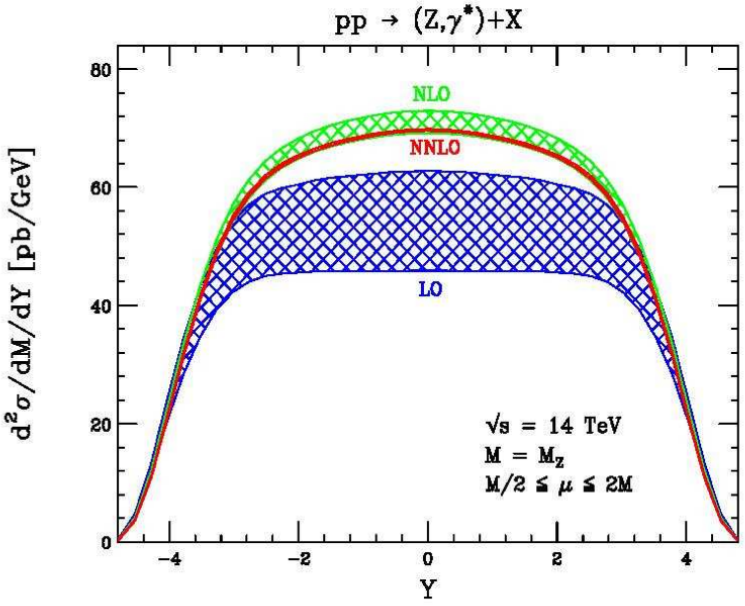

Figure 6.4: The rapidity distribution for $Z \to e^+e^-$ [78].

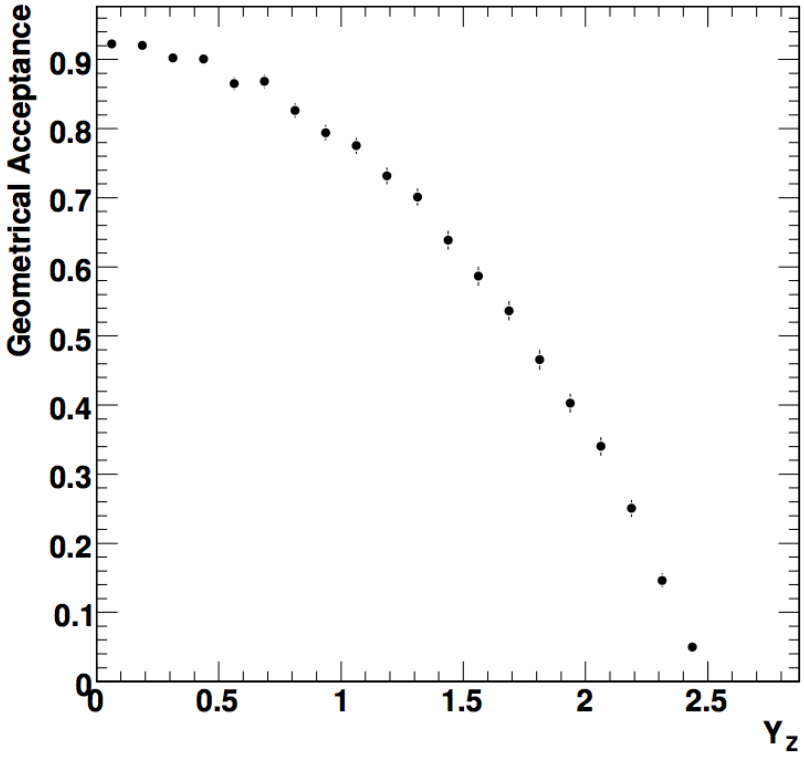

Figure 6.5: Geometrical acceptance for $\gamma^*/Z \to e^+e^-$ events as a function of Z rapidity, calculated from the fraction of the events in which both electrons fall within the CMS ECAL fiducial region.

6.2.1 Data Sample

The data used in the subsequent analysis were collected from May to September 2010. Application of basic beam, detector, and data-quality requirements resulted in a total integrated luminosities of 2.88 ± 0.32 pb^{-1}. The official JSON files [1] were used without modifications.

6.2.2 Monte Carlo Samples

Monte Carlo simulations were used in order to compare the data with theoretical predictions and to estimate the backgrounds from various physics processes.

For the relevant background two QCD samples and one electroweak sample are used:

- QCD EM Enriched, i.e., high p_T QCD events with an electromagnetic filter applied;
- QCD $b/c \to e$, i.e., high p_T QCD events with a transverse energy filter applied;
- $W \to e\nu$, where a jet fakes an additional electron.

Another potential background comes from $t\bar{t}$, with the t-quark decaying into Wb, as shown in Figure 6.6.

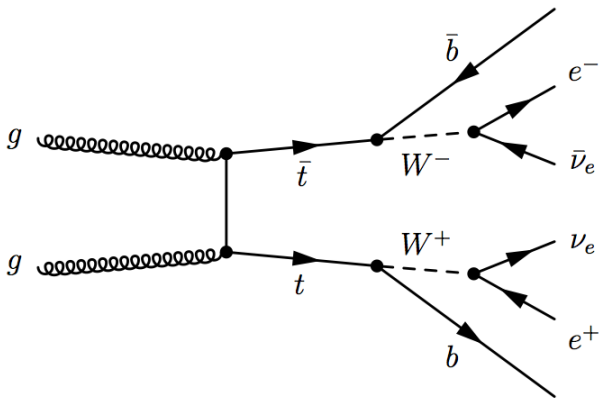

Figure 6.6: Feynman diagram for $gg \to t\bar{t}$ process, decay to Wb and the W's decaying leptonically into an electron-neutrino pair.

For the electroweak processes with W and Z production, both for signal and background events, samples produced with POWHEG interfaced with the PYTHIA parton-shower generator were used. For other backgrounds: $t\bar{t}$ events are studied with PYTHIA; EM-enriched QCD samples contain no $b/c \to e$ and decays, that are simulated in a separate sample.

[1] In CMS, files that describe which luminosity sections in which runs are considered good and should be processed are in the Java Script Object Notation (JSON) format.

Generator	Process	Kinematic cuts (in GeV)	σ (pb)	Events
POWHEG (+PYTHIA)	$\gamma^*/Z \to e^+e^-$	$m_{e^+e^-} > 20$	1631	> 1 M
POWHEG (+PYTHIA)	$W^+ \to e^+\nu$	no cuts	5825	$\sim$ 700k
POWHEG (+PYTHIA)	$W^- \to e^-\bar{\nu}$	no cuts	3954	$\sim$ 700k
PYTHIA	$t\bar{t}$	no cuts	94.3	500k
PYTHIA	EM-enriched QCD	$20 < p_T < 30$	1719150	30M
PYTHIA	EM-enriched QCD	$30 < p_T < 80$	3498700	40M
PYTHIA	EM-enriched QCD	$80 < p_T < 170$	134088	5M
PYTHIA	$b/c \to e$	$20 < p_T < 30$	108330	2M
PYTHIA	$b/c \to e$	$30 < p_T < 80$	138762	2M
PYTHIA	$b/c \to e$	$80 < p_T < 170$	9422	1M

Table 6.1: Summary of signal and background Monte Carlo samples as well as the generators used in the simulation.

