Alma Mater Studiorum - Università di Bologna

Facoltà di Scienze Matematiche Fisiche e Naturali
Dottorato di Ricerca in Fisica, XX Ciclo

PhD Thesis

Sensitivity of the Top quark mass measurement with the CMS experiment at LHC using t-tbar multijet simulated events

Dott. **Giuseppe Codispoti**

Advisor:
Chiar.mo Prof.
Andrea Castro

Co-advisor:
Chiar.mo Prof.
Paolo Capiluppi

PhD Coordinator:
Chiar.mo Prof.
Fabio Ortolani

FIS/01

Bologna, Italy, March 2008

Contents

Introduction

The CERN Large Hadron Collider will start soon its operations. Its design parameters for proton collisions, a centre-of-mass energy of 14 TeV and a luminosity of $\mathcal{L} = 10^{34}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, will enable to study fundamental constituents of matter and their interactions up to the TeV energy scale. Its goal is to check the the consistency Standard Model of particle physics and to explore alternative theories such as supersimmetry, looking for the way toward a unified theory.

The need to analyze very large statistics in pursuit of rare signal, together with the high interaction rate, requires fine granularity detectors, leading to several challenges for the realization and operation of the detector and the final analysis. Although the strong data reduction performed by the online trigger, a large amount of data will be produced, reconstructed to build analysis object and spread over the world, thus requiring an efficient distributed computing system.

The work here described comes after a strong contribution on the realization of the CMS computing system, which can be seen as a relevant part of the experiment itself. A physics analysis completes this road from Monte Carlo production and analysis tools to the final physics study which is the actual goals of the experiment.

The topic of physics work of this thesis is the study of $\mathrm{t\bar{t}}$ events in CMS. The top quark enters in many open fields in particle physics: its high mass is for instance very close to the scale of electroweak symmetry breaking and enters in loop correction for many physics observables of the Standard Model, such as the Higgs boson mass. An accurate knowledge of its properties gives thus a powerful instrument for the indirect estimate of many electroweak parameters as well as for the investigation of the still obscure mechanism of the symmetry breaking. Furthermore the multi-jet final states give an instrument for the comprehension of the detector itself, since LHC will produce about 8 million of $\mathrm{t\bar{t}}$ events per year yet at low luminosity.

The top quark decays almost exclusively via $t \rightarrow Wb$ and the final states depend then on the decay modes of the W boson: approximately 46% of $\mathrm{t\bar{t}}$ events, therefore, are fully hadronic, while about 44% and 10% of the events decay respectively semileptonically and dileptonically.

This work describes the analysis of $t\bar{t}$ decay in the fully hadronic channel, characterized by nominal six jets production ($t\bar{t} \rightarrow WWb\bar{b} \rightarrow qqqqb\bar{b}$) thus leading to a well defined topology. The event kinematics can be thus fully reconstructed an this is, together with the largest branching ratio, the main advantage of this channel. However, the large background from QCD multi-jet production makes the isolation of the signal rather challenging, while the initial and final state gluon radiation alter the nominal six jets topology.

Although the channel look unfavorable, using a suitable multi-jet trigger and an optimized kinematical selection, a good signal to background ratio can be achieved. In addition, the identification of b-jets applied both at the trigger level and after the kinematical selection helps to isolate clean top quark samples. Using then multi-jet samples compatible with the nominal six-jets topology it is possible to perform precise measurement, such as the reconstruction of the top quark mass.

The thesis is organized as follows.

Chapter 1 is dedicated to the description of the LHC machine and the CMS experiment, focusing on the trigger for the online selection.

Chapter 2 describes the offline part of the CMS experiment, i.e. the computing system, focusing on the tools for Monte Carlo events production and analysis, given the contribution on their realization and the strong usage for the analysis described.

Chapter 3 gives a short description of the Standard Model, its success and its open questions and how the top quark physics contributes to the final understanding of the particle physics.

Chapter 4 describes the tools provided by the CMS software for the analysis, e.g. the Monte Carlo generators and the High Level Objects as well as the b-tagging algorithm used in the analysis.

In Chapter 5 a multi-jet trigger selection is described evaluated on the basis of the CMS High Level Trigger requirements, relaxed by the consideration that some of the selection we made offline can be performed already online.

The multi-jet trigger is integrated in Chapter 6 with an optimized kinematical selection, based on the best statistical significance achievable; again part of the selection can be used also online to reduce the High Level Trigger output rate.

Chapter 7 describes the measurement of the top mass, using a comparison of event samples with the expected signal and background probability density functions through a likelihood maximization.

Chapter 1

The CMS detector at LHC

The Standard Model (SM) of particle physics is considered to be an effective theory up to the TeV energy scale. Even if it has so far been tested to good precision over many collider experiment, the nature of the electroweak symmetry breaking and the Higgs mechanism presumed to be responsible for it, still need to be experimentally proved.

The CERN **L**arge **H**adron **C**ollider (LHC) [1] main motivation is the study of this mechanism as well as the check of the overall SM consistency. More in general LHC experiments will explore the physics at the TeV energy scale, also exploring alternatives theories such as technicolour and supersimmetry and looking for the way toward a unified theory.

1.1 The Large Hadron Collider

The LHC [2] accelerating machine will provide both proton-proton (pp) and heavy-ion collisions. The design parameters have been chosen in order to study physics at the TeV energy scale.

Two key parameters characterize the overall machine performance and discovery power: the center of mass energy and the instantaneous *luminosity.*

The luminosity is a beam related variable which determines the interaction rate. It is defined as follows:

$$\mathcal{L} = \frac{\gamma f k_B N_p^2}{4\pi \epsilon_n \beta^*} F$$

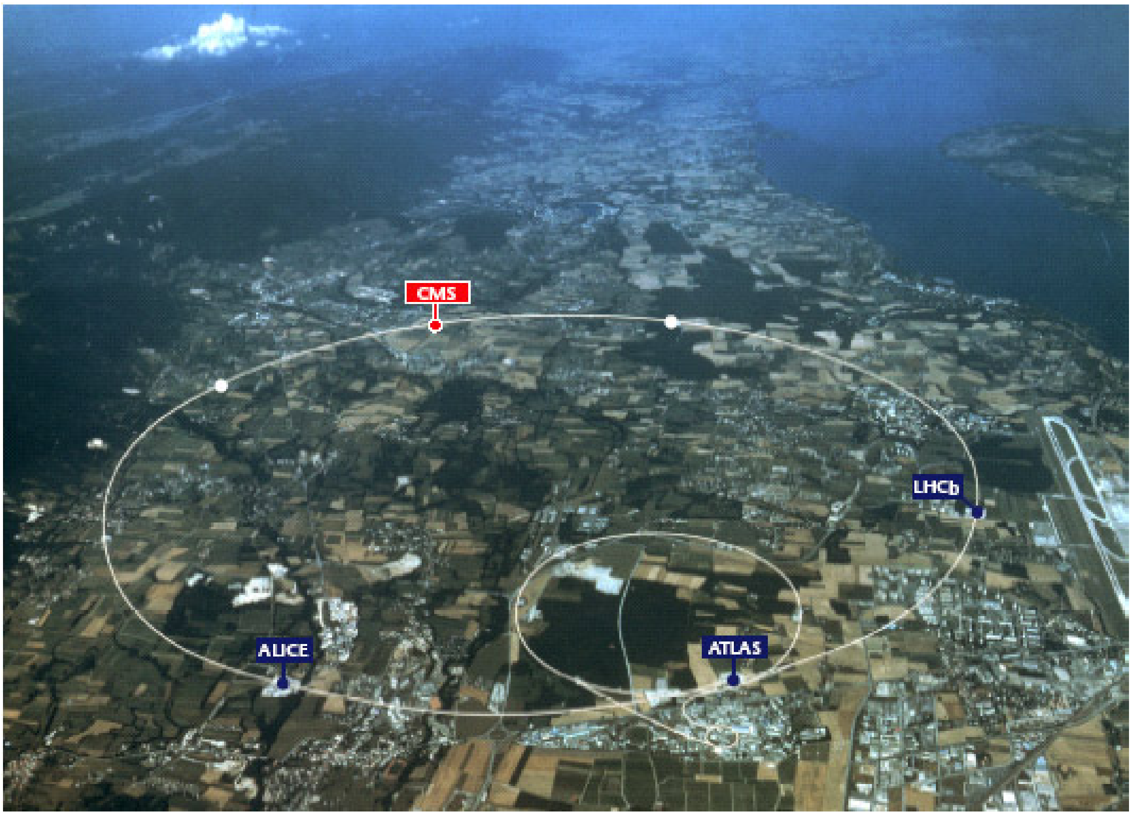

Figure 1.1: LHC location in the Geneva region.

where γ is the Lorentz factor,f is the revolution frequency, k_B is the number of bunches, N_p is the number of protons per bunch, η_n is the normalized transverse emittance, β^* is the betatron function at the interaction point and F is a reduction factor due to the crossing angle. The interaction rate R is indeed a function of the luminosity and the cross section of the event:

$$R = \sigma \times \mathcal{L}$$

The design luminosity for LHC is $\mathcal{L} = 10^{34}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, hundred times the one reached at the Tevatron at Fermilab. To give an example, the total inelastic cross section for the pp interaction is 80 mb[1], which, at the design luminosity brings to an expected events rate of 10^9 events/s.

[1] milli barn: barn (b) is a cross section measure unit used in particle physics, where $1\ b = 10^{-24}\,\mathrm{cm}^2$, corresponding to the typical cross section of a nuclear process.

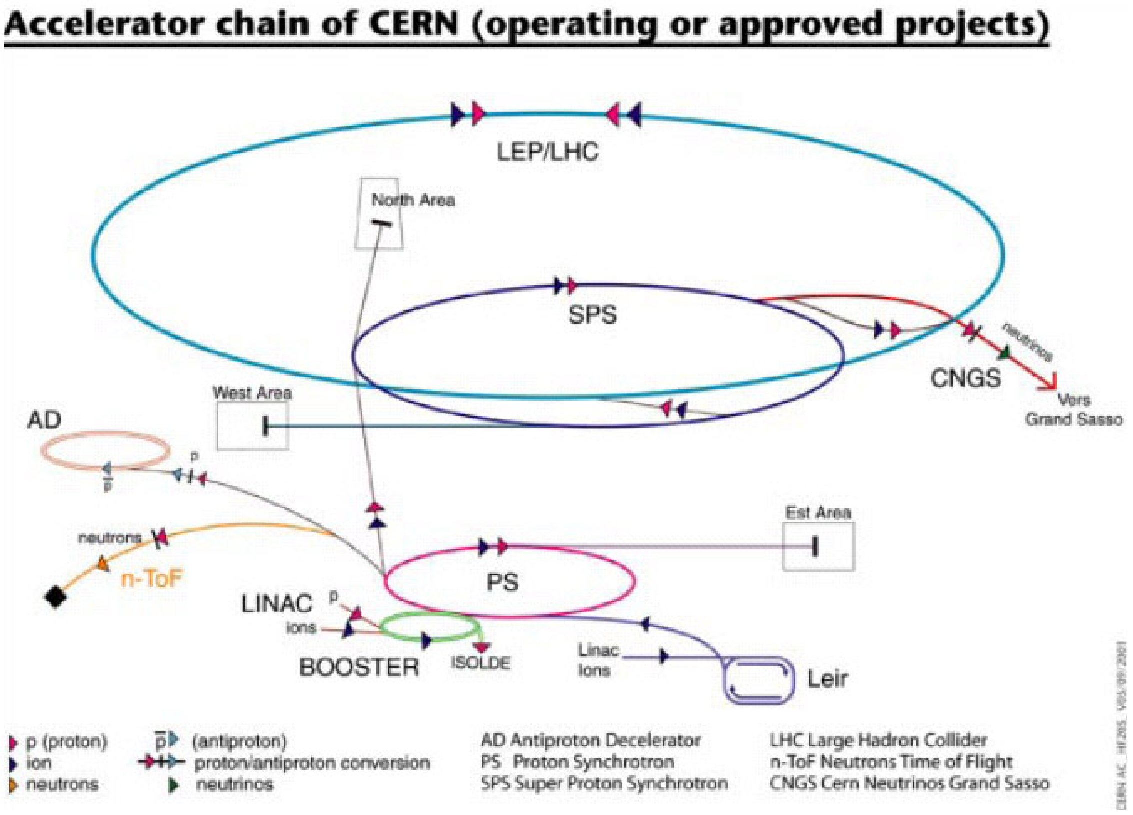
Accelerator chain of CERN (operating or approved projects)
LEP/LHC
North Area
SPS
CNGS
neutrinos
Vers
Grand Sasso
West Area
AD
p
p̄
neutrons
n-ToF
LINAC
p
ions
BOOSTER
ISOLDE
PS
Est Area
Linac
Ions
Leir
p (proton)
ion
neutrons
p̄ (antiproton)
proton/antiproton conversion
neutrinos
AD Antiproton Decelerator
PS Proton Synchrotron
SPS Super Proton Synchrotron
LHC Large Hadron Collider
n-ToF Neutrons Time of Flight
CNGS Cern Neutrinos Grand Sasso
CERN AC _HF205_ V05/09/2001

Figure 1.2: LHC injection system: protons are produced by ionization of hydrogen and injected with an energy of 750 keV in the linear accelerator (LINAC2) through a radiofrequence quadrupole where they reach the energy of 50 MeV, then in the BOOSTER, where they are rearranged in packets of approximatively 10^{11} protons of 1.4 GeV, then in the PS up to 25 GeV and finally in the SPS which will inject them at 450 GeV in the LHC.

