Top quark physics at the LHC with the CMS detector

Sergey Senkin

School of Physics

University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Science

Autumn 2013

Contents

1. The LHC and the CMS detector

1.1 The Large Hadron Collider

The LHC [1] is currently the largest and the most powerful particle accelerator ever built. It is installed in the 26.7 km tunnel that was originally constructed for the LEP accelerator in the 1980s. The tunnel lies at a depth of 45 m to 170 m underground between the Jura mountain and Lake Geneva, being the main part of the CERN accelerator complex.

The machine is designed to accelerate proton beams and provide collisions at a centre of mass energy of $\sqrt{s} = 14$ TeV. Unlike particle-antiparticle colliders, the LHC requires two rings with opposite magnetic dipole fields in order to maintain and collide two counter-rotating proton beams. Since the tunnel was originally designed for the electron-positron LEP, it has an internal diameter of 3.7 m which is not enough to install two separate independent rings. Therefore, a twin-bore magnet design was adopted [2], which resulted in substantial cost savings.

A schematic view of the LHC accelerator chain is shown in Figure 1.1. Initially, the protons are obtained by stripping orbiting electrons from hydrogen atoms. Then they are injected into the linear accelerator LINAC2 to reach the energy of 50 MeV and enter the Proton Synchrotron Booster (PSB). The booster accelerates them to 1.4 GeV and passes the beam to the Proton Synchrotron (PS) where the energy rises to 25 GeV. In the next step, protons enter the Super Proton Synchrotron (SPS) where they are accelerated to 450 GeV. Finally, the beam is transferred to the LHC in both clockwise and anti-clockwise directions where it takes about 20 minutes to reach the design 7 TeV energy (per beam).

[add the number of magnets, total energy stored, number of bunches, bunch spacing etc.?]

The LHC has four interaction points, providing collisions to four major experiments. Two of them, CMS and ATLAS, are multi-purpose high-luminosity experiments with a peak luminosity of $L = 10^{34}$ cm^{-2} s^{-1}. The other two experiments operate at low luminosities and have more specific physics goals: LHCb studies b-meson decays, and Alice is a dedicated heavy ion experiment.

The instantaneous luminosity of a collider can be calculated as

$$L = \frac{n_1 n_2 f}{4\pi\sigma_x\sigma_y}, \tag{1.1}$$

where n_1 and n_2 are the numbers of particles in each of the colliding bunches, f is the

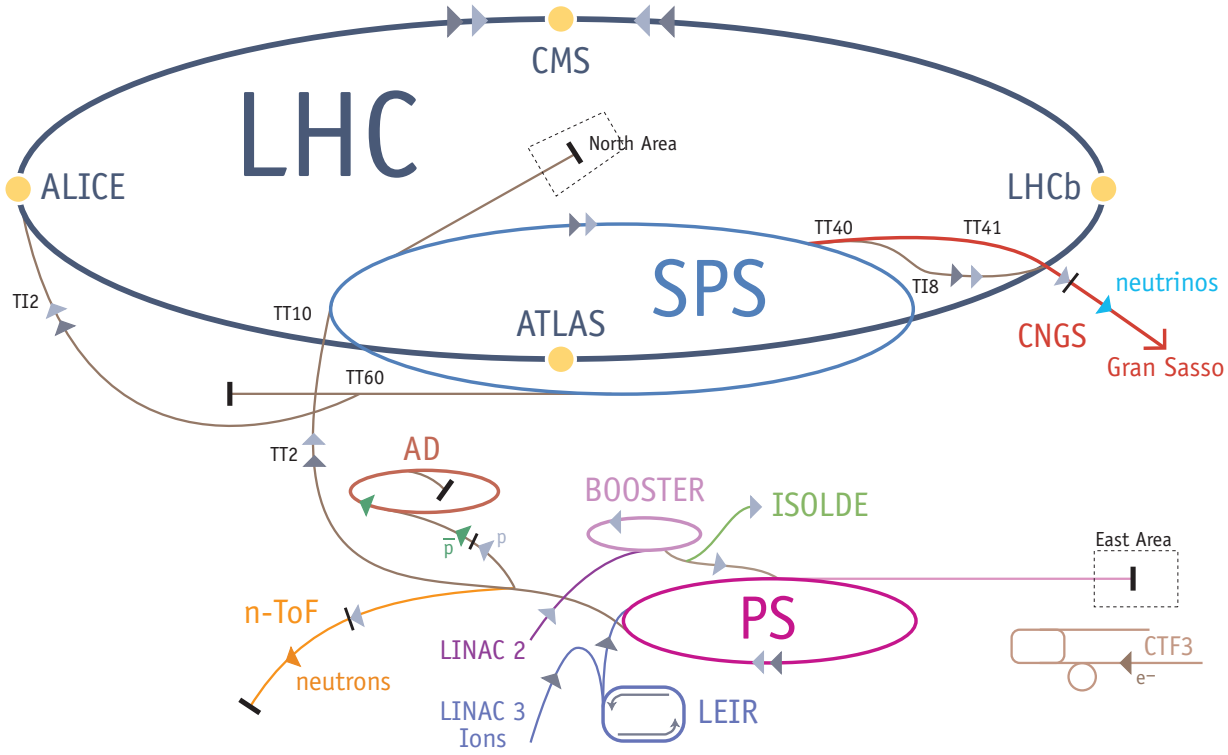

Figure 1.1: CERN accelerator complex.

revolution frequency, σ_x and σ_y are the horizontal and vertical beam sizes, assuming the two beams have the same size.

The number of events generated in the collisions per second is given by

$$N_{events} = L \times \sigma, \tag{1.2}$$

where σ is the cross section of the process under study.

[add the plots with cross sections and production rates?]

The LHC started operating on the 10th of September 2008, with the first beams fully circulating in both rings. However, only 9 days later a magnet quench occurred in two sectors of the tunnel, which was caused by an electrical fault due to a bad connection between two magnets. A consequent liquid helium explosion damaged a total of 53 superconducting magnets. Over a year was spent on repairs and tests, and the first collisions were recorded on the 23rd of November 2009 at a centre of mass energy of 0.9 TeV. The following few months showed the continuous ramp up of the beam energies up to 3.5 TeV per beam which was achieved on the 30rd of March 2010 when the LHC physics programme started.

Throughout the rest of 2010, the two general-purpose LHC experiments (CMS and ATLAS) recorded approximately 40 pb^{-1} of data, which resulted in the first measurements of various physics processes at the LHC. The following year became the main 7 TeV data-taking period, with about 5 fb^{-1} of data recorded by ATLAS and CMS. On the 5th of April 2012 the centre of mass energy was increased to 8 TeV, and July of 2012 marked the first major discovery of a new boson which was later shown to be consistent with the Standard Model Higgs boson, according to approximately 21.8 fb^{-1} of data recorded until early 2013. A long shut-down is planned for the following two years with various upgrades scheduled. The next physics run is expected in 2015 with the beam energy increased up to 6 or 7 TeV.

[add any upgrade details and distant future plans, like SLHC?]

1.2 The CMS Detector

The Compact Muon Solenoid [3] is a general-purpose detector designed to carry out precise measurements of the Standard Model and searches for physics beyond it. The primary design requirement was the ability to discover the nature of electroweak symmetry breaking, and the first observation of a Higgs boson was obtained in the Summer of 2012 [4].

The detector is installed at one of the LHC interaction points (Point 5) at about 100 m underground near the French village of Cessy, between the Jura mountains and Lake Geneva. The overall dimensions of the CMS detector are a length of 21.6 m, a diameter of 14.6 m and a total weight of 12 500 t.

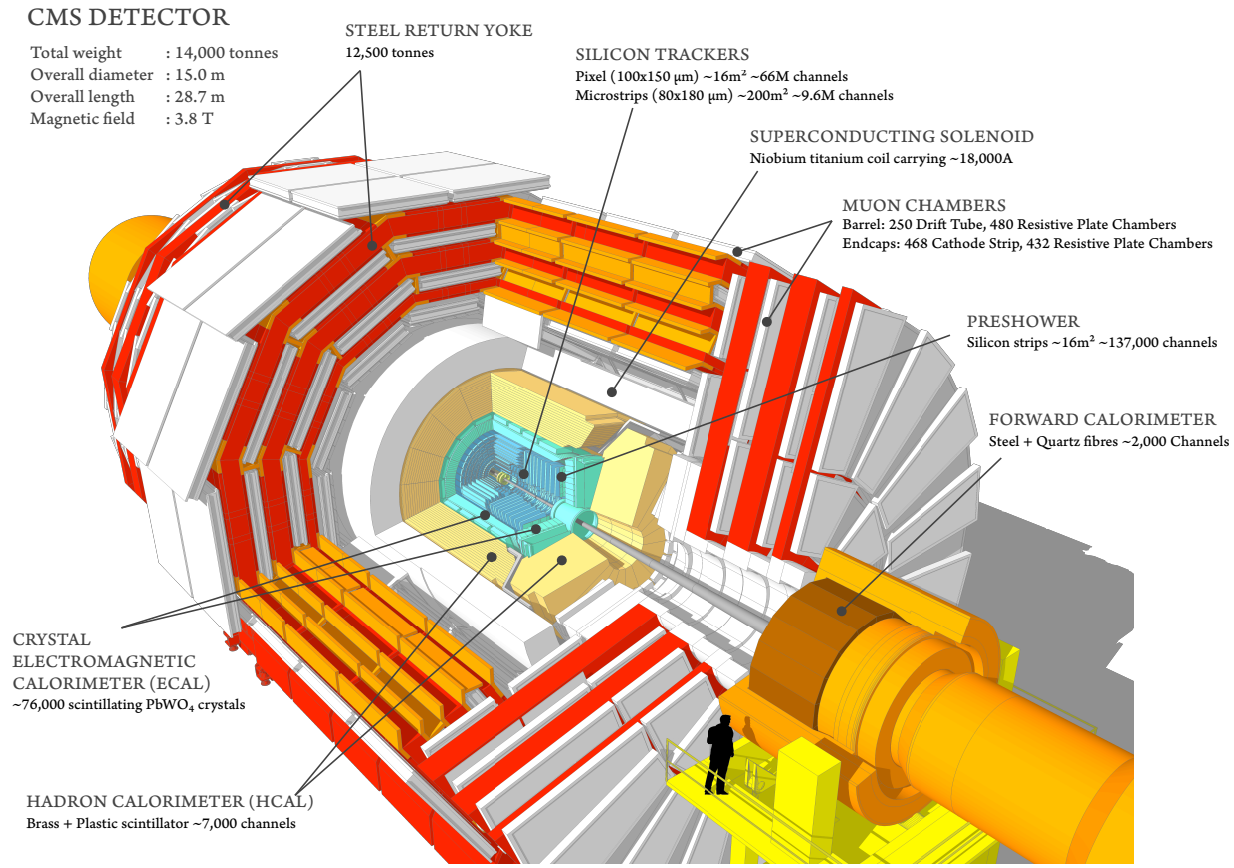

Figure 1.2: Sectional view of the CMS detector.

The sectional view of CMS is shown in Figure 1.2. In the centre of the detector, tracking and calorimetry systems are surrounded by the superconducting solenoid. On the outermost part of it the magnetic flux is returned through the iron yoke in which the muon system is also integrated. All the sub-systems are discussed in the following sections in more detail.