An overview of all signal and background processes considered and of the generators used for the simulation is given in Table 6.1. All signal and background samples were processed through the full GEANT4 detector simulation, reconstructed and passed through the same analysis chain as the data.

Using the $\gamma^*/Z \to e^+e^-$ sample of Table 6.1, the kinematic distributions related to the Z production ($M_{e,e}$, p_T, rapidity and pseudorapidity distributions of Z) are shown in Figure 6.7 till 6.9. For all plots, the generator level information was used. For all plots $M_{e,e} > 40$ GeV was required.

6.3 Electron Reconstruction and Identification

The reconstruction of electrons in CMS is based on hits in the silicon tracker and the energy deposits in the crystal electromagnetic calorimeter. The superclusters in the crystal calorimeter and in the silicon tracker are first reconstructed separately followed by a combination of both objects to measure their energy and momentum.

6.3.1 Electron Reconstruction

The reconstruction of electrons in CMS uses information from the inner tracking system and the electromagnetic calorimeter (ECAL). The inner tracker measures trajectories and vertex positions of electrons in the magnetic field, which determine their charge and momenta. The

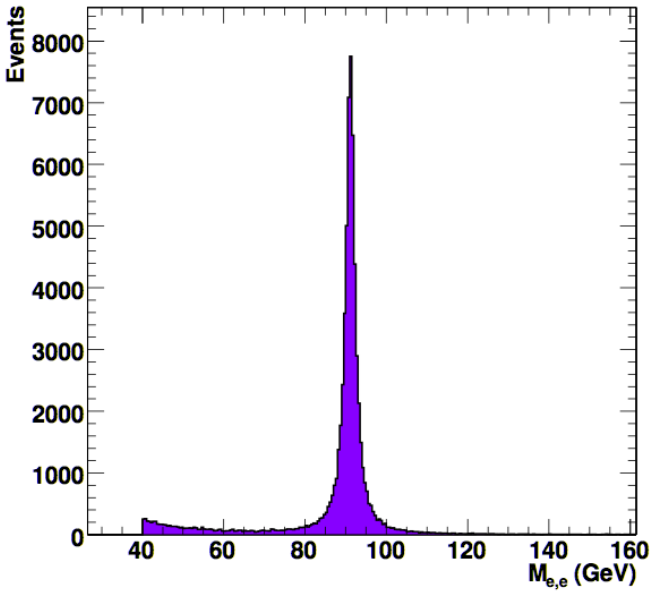

Figure 6.7: Distribution of $M_{e,e}$, at generator level, for events from the $\gamma^*/Z \to e^+e^-(M_{e,e} >$ 40GeV) sample.

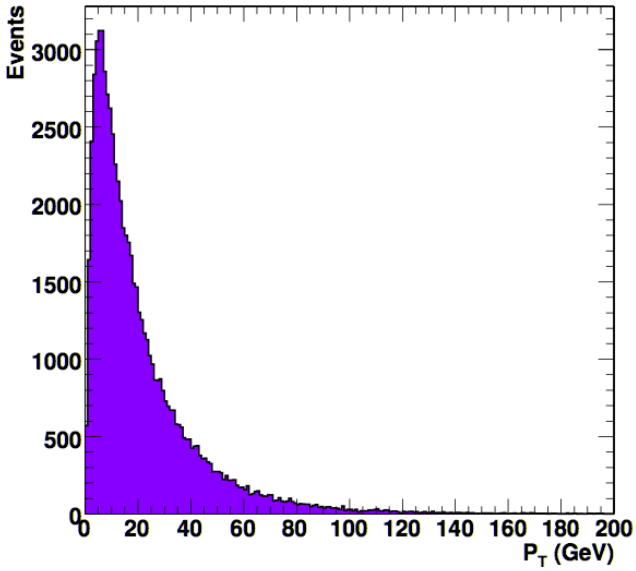

Figure 6.8: Distribution of Z p_T, for events from the $\gamma^*/Z \to e^+e^-(M_{e,e} > 40\text{GeV})$ sample.

electromagnetic calorimeter measures the position and the energy of electromagnetic showers deposited in ECAL. The combination of the tracking and calorimetric information also allows low p_T electrons to be measured and identified in the challenging kinematics and background conditions relevant for the Standard Model Higgs boson decays [40]. However, the measured

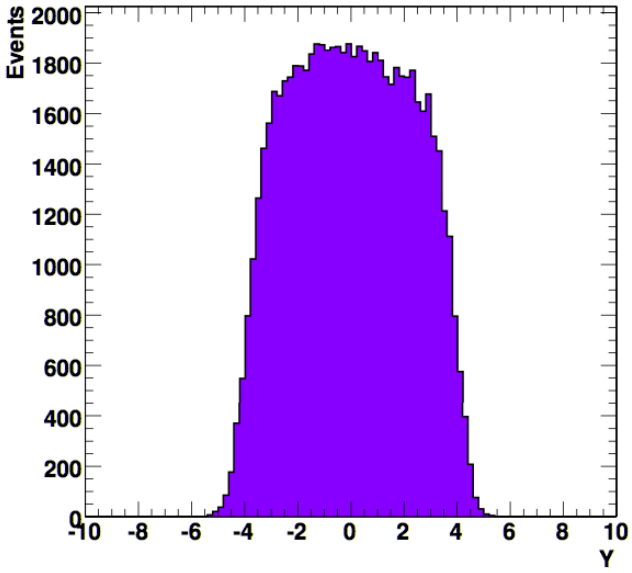

Figure 6.9: Distribution of Z rapidity, at generator level, for events from the $\gamma^*/Z \to e^+e^-(M_{e,e} > 40\text{GeV})$ sample.

resolution of the electron energy in ECAL is limited by the amount of tracker material that is distributed in front of ECAL, and by the presence of the strong magnetic field aligned with the collider beam axis.

The energy and momentum measurements are complementary: the uncertainty of the track-based momentum measurement is proportional to p, while the uncertainty of the calorimetric energy is proportional to $1/\sqrt{E}$. Thus the best estimation of the electron's energy is obtained from the combination of the tracker and calorimetric measurements (shown in Figure 6.10).

Although the inner tracking system is very useful for the electron reconstruction, the presence of the tracker material between the vertex and the ECAL poses a particular challenge for the energy measurements. As previously discussed (Chapter 3.3), high energy electrons predominantly lose energy through the bremsstrahlung emission.

The bremsstrahlung photons do not bend in the magnetic field while the electron does: resulting in the energy of the electron being spread in the azimuthal (ϕ) direction. To measure the initial energy of the electron by ECAL, a special 'clustering' reconstruction algorithm must be used to incorporate this energy spread.