The accelerator is designed to produce a 14 TeV center of mass energy, seven times the one reached in the most powerful collider realized before, the Tevatron. This value is limited by the geometrical size of the collider by:

$$p(\text{TeV}) = \frac{q}{e}0.3B(Tesla)R(km)$$

where q is the particle electric charge, B the magnetic field, $\text{R} \simeq 4.3\,\text{km}$ the accelerator radius. To reach a 7 TeV proton beam the magnetic field should be 5.4 T. This will be obtained practically through radio frequency cavities which will accelerate the particle bunch and bending stations of superconducting magnets of

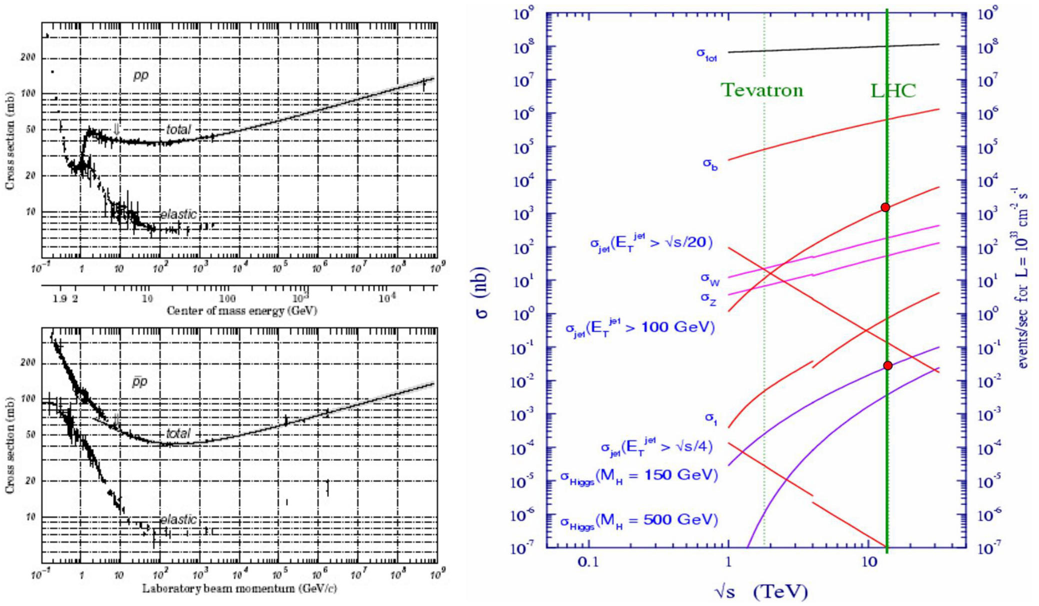

Figure 1.3: The pp cross section compared with the $p\bar{p}$ cross section: while at lower center of mass energy it is lower for pp interactions, at higher energies they are pretty equal, since the partons interactions are dominant. The second plot shows the cross section for some relevant process for both Tevatron and LHC at low luminosity.

8.33 T maintained at 2.1 K by superfluid helium.

From the above formula, it follows that the proton beams require two separate beam-pipes, unlike the $p\bar{p}$ acceleration. Anyway the first is preferred because at the LHC center of mass energy the cross section for pp is comparable with the $p\bar{p}$ one (figure 1.2), while the p production is faster and more efficient so that a high luminosity beam can be easily reached and maintained for longer times.

Moreover, protons are used instead of leptons to reduce the radiation loss, since this is proportional to the fourth inverse power of the mass. This allows accelerators of shorter radius and with smaller number of revolution to reach the design energy beam and in the LHC case, reusing the same LEP collider tunnel. The implication of this choice is that a particle pair production will not have the availability of the whole center of mass energy, but instead only the fraction related to the actually colliding constituents. Moreover, many hadronic final states from soft core interaction will overlap interesting hard-core events, leading to a hard analysis environment, but also to a wider energy range study.

Other relevant beam parameters are summarized in table 1.1 while the whole acceleration system is described in figure 1.2 and table 1.2.

Parameter		pp	Pb-Pb	
Energy per nucleon	E	7	2.76	TeV
Design luminosity	$\mathcal{L}$	10^{34}	10^{27}	$\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$
Bunch separation		25	100	ns
N. of bunches	b_B	2808	592	
N. particles per bunch	Np	1.15×10^{11}	7.0×10^{7}	
β-value at IP	β^*	0.55	0.5	m
RMS beam radius at IP	σ^*	16.7	15.9	μm
Luminosity lifetime	τ_L	15	6	h
Number of collisions/crossing	n_c	20	-	
Total n. of particles		3.1×10^{14}		
Bunch length (σ_z)		53		mm
Beam current		560		mA

Table 1.1: Main beam parameters of LHC.

Parameter	Value
Radius	$\sim 4.3km$
Dipole field	8.3 T
N. of magnetic dipoles	1232
N. of quadrupoles	520
N. of sextupoles	2×1232
N. of octupoles	1232

Table 1.2: Additional parameters of LHC.

Four detectors are installed in as many interaction points: ATLAS (**A** **T**oroidal **L**HC **A**pparatu**S**) and CMS (**C**ompact **M**uon **S**olenoid) are general purpose detectors, ALICE (**A** **L**arge **I**on **C**ollider **E**xperiment) will focus on the heavy ions physics and on the study of the quark-gluon plasma, and LHCb (**LHC** **b**eauty experiment) will study the CP violation in b-physics. Figure 1.2 shows where these detectors will be placed in the LHC ring.

The collider will operate in 2008 with a bunch spacing of 75 ns for the commissioning of LHC and the experiments, hopefully moving soon to 25 ns. Until the beam dump and collimation system are fully staged, the current will be limited to half the nominal value, so the luminosity will be limited to $\mathcal{L} = 2 \times 10^{33}$ cm^{-2} s^{-1} until the 2010 run. In the first full year of running, the integrated luminosity reachable should be around 5 fb^{-1}, collecting enough statistic for SM measurements and the Higgs discovery up to mass of 135 GeV/c^2.

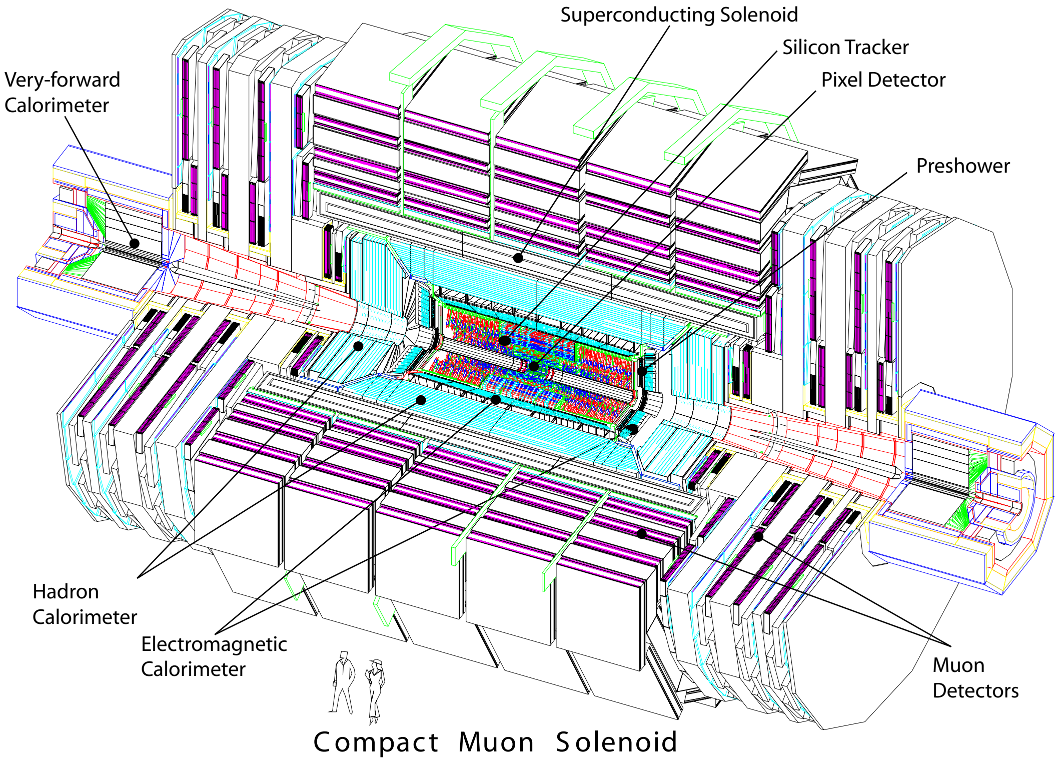

Figure 1.4: A CMS exploded view

1.2 The Compact Muon Solenoid detector

The **C**ompact **M**uon **S**olenoid (CMS) [3], [4], [5] is a general purpose experiment, mainly designed for pp interaction, but which will operate also in heavy ions mode. The experimental area is located 100 m underground near the French village of Cessy. The global layout is of a barrel, built of five slices, whose extremities are closed with four endcap wheels, for a total length of 21.6 m, a 14.6 m diameter and 12500 tons of weight. A high field superconducting solenoid surrounds a full silicon based inner tracking system, a homogeneous scintillating crystal based electromagnetic calorimeter and a hadron calorimeter, while the return yoke are instrumented with muon detectors and forward sampling calorimeters are used to extend the CMS spatial coverage and hermeticity.

The detector requirements can be summarized as follows:

- Good muon identification and momentum resolution over a wide range of momenta and angles , good dimuon mass resolution ($\approx 1\%$ at 100 GeV/c^2) and the ability to determine unambiguously the charge of muons with p $<$ 1 TeVc.
- Good charged particle momentum resolution and reconstruction efficiency in the inner tracker, with particular attention to an efficient triggering and offline tagging of τ and b jets, requiring a pixel detector near to the interaction point.
- Good electromagnetic energy resolution, diphoton and dilepton mass resolution ($\approx 1\%$ at 100 GeV/c^2, wide geometrical coverage, π^0 rejection and efficient photon and lepton isolation at high luminosities.
- Good missing transverse energy and dijets mass resolution, requiring hadron calorimeters with large hermetic geometric coverage and fine lateral segmentation.

The coordinate system is so defined: the x-axis points radially inward toward the center of LHC, the y-axis points vertically upward the z-axis is along the beam axis at the interaction point (the direction is toward the Jura mountains). The polar angle θ is measured from the z-axis ($0 \leq \theta \leq \pi$) while the azimuthal angle ϕ is measured from the x-axis in the x-y plane ($0 \leq \phi \leq 2\pi$). Instead of the angle θ is usually preferred the pseudorapidity

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

because, loosely speaking, particle production is constant as a function of rapidity: indeed the angular particle production decreases while we move far form the z-axis while a fixed η range increases its angular extension. The pseudorapidity is in fact the ultra-relativistic limit for the rapidity:

$$y = \tanh^{-1} \beta = \frac{1}{2}\ \ln\ \frac{1+\beta}{1-\beta} = \frac{1}{2}\ \ln\ \frac{E+p_z}{E-p_z}$$

and depends only on the polar angle of its trajectory, but not on the energy of the particle.

Since the total boost along the z-axis is null, topic values are related to the transverse x-y plane: the transverse momentum p_{T}, transverse energy $\mathrm{E_T}$ and the missing energy from the measured total transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$.

As discussed above, the expected event rate is about 10^9 events/s. An efficient online selection event, "trigger", must reduce the rate to 10 events/s; readout and selection are therefore complicated by the short time separation between two bunch crossings. Indeed, at the design luminosity, 20 hard core inelastic scatterings will arise every 25 ns, producing around 1000 charged particles which will be superimposed to an interesting event making hard its study. To avoid this, every detector must have fine granularity, good time response and low occupancy, which will result in millions of electronics channels with very good synchronization. The large flux of particles, estimated to be $1 \div 2\,\mathrm{kGy/year}$[2] at the design luminosity, requires radiation-hard detectors and front-end electronics.