The cylindrical shape of the CMS detector dictates using a cylindrical coordinate system, with the origin centred at the interaction point, the x-axis pointing towards the centre of the LHC ring, the y-axis pointing upwards and the z-axis pointing along the beamline in the anti-clockwise direction. The azimuthal angle ϕ is measured from the x-axis in the transverse $(x-y)$ plane and the polar angle θ is measured from the z-axis. The radial distance to the beamline is denoted by r. Pseudorapidity is defined as:

$$\eta = -\ln \tan \frac{\theta}{2}. \tag{1.3}$$

This implies that the particles moving in the transverse plane (perpendicular to the beamline) have a pseudorapidity of 0, whereas the beam direction has an infinite pseudorapidity. Considering the cylindrical shape of the detector, it has barrel and endcap regions, with the transition occurring at $\eta \sim 1.4$. The momentum and energy transverse to the beamline are denoted by p_T and E_T respectively; the imbalance of the energy measured in the transverse plane, called missing transverse energy, is denoted by $E_{\mathrm{T}}^{\mathrm{miss}}$.

1.2.1 Inner Tracking System

The tracking system lies in the heart of the CMS detector and is the closest to the interaction point where the particle flux has the highest value. This imposes demanding requirements on the configuration of the system. At design luminosity of $L = 10^{34}$ cm^{-2} s^{-1} with the bunch spacing of 25 ns, an average of 1000 particles from about 25 proton-proton interactions (pile-up vertices) is expected to traverse the tracker for each bunch crossing. However, up until the long shutdown a bunch spacing of 50 ns was used, which meant a higher number of protons in each bunch leading to approximately twice the number of pile-up vertices. Therefore, in order for the particle tracks to be identified reliably and separately for each bunch crossing, the tracker requires very fine granularity and fast response parameters. Another complication caused by the intense particle flux is the severe radiation damage, so the tracker has to be highly resilient in operating in the harsh environment for a reasonable lifetime.

To meet these requirements on granularity, response time and radiation resilience, the tracker design was chosen to be based on silicon detector technology. Although capable of meeting such conditions, this technology has a disadvantage of a high power density of

on-detector electronics. This implies the necessity of an efficient cooling system. Moreover, a large amount of dense material interacting with the particles leads to higher multiple scattering, bremsstrahlung, photon conversions and nuclear interactions. Therefore, there are complications in the reconstruction of the tracks, meaning some loss of efficiency and precision. This will be discussed in detail later on in the object reconstruction section.

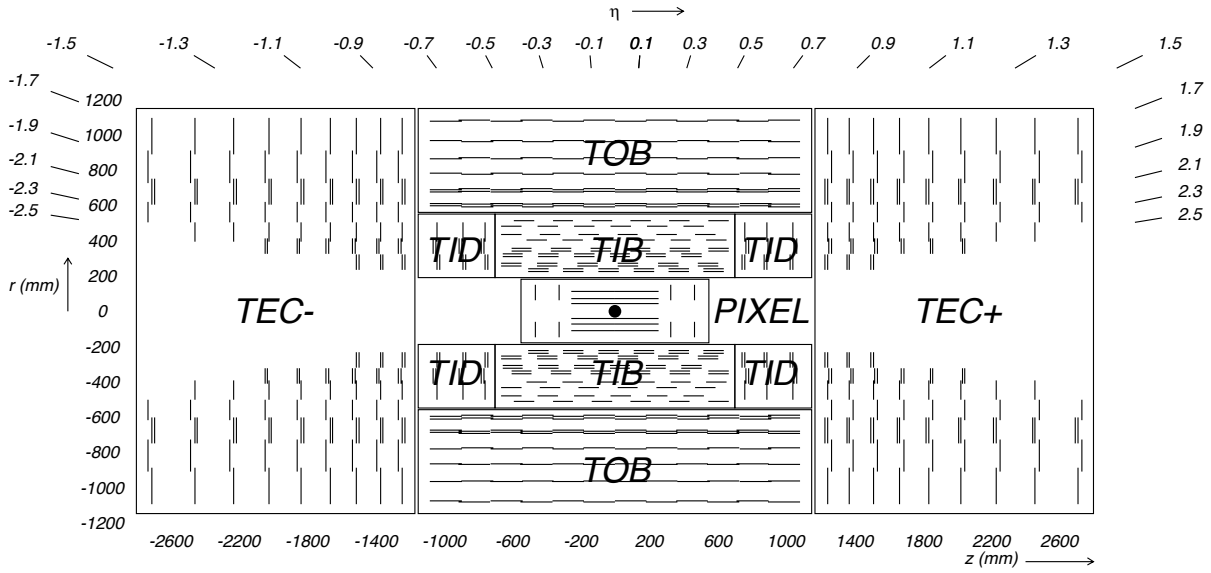

Figure 1.3: Cross-section of the CMS tracker system [3].

Figure 1.3 shows the overall layout of the tracking system. It consists of the inner pixel detector, located in the vicinity of the interaction point, and silicon strip tracker detectors: inner barrel and disks (TIB and TID), outer barrel (TOB) and endcaps (TEC). The geometrical acceptance of the tracker system goes up to $|\eta|< 2.5$. The outer radius of the CMS tracker reaches approximately 110 cm, and its total length is about 540 cm.

The pixel detector consists of three layers of pixel sensors at radii of 4.4 cm, 7.3 cm and 10.2 cm from the beamline in the barrel region. In addition there are two endcap disks on each side at $|z| = 34.5$ cm and 46.5 cm. The pixel size equals 100×150 µm^2 in $r\phi \times z$ coordinates. The pixel detector has 66 million pixels and the total area of about 1 m^2.

The silicon strip tracker consists of several layers of silicon microstrip detectors. It covers the region between 20 cm to 110 cm in radius and extends up to ±280 cm in the z direction. The Tracker Inner Barrel (TIB) is made out of 4 layers and the Tracker Outer Barrel (TOB) has 6 layers in it. The tracker endcaps (TEC) comprise 9 disks, and there are also the tracker inner disks (TID) that consist of 3 disks filling the gap between TIB and TEC as shown in

Figure 1.3. There are 9.3 million silicon strips covering the area of about 200 m^2. The silicon sensors' thickness varies between 320 and 500 µm and the strip pitch varies from 80 µm in the TIB to 180 µm in TOB and TEC.

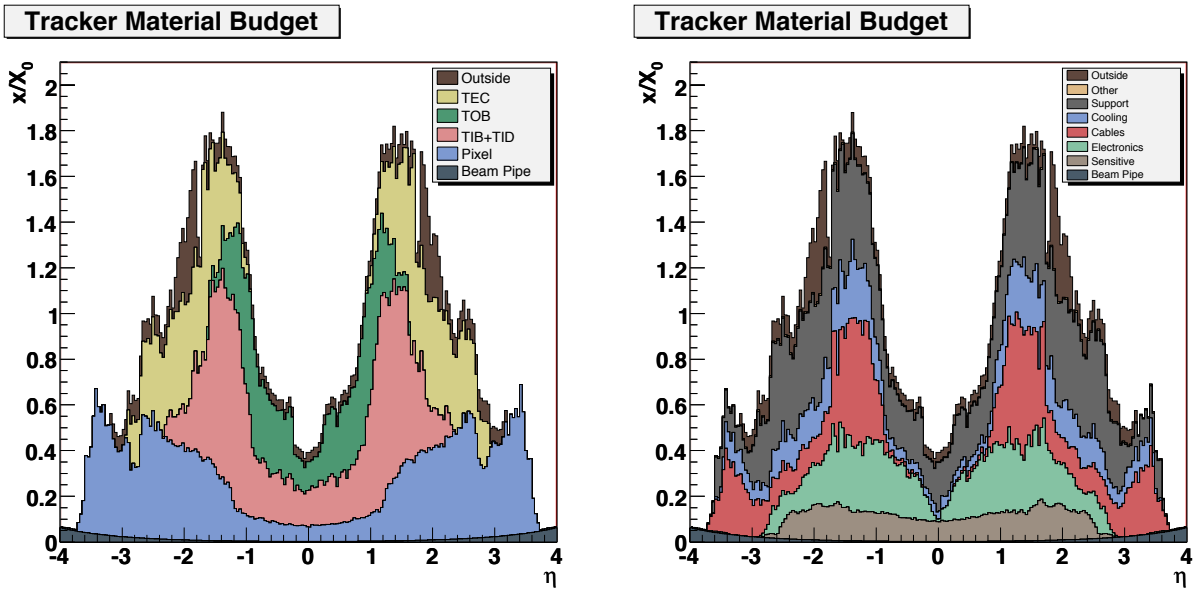

Figure 1.4: Material budget as a function of pseudorapidity η for the different sub-detectors of the tracker (left) and broken down into the functional contributions (right), in units of radiation length [3].

The silicon detectors of the tracker, the readout electronics and support structure form a considerable amount of material for the particles traversing from the interaction point. Figure 1.4 [3] shows the material budget of the CMS tracker in units of radiation lengths[1] (X_0). It grows from about 0.4 X_0 to 1.8 X_0 in the barrel region, and then decreases to about 1 X_0 in the endcaps. This causes a substantial conversion rate for photons and electrons in the tracker material; it also will be discussed in more detail in the electron reconstruction section.

[perhaps need to add the p_T resolution plots]

1.2.2 Electromagnetic Calorimeter

The next detector subsystem which is surrounding the tracker is the electromagnetic calorimeter, or ECAL. It is of a primary importance for the analyses described in this thesis,

[1] A material's radiation length is the mean distance over which a high-energy electron loses all but $1/e$ of its energy by bremsstrahlung; this is equal to 7/9 of the mean free path for pair production by a high-energy photon.

as it provides information for the electron and positron reconstruction. Combination of this information with that from the tracking system must ensure a precise measurement of electron position and momentum, and also sufficient background removal. It has to effectively distinguish the energy deposit shape of an electromagnetic particle from the one of a hadronic particle, which requires good segmentation and high resolution.

ECAL is a hermetic, high-granularity, high-resolution scintillating crystal calorimeter consisting of 61 200 lead tungstate ($PbWO_4$) crystals located in the central barrel region ($|\eta|< 1.479$), and 7324 crystals in each of the two endcaps ($1.479 <|\eta|< 3.0$). All crystals are followed by photodetectors reading and amplifying their scintillation: avalanche photodiodes (APD) are used in the barrel, and vacuum phototriodes (VPTs) are used in the endcaps. These different choices were caused by the configuration of the magnetic field and the expected level of radiation.

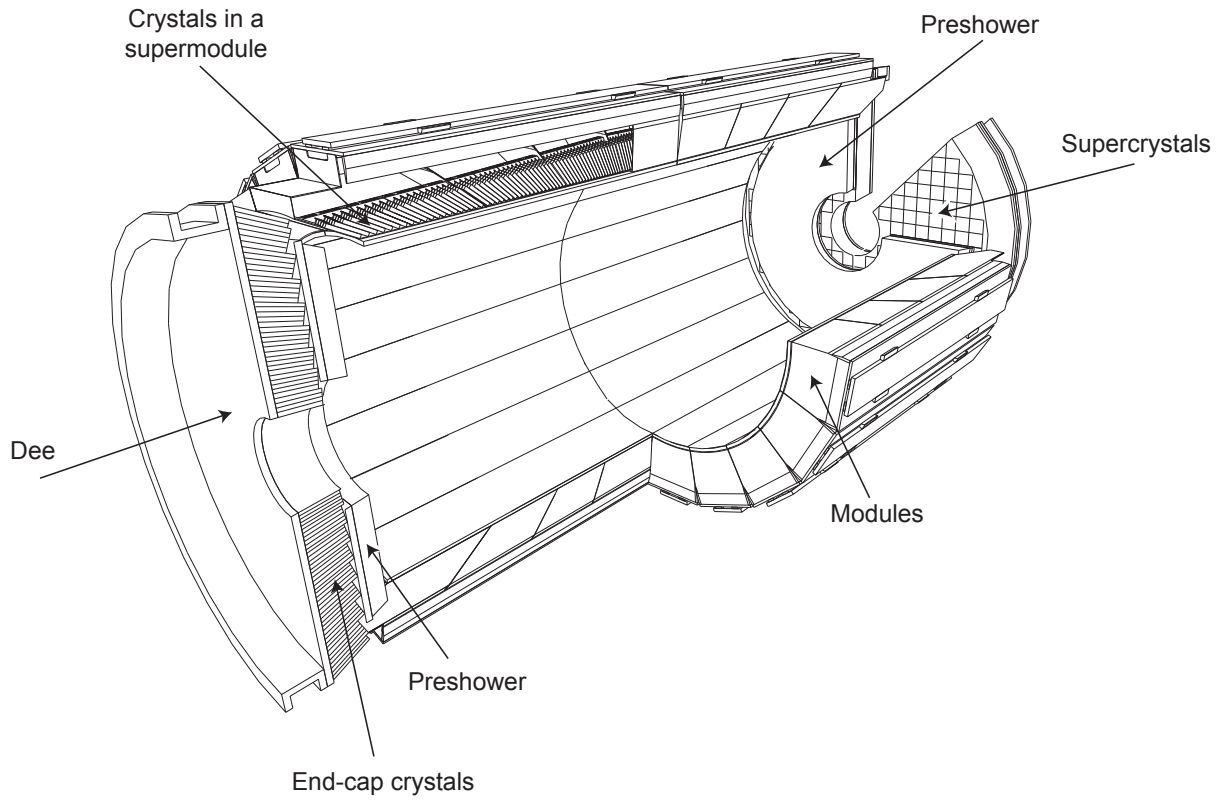

Figure 1.5: Layout of the CMS electromagnetic calorimeter [3].