Since an electron loses considerable energy in the tracker material, the parameters of the trajectory changes as the electron traverses the tracker. A special energy-loss modeling is therefore required in the reconstruction algorithms. Furthermore, the radiated photons have a significant

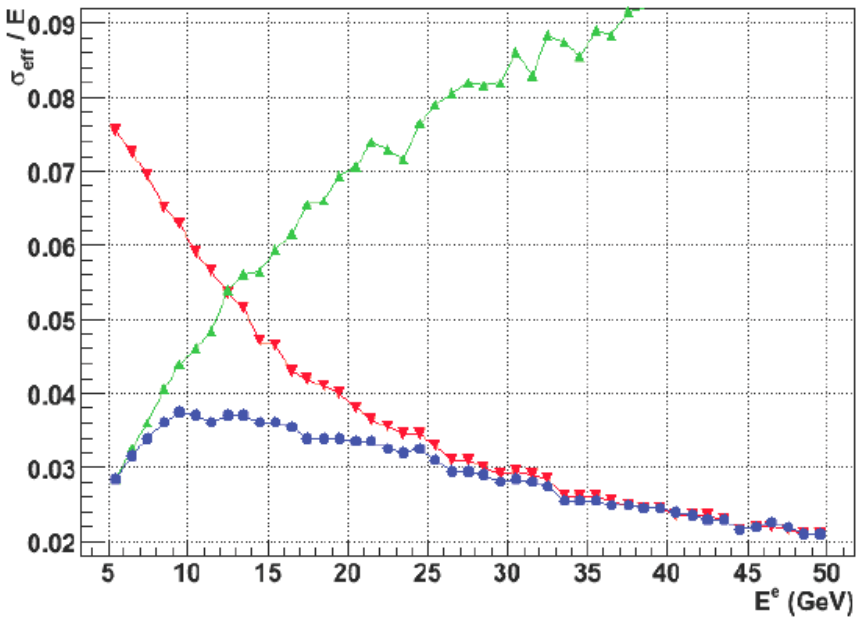

Figure 6.10: Measured electron energy resolution in the barrel as a function of the electron energy: tracker (green line), ECAL (red line) and combination of the two (blue line) [79].

probability of pair-converting into electrons and positions which will create their own 'hits' in the tracker, hindering track-finding and so impairing the momentum and charge measurement. The magnetic bending of the electrons and positrons makes the spread in the azimuthal angle. The optimization of the tracking algorithms minimizes this problem, and a dedicated method [80] for determining electron charge has been developed.

Energy Measurement in ECAL

The electromagnetic showers are narrow, e.g., from the measurements in the ECAL test beams, 25 crystals arranged in a 5×5 window contained 97% of the energy of electrons which struck the centre of the middle crystal. In the CMS experiment the energy of the incident electrons can in principle be reconstructed by summing the energies measured in these 25 crystals. However, such simple reconstruction can only be used for photons that are unconverted in the tracker material.

To reconstruct the energy of an energetic electron at the vertex, all radiated energies must be dynamically 'clustered': the crystals that have had energy deposited by an individual electromagnetic particle must be grouped. Two independent clustering algorithms are necessary due to the differing geometries of the ECAL barrel and endcaps of ECAL, though both define 'su-

perclusters' that reflect the narrow spread of energy in pseudorapidity and the wide spread in azimuthal angle due to the magnetic bending of electrons and positrons from the conversion of the bremsstrahlung photons. The extent of the spread in the η direction is essentially constant, while the ϕ extent varies. The algorithm is illustrated in Figure 6.11.

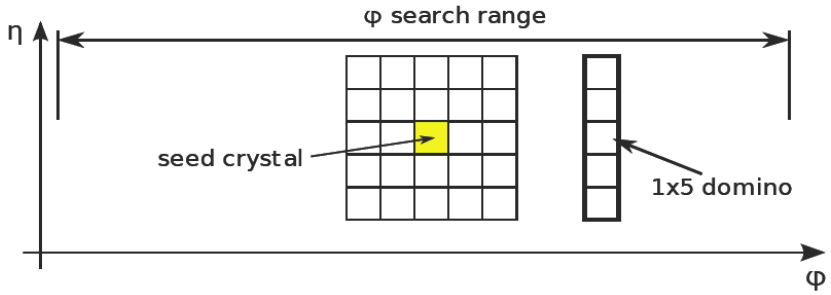

Figure 6.11: Illustration of the Hybrid algorithm, used in the barrel, which clusters the energy of the electrons that is well-contained in η, but spread in ϕ [80].

The Hybrid algorithm is used to measure the electron energy in the barrel. Dynamic clustering algorithms degraded energy resolution compared to fixed arrays such as 5×5 clusters. The Hybrid algorithm benefits from the $\eta - \phi$ geometry of the barrel by building clusters with fixed η width strips of 5 crystals, with only the ϕ extent of the cluster determined dynamically. Since in the endcaps the crystals are not arranged in an $\eta - \phi$ geometry, the hybrid algorithm can not be applied. Therefore, the multi 5×5 algorithm is used in the endcaps.

The ECAL endcaps are augmented by the preshower detector, which absorbs part of the energy of the incoming electrons before they interact with the crystals. To include this energy, an interpolation between the primary vertex and the ECAL superclusters is made. Any energy deposits found within a window around the intersection of these interpolations and the preshower are included in the corresponding supercluster energy.

The energy of the electron can be estimated by summing all energy deposits in the clustered crystals. However, this 'raw supercluster energy' does not agree with the truth energy of the electron. It must be corrected for a number of the effects in order to achieve an accurate measurement. These corrections factors (F) are applied as multiplicative factors:

$$E = F \sum_i G c_i A_i \tag{6.3.1}$$

where E is the corrected energy and $\sum_i G c_i A_i$ is the raw energy of the cluster.

Corrections are made for the following effects:

- The stepped front face of the ECAL barrel leads to lateral shower leakage. This is η dependent for the step depth increases with η: exposing more of the sides of the crystals and allowing more lateral leakage.
- Bremsstrahlung radiation leads to the energy of the electron being smeared and spread between several showers. The ECAL will have a different response to these showers, dependent on the fraction of the energy lost.

The distributions for uncorrected and corrected energies are plotted in Figure 6.12.