1.2.1 The Magnet System

The detector design and layout, and thus its name, are driven by the choice of the magnetic field for the measurement of the momentum of muons. The main requirement is the identification of muon final states from decays and the unambiguous determination of their sign up to 1 TeV/c which brings a moment resolution of $\Delta\mathrm{p/p} \approx 10\%$ at $\mathrm{p} = 1\,\mathrm{TeV/c}$. CMS chose a modestly-sized solenoid with a high field, where the length/radius ratio, allows a good momentum resolution also in the forward region, providing a particle bending only on the transverse plane. The bore of the solenoid accommodate a silicon inner tracker, the electromagnetic and the hadronic calorimeters, while the return field saturates four iron yokes, both in the barrel and in the endcap region, where muon chambers are integrated. The solenoid is made of four layers, built of Rutherford-type cable coextruded with high-purity aluminium, which act as thermal stabilizer, insulated by epoxy impregnation. The main working parameters are summarized in table 1.3.

1.2.2 The Tracker

The inner tracking system surrounds the interaction points, with a length of 5.8 m and a diameter of 2.5 m.

[2] kilo Gray: Gray Gy is the SI unit of absorbed radiation dose, $1Gy$ is the absorption of one joule of radiation energy by one kilogram of matter : $\mathrm{Gy} = 1\,\frac{\mathrm{J}}{\mathrm{kg}} = 1\,\mathrm{m}^2 \cdot \mathrm{s}^{-2}$.

Parameter	Value
Field	4 T
Inner Bore	6.3 m
Length	12.5 m
N. of Turns	2168
Operation Temperature	4.5 K
Current	19.5 kA
Stored Energy	2.6 GJ
Hoop stress	64 atm

Table 1.3: Parameters of the CMS superconducting solenoid.

It is designed to provide precise and efficient measurement of the trajectories of charged particles with transverse momentum above 1 GeV/c as well as precise reconstruction of secondary vertices and impact parameters for efficient identification of heavy flavours produced in many interesting physics channels. Together with the electromagnetic calorimeter and the muon system it has an important role in electron and muon identification respectively. It is heavily used in the high level trigger.

At LHC design luminosity, there will be about 1000 charged particle from more than 20 overlapping pp interaction every 25 ns. Therefore high granularity and fast response, as well as radiation hard technologies are required: the resulting high power density requires also an efficient cooling system. On the other hand the amount of material has to be kept to the minimum to reduce multiple scattering, Bremmsstralhung, photon conversion and nuclear interactions.

Following those requirements, the inner tracker was designed as follows, to fit with the estimated flux of charged particles:

- A pixel detector is placed close to the IP, at radii between 4.4 cm and 10.2 cm, where the particle flux is the highest. The pixel size is$\approx$ $100 \times 150\,\mu\text{m}^2$ giving a 10^{-4} occupancy per pixel per bunch crossing, while the almost square pixel shape allows achieving optimal vertex position resolution which results in about $10\,\mu$m from the r-ϕ measurement and about $20\,\mu$m for the z measurement. It is made of three barrel layers, and two forward layers for each endcap with a radius covering the range $-2.5 < |\eta| < 2.5$.

- An inner silicon microstrip detector is enabled to work in the intermediate region, 20 cm < r < 55 cm. The cell size is $10\,\text{cm} \times 80 \div 120\,\mu$m giving an occupancy of $\approx 2 \div 3\%$ per bunch crossing. This inner layer is made of four barrels and three disks per endcap covering $|z| < 65$ cm. The single point resolution varies from $23 \div 34\,\mu$m in r-ϕ and $230\,\mu$m in z.

- An outer silicon microstrip detector with a coarse cell size is placed in the outermost region $55\,\text{cm} < \text{r} < 110\,\text{cm}$, keeping the occupancy to $\approx 1\%$ per bunch crossing. The cell size is $25\,\text{cm} \times 120 \div 180\,\mu$m. It is made of three barrel layers and 9 forward layers for each endcap covering $|z| < 65$ cm.The single point resolution varies from $35 \div 52\mu$m in r-ϕ and $530\,\mu$m in z.

In the heavy-ion operations, the occupancy will increase to 1% in the pixel detector and less to 20% in the silicon microstrip detector, still allowing track reconstruction.

1.2.3 The Electromagnetic Calorimeter

The Electromagnetic Calorimeter (ECAL) is a hermetic homogeneous calorimeter made of lead tungstate ($PbWO_4$) scintillating crystals, 61200 in the barrel and 7234 in the endcaps.

The crystals have short radiation length ($\text{X}_0 = 0.89$ cm) and Moliere lengths (2.2 cm), fast time response (80% of the light is emitted within 25 ns) and they are radiation hard up to 10 Mrad. On the other hand, the low emitted light output, about $30\,\gamma$/MeV, leads to the usage of photodetector with intrinsic gain: silicon avalanche photodiodes (APD) in the barrel and vacuum phototriodes (VPT) in the endcaps. The APD response depends on the temperature, so thermal stability up to 0.1°C is required.

The barrel section has an inner radius of 129 cm and is structured in 36 supermodules covering each half the barrel length, corresponding to $0 < |\eta| < 1.479$. The crystals are quasi-projective, in the sense that their axes are tilted of 3° with respect to the nominal vertex position direction.

The endcaps are positioned at 314 cm from the vertex and cover a range of $1.479 < |\eta| < 3.0$. They are structured as 2 "Dees" consisting of semi-circular aluminium plates from which are cantilevered structures of 5×5 crystals. Like in the barrel, crystals point to the vertex but they are arranged in an x-y grid instead of η-ϕ grid.

A preshower device is placed over much of the pseudorapidity range whose active elements are 2 planes of silicon strips detectors with a pitch of 1.9 mm placed behind disks of lead absorber at depths of $2\,\text{X}_0$ and $3\,\text{X}_0$.

The energy resolution has been measured in the test beam, fitting a Gaussian function to the reconstructed energy distribution and parameterized as a function of the energy as follow:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2 \tag{1.1}$$

where S is the stochastic term, due to fluctuation in lateral shower containment, photostatistics and fluctuation in preshower absorber with respect to the measured energy, N the noise (electronics, digitization, pile-up), C a constant term due to non-uniformity of the longitudinal light collection, intercalibration errors and leakage of the back of the crystal. Typical values of those parameters, measured in the test beam, are: S = 2.8%, N = 0.12%, C = 0.30%.

1.2.4 The Hadron Calorimeter

The Hadron Calorimeter (HCAL) is placed between the ECAL and the magnet coil and so its design is strongly influenced by the magnet parameters. An important requirement is to minimize the non-Gaussian tails in the energy resolution and to provide a good containment and hermeticity for the $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement. Hence the HCAL design maximizes materials inside the magnet coil in terms of interaction lengths and is complemented by an additional layer of scintillators (Hadron Outer detector, HO) lining the outside of the coil.

The chosen absorber material is brass, for its short interaction length, ease to machine and being non-magnetic. The absorber structure is made of two brass plates bolted together to leave the space for the scintillator plates.

The active medium is made of plastic scintillator tiles read out with embedded wavelength-shifting fibres (WLS) which well fit with the relatively small space left.

Photodetection is made through multi-channel hybrid photodiodes. The overall assembly allows the HCAL to be built without uninstrumented or dead areas in ϕ. The gap through the endcap and the barrel is inclined at 35° and points away from the center of the detector.

The barrel part consists of 32 towers covering the region $-1.4 < |\eta| < 1.4$ resulting in 2304 towers with segmentation $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$. It is assembled in two half parts and read out as a single longitudinal sampling.

The Hadron Outer detector has the same coverage and contains scintillators with a thickness of 10 mm. The tiles are grouped in $30^\circ -$ sectors, matching the ϕ segmentation of the DT chambers. They serve as "tail catcher" for the hadron shower penetrating the rear layer of the calorimeter, increasing the effective thickness of the barrel over to 10 interaction lengths, reducing the tails in energy resolution function and improving the $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution.

Each Hadron Endcap consist of 14 η towers covering $1.3 < |\eta| < 3.0$ for a total of 2304 towers. The 5 outermost towers have a segmentation of 0.087 units in η and 5° in ϕ segmentation, while the 8 innermost have a 10° segmentation in ϕ while the segmentation in η varies from 0.09 to 0.035 moving to larger η.

The Hadron Forward calorimeter covers the region $3.0 < |\eta| < 5.0$. It is made of steel/quartz fibres. This region samples preferentially the neutral components of the hadronic shower, therefore the design is such to lead to narrower and shorter showers and is ideal for congested environment. The front face is located at 11.2 m from the IP. The absorber is 1.65 m thick, made of steel plates diffusion welded, with 1 mm^2 grooves where the fibres are inserted.

The granularity of the sampling of the three parts has been chosen such that the jet energy resolution as function of $\mathrm{E_T}$ is similar. The resolution of the $E_{\mathrm{T}}^{\mathrm{miss}}$ in QCD di-jets events with pile-up is given by $\sigma(\mathrm{E_T^{miss}}) \approx \sqrt{\Sigma \mathrm{E_T}}(\mathrm{GeV})$ without energy clustering corrections while the average $E_{\mathrm{T}}^{\mathrm{miss}}$ is given by $< \mathrm{E_T^{miss}} > \approx 1.25\sqrt{\Sigma \mathrm{E_T}}(\mathrm{GeV})$

1.2.5 CMS Muon Detectors

The muon measurement system characterize the whole CMS experiment.

Centrally produced muons are measured three times: in the inner tracker, after the coil and in the flux return.

The momentum of the muons using only the muon system is determinated by the muon bending angle at the exit of the magnet coil, taking the interaction point as the origin, which will be known to $\approx 220\,\mu$m. The resolution is dominated by the multiple scattering before the first station up to $p_{\mathrm{T}} \approx 200\,\mathrm{GeV/c}$, then by the chamber resolution. At lower p_{T} the inner tracker gives the best resolution, while at higher p_{T} the combination of the two systems improves the overall resolution.

The muon system is made of three types of gaseous detectors, whose coverage and technologies are driven by the large surface and the radiation environment.

The barrel region, $|\eta| < 1.2$, is characterized by a small neutron background, a low muon rate and small residual magnetic field and is covered by four layers of drift tube chambers. The endcap region, with opposite characteristics, is instead covered by four disks of three rings of cathode strip chambers up to $|\eta| < 2.4$. The same regions are covered by resistive plate chambers which provide a fast response, allowing a good bunch crossing identification, but a coarse spatial resolution.

The three detector systems operate within the first trigger level with independent and complementary sets of information.

The Drift Tube Chambers

The **D**rift **T**ube (DT) Chambers are gaseous ionization detectors ensuring a linear relationship between time and distance through a constant drift time.

A total of 250 DT chambers covers the barrel region with 4 layers positioned inside the magnet return yoke at radii of approximately 4.0, 4.9, 5.9 and 7.0 m from the beam axis.

The 5 wheels are divided into 12 sectors, each one covering 30° of azimuthal angle. Their position is such that a high p_{T} muon can hit at least three stations.

The three inner layers host 12 chambers, each one consisting of 12 planes of aluminium drift tubes: four r-ϕ planes sandwiching 4 z-planes. A fourth layer does not measure along z and is made just of two chambers at the top and two at the bottom for a total of 14 chambers per wheel.

A high-p_{T} muon crosses 4 DT chambers producing up to 44 points for the track reconstruction.

The maximum drift length is 2.0 cm and the single point resolution $\approx 200\,\mu$m. Each station is designed to give a muon vector with a precision better than $100\,\mu$m in position and 1 mrad in direction.

The Cathode Strip Chambers

Cathode **S**trip **C**hambers (CSC) are gaseous ionization detector working in avalanche mode. When a charged particle traverse a CSC, the ionization produces an avalanche and thus a charge on the anode wire and an image charge on a group of cathode strips. The signal over wires is fast and usable by the trigger, but leads to a coarse position resolution, while a center of gravity measurements over the cathode strips will lead to a more precise position measurement. Thus CSCs measure up to two sets of spatial coordinates

A total of 486 CSC complete the muon system in the endcap region. The CSC have trapezoidal shape and consists of 6 gas gaps having a plane of radial cathode strips and a plane of anode wires almost perpendicular to them. All CSCs, except those in the third ring of the first endcap, are overlapped in ϕ to avoid gaps in the muon acceptance. The innermost ring of each disk, except for the first one, hosts 18 chambers, all other rings host 36 chambers. The typical spatial resolution is about

200 μm, due to the strip measurement, while the angular resolution in ϕ is about 10 mrad.

The Resistive Plate Chambers

Resistive **P**late **C**hambers (RPC) work also in avalanche mode, providing a fast response, allowing a good bunch crossing identification, but a coarse spatial resolution. In the barrel region, the two first DT layers are sandwiched between two RPC, while the other two have only one RPCs layer placed in the innermost side. In the endcap region, 36 chambers are mounted in each endcap, used also to resolve ambiguities in the CSC in the first endcap station.

1.2.6 The Trigger System

The LHC provides proton-proton and heavy-ion collisions at high interaction rates. Focusing on the pp collisions, the beam crossing interval is 25 ns, corresponding to a crossing frequency of 40 MHz, which at the nominal design luminosity leads to approximately 20 simultaneous pp collisions for an interaction rate of about $\approx 10^9$ interactions/s.