The layout of the ECAL sub-detector is shown in Figure 1.5. An additional preshower detector is used in the endcap region to lower the required detector depth. Its principal

aim is to identify neutral pions in the endcaps, but it also helps to distinguish neutral pions and electrons from minimum ionising particles and improves the position determination of electrons and photons with high granularity.

Table 1.1: ECAL crystal characteristics

	Barrel	Endcaps
number of crystals	61 200	14 648
crystal cross-section in (η, ϕ)	0.0174×0.0174	varies
crystal cross-section at the front	22×22 mm^2	28.62×28.62 mm^2
crystal cross-section at the rear	26×26 mm^2	30×30 mm^2
crystal length	230 mm ($25.8X_0$)	220 mm ($24.7X_0$)

The main geometrical characteristics of the ECAL crystals are shown in Table 1.1. The choice of lead tungstate was driven by the constraints of the CMS design. It is a very dense material (8.28 g/cm^3) with a short radiation length of $X_0 = 0.89$ cm, which allows the calorimeter to fit inside the compact magnet. Lead tungstate also has a small Molière radius[1] of 2.2 cm, which allows a calorimeter with fine granularity. Finally, the crystals emit 80 % of their scintillation light in just 25 ns, however the light yield is relatively low. At 18 °C, about 4.5 photoelectrons per MeV are collected. The dependence of the light yield on temperature requires a cooling system capable of keeping the crystal temperature stable within ±0.05 °C to preserve energy resolution [5].

The energy-dependent resolution of the calorimeter can be parameterised as follows [3]:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2. \tag{1.4}$$

where S is the stochastic term, N is the noise term, and C is the constant term. Figure 1.6 shows the energy resolution measured using incident electrons, during the beam tests in 2004.

1.2.3 Hadron Calorimeter

The hadron calorimeter (HCAL) is the next sub-detector located mostly inside the solenoid and completing the CMS calorimetry system. It is essential for the measurement of hadron jets and missing transverse energy.

[1]The Molière radius R_μ is a characteristic constant of a material giving the scale of the transverse dimension of the fully contained electromagnetic showers initiated by an incident high energy electron or

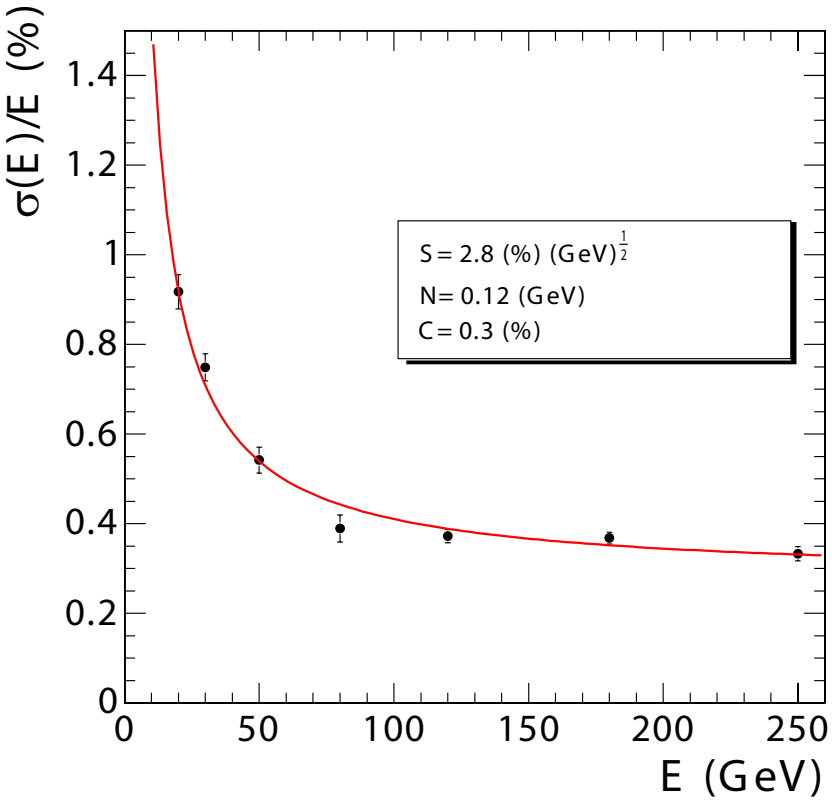

Figure 1.6: ECAL energy resolution derived from the test beam measurements as a function of deposited energy. The stochastic, noise, and constant contributions are shown [3].

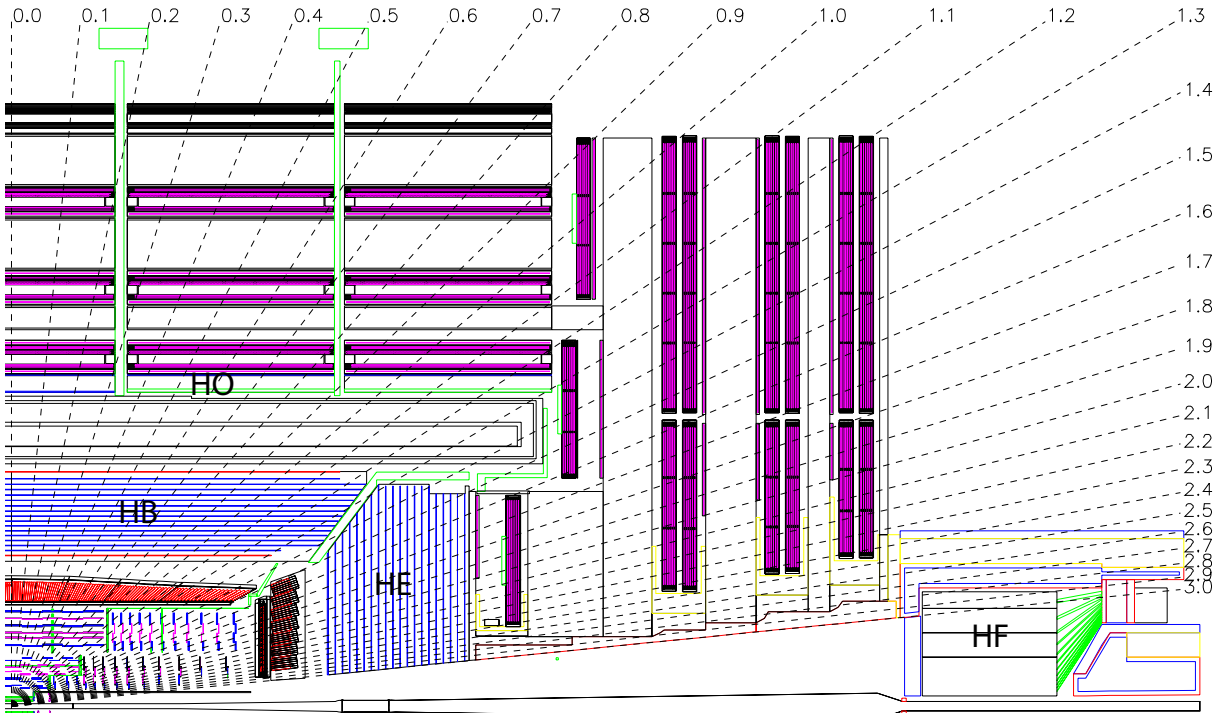

Figure 1.7: Longitudinal view of the CMS detector showing the locations of the hadron barrel (HB), endcap (HE), outer (HO) and forward (HF) calorimeters [3].

As shown in Figure 1.7, HCAL consists of four subsystems: the hadron barrel calorimeter (HB), the hadron endcap calorimeter (HE), the hadron outer calorimeter (HO) and the hadron forward calorimeter (HF). The barrel and endcap parts (HB, HE) cover the pseudorapidity range up to $|\eta|< 3.0$, and the forward part (HF) extends it to a total coverage of $|\eta|< 5.0$. HCAL surrounds ECAL from its outer limit of 1.77 m from the beamline, to the inner limit of the magnet coil at 2.95 m from the beamline. However, due to space limitations the barrel calorimeters do not contain complete hadronic showers, therefore an outer calorimeter (HO) was designed to measure the energy leakage. It is placed in the muon system just outside of the solenoid in the barrel region.

HCAL is a sampling calorimeter consisting of alternating layers of brass and stainless steel absorbers, and plastic scintillators as active elements. The choice of the absorber material was caused by its short hadronic interaction length and its property of being non-magnetic, which is crucial in the strong magnetic field of the CMS magnet. The scintillation light is guided by embedded wavelength-shifting (WLS) fibres. The light from the WLS is then transmitted via a network of clear fibres, arranged in read-out towers, to hybrid photodiodes (HPDs) [3].

Both HB and HE scintillators have a granularity of $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ for $|\eta|< 1.6$, and $\Delta\eta \times \Delta\phi = 0.17 \times 0.17$ for $|\eta|\geq 1.6$. The tower segmentation of the forward calorimeter (HF) varies from $\Delta\eta \times \Delta\phi = 0.175 \times 0.175$ at $|\eta|= 3.0$ to $\Delta\eta \times \Delta\phi = 0.3 \times 0.35$ at at $|\eta|= 5.0$. The HF is placed at about 11 m from the interaction point, and is essential to reconstruct very forward hadron jets. Together with HO, it provides the hermeticity of the calorimetry system, making it possible to measure the transverse missing energy to a reasonable precision.

[perhaps need to add the energy resolution plots]

1.2.4 Superconducting Magnet

The superconducting solenoid is a central feature of the CMS apparatus, essentially giving it its name. The magnet has a length of 12.5 m, diameter of 6.3 m and mass of 220 t. Although it was initially designed to sustain a uniform magnetic field of 4 T within the 5.9 m diameter free bore, operation at 3.8 T was chosen in order to increase the lifetime. The magnetic field is returned by a massive iron yoke. The main parameters of the CMS magnet are shown in Table 1.2.

The large bending power of the solenoid is required to bend the tracks of high energy charged particles to an extent where good momentum resolution is achieved. The design

photon. It is defined as the radius of a cylinder containing an average of 90 % of the shower's energy deposition.

Table 1.2: Parameters of the CMS superconducting solenoid [5] [6].

Field	3.8 T
Inner Bore	5.9 m
Length	12.5 m
Number of Turns	2168
Current	18 160 kA
Stored energy	2.3 GJ

requirement for the strength of the magnetic field was the ability to unambiguously determine the sign of the electric charge for muons with a momentum of ≈ 1 TeV/c [5].