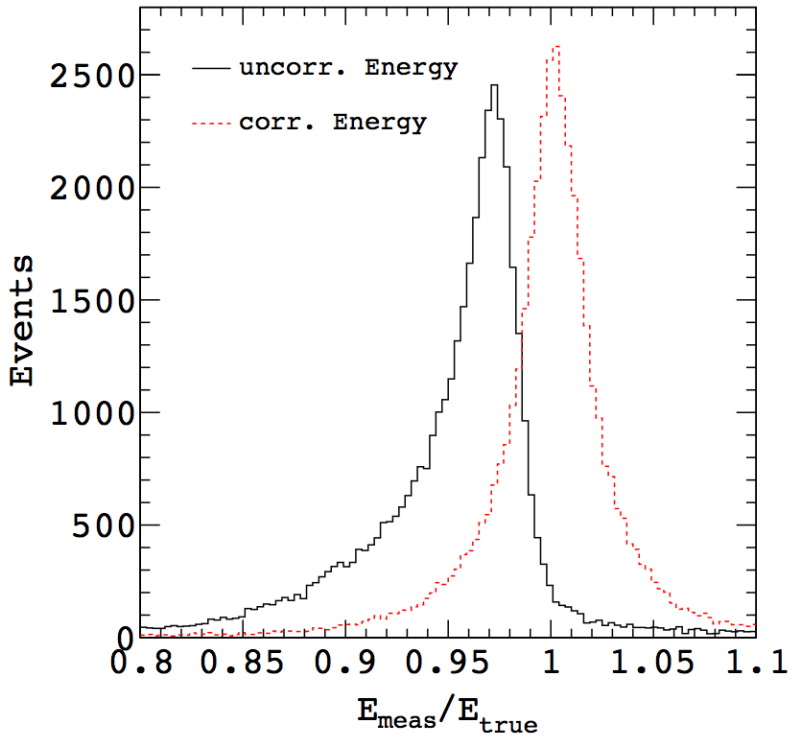

Figure 6.12: Distributions for uncorrected supercluster energy (black line) and corrected (red dashed line).

6.3.2 Identification and Isolation

Isolation variables for electrons are defined for the three sub-detectors as follows:

- $I_{\text{ECAL}} = \sum E_T(\text{ECAL})$;
- $I_{\text{HCAL}} = \sum E_T(\text{HCAL})$;
- $I_{\text{trk}} = \sum E_T(\text{tracks})$.

The sums are performed for objects falling within a cone $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} < 0.3$ around the electron candidate. The energy deposits and the track associated to the electron candidate

are excluded from the sums. The relative combined isolation variable is defined: $I^{rel}_{comb} = (I_{\rm ECAL} + I_{\rm HCAL} + I_{\rm trk})/p_T$.

In the analysis, only ECAL-driven electrons are considered. The ECAL fiducial region is defined by $1.566 < |\eta_{SC}| < 2.5$ and $|\eta_{SC}| < 1.4442$, where the η_{SC} is the pseudorapidity of the supercluster. These requirements exclude the barrel/endcap transition region and the first ring of endcap trigger towers. These regions are partially shadowed by cables between the barrel and endcap. A supercluster is considered to be within ECAL acceptance if it is reconstructed within the ECAL fiducial region and if it has $E_T > 20$ GeV. An electron is considered to be within ECAL acceptance if its associated supercluster is within ECAL acceptance. All the global electron efficiencies are normalized to superclusters within ECAL acceptance.

Electron identification is based on cuts on cluster shape covariance ($\sigma_{i\eta i\eta}$, the width of the EM shower normalized to units of crystals), on track-cluster matching variables ($\Delta\phi_{in}$ between supercluster position and track direction at vertex extrapolated to ECAL assuming no radiation; $\Delta\eta_{in}$ between supercluster position and track direction at vertex extrapolated to ECAL assuming no radiation), the ratio of energy in HCAL behind supercluster to supercluster energy (H/E). Photon converted electrons are rejected mostly by the requirement that the electron track must have no missing tracker hits before the first hit in the reconstructed track assigned to the electron. Electrons are rejected when a partner track is found which is consistent with a photon conversion, based on the opening angle and on the separation in the transverse plane and at the point at which the electron and partner tracks are parallel (Dcot and Dist). Electron isolation is based on cuts on the three isolation variables ($I_{\rm HCAL}/E_T$, $I_{\rm ECAL}/E_T$, $I_{\rm trks}/E_T$).

The electron selection working point WP80 are used in the analysis. WP80 is obtained by optimizing simultaneously identification and isolation criteria in the Monte Carlo simulation, and giving approximately 80% selection efficiency. The values of the cuts for WP80 are listed in Table 6.2.

6.4 $Z \to e^+e^-$ Signal Extraction

Selection Requirements

The selection of electrons is required to pass stringent electron identification criteria (WP80). The invariant mass of the electron pair is required to be within a window around the mass of the Z boson, ensuring a very high purity electron sample. Estimated backgrounds, mostly from QCD multi-jet processes, are less than 1%. Thus, the cut-based selections are used as follows:

- Two electrons satisfying the WP80 selection;

	Barrel	Endcap
I_{HCAL}/E_T	0.10	0.025
I_{trk}/E_T	0.09	0.04
I_{ECAL}/E_T	0.07	0.05
Missing hits $\leq$	0	0
Dcot	0.02	0.02
Dist	0.02	0.02
$\sigma_{i\eta i\eta}$	0.01	0.03
$\Delta\eta_{in}$	0.004	no cut
$\Delta\phi_{in}$	0.06	0.03
H/E	0.04	0.025

Table 6.2: The values of the cuts for WP80.

- the di-electron mass must satisfy $60 < M_{e,e} < 120$ GeV.

Using the data sample of 2.88 pb^{-1}, a total of 677 events were selected.

Ultimately the most performant selection should be obtained using multi-variant techniques, likelihood fits etc. However, cut-based selections can provide a useful tool to understand the data and make comparison with MC simulation. The advantages of 'Simple Cuts' are:

- Cut inversion (used in many data driven signal extraction and background subtraction methodologies) is simple;
- low statistics is sufficient for efficiency measurement;
- it is simple to cleanly separate the e-ID, isolation and conversion rejection requirements, to study their respective effect on the data selection and quality.

Electron Energy Scale

The di-electron invariant mass spectrum for the selected sample with the WP80 selection is shown in Figure 6.13 along with the predicted distribution. Because the actual energy response for each of the crystals is different, the data exhibit a mismatch of the mass scale of about 1.5 GeV relative to the simulation.

The energy scale and energy resolution correction factors are estimated as follows.