Since it is impossible to store and process the large amount of data associated with the resulting high number of events, the trigger system must achieve a drastic rate reduction, a factor of about 10^6. The rate is reduced in two steps called Level-1 Trigger (L1T) and High-Level Trigger (HLT), respectively.

The Level-1 Trigger has regional components, mainly calorimeters and muon triggers which build candidate objects, and global components which use them to take the decision to accept or reject an event.

The High-Level Trigger is instead software, executed in an online farm.

The Level-1 Trigger

Level-1 [6] triggers are made of custom hardware processors. They involve calorimetry and muon systems, as well as some correlation of information between these systems. The decision is based on the presence of “trigger primitive” objects, such as photons, electrons, muons and jets above certain $\mathrm{E_T}$ or p_T threshold and global sums of $\mathrm{E_T}$ and $E_\mathrm{T}^\mathrm{miss}$. The L1T uses coarsely segmented data from the calorimeters and the muon system, while holding the high-resolution data in pipelined memories in the front-end electronics.

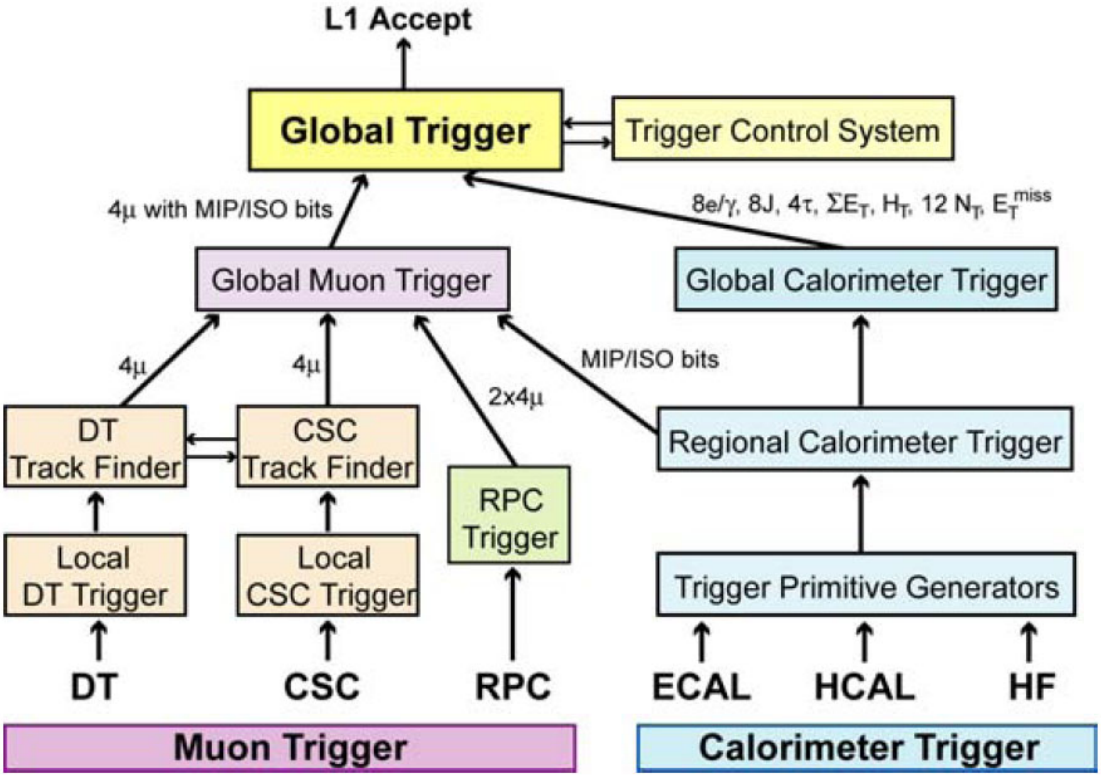

Figure 1.5: The architecture of the L1T

The design Level-1 rate value is 100 kHz, set by the average time required to transfer full detector information through the readout system, which translates in practice to a calculated maximal output rate of 30 kHz, assuming an approximate safety factor of three for simulation uncertainties as well as beam and detector conditions not included in the simulation programs. At the startup it will be reduced to 50 kHz which leads to an estimated rate of 16 kHz .

For reasons of flexibility the L1T hardware is implemented in Field Programmable Gate Arrays (FPGA) technology where possible, but custom Application Specific Integrated Circuits (ASICs), semi-custom and gate-array ASICs and programmable memory lookup tables (LUT) are also widely used where speed, density and radiation resistance requirements are important.

The size of the detector imposes a transit time for signal from the front-end electronics to reach the cavern allocating the Level-1 trigger logic and go back to the front-end electronics. The total time for CMS is 3.2 μs, where the time allocated by the Level-1 calculation is less than $1\,\mu$s. During this time, the high-resolution data are held in pipelined memories.

The L1T has local, regional and global components. At the bottom end, the Local Triggers, also called Trigger Primitive Generators (TPG), are based on energy deposits in calorimeter trigger towers and track segments or hit patterns in muon chambers, respectively.

Regional Triggers combine their information and use pattern logic to determine ranked and sorted trigger objects such as electron or muon candidates in limited spatial regions. The rank is determined as a function of energy or momentum and quality, which reflects the level of confidence attributed to the L1 parameter measurements, based on detailed knowledge of the detectors and trigger electronics and on the amount of information available.

The Global Calorimeter and Global Muon Triggers determine the highest-rank calorimeter and muon objects across the entire experiment and transfer them to the Global Trigger, the top entity of the Level-1 hierarchy. The latter takes the decision to reject an event or to accept it for further evaluation by the HLT. The decision is based on algorithm calculations and on the readiness of the sub-detectors and the DAQ, which is determined by the Trigger Control System (TCS).

The Level-1 Accept (L1A) decision is communicated to the sub-detectors through the Timing, Trigger and Control (TTC) system. The architecture of the L1T is described in Fig. 1.5. The L1T has to analyze every bunch crossing.

Calorimeters Trigger

For triggering purposes the calorimeters are subdivided in trigger towers. Level-1 trigger primitives are calculated in the readout boards. The Trigger Primitive Generators (TPG) make up the first or local step of the Calorimeter Trigger pipeline, by summing the transverse energies measured in ECAL crystals or HCAL readout towers to obtain the trigger tower $\mathrm{E_T}$ and attach the correct bunch crossing number. In the region up to $|\eta| = 1.74$ each trigger tower has an $(\eta,\phi\ )$-coverage of 0.087×0.087. Beyond that boundary the towers are larger.

The TPGs are transmitted through high-speed serial links to the Regional Calorimeter Trigger, which determines regional candidate electrons/photons, transverse energy sums, both electromagnetic and hadronic in each tower, τ-veto bits and information relevant for muons in the form of minimum-ionizing particle (MIP) and isolation (ISO) bits. A trigger region consists of 4×4 trigger towers except in HF where a region is one trigger tower.

The Global Calorimeter Trigger determines jets, the total transverse energy, the missing transverse energy, jet counts, and H_T (the scalar transverse energy sum of all jets above a programmable threshold). It also provides the highest-rank isolated and non-isolated e/γ candidates across the entire detector..

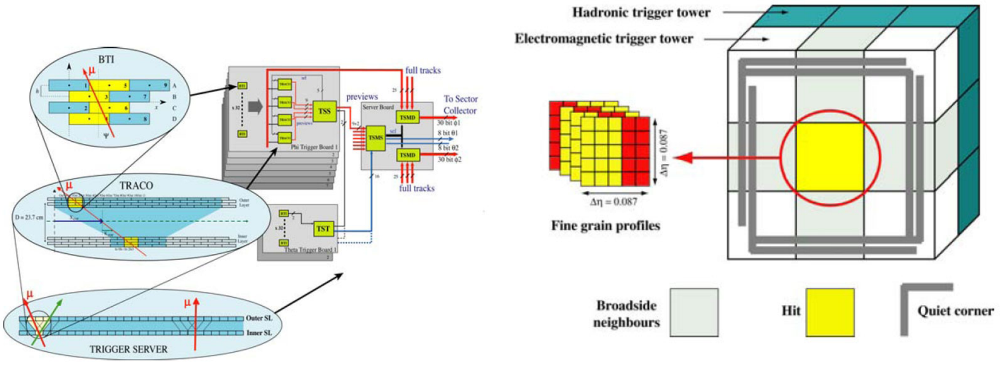

Figure 1.6: A schematic view of the DT muon trigger and e/γ algorithm in the calorimeters trigger.

Muon Trigger

The purpose of the Level-1 muon trigger of the CMS experiment is to identify muons, assign them to a particular beam crossing, and determine their transverse momenta and location.

The Level-1 muon trigger is organized into subsystems representing the 3 different muon detectors: the DT trigger in the barrel, the CSC trigger in the endcap and the RPC trigger covering both barrel and endcap. The Level-1 muon trigger also has the Global Muon Trigger (GMT) that combines the trigger information from the DT, CSC, and RPC muon subsystems, as well as from the calorimeter subsystem, and sends it to the Level-1 Global Trigger.

Each of the Level-1 muon trigger subsystems has its own trigger logic. The DT and CSC electronics first process the information from each chamber locally, delivering a trigger primitive vector (position, direction, bunch crossing, and quality) per muon per station. Trigger primitives from different stations are collected by the Track Finders (TF), which build them into tracks and assign a transverse momentum value to each. Therefore, the TF plays the role of a regional trigger. The DT and CSC Track Finders exchange track segment information in the pseudorapidity region where these systems overlap. Up to 4 best (highest p_{T} and quality) muon candidates from each subsystem are selected and sent to the GMT.

In the case of the RPC, there is no local processing apart from synchronization and cluster reduction. Hits from all stations are collected by the Pattern Comparator Trigger logic. If they are aligned along a possible muon track,a candidate is formed and a p_{T} value is assigned. Found muon candidates are ranked based on their quality and p_{T}; up to 4 best candidates from the barrel and 4 from the endcaps are sent to the GMT.

The Global Muon Trigger attempts to correlate the DT and CSC muon candidates with the RPC candidates. Bits delivered by the calorimeter trigger are used to determine if these muons are isolated. The final ensemble of muons is sorted based on their quality, correlation, and p_{T}, and the 4 best muons are then transmitted to the Global Trigger. Finally, transverse momentum thresholds are applied by the Global Trigger for all trigger conditions.

Global Trigger

The Global Trigger takes the decision to accept or reject an event at L1 based on trigger objects delivered by the GCT and GMT. These objects consist in candidate-particle, such as e/γ (isolated and non-isolated), muons, central and forward hadronic jets, as well as global quantities: total and missing transverse energies, the scalar sum (H_T) of the transverse energies of jets above a programmable threshold, and twelve threshold-dependent jet multiplicities. Objects representing particles and jets are ranked and sorted. Up to four objects are available. They are characterized by their p_{T} or $\mathrm{E_T}$, (η,ϕ)-coordinates, and quality. For muons, charge, MIP and ISO bits are also available.

The core of the GT is the Global Trigger Logic (GTL) stage, in which algorithm calculations are performed. The most basic algorithms consist of applying p_{T} or $\mathrm{E_T}$ thresholds to single objects, or of requiring the jet multiplicities to exceed defined values. Since location and quality information is available, more complex algorithms based on topological conditions can also be programmed into the logic. The results of the algorithm calculations are sent to the Final Decision Logic (FDL) in the form of one bit per algorithm. Finally, the Global Trigger Front-end (GTFE) board collects the GT data records, appends the GPS event time received from the machine, and sends them to the data acquisition for read-out.

The High Level Trigger

The HLT [7], [8] has access to the complete read-out data and can therefore perform complex calculations similar to those made in the analysis off-line software if required for specially interesting events. To achieve the maximum flexibility HLT algorithms will evolve with time and experience.

Commodity computer processors make subsequent decisions using more detailed information from all the detectors in more and more sophisticated algorithms that approach the quality of the final reconstruction

Upon the receipt of a Level-1 trigger, after a fixed time interval of about $3.2\,\mu$s, the data are transferred from the pipelines to front-end readout buffers. After further signal processing, zero suppression and/or data compression, the data are placed in dual-port memories for access by the DAQ system. Each event has a size of 1.5 MB and is contained in several hundred front-end readout buffers. Through the event building "switch", data from a given event are transferred to a processor. Each processor runs the same HLT software code to reduce the output rate to $10 \div 100$ Hz for mass storage.

The use of a processor farm for all selection beyond Level-1 trigger allows maximum benefit from the evolution of computer technology. The HLT flexibility relies not only on the hardware, but also in a complete freedom in the selection of data to access as well as in the sophistication of the algorithms.

Various strategies guide the HLT code development. For instance, whenever is possible, only objects and regions of the detector that are actually needed are reconstructed; since events are to be discarded as soon as possible. This leads to the idea of partial reconstruction and to the notion of many virtual trigger levels, e.g. calorimeter and muon information are used, followed by the use of the tracker pixel data and finally the full event information.