The solenoid coil is constructed from four layers of superconducting high-purity niobium-titanium cable co-extruded with pure aluminium, which acts as a thermal stabiliser. The cold mass is cooled down to 4.5 K by liquid helium. If a fast discharge happens (e.g. caused by a magnet quench), about 3 days are necessary to re-cool the coil.

1.2.5 Muon System

The last sub-detector placed on the outermost part of CMS is the muon system. Since the muons are the most penetrating particles detectable by CMS, they have the cleanest signature and play an important role in many physics analyses. Due to their ability to travel through the many layers of the calorimeters, muons are relatively easy to identify and separate from the background.

The layout of the CMS muon system is shown in Figure 1.8. It consists of the drift tubes (DT), cathode strip chambers (CSC) and resistive plate chambers (RPC). The entire system surrounds the solenoid and covers the pseudorapidity region of $|\eta| < 2.4$.

The drift tubes are located in the barrel region ($|\eta| < 1.2$). Consisting of four stations, they form concentric cylinders around the beam line; there are 250 drift chambers with about 172 000 sensitive wires in total. When a muon passes through the volume, it knocks electrons off the atoms of the gas, which then follow the electric field and reach the positively-charged wires, providing information on the muon's position. The chambers are filled with the gas mixture of 85 % Ar and 15 % CO_2, where the muon drift time does not exceed 380 ns. Although this value is bigger than the typical bunch crossing time (25 or 50 ns), it is sufficient because of the small muon rate in this region.

In the endcaps, the cathode strip chambers cover the pseudorapidity region of $0.9 < |\eta| < 2.4$. Each of 468 CSCs is a trapezoidal multi-wire proportional chamber consisting of 6 gas gaps with a plane of radial cathode strips and a plane of anode wires which are roughly per-

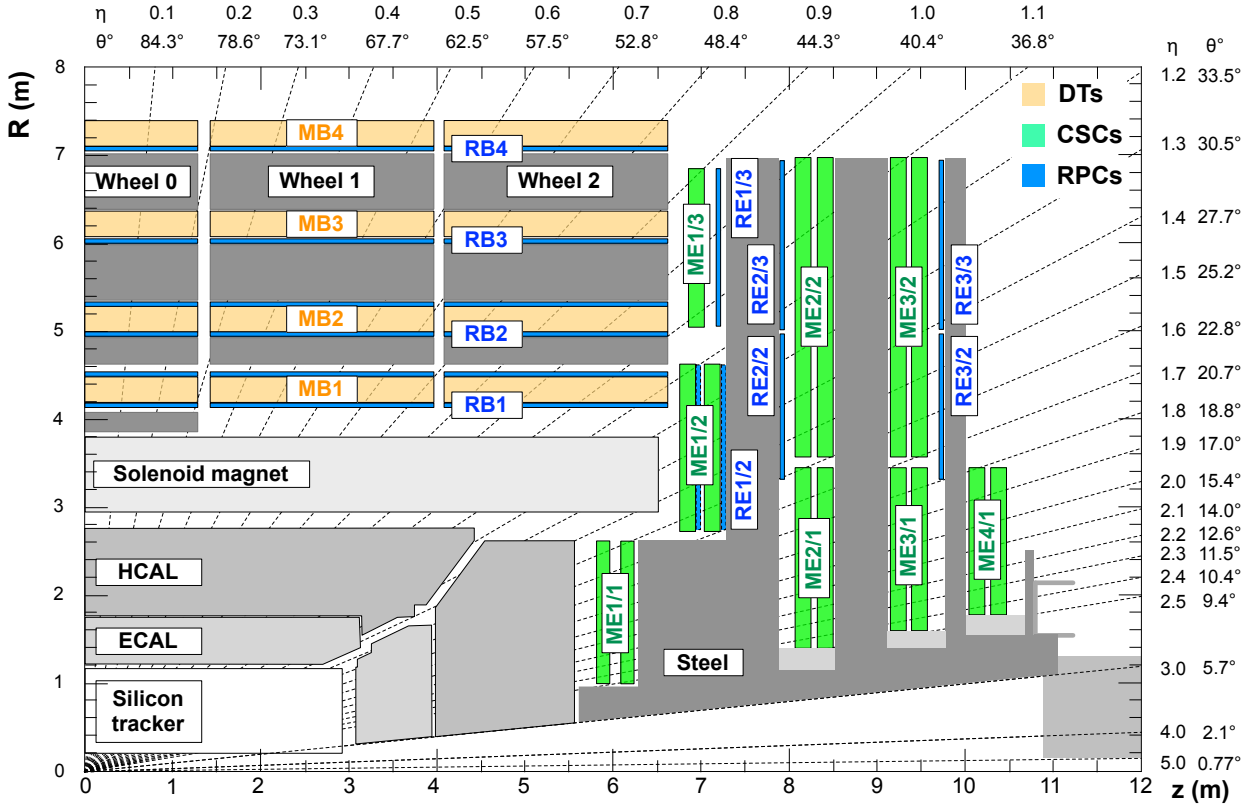

Figure 1.8: Layout of one quarter of the CMS muon system. Four drift tube (DT, in light orange) stations are labeled MB (muon barrel) and the cathode strip chambers (CSC, in green) are labeled ME (muon endcap). Resistive plate chambers (RPC, in blue) are in both the barrel and the endcaps of CMS, where they are labeled RB and RE, respectively.

pendicular. A charged muon traversing each plane of a chamber causes gas ionisation and a subsequent electron avalanche which produces a charge on the anode wire and an image charge on the cathode strips. The gas used in CSCs is a mixture of Ar, CO_2 and CF_4.

The resistive plate chambers system is complementary to both DT and CSC systems, and is located in both barrel and endcap regions ($|\eta|< 2.1$). RPCs also operate in avalanche mode with a gas mixture of $C_2H_2F_4$, C_4H_{10} and SF_6, and due to an excellent time resolution of about 1 ns they provide fast information for triggering. The spacial resolution is, however, quite limited ($\approx$ 1 cm, compared to $\approx$ 100 µm for DTs and CSCs).

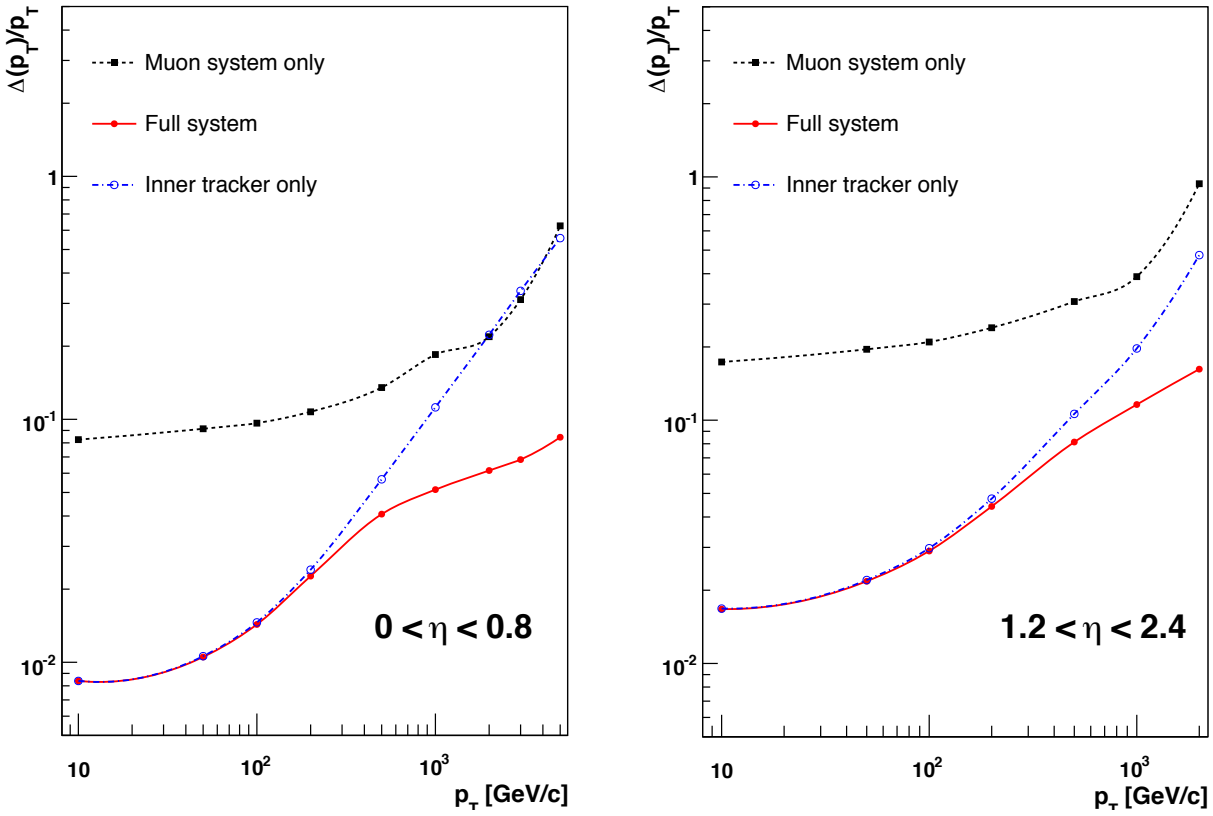

Figure 1.9: The muon transverse momentum resolution as a function of the transverse momentum (p_T) using the muon system only (black), the inner tracking only (blue), and both (red), in regions of $|\eta|< 0.8$ (left) and $1.2 <|\eta|< 2.4$ (right) [3].

The muon momentum is measured in both the tracker and the muon system. As it can be seen on Figure 1.9, both sub-systems contribute to the momentum resolution at different p_T values. This happens due to the difference in the magnetic field and detector technology. For low-p_T muons, the best momentum resolution is obtained in the tracker, whereas in

the high-p_{T} region the muon system provides a significant improvement. Therefore, by using information from both the silicon tracker and the muon chambers (i.e. reconstructing the "global muon"), the momentum resolution is improved in the whole p_{T} region up to a ≈ 1 TeV/c level.

1.2.6 Trigger and Data Acquisition

At design LHC luminosity of $L = 10^{34}$ cm^{-2} s^{-1}, approximately 25 collisions are expected to occur at each crossing of the proton bunches. The bunch spacing of 25 ns corresponds to a crossing rate of 40 MHz. Since every event produces ~ 1 MB of raw data, it corresponds to a total data production of 40 TB s^{-1}. Attempting to store all of this data is clearly beyond the available technology. Moreover, only a fraction of events contain hard scattering processes that are of interest, therefore an effective trigger system had to be implemented.

The CMS trigger is a two-level system, consisting of two independent parts: the Level-1 (L1) trigger and the High-Level Trigger (HLT). The L1 trigger is a hardware system implemented in programmable electronics residing partly on detector, and partly in the underground control room located at approximately 90 m from the experimental cavern. The maximum latency between the collision and the L1 accept decision received by front-end electronics is 3.2 µs. During this amount of time, the complete event information is buffered in pipelined memories on the detector. The only information used for the L1 trigger decision is that from the muon system and the calorimetry. Since the reconstruction of tracks exceeds the time scale required for the L1 decision, the tracker information can't be used. The L1 trigger reduces the event rate from ~ 40 MHz to ~ 100 kHz, corresponding to a data flow of about 100 GB s^{-1}. These events are fed into the HLT system.

The High-Level Trigger is a software system implemented in a single CPU farm, sometimes referred to as the "Event Filter Farm". Having access to the full event information, customised algorithms of increasing complexity are used which results in a highly flexible trigger system. The event rate is reduced down to ~ 300 Hz, with the final data rate of approximately 300 MB s^{-1} being stored on a large disk cache at the experimental site (the Storage Manager) and later on transferred to CERN Tier 0 for further processing (see Section 1.3).