A 2D grid of scale factors are applied to EB and EE electrons in the $Z \to e^+e^-$ simulation. For each node of the grid the negative log likelihood (NLL) of the data was calculated for that MC distribution resulting in an estimation of the two correction factors, their errors and correlations,

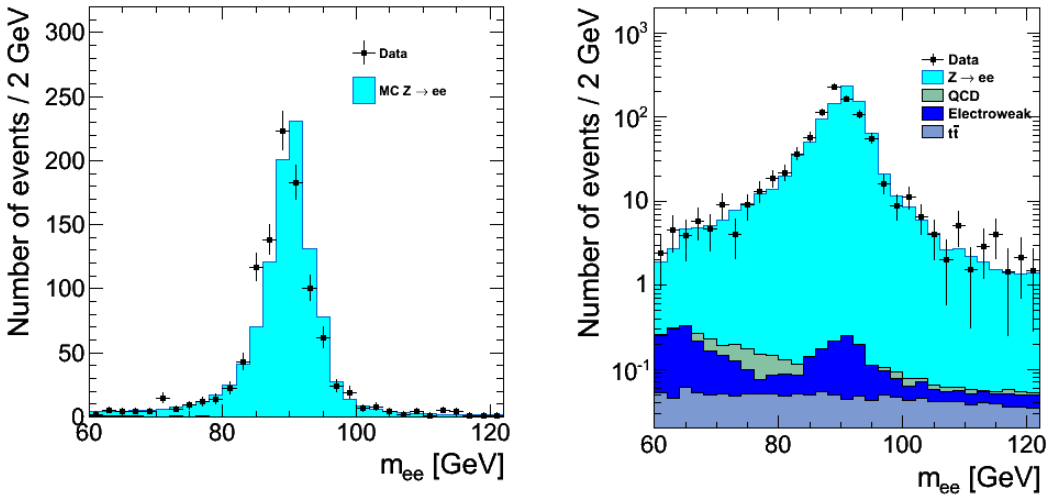

Figure 6.13: $Z \to e^+e^-$ signal in linear scale (left) and logarithmic scale (right). The points represent the data, and the histograms, the expected distribution from simulations normalized to 2.9 pb^{-1} and NNLO cross sections calculations. Backgrounds from Standard Model processes (QCD, Electroweak (EWK), $t\bar{t}$) are negligible and cannot be seen on the linear scale plot.

by fitting a 2D parabola in the vicinity of the node with minimum value of the NLL. The EB scale is derived from the sample of Barrel-Barrel events; the EE factors are derived from the sample of Barrel-Endcap events.

For the WP80 selection I obtained the following values:

- $\text{scale}_{\text{EB}-\text{WP80}} = 1.008 \pm 0.002$, energy scale correction in EB with WP80,
- $\text{scale}_{\text{EB}-\text{WP80}} = 1.024 \pm 0.003$, energy scale correction in EE with WP80,
- $\text{resol}_{\text{EB}-\text{WP80}} = 0.81 \pm 0.15$, additional smearing in EB with WP80,
- $\text{resol}_{\text{EB}-\text{WP80}} = 0.62 \pm 0.32$, additional smearing in in EB with WP80.

Applying these scale factors to electrons in the EB and EE result in the invariant mass spectrum shown in Figure 6.14. The agreement between data and MC has improved and is reasonably good.

6.5 Summary

The process of $Z \to e^+e^-$ can be used to study the performance of the CMS detector and to monitor the data quality. Based on the detail Monte Carlo simulation, electrons are selected using the information in the tracker and ECAL in CMS. The Hybrid algorithm is used to calculate the

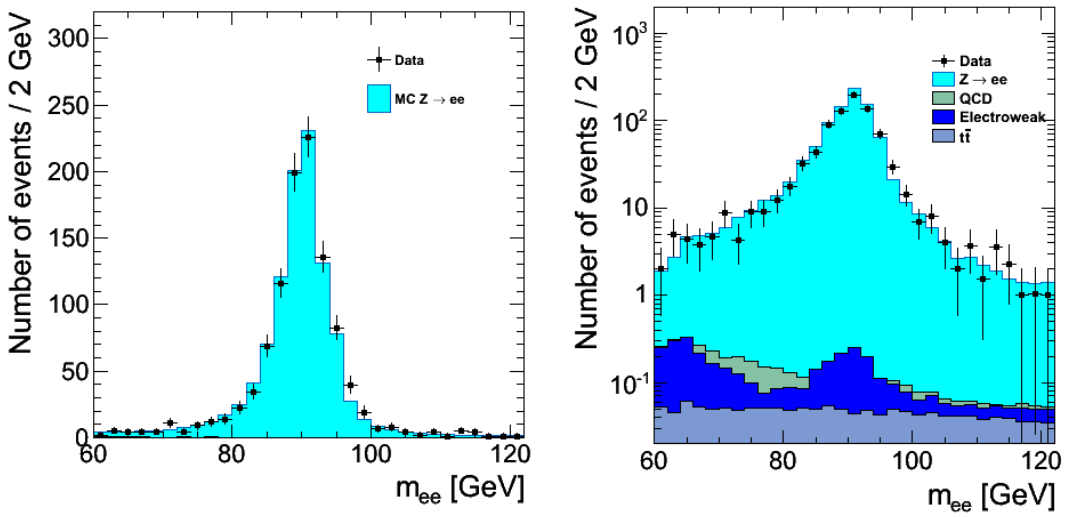

Figure 6.14: Same as Figure 6.13 after applying energy scale correction factors to electrons in the EB and EE in the data. Six additional events are selected in the signal mass window.

electron energy in the barrel. The multi 5×5 algorithm is used to calculate the electron energy in the endcaps. The electron selection working point WP80 is used to identify electrons. The events of $Z \to e^+e^-$ are selected by two charged isolated electrons with the invariant mass within 60 to 120 GeV. A 2D grid of energy correction scale factors are applied to EB and EE electrons. The invariant mass distribution of the $Z \to e^+e^-$ after correction is in good agreement with the Monte Carlo predictions and the backgrounds are very small. One can conclude that the quality of data collected with the CMS detector during the period is very good.

7 Summary

The Large Hadron Collider (LHC) started operation at the beam energy of 900 GeV in November 2009 and at the beam energy of 3.5 TeV in March 2010. The LHC commissioning went smoothly, and the luminosity was rapidly increasing. The detector performance was excellent and exceeded the expectations.

The Worldwide LHC Computing Grid (WLCG) is a global collaboration to provide a Grid infrastructure for the unprecedented distributed data storage and analysis demands from the LHC experiments, including CMS. The basic conceptions and structures of WLCG have been discussed in Chapter 4. The performance of WLCG fulfills the requirements from the LHC experiments during the period of the data taking covered in this thesis.

As a major part of the IT contribution to the CMS collaboration from Switzerland, a high-performance Tier-2 center for CMS was set up at CSCS and a CMS Tier-3 center was set up at PSI. I participated in the construction, software deployment and configuration, commissioning and operation of the Swiss Tier-2 and the Tier-3 respectively.

The computing environment deployment and configuration has been presented. In particular, during the commissioning and data challenges since 2008, many problems related to the evolutional LCG Middleware and the comprehensive dCache storage system were solved.

Valuable experience was gained for the preparation and data taking. The Swiss CMS Tier-2 showed high usability and good performance during the commissioning and operation phase, especially with big improvements of the gLite middleware and dCache. Another major work is related to the troubleshooting of the CMS data placement and the file transfer system PhEDEx at the Tier-2. The problematic links managed by PhEDEx between the Swiss Tier-2 and other Tier-centers were investigated and improved for an efficient data transfer for CMS.