Chapter 2

The CMS Computing

A computing system needs to be realized, together with actual detectors, to operate the CMS experiment, convert raw data in physics object and make any physics evaluation and discovery.

The physics analysis comes as the final goal, after the computing system is built, a software environment for the events modelization and analysis is set up, a Monte Carlo production system is created and the analysis tools are provided to the physicist.

In the very same way the physics work contained in this thesis comes after a significant amount of work on the computing system, including a contribution to the Monte Carlo production tools and analysis tools. This enabled the working environment not only for the specific analysis hereafter described, but also for any other analysis in the CMS collaboration.

The following sections will describe the CMS computing model that is based on a distributed environment, a short review on the application framework allowing both Monte Carlo simulation and analysis to deal with finite high level objects and focuses on the contribution given on the development of the tools for the end user.

2.1 The CMS Computing infrastructure

The CMS software and computing system [9] covers a broad range of activities:

- design, evaluation, construction and calibration of the detector;

- storage, transfer, access, reconstruction and analysis of data;

- production and distribution of simulated data
- access to conditions and calibration information and other non-event data;
- support of a distributed infrastructure for physicist's work.

The components of the computing system can be summarized as follows:

- Computer centers, managing and providing access to storage and CPU resources
- a distributed databases system allowing access to non-event data
- underlying generic Grid services giving access to distributed computing resources
- a set of computing services, providing tools for transferring, locating, and processing large collection of events
- an event data model and corresponding application framework

The system has been designed to be modular, made of loosely coupled components with well-defined interfaces, and with emphasis on scalability to very large event samples. It has also been taken into account that, during the lifetime of the system, several generations of underlying hardware and software and change of personnel may occur.

The whole infrastructure is periodically tested through "Data Challenges" of increasing size and complexity, approaching to the realistic final environment.

2.1.1 The Tiered architecture: Computing centers

The requirement to analyze very large statistics dataset in pursuit of rare signal, coupled with the fine granularity of the CMS detector, implies a volume of data without precedent in scientific computing. This requires a system of large scale, supporting efficient approaches to data reduction and event reconstruction. The storage, networking and power needed are well more of the reasonably reachable in a central computing system. Therefore a highly distributed computing model is required for CMS operations.

The resulting architecture is made of Tiered Centers:

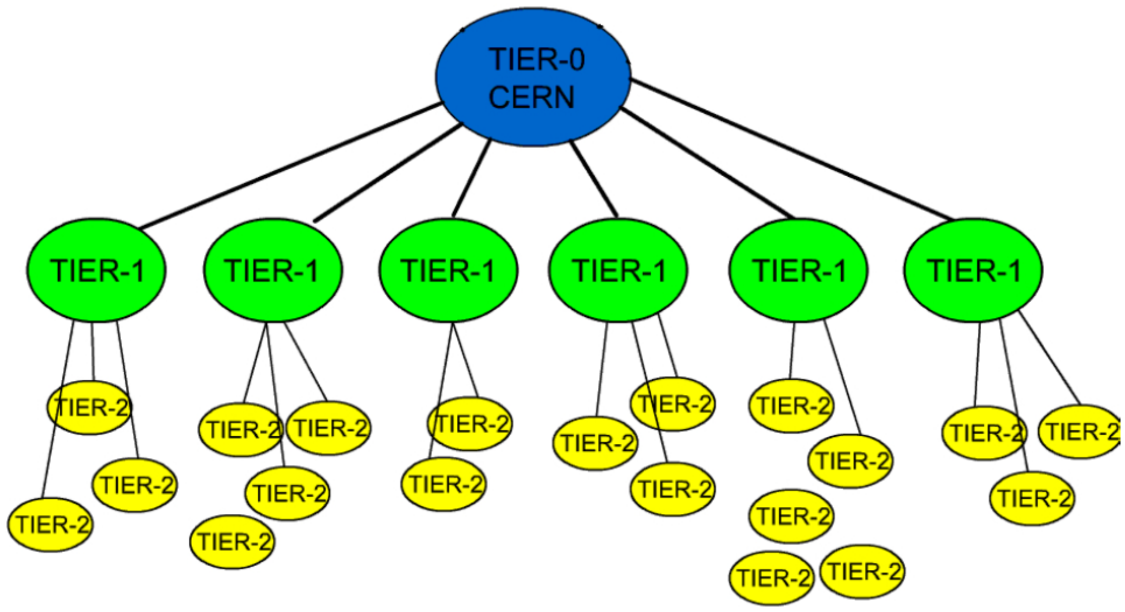

Figure 2.1: The Tier architecture.

- a single Tier-0 center at CERN, the online farm near the experiment, responsible also for first reconstruction and storage of both raw and reconstructed data. It hosts also the CERN Analysis Facility (CAF) which combines flexible CPU resources with rapid access to the entire CMS dataset. It supports fast turn-around analysis when required and specialized functions related to the operation of the detector, such as calibration and performance monitoring.

- eight Tier-1 centers at national computing facilities, responsible for data-intensive processing services and raw, simulated and processed data permanent storage

- about 25 Tier-2 centers at institutes for the user analysis over data replicated from Tier-1 centers and Monte Carlo production.

Further Tier levels can be considered as part of CMS although not included in the the baseline resources of CMS. Tier-3 sites are for instance foreseen, relatively small computing installations that serve the needs of a local institution's user community, giving a potentially significant contribution mainly as analysis facilities.

Data are spread over a number of centres following the physical criteria given by their classification. Replication of data is driven more by the need of optimizing the access to most commonly accessed data than by the need to have data “close to home”. Furthermore Tier-2 centres support users not only on a geographical basis but mainly on a physics-interest basis.

2.1.2 Data Format

CMS makes use of several event formats with different levels of detail and precision, produced with different processing:

- RAW events contain the full recorded information from the detector, plus a record of the trigger decision and other metadata. An extension of this format is used to store the output of Monte Carlo simulation tools, providing more information about the event generation.

- RECO events are obtained applying specific detector reconstruction algorithms and compression algorithms, including detector-specific filtering and correction of digitized data, primary and secondary vertex reconstruction, tracking and particle identification. The resulting events contain high level physics objects plus the subset of reconstructed hits and clusters used to reconstruct them, enough to allow subsequent application of calibrations and re-reconstruction, since basic improvements will require re-reconstruction at least once per year.

- Analysis Object Data (AOD) is a compact analysis format designed to allow a wide range of physics analysis whilst occupying sufficiently small storage, so that very large event samples can be stored in many centers. AOD events contain also additional information required to allow kinematic refitting.

- Non-Event Data are required in order to interpret and reconstruct events. There will be four kinds of non event data for CMS:

  - construction data, including all information about the sub-detector construction up to the start of integration;
  - equipment management data such as detector geometry and location as well as information about electronic equipment;
  - configuration data, comprising the sub-detector-specific information needed to configure the front-end electronic;

– conditions data, including calibrations, alignments and detector status information. They are produced both by online and offline applications and used by HLT, subsequent reconstruction and analysis.

The typical size and expected rate for each format are shown in table 2.1

name	size	MC size	rate	custodial site
RAW	1.5 MB/evt	2 MB/evt	4.5 Pb/year	full set T0/ spread over T1
RECO	0.25 MB/evt	0.4 MB/evt	2.1 Pb/year	spread over T1
AOD	50 kB/evt		2.6 Pb/year	full set T1/spread over T2

Table 2.1: Typical size and expected rate for each data format.

2.1.3 Event Data Flow

The process of data reduction and analysis takes place in several steps, typically carried out at different computer centers.

The CMS DAQ system writes DAQ-RAW events to the High-level Trigger farm input buffer. The HLT farm writes RAW events at a rate of 150 Hz, classified in $\mathscr{O}(50)$ primary datasets depending on their trigger history, with a predicted overlap of less than 10%. The primary dataset definition is immutable. An additional "express-line" is also written, for events that will be reconstructed with high priority. The primary datasets are grouped into $\mathscr{O}(10)$ online streams in order to optimize their transfer to the Offline farm and the subsequent reconstruction process. The data transfer from HLT to the Tier-0 farm must happen in real time at a sustained rate of 225 MB/s.

The first event reconstruction is performed without delay on the Tier-0 farm which writes RECO events. RAW and RECO versions of each primary dataset are archived on the Tier-0 and transferred to a Tier-1 which takes custodial responsibility for them.

Basic improvements in the software, as well as better knowledge of calibration and alignment of the detector will require re-reconstruction at least once per year, performed at the Tier-1 centres. They also produce AOD through bulk filtering and selections (*skimming*). Further skimming of RAW, RECO and AOD data at the Tier-1 centres will be triggered by Physics Groups requests and will produce custom versions of AOD. Only very limited analysis activities from individual users are foreseen at the Tier-1 centre.

Non event data will be stored in an online database which is directly connected to the detector and makes available configuration data to the detector and receives conditions data from the Detector Control System. An offline version has the master copy of non event data.

Data needed for analysis, reconstruction, and calibration activities are replicated at the various CMS computing centres.

The data flow is summarized in figure 2.2.

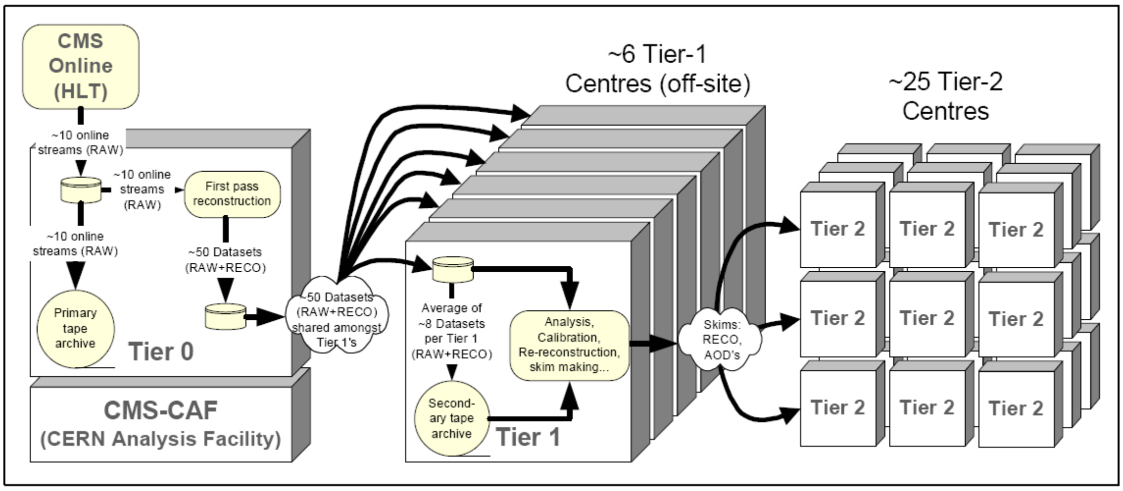

Figure 2.2: Data flow.

2.1.4 The underlying Grid Infrastructure

LHC chose a novel globally distributed model for data storage and analysis instead of the traditional approach of centralizing all of this capacity at one location near the experiments. This approach provides several key benefits:

- the significant costs of maintaining and upgrading the necessary resources for such a computing challenge are more easily handled in a distributed environment, where individual institutes and participating national organizations can fund local computing resources and retain responsibility for these, while still contributing to the global goal. Commodity hardware can be used for the purpose, reducing costs and allowing the possibility to take advantage of the rapidly evolving technologies.

- in a distributed system there are no single points of failure. Multiple copies of data and automatic reassigning of computational tasks to available resources ensures load balancing of resources and facilitates access to the data for all the scientists involved, independent of geographical location. Spanning all time zones also facilitates round-the-clock monitoring and support.

Of course, a distributed system also presents a number of significant challenges. These include ensuring adequate levels of network bandwidth between the contributing resources, maintaining coherence of software versions installed in various locations, coping with heterogeneous hardware, managing and protecting the data so that they are not lost or corrupted over the lifetime of the LHC, and providing accounting mechanisms so that different groups have fair access, based on their needs and contributions to the infrastructure.

The globally distributed model is realized through the computational Grid approach. The Grid [10] idea was introduced in the 1990s by Ian Foster and Carl Kesselman. It refers to a flexible, secure, coordinated resource sharing among dynamic collection of individuals, institutions and resources [11], referred as *Virtual Organizations*. The word Grid is used by analogy with the electric power grid, which provides pervasive access to electricity and has had a dramatic impact on human capabilities and society. Many people believe that the computational Grid will have a similar transforming effect, allowing new classes of applications to emerge.

WLCG

The integration of the resources in CMS computing centers into a single coherent system relies upon Grid middleware which presents a well known interface to storage and CPU facilities at each LCG site.

The World-wide LHC Computing Project (WLCG) Project [12] will implement a Grid to support the computing models of the LHC experiments.