Since the start of the LHC running, the operating conditions have been changing drastically. During the start-up year of 2010, the instantaneous luminosity went up from about 10^{27} cm^{-2} s^{-1} to approximately 0.2×10^{32} cm^{-2} s^{-1}. In 2011 the luminosity ramped up to a factor of 20 above that of 2010, reaching approximately 4×10^{33} cm^{-2} s^{-1}. This required a lot of continuous effort to control the trigger rates at reasonable level, whilst also keeping its efficiency acceptable. In 2012 the luminosity was more stable, peaking

at $\approx 7.6 \times 10^{33}$ cm^{-2} s^{-1} which is just a factor of 2 above the 2011 values. However, it still came as a challenge because of the impact of pile-up. At a bunch spacing of 50 ns and increased centre-of-mass energy of 8 TeV, the average number of pile-up vertices nearly doubled comparing to that in 2011, which required a major CPU extension and implementation of sophisticated PU mitigation techniques at the HLT level. The author's contribution to the HLT development of the trigger paths important for top physics is described in Chapter 2.

1.3 Computing

The vast amounts of data delivered by the CMS detector impose high requirements on the offline computing system. During 2010–2012 operation, CMS collected ~ 10 PB of raw data per year. Including Monte Carlo simulations, reconstructed data and analysis skims, the total annual amount of data essentially doubles. To handle the distributed storage and processing of this data, not just for CMS but for the entire high energy physics community using the LHC, a worldwide LHC computing grid (WLCG) has been put in place.

WLCG is a global collaboration of more than 150 computing centres in about 40 countries. The grid has a tiered architecture, comprising 4 tiers with different resources and services. The first one, Tier 0, is based at CERN and is responsible for data-taking. It accepts raw data from the data acquisition system and repacks it into primary datasets according to the trigger information. The raw data is archived to tape, and is also prompt-reconstructed (within 48 hours) before being distributed to the Tier 1 (T1) centres around the world. There are 8 T1 sites based at large national laboratories in collaborating countries (e.g. RAL in the UK and FNAL in the US). Each of the T1 centres is used for large-scale centrally organised data-processing activities. The data is then distributed in the reduced format (see Section 1.3.1) to a more numerous set of Tier 2 centres, typically located at collaborating universities. Each of these centres is used for the grid-based analysis and Monte Carlo simulation for the whole experiment, as well as local services for groups maintaining them. The last stage of computing system, Tier 3, is meant solely for the local institution's user analysis.

1.3.1 Event Data Model

In the basis of the CMS Event Data Model lies the concept of an event, which is physically a result of a single collision in the LHC. From a software point of view, the event is a C++ object container storing raw data from a single readout of detector electronics (e.g. hits in various sub-detectors), as well as reconstructed data which is based on this information, such as tracks, clusters and physics objects. All these C++ objects are stored in ROOT format [7].

The EDM makes use of three main data formats, based on different levels of detail and precision:

- RAW format, containing full information from the detector as well as L1 and HLT trigger decisions, with the event size of ~ 1.5 MB.
- RECO (reconstructed data) format, which is obtained from raw data by application of pattern recognition and compression algorithms. This data includes reconstructed detector hits, clusters and physics objects (electrons, muons, etc.). The typical event size is ~ 250 kB.
- AOD (Analysis Oriented Data) format, produced by filtering the RECO data from the reconstructed detector objects, leaving just the high-level physics objects required for analysis. The event size is reduced down to ~ 50 kB.

The RECO and AOD data are analysis-ready data formats, produced centrally and used by many physics analysis groups. However, further simplification of the data is also a common practice. By transforming the C++ objects produced by CMS software into plain basic types or vectors of them, only including the analysis-specific content, the event size can be reduced down to ~ 3 kB level depending on the needs of a particular analysis. This data format is often referred to as private "ntuples", and it requires specific analysis software capable of restructuring the data into user-defined classes. By following this approach, the analysis can be run locally and generally much faster than processing the RECO or AOD data. However, it requires "ntuplising" this data every time when new centrally-recommended physics objects or corrections are produced.

1.3.2 Analysis Software

Both of the analyses described in this thesis use the CMS software framework[1] (CMSSW), as well as Bristol Analysis Tools[2] (BAT). The differential cross section analysis also uses an additional level of python scripts for post-processing[3].

CMSSW is the key CMS software framework built around the Event Data Model (see Section 1.3.1). The framework is essential for purposes of Monte Carlo simulation, detector calibration and alignment, as well as data reconstruction and analysis. CMSSW has a modular architecture, consisting of one configurable executable (cmsRun) and a large set of plug-in modules that contain all the code needed for event processing (reconstruction algorithms, calibration, etc.). Different versions of CMSSW were used for different analyses:

[1] http://cms-sw.github.io/cmssw
[2] https://github.com/BristolTopGroup/AnalysisSoftware
[3] https://github.com/BristolTopGroup/DailyPythonScripts

- `CMSSW_4_2_8` for the top mass analysis on 2011 data;
- `CMSSW_4_4_4` for the missing transverse energy analysis on 2011 data;
- `CMSSW_5_3_9` for the top cross pair cross section analysis on 2012 data.

Corresponding versions were used to produce ntuples for processing by BAT, which was used to read the data, apply selections, calculate high-level variables and to create various histograms of distributions. BAT was originally started in 2010 by Dr. Lukasz Kreczko for the needs of the Bristol top group, later on also developed by the author and other researchers from Bristol and affiliated top groups. Like CMSSW, this framework has a modular structure, with its classes falling in four main categories:

- readers, for translating plain data types from ROOT files into C++ objects;
- RECO objects, i.e. output of the readers (physical objects like leptons, jets and its collections);
- selections, for application of event selections;
- analysers for creating histograms, applying selections, algorithms, and filling histograms.

All analysers are independent from each other, making the analysis chain stable and reliable. The final set of python scripts is used to prepare the histograms, perform fitting and unfolding procedures (in case of cross section analysis), and producing final tables and plots. Rootpy[1] package was used to access ROOT libraries in python interface, and matplotlib[2] was used to create plots.

1.4 Object Reconstruction

Most CMS analyses, including the ones described in this thesis, adopt a reconstruction technique called Particle Flow (PF) [8]. This algorithm is used to obtain a global event description at level of individually reconstructed particles by means of combining information coming from all sub-detector systems. The ultimate goal is to determine type, energy and momentum of all the particles in the event with highest possible precision and in the most optimal way. The types of these particles include electrons, muons, charged hadrons, neutral hadrons and photons. All these particles are then used to reconstruct jets (Section 1.4.3), missing transverse energy (Section 1.4.1) and tau leptons from their decay products.

[1] http://rootpy.org/
[2] http://matplotlib.org/

1.4.1 Electron Reconstruction

The reconstruction of the $t\bar{t}$ pair with an electron in the final state imposes high requirements on the electron identification and its energy-momentum measurement, precision of which is of major importance for both top mass and $t\bar{t}$ cross section measurements.

Although the CMS detector is equipped with highly accurate ECAL and tracker systems, electron identification and reconstruction is still a challenging task due to the large amount of tracker material (see Section 1.2.1). This results in a significant Bremsstrahlung photon emission, which often causes an ECAL energy deposit to be widely spread in azimuthal direction because of the high magnetic field. Therefore, dedicated algorithms were developed in order to collect all Bremsstrahlung energy deposits in the calorimeter (Bremsstrahlung recovery), and also to take into account the kinks in the electron trajectory caused by photon emissions.

Electron reconstruction in CMS has following distinct stages: seeding, track finding, pre-identification, Bremsstrahlung recovery, track-cluster linking and final identification. Historically, the original seeding algorithm was designed and optimised for isolated high-p_{T} electrons. This approach starts from ECAL clusters, and therefore is called the 'ECAL-driven' seeding. It is based on the property of the ECAL energy deposits to have narrow width in the η coordinate, and to be widely spread in ϕ (azimuthal direction) like it was mentioned above. The electron and all the associated Bremsstrahlung energy deposits form a single "super-cluster", and the ability to correctly determine it affects the overall performance of this method. Only super-clusters with transverse energy above 4 GeV are taken into account. Super-clusters are then matched to pairs or triplets of hits in the inner tracker layers, forming the track seeds on which the electron tracks are built upon.

The performance of the ECAL-driven method is not very well suited for non-isolated and low-p_{T} electrons. This occurs mainly due to the fact that the super-cluster position and energy can be highly biased by the impact of overlapping particles, especially if the electron happens to be within a jet and therefore non-isolated. Also, high track multiplicity complicates the backward propagation from a super-cluster, because it can be consistent with a number of track seeds corresponding to other particles. To minimise the number of these fake seeds, the ratio between the HCAL and ECAL energy deposits (H/E) is required to be smaller than 0.15. The HCAL towers used in the calculation of this ratio are taken within a cone of $\Delta R = 0.3$ behind the super-cluster position. Although this helps to keep the fake seed rate under control, the efficiency for non-isolated electrons becomes rather limited. As for the low-p_{T} electrons, the wider azimuthal spread of Bremsstrahlung photons leads to poorer reconstruction of the super-cluster, biasing its position and therefore preventing the

efficient matching with a track seed.

Within the particle flow method, efficient reconstruction of non-isolated and low-p_{T} electrons is particularly important since it affects the reconstruction of jets and missing transverse energy. Therefore, a different ('tracker-driven') seeding algorithm is used, which starts from reconstruction of tracks. The baseline of the CMS track reconstruction is the Kalman filter (KF) [9], which is a linear least-squares estimator based solely on Gaussian probability density functions. It is particularly suitable for muon reconstruction since it is dominated by multiple Coulomb scattering and its impact is well modelled by Gaussian fluctuations. However, this approach usually fails for electrons because Bremsstrahlung photon emission is highly non-Gaussian. To accommodate for the resulting kinks in the electron trajectory, the Gaussian-Sum Filter (GSF) [10] is used, which is essentially a non-linear generalisation of the Kalman Filter. In this method, Bremsstrahlung energy loss is modelled by a Gaussian mixture, therefore GSF track fit provides a better estimate for the inner and outer track momentum comparing to the KF algorithm. The downside of this approach is its high CPU usage, which means it can be run on a limited number of seeds.

The GSF tracks are reconstructed upon all ECAL-driven seeds. In case of the 'tracker-driven' seeds, a pre-identification based on high-purity KF tracks has been adopted. This procedure starts with the tracks reconstructed with very tight criteria, thus decreasing the fake rate yet compromising on tracking efficiency. Then an iterative-tracking strategy is carried out by means of removing hits unambiguously assigned to tracks from the previous iteration, and also progressively relaxing track seeding criteria. This approach leads to both high efficiency and low fake rate, which is crucial for low-p_{T} and non-isolated electrons.

In the next step, track-cluster matching has to be performed. In case if electron has negligible Bremsstrahlung emission, the track is well reconstructed with the KF algorithm all the way to the ECAL internal surface, where the closest cluster is matched to the track. The corresponding cluster energy is compared with the track momentum, and if the ratio (E/p) is close to unity, the track is selected. On the contrary, if the electron experiences a significant Bremsstrahlung emission, other track characteristics have to be exploited. In this case a selection based on the number of hits in the tracker and the χ^2_{KF} of the KF fit is applied before running a GSF refit. Finally, the number of hits, the GSF refit χ^2_{GSF}, $\chi^2_{\mathrm{KF}}/\chi^2_{\mathrm{GSF}}$ ratio, the energy loss measured by the track and the quality of the ECAL cluster-track matching are fed into a multivariate analysis using a Boosted Decision Trees (BDT) estimator.