With the experience gained from the Swiss Tier-2, the scheme of the Swiss CMS Tier-3 was optimized for the end-user analysis. The software deployment and configuration were carried out smoothly in September 2008. With the advanced local batch system SGE and NFS share-based filesystem, the program development, submitting and debug of the CMS analysis jobs on the User Interface of the Tier-3 are much easier than that over Grid. A large fraction of analysis jobs

from Swiss CMS groups was carried out on the Swiss Tier-3. The Tier-3 has shown excellent performance and usability since the production phase started in November 2008.

CRAB is the official utility to create and submit CMSSW jobs to WLCG. Since the CRAB did not support our SGE batch system when we were setting up the Swiss Tier-3, I programmed a new SGE scheduler module to adapt CRAB into the Tier-3 local analysis environment. The CRAB on the Tier-3 provides a friendly and uniform way for Tier-3 users to submit CMSSW jobs to the Tier-3 SGE batch system or WLCG.

The process of $Z \to e^+e^-$ has been used for the performance study of the CMS detector and to monitor the data quality. A large data sample of $Z \to e^+e^-$ is obtained for the data between May and September 2010. The $Z \to e^+e^-$ events are selected requiring two oppositely charged isolated electrons. The backgrounds are very small. The invariant mass distribution is in good agreement with the Monte Carlo prediction. One can conclude that the CMS data quality of the period is very good.

The integrated luminosity of 1 fb^{-1} of the data collected at the center mass energy of 7 TeV is expected by the end of 2011. With this large data sample, a very large sample of $Z \to e^+e^-$ events can be used to study the performance of the electromagnetic calorimeter in detail. Moreover, at the design luminosity of the LHC of 10^{34} $cm^{-2}s^{-1}$, one could monitor the data quality with the data sample of $Z \to e^+e^-$ 'online'.

Glossary

BOSSLite The BOSSLite is a python implementation of Batch Object Submission System for CMS. 115, 117

CE Computing Element (CE) is a batch queue to a centrally managed farm of computers (Worker Nodes) than can run GRID jobs. 63, 64, 67, 83, 98, 99, 102, 105, 108

CERN European Organization for Nuclear Research (CERN) is one of the world's larges centers for scientific research in fundamental physics. 5

CHIPP Swiss Institute of Particle Physics. 91

CMS VObox CMS VObox is a server provides a collection of services for CMS computing. 103–105

CMSSW CMSSW is the current official CMS simulation and reconstruction software framework. 93

CRAB The CMS Remote Analysis Builder (CRAB) is a utility to create and submit CMSSW jobs to CMS distributed computing resources. 115

Cron Cron is a time-based job scheduler in Unix-like computer operating systems. 103, 114

CSCS Swiss National Supercomputing Centre (CSCS) is an autonomous unit of the Swiss Federal Institute of Technology in Zurich (ETH Zurich). 91, 121

DBS The CMS Dataset Bookkeeping System (DBS) is a database and user API that indexes event-data data for the CMS Collaboration. The primary functionality is to provide cataloging by production and analysis operations and allow for data discovery by CMS physicists. 115

dCache dCache is a disk pool management system with a SRM interface, jointly developed by DESY and Fermilab. 93, 95, 99, 100, 113

DLS The Data Location Service (DLS) is part of the CMS Data Management system and provides a means to locate replicas of data in the distributed computing system. 115

FroNTier FroNTier is a simple web service approach providing client HTTP access to a central database service. 103–105

GridFTP The standard protocol for Grid based file transfers. 100

LEP The Large Electron-Positron Collider (LEP) was used from 1989 until 2000 at CERN. To date, LEP is the most powerful accelerator of leptons ever built. 6, 7, 12, 29

LHC The Large Hadron Collider (LHC) is a gigantic particle accelerator built at CERN, where it spans the border between Switzerland and France about 100 m underground.. 5–8, 12, 13, 16, 18–20, 24–26, 29

LRMS Local Resource Management System. 99

OGF OGF is an open community committed to driving the rapid evolution and adoption of applied distributed computing. 57

PBS PBS is a computer software that performs job scheduling originally developed by MRJ for NASA in the early to mid-1990s. 98

PhEDEx Physics Experiment Data Export (PhEDEx) provides the data placement and the file transfer system for the CMS experiment. 103, 105, 106, 108, 114

pNFS Parallel NFS (pNFS) is a part of the NFS v4.1 standard that allows clients to access storage devices directly and in parallel.. 97

PSI Paul Scherrer Institute is a multi-disciplinary research institute which belongs to the Swiss ETH-Komplex covering also the ETH Zurich and EPFL. 91, 121

RFIO Remote File I/O (RFIO) is one of the components that make up the CERN Advanced Storage Manager (CASTOR). RFIO implements a remote version of most standard POSIX calls like open, read, write, lseek and close using a very light weight protocol. 100

Runtime In computer science, the qualifier runtime or execution time refers to a single installation of a given software or computer program on a single computer. 116, 117, 119

SE A Storage Element provides uniform access to storage resources. It could be simply a disk servers, large disk arrays or Mass Storage System such as dCache or Castor. 62, 63, 105

SGE Sun Grid Engine (SGE) is an open source batch-queuing system, developed and supported by Sun Microsystems. 112, 115, 118, 122, 142

SM The Standard Model is currently accepted and experimentally well-tested theory of electromagnetic, weak and strong interactions. 5, 12, 13, 18, 19, 21

SRM Storage Resource Manager (SRM) is a middleware module provides management services for the storage resource and provides capabilities like transparent migrations from disk to tape, file pinnings, reservations, etc. 100

TORQUE Terascale Open-Source Resource and QUEue Manager (TORQUE) is an open-source distributed resource manager providing control over batch jobs and distributed worker nodes. 98

UI User Interface is the access point to the Grid. From a UI, a user can be authenticated and authorized to use the Grid resources. 121

UML Unified Modeling Language (UML) is a standardized general-purpose modeling language in the field of software engineering. 117, 119

VO Virtual Organization (VO) is an organization, typically an experiment, that collectively run jobs on the grid. It is managed using VOMS (Virtual Organization Membership Service). 91, 102

W3C The World Wide Web Consortium (W3C) is an international community. Its primary activity is to develop protocols and guidelines for the Web. 57

Web Service Web service is online service whose public interfaces and bindings are defined and described using XML. 57, 58

WLCG Worldwide LHC Computing Grid is a global collaboration to build and maintain a data storage and analysis infrastructure for the LHC at CERN. 7, 92

WMS The Workload Management System (WMS) comprises a set of Grid middleware components responsible for the distribution and management of tasks across Grid resources. 117

Xen Xen is a virtual-machine manager. It allows several guest operating systems to execute on the same computer hardware concurrently.. 97

XML eXtensible Markup Language (XML) is a set of rules for encoding documents in machine-readable form. 58

List of Tables

List of Figures

Bibliography

[1] G. Hinshaw et al., *Five-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Data Processing, Sky Maps, and Basic Results*, The Astrophysical Journal Letters, 180(2009):225–245.