The WLCG Project will collaborate and inter-operate with other major Grid development projects, network providers and production environments around the world. It is for instance strictly coupled with the EGEE (Enabling Grids for E-SciencE) [13] project, which has the goal of deploying a robust Grid infrastructure. It interconnects a large number of sites in 34 countries around the world, integrating several national and regional Grid initiatives in Europe. Of fundamental importance is the relationship with the Open Science Grid (OSG) [14], a national production computing grid infrastructure for large scale science, built and operated by a consortium of U.S.A. universities and national laboratories.

The Grid middleware allows single sign-on on a User Interface (UI), the user submit jobs through the middleware Workload Management System accessing batch queues in all CMS centres, while the Computing Element (CE) interface allows to access them transparently. Automatic mechanisms enable installing, configuring and verifying CMS software at remote sites.

The Storage Element (SE) allows remote access to storage resources. A standard interface hides the complexity and the peculiarities of the underlying storage system, presenting to the user a single logical file namespace where CMS data are stored.

2.1.5 CMS Computing services

A number of CMS-specific computing services operate on top of the generic Grid layer, facilitating higher-level data and workload management functions. These services require CMS-specific software agents to run at some sites in addition to generic Grid services. CMS provides also user interfaces to the Grid for analysis job submission and monitoring and tools for automated steering of large-scale data production an processing.

Workload management

CMS can take advantage of the Grid infrastructure since a typical computing task can be easily split in many independent processes without needing to intercommunicate. Sending process over the Grid allows to run them in parallel over multiple machines reducing the total execution time, taking advantage of bulk operations provided by the middleware. Outputs from the different jobs needs then to be merged together for an efficient access, transferred to the destination sites and tracked for the user analysis from the physics community, which is as well distributed all around the world.

Processing and analysis of data at sites is typically performed by submission of batch jobs to a remote site via the Grid Workload Management System (WMS). A standard job wrapper performs the necessary setup, executes the user application analyzing the data present on local storage at the site, arrange for any produced data to be made accessible via the Grid data management tools and provides logging information. This process is supported by several CMS-specific services.

A lightweight job bookkeeping and monitoring system allows user to track, monitor and retrieve output from jobs submitted to and executing at remote sites. The system also provides an uniform interface to a variety of Grid-based and local batch-system based submission tools.

In addition, a suite of software distribution tools provides facilities for automated installation of standard CMS applications and libraries at remote sites.

The current implementation of the analysis and Monte Carlo tools is described later in Section 2.2 .

Data management

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

The connection between logical datasets and physical files is provided by a catalogue system, the Dataset and Bookkeeping System (DBS), which provides tools for cataloguing and describing event data. A second catalogue system, the Data Location Service (DLS), provides the mapping between file blocks and sites at which they are located, taking into account the possibility of replicas at multiple sites. The actual location of files is known only within the site itself through a Local File Catalogue. CMS applications know only about logical files and rely on this local service to have access to the physical files.

Data transfers are never done as direct file copy by individual users. The data transfer and placement system is responsible for the physical movement of file-blocks between sites and is currently implemented by the Physics Experiment Data Export (PhEDEx) [15] system. This system schedules, monitors and verifies the movement of data in conjunction with the storage interface at CMS sites, ensuring optimal use of the available bandwidth.

2.1.6 The CMS SoftWare

The overall collection of CMS software needed by simulation, calibration, alignment and reconstruction is referred as CMSSW. It has the responsability to process and select events from the High Level Trigger Farm, implementing calibration and alignment strategies, ensure tracking and reproducibility of the reconstruction results, simplify and standardize the way physicist develop reconstruction algorithms and facilitate the interactive analysis. It is used in both offline and online context.

The adopted methodology is object-oriented programming, based primarily on the C++ programming language.

The architecture consists of:

- an application framework customizable for each of the computing environments;
- physics software modules with clearly defined interfaces that can be plugged into the framework at runtime without a direct reciprocal interaction
- a service and utility toolkit that decouples the physics modules from details of event I/O, user interface and other environmental constraints

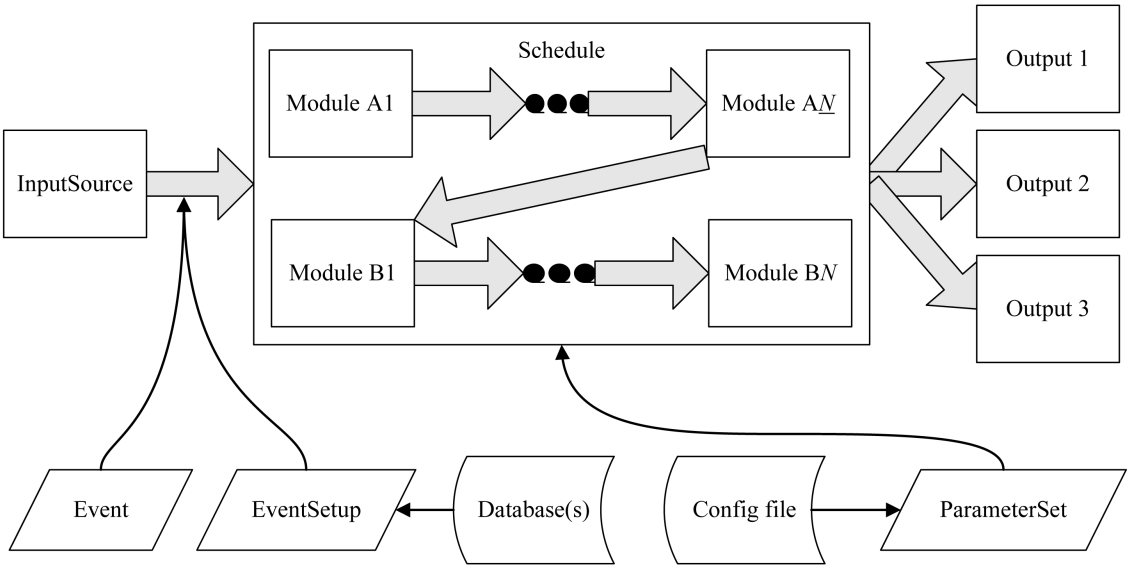

Figure 2.3: Components of the CMS Framework and Event Data Model.

The framework defines the top level abstractions, the behavior and collaboration patterns among the physics modules and comprises a set of classes for specific CMS concepts like detector components and event features and a control policy that orchestrate the instances taking care of the flow of control module scheduling in/out etc.

The central concept of the data model is the *Event* which provides access to the recorded data from a single triggered bunch-crossing and to data derivated from it. It may include raw digitized data, reconstructed products or high-level objects for real or simulated bunch crossings. The Event also contains information describing the origin of the raw data and the provenance of all derived data products, to unambiguously identify how each event contributes to the final analysis and includes a record of the software configuration and conditions/calibration setup used to produce each new data product (EventSetup). Some of the functionalities of the framework, such as mathematical and statistical tools and the I/O system, are implemented through ROOT [16], a popular interactive tool for analysis. Persistent Events are stored as rootfiles.

The Event is used by a variety of physics modules which may read data from it, or add new data recording their provenance. Each module performs a well-defined function related to selection, reconstruction or analysis. Several module types exists, with a specialized interface:

- event data producers, used in triggering, reconstruction and simulation, which add data into the Event ;
- filters, used in the online triggering and selection;
- analyzers, producing summary information and histograms from an event, but not modify the event data;
- input and output modules for DAQ and disk storage.

Modules are insulated from the computing environment, are executed independently and communicate through the Event, so that they can be developed and tested independently.

A CMS application is made of one or more ordered sequences of modules through which each module must flow along with the configuration for each module. The framework configures the modules, schedules their execution and provides access to global services an utilities.

2.2 CMS tools for analysis and Monte Carlo production

The Workload management section refers to three main tools, used as interface to automate the Monte Carlo production and the user Analysis job management. Given my strong contribution on their design, implementation and maintenance they are described in a dedicated section.

2.2.1 Monte Carlo production tools

The Monte Carlo production is a crucial aspect both for detector studies and physics analysis. For that reason, a large amount of simulated data is required. Monte Carlo production is thus a large scale process even if its management is basically centralized. It is characterized by a complex workflow, consisting of multiple steps: physics event generation, detector response simulation, signal digitization, and complete event reconstruction. The resulting output files need to be stored and accessed efficiently. Although the production jobs are parallelized, the single job output needs to be merged and tracked for efficient access.

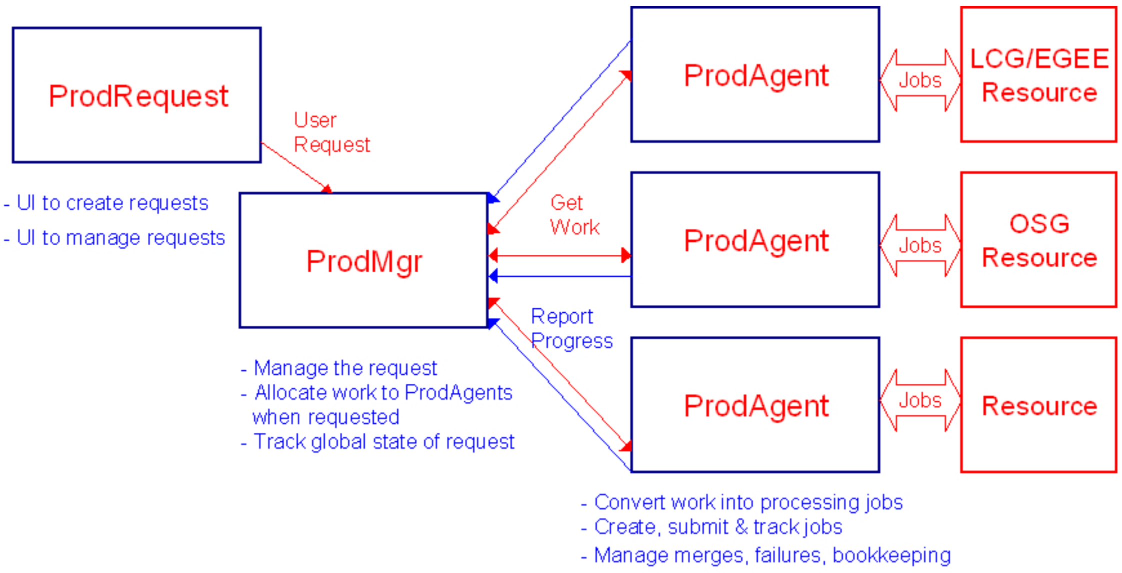

Figure 2.4: The Monte Carlo Production system.

The main goal of the CMS Production System is to automate as much as possible the whole production chain, allowing easy operation and maintenance. On the other hand it has to deal efficiently with a large number of jobs and sites in a distributed environment.

The Production system takes care of managing requests from users, breaking them into jobs, performing submission to the the Grid, tracking the jobs, handling the errors and performing job resubmission when needed.

The system consists of three major components:

- Production Request (ProdRequest), which takes care of the physics groups requests.

- Production Manager (ProdMgr), provides the accounting functionalities of the system, keeping track of request progress and distributing the work among ProdAgent instances.

- Production Agent (ProdAgent), deployed in several instances, take care of the job set up, submission and tracking: it is made of several autonomous components, communicating through asynchronous messages. Delayed and queues messages enables the ProdAgent to deal with CMS catalogs, transfer system and other third part components even when they are offline for local failures or maintenance shut off. The system does automatically job preparation, submission, tracking and possible resubmission as well as resource monitoring, job queuing, job distribution according to the available resources, data merging, data registration into the data bookkeeping, data location systems, data transfer and placement systems.

Complementary monitoring systems, usually shared with the analysis tools, are used to track down potential problems.

2.2.2 User Analysis tools

The management of user data analysis is not planned and centralized, as the Monte Carlo Production can be, since it has to cover many different use cases characterized by a rapid change of parameters and applications.

The distribution of data over many computing centers located in many different countries may results in additional complication for the physics analysis user. To hide the complexity of the whole system to the end user, a set of high level tools have been developed. They consist mainly of and end user tool, CMS Remote Analysis Builder (CRAB) and additional monitoring tools.

CRAB is a specific tool designed and developed by the CMS collaboration allowing easy access to distributed data, preparation and submission of jobs to the Grid infrastructure, monitoring the job status and retrieval of the output.

The tool main feature is the possibility to distribute and parallelize the local CMS batch data analysis process over different Grid environments without any specific knowledge of the underlying computational infrastructures. It allows transparent usage of EGEE and OSG infrastructures. CRAB interacts with the local user environment, the CMS Data Management services and the Grid middleware.