Both tracker-driven and ECAL-driven seeds are used to obtain the GSF track collection of electron candidates. In the particle flow algorithm, it is necessary to link both electron and Bremsstrahlung energy deposits to the GSF track. A super-cluster is linked to a track if

the extrapolated position from the outermost tracker measurement is within the boundaries of one of the ECAL cells at the expected depth of the electron shower maximum. The preshower-ECAL and ECAL-HCAL links are made in a similar way.

Another important particle flow procedure, also driven by GSF tracks, is Bremsstrahlung recovery. In order to reconstruct an electron with correctly assigned energy and momentum, it is crucial to identify all energy deposits from Bremsstrahlung photons, thus forming a super-cluster. This procedure is carried out for each tracker layer by computing a straight-line extrapolation tangent to the track, up to the calorimeter. To determine a Bremsstrahlung photon, track-cluster linking is performed as described above. To limit the charged hadron contamination, clusters already assigned to KF tracks are not included in the calculation. Also, the distance in η coordinate between the extrapolation and the cluster is required to be smaller than 0.015, which helps to reduce the neutral particles background.

1.4.1.1 Electron Identification

Electron reconstruction in CMS is based on a characteristic signature that electrons leave in the tracker and calorimetry systems. However, other objects like charged hadrons, jets or photon conversions can produce very similar signatures and therefore may be reconstructed as electrons. Therefore, in order to distinguish these "fake" electrons from "real" ones, a further selection has to be applied. This procedure is referred to as electron identification (or electron ID).

Initially the electron identification is performed at the final stage of the electron reconstruction process. The working points of the cuts applied are selected to be loose enough in order to satisfy most CMS analyses requirements. Afterwards, more specific (tighter) ID cuts are applied for each individual analysis, defining the working point in the trade-off between selection efficiency and fakes contamination. This will be discussed separately in the selection description for each analysis in corresponding chapters.

There are several electron identification algorithms used by various CMS analyses, and four of them are used in the analyses described in this thesis: simple cut-based (SCB ID), cuts in categories (CiC ID), particle flow (PF ID) and multi-variate analysis (MVA ID).

Simple cut-based identification is used in the High-Level Trigger, and therefore has to be as simple, fast and robust as possible. Cuts are applied on the following variables:

- H/E, i.e. the ratio of hadronic energy of the HCAL towers centred at the super-cluster position of the electron in a cone of radius $\Delta R = 0.15$, and the super-cluster electromagnetic energy;

- $\Delta\eta_{\text{in}}$, i.e. the difference in η between the extrapolated track position and η of the super-cluster;

- $\Delta\phi_{\text{in}}$, i.e. the difference in ϕ between the extrapolated track position and ϕ of the super-cluster;

- $\sigma_{i\eta,i\eta}$, i.e. cluster shape of the electron energy in a 5×5 block of crystals around the seed crystal (the one with the highest energy) [11]:

$$\sigma_{i\eta,i\eta} = \sum_{5\times5 \text{ crystals}} (\eta_i - \eta_{\text{seed cluster}})^2 \frac{E_i}{E_{\text{seed cluster}}} \tag{1.5}$$

Although this identification method has an advantage of its simplicity, it does not show the best signal efficiency and background rejection. One of the more complex methods is the CiC ID, which exploits the categorisation of electrons. It is optimised to select electrons from different sources (W, Z and J/ψ decays) and reject fakes from jets or conversions. In order to achieve higher efficiency, electron candidates are split into categories with different signal to background ratio, allowing to better tune the working points of the cuts. The first step in categorisation is done by division between barrel and endcap regions of the detector. Since the properties of both the tracker and calorimetry systems differ significantly in these two regions, different cuts are applied. The second step is based on the following observables:

- f_{brem}, or measured bremsstrahlung fraction, defined as:

$$f_{\text{brem}} = \frac{p_{\text{in}} - p_{\text{out}}}{p_{\text{in}}} \tag{1.6}$$

  where p_{in} is the initial track momentum at the vertex and p_{out} is the track momentum at the last hit;

- E/p, which is the ratio of the super-cluster energy and the initial track momentum.

Three categories of electron candidates are distinguished [12]:

- "Low-Brem": $0.9 < E/p < 1.2 - f_{\text{brem}} < 0.12$ (barrel), $0.82 < E/p < 1.22 - f_{\text{brem}} < 0.2$ (endcap), the fake-like region with high number of both real and fake electrons;

- "Bremming": $0.9 < E/p < 1.2 - f_{\text{brem}} > 0.12$ (barrel), $0.82 < E/p < 1.22 - f_{\text{brem}} > 0.2$ (endcap), the electrons-like region with a little contamination from fakes;

- "Bad-Track": remaining regions with a low number of real electrons.

The third and the final step in categorisation is based simply on the electron transverse energy, with the lower threshold taken as 10 GeV. In each category, the selection is performed on the basic ID variables already mentioned above for the simple cut-based identification. On top of these variables, isolation and conversion rejection variables are also used, they are discussed in Section 1.4.1.2 and Section 1.4.1.3, respectively.

The cuts within CiC ID are applied in order to maximise the signal to background ratio. Depending on the needs of different analyses, nine levels of cut severity are implemented: *VeryLoose, Loose, Medium, Tight, SuperTight, HyperTight(1-4)*. Each step decreases the fake rate by about a factor of two for electrons with $E_{\mathrm{T}} > 20$ GeV [12]. The CiC ID is used as a primary electron identification method in the Top Mass analysis (Chapter **??**).

Particle flow ID is the final step of the particle flow electron reconstruction. Following the particle flow concept, is uses information from all the CMS sub-detectors obtained in previous reconstruction steps to build new observables for electron identification. The complete list of PF ID observables is given below:

- p_{T} and η of the GSF track;
- GSF $\sigma_{p_{\mathrm{T}}}/p_{\mathrm{T}}$, transverse momentum resolution of the GSF track;
- $\#\mathrm{hits}_{\mathrm{KF}}$, number of reconstructed KF track hits;
- χ^2_{GSF} and χ^2_{KF}, GSF and KF goodness-of-fits;
- $\Delta\eta$: distance in η between the position of the cluster and the extrapolated position of the GSF track;
- $\sigma_{\eta\eta}$, cluster shape (Equation 1.5) of the ECAL cluster linked to the GSF track;
- $H/(H+E_{\mathrm{e}})$, hadron fraction of the shower, where H is the energy of the hadron cluster linked to the GSF track;
- $(E_{\mathrm{e}}+\sum E_{\gamma})/p_{\mathrm{in}}$, ratio between the super-cluster energy and the inner track momentum;
- $E_{\mathrm{e}}/p_{\mathrm{out}}$, ratio between the electron cluster energy and the track outer-momentum;
- $\sum E_{\gamma}/(p_{\mathrm{in}} - p_{\mathrm{out}})$, the ratio between the Bremsstrahlung photon energy as measured by ECAL and by the tracker;

- *EarlyBrem*, flag of $(E_{\mathrm{e}}+\sum E_{\gamma}) > p_{\mathrm{in}}$ inequality, corresponding to an electron emitting an "early" Bremsstrahlung photon, i.e. before it has crossed at least three tracker layers;
- *LateBrem*, flag of $E_{\mathrm{e}} > p_{\mathrm{out}}$ inequality, corresponding to an electron emitting a "late" Bremsstrahlung electron, when the ECAL clustering is not able to disentangle the overlapping electron and photon showers;
- $f_{\mathrm{brem}} = (p_{\mathrm{in}} - p_{\mathrm{out}})/p_{\mathrm{in}}$, Bremsstrahlung fraction (Equation 1.6).

All these variables are combined into a single discriminator by a multivariate analysis technique (BDT method), which has been trained on signal and background Monte Carlo samples. PF ID is a relatively loose identification method, since it has to satisfy the needs of all analyses using particle flow collections.

Finally, MVA electron identification is used in top cross sections analysis on 2012 data. It is another multivariate analysis technique, optimised to select isolated electrons from W and Z decays. The variables used in MVA ID are also combined into a single discriminator, they are largely similar to the ones used in PF ID:

- p_{T} and η of the GSF track;
- $\#\mathrm{hits}_{\mathrm{KF}}$, number of reconstructed KF track hits;
- χ^2_{GSF} and χ^2_{KF}, GSF and KF goodness-of-fits;
- $\Delta\eta$: distance in η between the position of the cluster and the extrapolated position of the GSF track;
- $\Delta\phi$: distance in ϕ between the position of the cluster and the extrapolated position of the track;
- $\Delta\eta_{vtx}$: distance in η between the position of the cluster and the position of the GSF track at vertex;
- $\Delta\phi_{vtx}$: distance in ϕ between the position of the cluster and the position of the GSF track at vertex;
- $\sigma_{\eta\eta}$, cluster shape in η;
- $\sigma_{\phi\phi}$, cluster shape in ϕ;
- η width of the super-cluster;

- ϕ width of the super-cluster;
- H/E, ratio of HCAL and ECAL cluster energy;
- $E_{\text{super-cluster}}/p_{\text{T}}$, ratio between the super-cluster energy and the track momentum;
- $1/E_{\text{super-cluster}} - 1/p_{\text{T}}$, difference between inverse super-cluster energy and inverse track momentum;
- $E_{\text{e}}/p_{\text{out}}$, ratio between the electron cluster energy and the track outer-momentum;
- $1 - E_{1\times5}/E_{5\times5}$, where $E_{1\times5}$ is the energy in the central 1×5 strip of the 5×5 electron cluster and $E_{5\times5}$ its total energy;
- $E_{3\times3}/E_{\text{super-cluster, raw}}$, ratio of the energy of a cluster of 3×3 and the uncorrected (raw) energy of the super-cluster;
- $E_{\text{PS}}/E_{\text{super-cluster, raw}}$, ratio of the energy in the preshower detector and the raw super-cluster energy (only in the endcap region).

1.4.1.2 Electron Isolation

Isolation is an observable that allows to distinguish prompt electrons (i.e. the ones from W and Z decays) from jets faking electrons and electrons within jets. It is essentially a measure of activity around the particle. In CMS, there are two different ways of quantifying isolation: detector-based and particle-based. The detector-based isolation is defined separately for each detector sub-system (tracker, ECAL and HCAL):

- Tracker isolation, calculated as the sum of transverse momenta of all track within a cone of $\Delta R = 0.3$ around the electron, excluding the electron momentum itself;
- ECAL isolation, i.e. the sum of transverse energy of all ECAL clusters within a cone of $\Delta R = 0.3$ around the super-cluster position. The footprint of the original electron is also removed;
- HCAL isolation, defined as the sum of transverse energy of all HCAL towers within a cone of $\Delta R = 0.3$ centred at the super-cluster position.

Particle-based isolation exploits the particle flow information: it is calculated as the sum of transverse energy of all PF particles in the cone of $\Delta R = 0.3$ around the electron. Both detector-based and particle-based isolation definitions are often normalised to the electron

transverse momentum (or energy in case of calorimeter isolation) in order to improve signal efficiency. The normalised sum of the tracker, ECAL and HCAL isolation variables is referred to as detector-based relative isolation (or RelIso), similarly normalised particle-based isolation is called PF RelIso. These quantities are crucial in various methods to estimate the QCD background contribution for many analyses, and will be used for both electrons and muons throughout this thesis.

1.4.1.3 Identification of photon conversions

Interacting with detector material, photons can convert into electron-positron pairs. Due to the large amount of material budget in the tracker (Figure 1.4), especially in the end-cap region, there is a high chance of conversions to happen. The resulting electrons can successfully fake prompt signal electrons, passing all the identification criteria and appearing isolated if the initial photon was isolated, too. Therefore conversions constitute a large proportion of the QCD background to top signal. The electron-positron pair may not be symmetrical in transverse momenta, therefore a veto on a second electron is not sufficient to reject such electrons and other conversion identification criteria are necessary.