[2] F. Cooper et al., *Supersymmetry and quantum mechanics*, Technical Report hep-th/9405029, LA-UR-94-569, Los Alamos Nat. Lab., Los Alamos, NM, 1994.

[3] O. S. Bruening et al., *LHC Design Report*, CERN-2004-003-V-1, CERN, 2004.

[4] *Proposal for building the LHC computing environment at CERN*, (2001)(CERN/2379/Rev.).

[5] CMS Collaboration, *WLCG Memorandum of Understanding*, Technical Report CERN-C-RRB-2005-01, 2006.

[6] F. J. Hasert et al., *Observation of neutrino-like interactions without muon or electron in the Gargamelle neutrino experiment*, Nucl. Phys. B, 73(1974)(1):1–22.

[7] UA1 Collaboration, *Experimental observation of isolated large transverse energy electrons with associated missing energy at* $\sqrt{s} = 540\ GeV$, Phys. Lett. B, 122(1983)(CERN-EP-83-13):103–116.

[8] UA1 Collaboration, *Experimental observation of lepton pairs of invariant mass around 95* GeV/c^2 *at the CERN* $p\overline{p}$ *collider*, (1983).

[9] UA2 Collaboration, *Evidence for* $Z^0 \to e^+e^-$ *at the CERN* $\overline{p}p$ *collider*, Phys. Lett. B, 129(1983)(CERN-EP-83-112):130–140. 21 p.

[10] UA2 Collaboration, *Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN* $\overline{p}p$ *collider*, Phys. Lett. B, 122(1983)(CERN-EP-83-25):476–485. 15 p.

[11] T. Berners-Lee and M. Fischetti, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*, Harper San Francisco, 1999.

[12] *Statement concerning CERN W3 software release into public domain*, http://tenyears-www.web.cern.ch/tenyears-www/Welcome.html, 1993.

[13] Maurice de Kunder, *The size of the World Wide Web*, http://www.worldwidewebsize.com/, 2010.

[14] M. L. Mangano and G. Altarelli, *CERN workshop on Standard Model physics (and more) at the LHC*, CERN-2000-004, CERN, Geneva, 2000.

[15] CMS Collaboration, *Detectors at LHC*, Phys. Rep., 403-404(2004):401–434.

[16] S. L. Glashow, *Partial-symmetries of weak interactions*, Nuclear Physics, 22(1961)(4):579 – 588.

[17] S. Weinberg, *A model of leptons*, Phys. Rev. Lett., 19(1967)(21):1264–1266.

[18] A. Salam, *Proc. of the 8th Nobel Symposium on 'Elementary Particle Theory, Relativistic Groups and Analyticity'*, (1969).

[19] R. Davis et al., *Search for neutrinos from the sun*, Phys. Rev. Lett., 20(1968):1205–1209.

[20] Super-Kamiokande Collaboration, *Evidence for oscillation of atmospheric neutrinos*, Phys. Rev. Lett., 81(1998)(8).

[21] K2K Collaboration, *Indications of neutrino oscillation in a 250 km long-baseline experiment*, Phys. Rev. Lett., 90(2003)(4):041801.

[22] KamLAND Collaboration, *First results from KamLAND: Evidence for reactor antineutrino disappearance*, Phys. Rev. Lett., 90(2003)(2):021802.

[23] *Standard Model of elementary particles*, http://en.wikipedia.org/, 2006.

[24] D. H. Perkins, *Introduction to High Energy Physics*, Cambridge University Press, 2000.

[25] B. Kilminster, *Higgs searches at the Tevatron*, ICHEP, 2010.

[26] N. Arkani-Hamed et al., *The hierarchy problem and new dimensions at a millimeter*, Physics Letters B, 429(1998)(3-4):263 – 272.

[27] CMS Collaboration, *CMS physics: Technical Design Report, v.2: Physics performance*, CERN, Geneva, 2006.

[28] T. C. Petersen, *Precision measurement of W mass at LHC*, Nucl. Phys. B, Proc. Suppl., 177-178(2008):32–35.

[29] F. Abe et al., *Observation of Top Quark Production in $p\bar{p}$ Collisions with the CDF Detector at Fermilab*, Phys. Rev. D50, 73(1995).

[30] A. Onofre, *Top quark physics with ATLAS and CMS*, Technical Report ATL-PHYS-PROC-2009-068. ATL-COM-PHYS-2009-344, 2009.

[31] M. Spira and P. M. Zerwas, *Electroweak symmetry breaking and Higgs physics*, (1997)(hep-ph/9803257. CERN-TH-97-379. DESY-97-261):70.

[32] Particle Data Group, *Review of particle physics*, J. Phys. G, 075021(2010).

[33] L. J. Hall and M. Suzuki, *Explicit R-parity breaking in supersymmetric models*, Nuclear Physics B, 231(1984)(3):419 – 444.

[34] D. Feldman et al., *The stueckelberg Z prime at the LHC: discovery potential, signature spaces and model discrimination*, Journal of High Energy Physics, 2006(2006)(11):007.

[35] M. Bajko et al., *Report of the Task Force on the Incident of 19th September 2008 at the LHC*, Technical Report LHC-PROJECT-Report-1168, 2009.

[36] C. Carli, *Chamonix 2009 workshop on LHC performance*, CERN-ATS-2009-001, 2009.

[37] M. Lamont, *LHC: status and commissioning plans*, (2009)(arXiv:0906.0347).

[38] L. Rossi et al., *Magnets repair for 3-4 sectori*, in *Proceeding of Chamonix 2009 Workshop on LHC Performance*, 2009.

[39] S. Myers, *Report on the LHC*, ICHEP, 2010.

[40] CMS Collaboration, *CMS physics: Technical Design Report, V.1: Detector performance and software*, CERN-LHCC-2006-001, 2006.

[41] The TOTEM Collaboration, *The TOTEM experiment at the CERN large hadron collider*, Journal of Instrumentation, 3(2008)(08):S08007.

[42] CMS Collaboration, *The CMS magnet project: Technical Design Report*, CERN-LHCC-97-010, CMS-TDR-001, CERN, 1997.

[43] CMS Collaboration, *The CMS experiment at the CERN LHC*, J. Instrum., 3(2008):S08004.

[44] V. Karimaki, *The CMS tracker system project: Technical Design Report*, Geneva, 1997.

[45] CMS Collaboration, *The CMS tracker: addendum to the Technical Design Report*, CERN-LHCC-2000-016, CMS-TDR-005-add-1, CERN, 2000.

[46] T. S. Virdee and R. Cousins, *The status of the CMS experiment at the LHC*, PoS, EPS-HEP2009(2009):6.

[47] CMS Collaboration, *Commissioning of the particle-flow event reconstruction with the first LHC collisions recorded in the CMS detector*, (CMS PAS PFT-10-001).