Following the analysis model, the user runs interactively over small data samples in order to develop and test his code, using CMS analysis framework. A CRAB

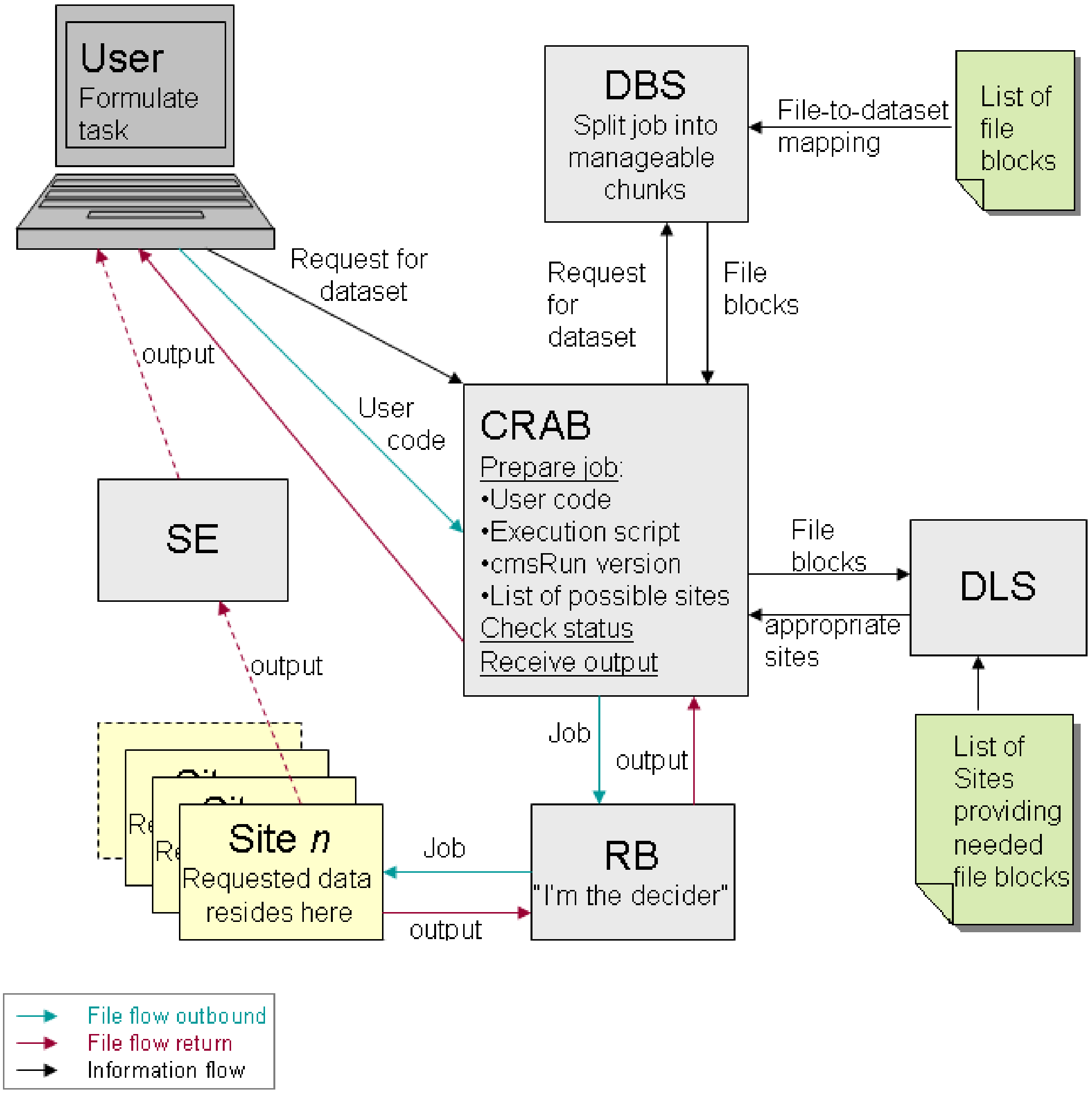

Figure 2.5: Crab Workflow.

installation stands over the Grid User Interface, providing the user with few friendly commands. More in details, the CRAB workflow is factorized as follows:

- the Data Discovery step, through the interaction with the CMS data management infrastructure (DBS and DLS)
- the interaction whit the software framework, to allow the reproduction of the

test environment in the remote resources. This involves also the user code packaging, e.g. the user libraries, the framework modules, or even an user executable, data and configuration files;

- a set of Grid specific actions, such as: the creation of the Grid job description for the resource matchmaking and job submission; resources discovery; submission; monitoring; job cancellation, resubmission, when required, and post mortem functionalities to track down execution failures at any level of the whole infrastructure; output retrieval and user output handling, e.g. copying back through the Grid Workload Management, copy to a generic Storage Element or a tape server.

The job parameters are defined through a configuration file: the dataset to be accessed, the CMSSW specific configuration file, the task splitting parameters which brings to many actual jobs, the produced output handling.

CRAB has been in production and in routine use singe Spring 2004 and extensively used during studies to prepare the CMS Physics Technical Design Report and the data challenges and the Magnet Test Cosmic Challenge, generating thousand of jobs per day at peaks rates.

A CRAB server implementation

CRAB was initially a standalone interface to the CMS computing infrastructure. Although the Grid implementation has the functionality to allow a flexible usage from the experiments, some functionalities are CMS specific enough to be not included in the Grid middleware and, at the same time they are heavily used by the generic job that they can be implemented in an intermediate server structure which takes care of the job during his lifetime, from the submission to the output retrieval, leaving to the final user just the preparation step and the final access for the results evaluation.

These are the main motivations for a new client-server structure: the same user interface, but placed in front of an intermediate server which automates as much as possible the whole analysis workflow and improves the scalability of the system providing a better job distribution and management.

The server implementation remains completely transparent to the end user, completely hidden by the same standalone interface, and easy to be maintained, reusing where possible the ProdAgent experience and infrastructure and at the same time contributing to its improvement and integrating the missing functionalities required by the flexibility requirement of the analysis tasks.

2.2.3 BOSS (Batch Object Submission System)

Both ProdAgent and CRAB needs to perform actions over the different Grid infrastructures (e.g. EGEE and OSG) as well as on local batch systems.

BOSS, Batch Object Submission System, was designed to cope with both local and grid submissions for Monte Carlo productions and analysis tasks, providing also a logging and bookkeeping system and extra monitoring tools.

The information is persistently stored in a relational database (right now MySQL [20] or SQLite [21]) for further processing. In this way the information that was available in the log file in a free form is structured in a fixed-form that allows easy and efficient access. The database is local to the user environment and is not requested to provide server capabilities to the external world: the only component that interacts with it is the BOSS client process.

BOSS can log not only the typical information provided by the batch systems (e.g. executable name, time of submission and execution, return status, etc.), but also information specific to the job that is being executed (e.g. dataset that is being produced or analyzed, number of events done so far, number of events to be done, etc.). This is done by means of user-supplied filters: BOSS extracts the specific user-program information to be logged from the standard streams of the job itself filling up a fixed form journal file to be retrieved and processed at the end of job running via the BOSS client process.

The scheduler interface

BOSS interfaces to a local or grid scheduler (e.g. LSF, PBS, Condor, WLCG, etc.) through a set of plugins provided by the system administrator, using a predefined interface. They can be written using any script/programming language, since they are accessed through standard Inter Process Communication. This allows hiding to the upper layers its implementation details, in particular whether the batch system is local or distributed. The interface provides the capability to register, un-register and list the schedulers. BOSS provides an interface to the local scheduler for the operations of job submission, deletion, querying and output retrieval. At output retrieval time the information in the database is updated using information sent back with the job.

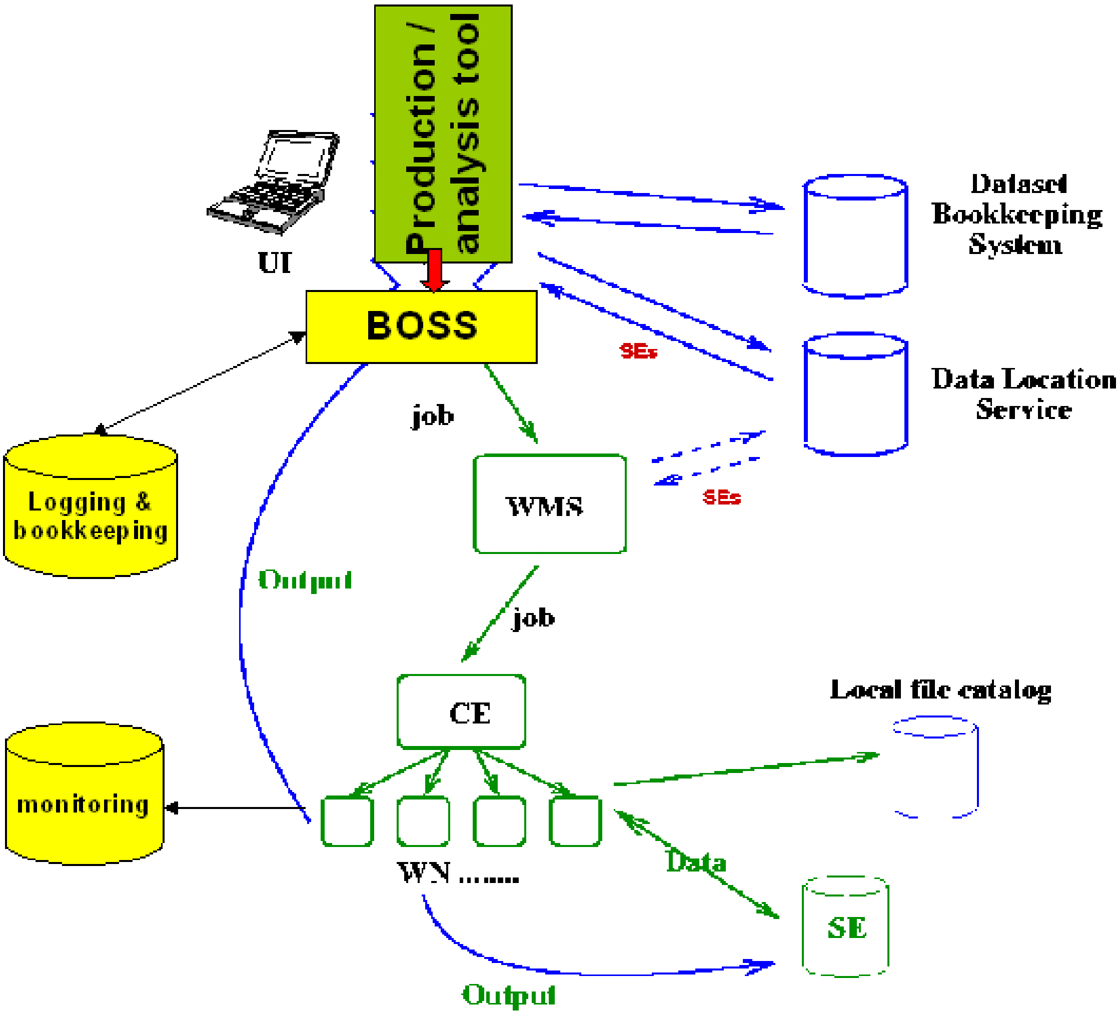

Figure 2.6: BOSS in the analysis/production system.

The Real Time Monitoring System

BOSS provides also an optional run-time monitoring system that, working in parallel to the logging system, collects information while the computational program is still running, and presents it to the upper layers through the same interface. The real-time information sent by the running jobs are collected in a separate database server. The same real-time database server may support more than one BOSS database. The information in the real-time database server has a limited lifetime: in general it is deleted after that the user has accessed it, and in any case after successful retrieval of the journal file. It is not possible to use the information in the real-time database server to update the logging information in the BOSS database once the journal file for the related job has been processed. The run-time monitoring is made through a pair client-updater registered as a plug-in module: they are the only components that interact with the real time database. The real-time updater is a client of the real-time database server: it sends the information of the journal file to the server at pre-defined intervals of time. The real-time client is a tool used by BOSS to update his database using the real-time information.

The user interface

The interface with the end-user is made through:

- a command line, kept as similar as possible to the one of the previous versions; it is the minimal way to access BOSS functionalities to give a straightforward test and training instrument;
- C++ and Python API, used by both ProdAgent and CRAB that are python programs; .

BOSS is designed to deal with complex workflows, since user programs may be chained together to be executed by a single batch unit (job). The relational structure supports not only multiple programs per job (program chains) but also multiple jobs per chain, in the event of job resubmission. Homogeneous jobs, or better ”chains of programs”, may be grouped together in tasks, e.g. as a consequence of the splitting of a single processing chain into many processing chains that may run in parallel.

The description of a task is passed to BOSS through an XML file, since it can model its hierarchical structure in a natural way.