The simplest method is based on counting the number of missing hits in the tracker. Conversions are most likely to happen at some distance from the interaction point, essentially anywhere between the point of photon production and the end of the tracker. Therefore electrons produced in such conversions are likely not to traverse through all the pixel layers. To separate these electrons from the prompt ones, a cut on the number of missing layers can be used. However, this approach fails if conversion occurs in the beam pipe, or if the reconstructed electron track is paired with unrelated hits in the tracker, making it look like a prompt electron.

A slightly more sophisticated method is a partner track method. It is based on geometrical cuts on *dist* and *dcot* variables, which refer to the distances between tangent points of any two tracks in the $r-\phi$-plane and $r-z$-plane, respectively:

$$dist = ||\vec{r}_1 - \vec{r}_2| - r_1 - r_2| \tag{1.7}$$

$$dcot = \left| \frac{1}{\tan\theta_1} - \frac{1}{\tan\theta_2} \right| \tag{1.8}$$

Here $\vec{r}_{1,2}$ are radial vectors of the two tracks, $r_{1,2}$ are the track radii and $\theta_{1,2}$ – angles between the tracks and the beam pipe. Apart from the cuts on *dist* and *dcot* variables, tracks should have opposite charge in order to be considered to come from a photon conversion.

These two methods for conversion identification were used in the top mass analysis on

2011 data. For the top cross sections analysis on 2012 data, along with the number of hits method, a more advanced technique called vertex fit was used. It is essentially a full vertex fit of all pairs of tracks, with the selection being made on the the fit probability. The method benefits from the full use of track uncertainties and covariances, combining all information into a single discriminator. However, increased complexity of this technique makes it more CPU-intensive than the partner track method.

1.4.2 Muon Reconstruction

A good muon reconstruction and identification was one of the main CMS design requirements. Initially, muons are reconstructed independently in the tracker ("tracker tracks") and in the muon system ("standalone muon tracks") using the Kalman Filter technique [9]. Based on these objects, two different reconstruction methods are used: [13]

- Global muon reconstruction. Each standalone muon track reconstructed in the muon system is matched with the tracker track by propagating onto a common surface. A global fit is performed on the combined collection of hits from both tracks, and the resulting muon is referred to as a *global muon*.

- Tracker muon reconstruction. All tracker tracks above certain threshold ($p_{\mathrm{T}} > 0.5$ GeV, $p > 0.5$ GeV) are extrapolated to the muon system, taking into account possible energy losses, magnetic field and multiple Coulomb scattering. If the track matches at least one muon segment in the muon system, i.e. a short track stub made of DT or CSC hits, it is considered a *tracker muon*.

Global muon reconstruction has a high efficiency for high-p_{T} muons, penetrating through more than one muon station. On the contrary, tracker muon reconstruction is more efficient for low-p_{T} muons ($p_{\mathrm{T}} < 5$ GeV), as it requires just a single muon segment. Since the $\mathrm{t\bar{t}}$ analyses described in this thesis have a semileptonic signature with exactly one energetic muon in the final state (in case of the muon channel), only the global muon reconstruction method is used, as its momentum resolution at high transverse momenta benefits from both the tracker and the muon system (Figure 1.9).

Following the reconstruction, the quality of the muon objects is verified by applying identification criteria. This is done in order to suppress hadronic punch-through, muons from decays in flight and cosmic muons. Selection is applied on the following observables:

- normalised χ^2 (χ^2/number of degrees of freedom) of the global muon fit;

- number of muon chamber hits in the global muon fit;

- number of muon stations with muon segments;
- transverse impact parameter d_{xy} (closest approach of the track to the primary vertex);
- longitudinal distance d_z of the tracker track w.r.t. the primary vertex;
- number of hits in the pixel detector;
- number of hits in the tracker layers.

The muon identification also exploits the CMS particle flow event reconstruction, making use of information coming from all sub-detectors. Particle flow muons (PF muons) are identified by imposing selection on all muon candidates reconstructed with the global muon method. This selection was optimised for identification of muons in jets, minimising the fake rate from misidentified charged hadrons, which is crucial for correct reconstruction of jets and missing transverse energy (described in more detail in the following sections). Depending on the analysis specifics, isolation requirements can also be applied; isolation definitions closely follow the ones for electrons (Section 1.4.1.2).

1.4.3 Jet Reconstruction

Hadronisation of quarks and gluons leads to production of narrow cones of particles moving in approximately one direction, called jets. This happens due to colour confinement, as particles carrying a colour charge cannot exist in free form, they have to fragment into hadrons before they can be detected directly. Therefore, to measure the initial parton's momentum and energy, all these particles must be combined into jets.

The CMS particle flow algorithm implies reconstruction of individual particles (charged and neutral hadrons, electrons, muons, photons) before combining (or clustering) them into jets. A few different jet clustering techniques exist, but the one used predominantly in CMS and exclusively in this work is anti-k_t algorithm [14], which defines the distance between constituent particles as:

$$d_{ij} = \min\left(\frac{1}{k_{t,i}^2}, \frac{1}{k_{t,j}^2}\right)\frac{\Delta_{i,j}^2}{R^2} \tag{1.9}$$

where $\Delta_{i,j}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, $k_{t,i/j}$, $y_{i/j}$ and $\phi_{i/j}$ are respectively transverse momenta, rapidities and azimuth angles of particles i/j, and R is the radius parameter.

The clustering proceeds by identifying the smallest distances between particles and recombining them until all jets are formed and no particles are left. An event typically has a few well-separated hard (high-p_{T}) particles and a large amount of soft (low-p_{T}) particles. The distance d_{1i} between a hard particle 1 and a soft particle i will be fully determined

by the transverse momentum of the hard particle and the Δ_{1i} separation. On the other hand, distance between soft particles with similar separation will be substantially larger. Therefore, soft particles tend to cluster around the hard ones before they cluster amongst themselves. On the output, the anti-k_t algorithm forms conical jets with boundaries resilient to soft radiation.

1.4.3.1 Jet Energy Corrections

Due to non-linear and non-uniform response of the calorimetry systems, jets reconstructed using detector inputs typically have energies different to the ones of corresponding Monte Carlo particle jets (or generator jets), reconstructed by clustering the four-momenta of all stable particles generated in Monte Carlo simulation. Therefore some mapping procedure is necessary. Corrections applied to reconstructed jets in order to translate the measured energy to the true particle (or parton) energy are referred to as jet energy corrections, or JEC.

CMS has adopted a factorised approach of applying jet energy corrections, meaning that each correction level takes care of a different effect. The set of corrections is applied sequentially, i.e. the output of each step is the input to the next one. Essentially, each level of correction is a scaling of a jet four-momentum, with a scale factor depending on various parameters of the jet, typically pseudorapidity and transverse momentum. Currently, the correction levels go as follows:

- L1 Offset correction;
- L2 Relative (η) correction;
- L3 Absolute (p_{T}) correction;
- L4 EMF (electromagnetic energy fraction) correction;
- L5 Flavour correction;
- L6 UE (underlying event) correction;
- L7 Parton correction.

The goal of the L1 correction is subtraction of pile-up and electronic noise contributions from the jet energy. The energy from pile-up vertices can be deposited in calorimeters since the additional proton-proton collisions occur close enough in time to the hard scattering process. Electronic noise in calorimeter readouts also creates additional energy offset which

needs to be corrected for. The scale factors for L1 correction are derived using data-driven methods (zero-bias collisions).

The relative L2 correction is designed to flatten the jet response in pseudorapidity. These corrections are extracted with respect to the barrel region in bins of p_{T}, and therefore are uncorrelated with the following L3 corrections. Both Monte Carlo and data-driven (dijet balance) methods are used to derive the L2 scale factors. Once a jet is corrected for η dependence, it is corrected back to the particle level by applying absolute L3 correction. The goal of this correction is to flatten the jet response in transverse momentum. Derivation of L3 scale factors is done by using Monte Carlo truth information, or data-driven Z/γ+jet balance techniques.

The L4 EMF correction takes care of variations in jet response with respect to electromagnetic energy fraction, i.e. the fraction of energy deposited in ECAL by hadrons. Although it has been shown that EMF-dependent correction in three parameters (p_{T}, η, EMF) can improve the observed jet resolution, this correction is optional for most CMS analyses and hasn't been used in this work.

Another optional correction is the L5 flavour correction, which is intended to correct the jets depending on the flavour of initiating partons, i.e. light quarks, heavy (b and c quarks), or gluons. Jets originating from heavy quarks are different from light quark jets in a few aspects. The average number of charged hadrons within b-jets is higher than the one for the light jets, although the average momentum of charge hadrons is smaller. Moreover, b-hadrons have a larger branching fraction into semileptonic decays with neutrinos in the final state that cannot be detected. All these factors may lead to a substantially smaller charged hadron energy fraction comparing to the light jets, therefore the average calorimeter energy response is lower for the b-jets. These effects are meant to be taken care of by the L5 correction, which is derived either from Monte Carlo or data-driven $\mathrm{t\bar{t}}$ events.

Underlying event [15] refers to all the activity in the proton-proton collision apart from the process of interest, i.e. the hard scattering process. It includes particles coming from additional parton interactions (or multiple scatterings) and beam remnants. Depending on the analysis goals, underlying event may also include initial and final state radiation (ISR and FSR) which represent the soft gluon radiation before and after the hard scattering process, respectively. Underlying event activity results in production of additional soft jets which can bias the jet energy measurement. Optional L6 UE correction attempts to mitigate such bias. However, it was not used in the analyses described in this thesis, since underlying event is an intrinsic part of proton-proton interactions effectively modelled by Monte Carlo simulation. In particular, the top cross sections analysis is designed to be very sensitive to

such higher-order processes, and not taking them into account may result in compromising the discriminating power between Monte Carlo generators.

Finally, L7 parton correction attempts to correct the jet energies back to the parton level. It is also optional and was only used in the top mass analysis. The correction factors were derived from Monte Carlo simulations by comparing the generator jets momenta to their matched partons.

Due to the fact that CMS simulation is not perfectly tuned to the data yet, additional residual L2L3 corrections are applied in order to achieve better agreement between data and simulation. It is essentially a small residual η- and p_{T}-dependent calibration applied exclusively to data, which will remain in place until CMS develops a perfectly tuned simulation reproducing the data features out of the box.

1.4.3.2 Particle Flow Jet Identification

To ensure the quality of the jets, a final identification criteria are applied to all jet objects. Particle flow jet identification (or PF jet ID) is used to reduce the noise and rate of electrons reconstructed as jets. The cuts are applied on the following observables :

- number of constituent particles;
- NHF (neutral hadron energy fraction);
- CHF (charged hadron energy fraction);
- NEF (neutral electromagnetic energy fraction);
- CEF (charged electromagnetic energy fraction).

A rather loose identification cuts (referred to as "loose PF jet ID") were used in this work: a requirement of more than one constituent particle in the jet, NHF$<$ 0.99, CHF$<$ 0.99, NEF$>$ 0 and CEF$>$ 0 in pseudorapidity region of $\eta < 2.4$.

1.4.3.3 b-tagging

The identification of jets originating from b-quarks, or b-tagging, is one of the most important tools in top quark physics, capable of significantly decreasing the background contamination of signal processes. A variety of b-tagging algorithms have been developed by CMS [16], and the one that was used in this work is called combined secondary vertex (CSV) algorithm. It is based on the fact that b-hadrons have a significant lifetime ($\sim 10^{-12}$ s) and can travel a distance of a few centimetres before decaying. Therefore jets originating from b-quarks are likely to have a secondary vertex located at a considerable distance from the

primary vertex, which can be used as an efficient discriminator between light jets and b-jets. In order to maximise its efficiency, apart from the secondary vertex information the CSV algorithm also exploits the track-based lifetime information. The following variables are used in the algorithm:

- number of tracks in the jet;
- number of tracks at the secondary vertex;
- secondary vertex category;
- secondary vertex invariant mass;
- ratio of the total track energy at secondary vertex with respect to all tracks in the jet;
- pseudorapidities of tracks at secondary vertex with respect to the jet axis;
- impact parameter significance (i.e. ratio of IP to its uncertainty) of the first track that increases the invariant mass above the charm threshold of 1.5 GeV (tracks are ordered by IP significance; the mass of the system is recalculated after adding each track);
- impact parameter significances of each track in the jet.