[48] *CMS e-commentary for 2010 LHC beams*, http://cms.web.cern.ch/cms/News/e-commentary/cms-e-commentary10.htm, 2010.

[49] G. Cerminara, *Operation of the CMS detector with first collisions at 7 TeV at the LHC*, ICHEP, 2010.

[50] I. Bird et al., *LHC computing Grid: Technical Design Report*, Technical Report CERN-LHCC-2005-024, 2005.

[51] I. Foster, *What is the grid? a three point checklist*, GRID today, 1(2002)(6):32–36.

[52] I. Foster et al., *The physiology of the Grid: An open grid services architecture for distributed systems integration*, 2002.

[53] S. Tuecke et al., *Open Grid Services Infrastructure (OGSI)*, (2003).

[54] *The Globus Alliance*, http://www.globus.org, 2010.

[55] C. Grandi et al., *CMS Computing Model: The CMS Computing Model RTAG*, Technical Report CMS-NOTE-2004-031. CERN-CMS-NOTE-2004-031. CERN-LHCC-2004-035. LHCC-G-083, CERN, Geneva, 2004.

[56] V. Innocente et al., *CMS software architecture. software framework, services and persistency in high level trigger, reconstruction and analysis*, Comput. Phys. Commun., 140(2001)(1-2):31–44.

[57] R. Brun et al., *ROOT: An object oriented data analysis framework*, NUCL INSTRUM METH A, 389(1997)(1-2):81 – 86.

[58] W. Tanenbaum, *A ROOT/IO Based Software Framework for CMS*, Technical Report cs.DB/0306034, 2003.

[59] G. L. Bayatyan and D. Negra, *CMS computing: Technical Design Report*, CERN, 2005.

[60] R. Egeland et al., *PhEDEx Data Service*, Technical Report CMS-CR-2009-071. CERN-CMS-CR-2009-071, CERN, Geneva, 2009.

[61] B. Blumenfeld et al., *CMS conditions data access using FroNTier*, Journal of Physics: Conference Series, 119(2008)(7):072007.

[62] D. Spiga et al., *Crab: the cms distributed analysis tool development and design*, Nucl. Phys. B, Proc. Suppl., 177-178(2008):267–268.

[63] H. B. Newman et al., *Monalisa : A distributed monitoring service architecture*, in *Proceedings of CHEP03*, 2003.

[64] A. Sciaba et al., *The usage of the glite Workload Management System by the LHC experiments*, (2007).

[65] *Site status for the CMS sites*, http://dashb-ssb.cern.ch/dashboard/request.py/siteviewhome, 2010.

[66] P. Mendez Lorenzo et al., *The WLCG common computing readiness challenge: CCRC'08*, 3rd EGEE User Forum, 2008.

[67] C. Collaboration, *Cms data processing workflows during an extended cosmic ray run*, Journal of Instrumentation, 5(2010)(03):T03006.

[68] G. TONELLI, *CMS status and highlights*, ICHEP, 2010.

[69] C. Grab, *Chipp computing board report*, http://www.chipp.ch/chipp-meet-computing-board.html, 2008.

[70] J. Henning, *SPEC cpu2000: measuring cpu performance in the new millennium*, Computer, 33(2000)(7):28 –35.

[71] G. Staples, *TORQUE resource manager*, in *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, SC '06, ACM, 2006.

[72] B. Bode et al., *The portable batch scheduler and the maui scheduler on linux clusters*, in *Proceedings of the 4th annual Linux Showcase & Conference - Volume 4*, 27–27, USENIX Association, 2000.

[73] *SWITCH*, http://switch.ch/, 2010.

[74] Python Software Foundation, *Python programming language*, http://www.python.org/, 2010.

[75] M. Dittmar et al., *Towards a precise parton luminosity determination at the cern lhc*, Phys. Rev. D, 56(1997)(hep-ex/9705004. ETHZ-IPP-P-97-01. 11):7284–7290.

[76] M. Malberti, *W and Z at the LHC*, Technical Report CMS-CO-2009-164. CERN-CMS-CO-2009-164, CERN, Geneva, 2009.

[77] A. Martin et al., *Parton distributions and the LHC:W and Z Production*, The European Physical Journal C, 14(2000)(1):133.

[78] C. Anastasiou et al., *High-precision qcd at hadron colliders: Electroweak gauge boson rapidity distributions at next-to-next-to leading order*, Phys. Rev. D, 69(2004)(9):094008.

[79] D. J. A. Cockerill, *The CMS Electromagnetic Calorimeter at the LHC*, Technical Report CMS-CR-2008-082. CERN-CMS-CR-2008-082, CERN, Geneva, 2008.

[80] S. Baffioni et al., *Electron reconstruction in CMS*, Technical Report CMS-NOTE-2006-040. CERN-CMS-NOTE-2006-040, CERN, Geneva, 2006.

Acknowledgements

Working in the field of High Energy Physics involves very large international collaborations, adding an important social and cultural dimension to the scientific research. Therefore, I am grateful to many people for their help and support during my PhD study.

First and foremost I wish to express my indebtedness to Prof. Felicitas Pauss, who gave me the opportunity to write this thesis under her supervision. It would have been impossible to finish my thesis if it was not for her patient guidance, encouragement and advice. I have been extremely lucky to have a supervisor who cared so much about my thesis, and who spent a lot of time to help make this thesis better.

My special thanks go to Prof. Günther Dissertori who accepted to act as co-examiner for this thesis.

Next I would like to thank Dr. Derek Feichtinger. He gave me lots of practical advice concerning Grid Computing. I also would like to thank Dr. Michael Dittmar, for the guidance and advice on the CMS data analysis and physics studies.

I would like to thank my colleagues of the Institute of Particle Physics at ETH Zurich and the CMS Collaboration. They have been very kind and supportive.

Last, but certainly not least, I would like to sincerely thank my family, who have very much supported me all these years.

Curriculum Vitae

Personal Data

Name	Zhiling CHEN
Date of Birth	31 July 1978
Place of Birth	Wuhan, Hubei Province, China
Nationality	Chinese

Education

2006 – 2010	Doctoral studies at ETH Zürich Research performed with the CMS experiment at CERN's LHC
2003 – 2006	Master degree in Computer Applied Technology Master thesis: *Deployment of LCG sites and Integration of Portlet–based Grid Portal* at the Graduate School of the Chinese Academy of Sciences in Beijing, China Design and implementation of a Portlet–based Grid Portal
1996 – 2000	Bachelor of Science in Atmospheric Physics Bachelor thesis: *Mathematical Evaluation System of Atmosphere Environment* at Nanjing University in Nanjing, China
1990 – 1996	October First School (Middle and High School) in Beijing, China
1985 – 1990	YuQuan Primary School in Beijing, China

Professional Experience

2003 – 2006	Chief technical manager of the Distance Education System of the Chinese Academy of Sciences in Beijing, China
2001 – 2003	Engineer of the Distance Education System of the Chinese Academy of Sciences in Beijing, China