All these observables are combined into a single discriminator, the "medium" working point of which is used in this work. It provides ~ 70 % b-tagging efficiency for mis-tag rate of approximately 1 % [16].

1.4.4 Missing Transverse Energy

Due to the energy and momentum conservation, the sum of transverse momenta and energy of all particles in the final state of proton-proton collisions is expected to be zero. However, some particles can escape the detector without being reconstructed, therefore creating the imbalance in transverse momentum which is referred to as missing transverse energy, defined as:

$$\vec{E}_{\mathrm{T}}^{\mathrm{miss}} = -\sum_{i} \vec{p^{i}}_{\mathrm{T}} \tag{1.10}$$

where $\vec{p^{i}}_{\mathrm{T}}$ are the transverse momentum vectors of all reconstructed particles. The modulus of $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}$ vector is denoted by $E_{\mathrm{T}}^{\mathrm{miss}}$.

Accurate reconstruction of missing transverse energy is crucial for precise measurements of Standard Model processes with neutrinos in the final state. Top quark pair semileptonic decay is one of such processes, therefore analyses covered in this thesis implicitly (top mass)

or explicitly ($t\bar{t}$ cross section with respect to $E_{\mathrm{T}}^{\mathrm{miss}}$-related variables) rely on efficient reconstruction of $E_{\mathrm{T}}^{\mathrm{miss}}$. Misidentification and misreconstruction of any visible particles in the event contribute to $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement, therefore it is a rather demanding task.

Just like in the case of leptons and jets, particle flow was used for $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction: in equation 1.10 the transverse momentum vectors are of the particles reconstructed using particle flow algorithm. This procedure gives so-called raw $E_{\mathrm{T}}^{\mathrm{miss}}$ on the output. The raw $E_{\mathrm{T}}^{\mathrm{miss}}$ is systematically different from true $E_{\mathrm{T}}^{\mathrm{miss}}$, which denotes the transverse momentum carried by invisible particles. This happens mainly due to non-compensating nature of the calorimeters, effects of pile-up, noise, etc. Therefore, a set of corrections is applied:

- Type-0, which corrects $E_{\mathrm{T}}^{\mathrm{miss}}$ for pile-up;
- Type-I, a propagation of jet energy corrections (Section 1.4.3.1) to $E_{\mathrm{T}}^{\mathrm{miss}}$;
- xy-shift correction, reducing the $E_{\mathrm{T}}^{\mathrm{miss}}$ ϕ modulation.

The causes of systematic $E_{\mathrm{T}}^{\mathrm{miss}}$ ϕ modulation include detector misalignment, beam spot displacement, inactive calorimeter cells and anisotropic detector response. xy-shift correction mitigates these effects, making the measured $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution closer to true $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution which is flat in ϕ because of the rotational symmetry of the collisions.

All these corrections were applied to missing transverse energy in the analyses described in this work. However, the xy-shift correction was not applied in the top mass analysis, since it was not available at the time. As this analysis is not particularly sensitive to $E_{\mathrm{T}}^{\mathrm{miss}}$, the ϕ modulation is not expected to affect the top mass measurement and its resolution.

1.5 Summary

In this chapter, the Large Hadron Collider (LHC) has been introduced to the reader. The CMS experiment including all its subsystems has been described in detail, the overview of the CMS computing model and analysis software have been shown. Reconstruction and identification methods of various analysis objects including particle flow algorithm have been discussed in detail.

2. High level trigger development for Top Physics

The LHC is often referred to as a top quark factory, producing a $t\bar{t}$ pair nearly each second of its nominal operation. While the production rate of ≈ 1 Hz seems manageable in terms of recording the data, it is significantly complicated by background processes with similar signatures occurring at much higher rates.

The trigger is the starting point of any physics event selection process, and therefore is clearly important for any physics analysis. As it was mentioned in Section 1.2.6, the CMS L1 trigger rate is limited to ~ 100 kHz. In order to meet the data recording constraints of approximately 300 $\mathrm{MB\,s^{-1}}$, this rate is further reduced down to ~ 300 Hz, which is done by the HLT system. The total rate budget has to be shared between various physics analysis groups (e.g. Top, Higgs, Exotica, etc.). Corresponding allocations are determined by CMS trigger coordination according to CMS physics goals, and can be a matter of serious debate.

During the LHC operation in 2011 and 2012 under conditions of gradual increase of instantaneous luminosity and pile-up, but very limited rate budget, trigger developers constantly tackled the challenge of finding the best compromise between growing rates and maintaining reasonable signal acceptance. While the simplest approach is tightening the cuts on physical quantities like lepton or jet transverse momenta, it is not favourable since it lowers the number of stored signal events and decreases the phase space which is crucial for new physics searches as well as Standard Model precision measurements. Therefore, development of more efficient algorithms allowing to keep high level of acceptance for signal events, whilst effectively rejecting background events, is the most preferable solution. This can often be achieved by increasing the level of approximation of the online (HLT) object reconstruction, making the algorithms closer to their sophisticated offline counterparts. However, it leads to a higher execution time, which is limited by computing resources available for HLT reconstruction. Hence, the CPU timing is another major constraint faced by the HLT developers.

This chapter covers my contribution to development and verification of High-Level Triggers for top physics with semileptonic signature, where one of the W bosons decays into an electron and a neutrino. Trigger efficiency measurement, CPU timing studies, validation of jet energy corrections and pile-up subtraction applied at the HLT level are discussed in

relevant sections of this chapter.

2.1 Level-1 triggers

The L1 trigger budget of ~ 100 kHz is shared between several L1 trigger "seeds", corresponding to different physics objects being present in the event. For top physics with a single electron in the final state, the following L1 seeds are used: L1_SingleEG_18, L1_SingleEG_20 and L1_SingleEG_22.

The L1 trigger decisions are used as an input to the HLT system, described in the following section.

2.2 High-level triggers for top physics

The High-Level Trigger [17] is the crucial part of CMS event selection process. As it was mentioned in Section 1.2.6, it is based on software algorithms running on the Event Filter Farm, i.e. a large cluster of commercial CPUs. The HLT reconstruction, often referred to as online reconstruction, is implemented in the same software framework (CMSSW) which is used for offline reconstruction, and the algorithms can be very similar. However, the key difference between online and offline reconstruction is the running time. Since the HLT selection has to be performed in real time, it imposes a significant constraint on computing resources, enforcing to compromise on robustness and efficiency of online algorithms. With an exception of small samples for performance monitoring, data rejected by the HLT is lost irrevocably. Therefore, correct and efficient operation of the HLT is of major importance for CMS physics programme.

Modular structure of CMSSW provides high flexibility in use of selection and reconstruction algorithms, allowing their continuous optimisation according to changes in physics needs and data-taking conditions. Various modules account for reconstruction of different physics objects and their matching with L1 objects, filtering, logging, monitoring, etc. All these modules are grouped into so-called trigger "paths", ultimately giving trigger decisions on whether to accept an event or not. A typical example of a trigger path used for top physics is an electron-plus-jets trigger, which requires the presence of isolated electron and at least three energetic jets in the event.

A set of trigger paths is combined in the HLT configuration, referred to as the HLT menu or table [18]. Since the start of data-taking in 2010, the CMS trigger coordination adopted a "trigger train" model, implying the schedule of regular deadlines for updates in the trigger menu. These deadlines are usually imposed according to changes in instantaneous luminosity or other alterations in data-taking conditions. In order for a trigger path to be included in the trigger menu, it has to be implemented in the HLT configurations database, validated by measuring the trigger rate and signal efficiency, tested for CPU timing and finally approved by the trigger studies group.

As it was mentioned before, the top quark pair decay covered in this work has a semileptonic signature, containing an electron, at least four jets and a neutrino in form of $E_{\mathrm{T}}^{\mathrm{miss}}$. Due to the nature of the decay, all these objects are highly energetic and can be triggered on with rather high thresholds on transverse energy. During the start-up year of 2010 when the instantaneous luminosity went up from $\sim 10^{27}$ cm^{-2} s^{-1} to $\sim 10^{31}$ cm^{-2} s^{-1}, top analyses with the signature of interest used the single electron trigger, requiring just one electron with certain isolation and transverse momentum criteria. However, extrapolating the trig-

ger rates on higher luminosities foreseen in 2011, it became obvious that the rate of single electron trigger would quickly become unmanageable unless selection is tightened to limits unacceptable for top physics. Therefore, a solution of electron-plus-jets trigger was exploited.

2.3 Trigger rate and efficiency estimate

2.4 JEC at HLT validation

2.5 CPU timing studies

2.6 Summary

References

[1] L. Evans and P. Bryant. LHC Machine. *JINST*, 3(08):S08001, 2008. 1

[2] J.P. Blewett. 200 GeV intersecting storage accelerators. In *Proceedings of The 8th International Conference on High-Energy Accelerators*, page 501, Geneva, Switzerland, 1971. CERN. 1

[3] The CMS Collaboration. The CMS experiment at the CERN LHC. *JINST*, 3(08):S08004, Aug 2008. 3, 6, 7, 8, 9, 10, 11, 12, 15

[4] The CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30–61, 2012. 3

[5] The CMS Collaboration. *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical Design Report CMS. CERN, Geneva, 2006. 9, 13

[6] The CMS Collaboration. Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays. *Journal of Instrumentation*, 5(03), 2010. 13

[7] Rene Brun and Fons Rademakers. ROOT – An object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(12):81 – 86, 1997. New Computing Techniques in Physics Research V. 17

[8] The CMS Collaboration. Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET. Technical Report CMS-PAS-PFT-09-001, CERN, Geneva, Apr 2009. 19

[9] R. Frühwirth. Application of Kalman filtering to track and vertex fitting. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 262(23):444 – 450, 1987. 21, 28

[10] W Adam, R Frühwirth, A Strandlie, and T Todorov. Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC. *Journal of Physics G: Nuclear and Particle Physics*, 31(9):N9, 2005. 21

[11] S. Baffioni, C. Charlot, F. Ferri, D. Futyan, P. Meridiani, I. Puljak, C. Rovelli, R. Salerno, and Y. Sirois. Electron reconstruction in CMS. *The European Physical Journal C*, 49(4):1099–1116, 2007. 23

[12] The CMS Collaboration. "Cuts in Categories (CiC) Electron Identification". https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideCategoryBasedElectronID, 2011. [Online; accessed 04-September-2013]. 23, 24

[13] The CMS collaboration. Performance of CMS muon reconstruction in pp collision events at s = 7 TeV. *Journal of Instrumentation*, 7(10):P10002, 2012. 28

[14] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063, 2008. 29

[15] Rick D. Field. The underlying event in hard scattering processes. *eConf*, C010630:P501, 2001. 31

[16] The CMS collaboration. Identification of b-quark jets with the CMS experiment. *Journal of Instrumentation*, 8(04):P04013, 2013. 32, 33

[17] The CMS collaboration. The CMS High Level Trigger System. In *2007 15th IEEE-NPSS Real-Time Conference*. Institute of Electrical and Electronics Engineers, Apr 2007. 37

[18] The CMS collaboration. Commissioning of the CMS High Level Trigger. *Journal of Instrumentation*, 4(10):P10005, 2009. 37