The BOSS wrapper system

The process submitted to the batch scheduler is the BOSS job wrapper. All interactions of the batch scheduler to the user process pass through the BOSS wrapper. The BOSS job wrapper starts the chosen chaining tool, and optionally the real-time updater. An internal tool for chaining programs linearly is implemented in BOSS but in future external chaining tools may be registered to BOSS so that more complex chaining rules may be requested by the users. BOSS will not need to know how they work and will just pass any configuration information transparently down to them. The chaining tool starts a BOSS “program wrapper” for each

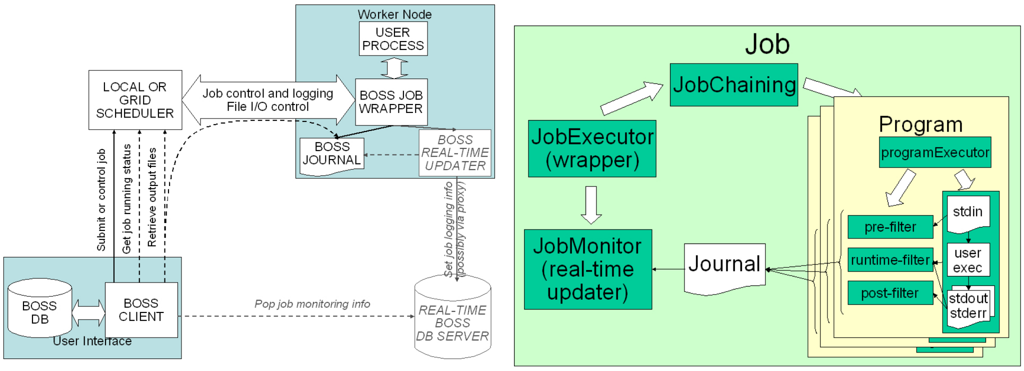

Figure 2.7: The BOSS workflow: the left picture shows the detailed workflow of a job in the execution host.

user program. The program wrapper starts all processes needed to get the run-time information from the user programs into the journal file. This program wrapper is unique and it has to be started passing only one parameter, the program id. The BOSS client determines finished jobs by a query to the scheduler. It retrieves the output for those jobs and uses the information in the journal file to update the BOSS database. The BOSS client pops the information about running jobs from the real-time database server through the client part of the registered Real Time Monitor plug-in. It also deletes from the server the information concerning jobs for which the BOSS database has already been updated using the journal file. The information extracted from the real-time database server may be used to update the local BOSS database or just to show the latest status to the user.

The usage of BOSS in the CMS computing

First BOSS implementations have been successfully used by most CMS Regional Centers for managing Monte Carlo data productions since 2002. Furthermore in fall 2002 it has been used in a prototype of the CMS production system deployed on the European DataGrid test bed demonstrating its ability to be used also in a grid environment.

BOSS was then used to finalize the first CRAB release. The current implementation was used since the very begin in ProdAgent, taking complete care of the grid interaction, while its database was the main source for job tracking and monitoring. In the meanwhile it was integrated in the new CRAB releases, reducing the effort for CRAB code maintainability and increasing its scalability.

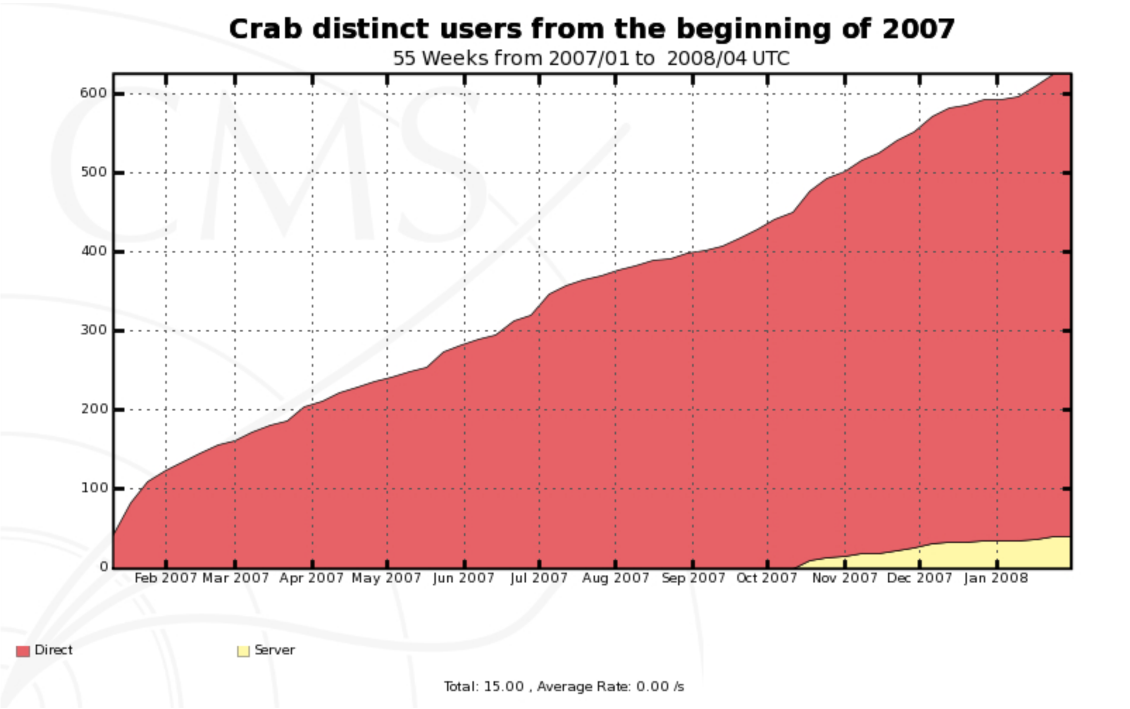

Figure 2.8: Crab usage in 2007 in term of user number.

The BOSS team was in charge also for the integration with both tools as well as for the first test of the massive submission to the new Grid infrastructure, managing scalability issues, continuously evolving for new incoming requirements and hiding to the upper layers problems of the first Grid implementation and issues related to new scheduler integrations.

As part of both CRAB and ProdAgent, it has been used in the part years for Monte Carlo production and analysis, and to integrate and test the new middleware from the EGEE project, gLite [22] . Being it the front-end tool, it has successfully used to reduce the impact of middleware as well as of many other system changes such as in the operating systems used in the collaboration resources.

2.2.4 Usage and status of the computing system

The computing system has to be ready for use with full functionality and reliability from the start of LHC data taking.

The system is in continuous evolution in terms of hardware infrastructure as well as the services to be delivered and the tools for their usage.

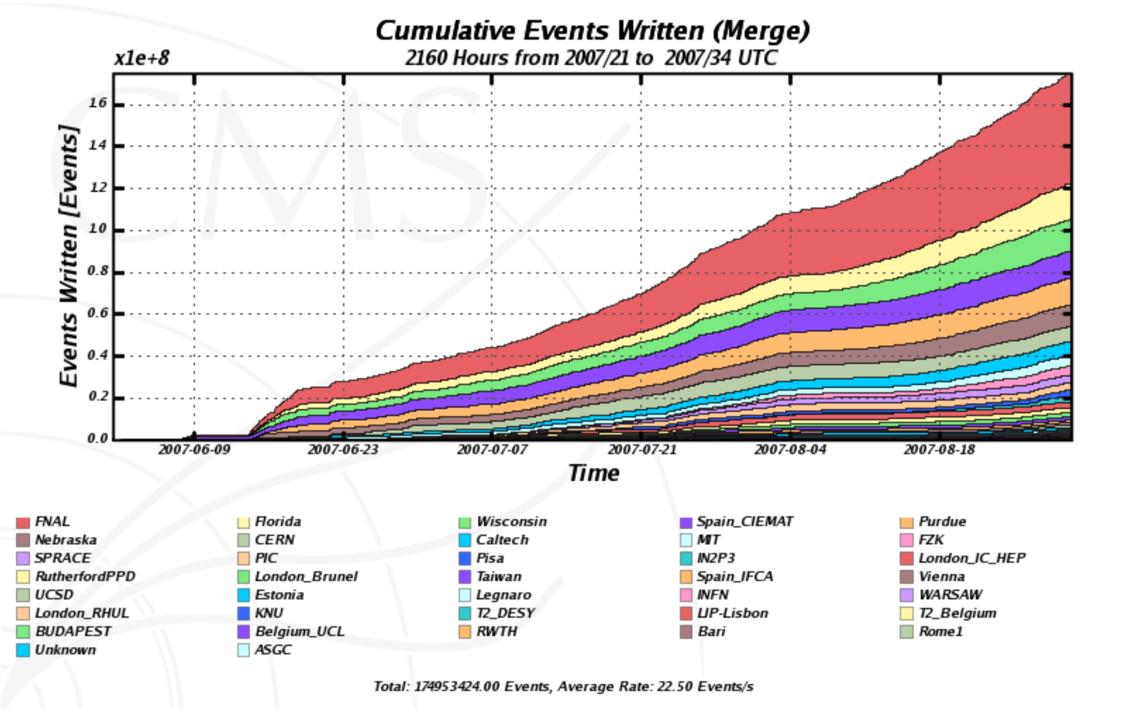

Figure 2.9: Monte Carlo production distribution over CMS sites.

The test of the infrastructure is tightly linked to other CMS activities as production and analysis of the simulated data needed for studies on detector, trigger and DAQ design and validation, and for physics system set-up.

An idea on how the full system is currently used is given in figures 2.8 and 2.9, showing respectively the increasing number of CRAB distinct users in 2007 and the cumulative number of events produced and distributed over CMS sites in the Summer 2007 time slot.

This continuous evolution is periodically tested with a series of increasing scale full-system tests, “Data Challenges” exercising all available components in a realistic way. They comprise the simulation, done as realistically as possible, of data (events) from the detector, followed by the processing of that data using the software and computing infrastructure that will, with further development, be used for the real data when the LHC starts operating.

Part of the plan of the Challenges has already been executed, and has provided useful feedback. The evaluation of the results of the Challenges and the implementation of the suggestions coming from this evaluation will provide an important contribution towards reaching the full readiness of the CMS computing system on schedule.

The tools described in this section played an important role in the two last challenges: the Computing Software and Analysis Challenge in 2006 (CSA06) and 2007 (CSA07). The Computing Software and Analysis Challenge in 2006 (CSA06) was a test at the 25% of the full system scale, including several workflow elements: event reconstruction at the CERN Tier-0 center; data distribution to Tier-1's for archiving and data serving purposes; data skimming, driven by CMS physics groups and reconstruction at Tier-1's for archiving, serving of re-processed data to Tier-2's and Grid submission of physics analysis jobs. CSA07 challenge was a test at the 50% scale started in July 2007.

The CMS distributed computing system is reaching the scale required by the LHC start-up in 2008, when a last challenge will bring to the first data taking run.

Chapter 3

Top quark physics

The top quark was discovered at Fermilab in 1995. Although its discovery completes the three-generation structure of the Standard Model, it opens also many questions and a new field for understanding the particle physics.

The top mass, about 35 times larger than the mass of the b quark, is very close to the scale of electroweak symmetry breaking, raising many questions about the actual role of the heavier quark in the Higgs mechanism. Unlike others fermion masses, it enters in loop correction giving quadratic contributions. Thus the accurate knowledge of the top mass, gives a powerful instrument for the indirect estimate of many electroweak parameters.

Furthermore, an accurate study of the top quark and eventual anomalies on its production and decays could suggest the presence of eventual lighter particles, as well a as non standard couplings with other particles, leading to an first evidence of new physics beyond the Standard Model.

There are so many reasons for an accurate study of the top quark and in particular for its mass measurement, which is the goal of this thesis.

In this chapter, the Standard Model framework is briefly discussed to introduce the top quark physics and the contribution that can be given by the CMS experiment.

3.1 The Standard Model

The Standard Model constitutes one of the most successful achievements in modern physics. It provides a very elegant theoretical framework, which is able to describe the known experimental facts in particle physics with high precision.

It is a gauge theory based on the symmetry group $\mathrm{SU(3)_C} \otimes \mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$, where C refers to the associated colour charge, L to the left-handedness, Y to the hypercharge. The structure of the fundamental interactions arises by requiring the invariance of the Lagrangian under local gauge transformations.

Interactions are described via the exchange of the corresponding spin-1 gauge field: eight massless gluons, forming a colour octet, and one massless photon, respectively for the strong and electromagnetic interactions, and three massive bosons $\mathrm{W}^{\pm}$ and Z for weak interaction. As well as the "real" fermionic matter these bosons are considered pointlike and structureless. The fermionic matter content is given by the known leptons and quarks organized in three-fold family structure:

$$\begin{bmatrix} \nu_e & u \\ e^- & d' \end{bmatrix}, \qquad \begin{bmatrix} \nu_\mu & c \\ \mu^- & s' \end{bmatrix}, \qquad \begin{bmatrix} \nu_\tau & t \\ \tau^- & b' \end{bmatrix}$$

where each quark appears as colour triplets, while the flavour eigenstate q_d' is given by a mix of mass eigenvalues through the Cabibbo-Kobayashi-Maskawa (CKM) matrix: $q_{dj}' \equiv \sum_j V_{ij} q_{dj}$. To any of them corresponds an antiparticle with equal mass but with opposed quantum number such as electric charge, colour, isospin.

Each of the family above is formed of a $\mathrm{SU(2)_L}$ left-handed doublet and $\mathrm{SU(2)_L}$ right-handed singlet:

$$\begin{bmatrix} \nu_l & q_u \\ l^- & q_d \end{bmatrix} \qquad \equiv \qquad \begin{pmatrix} \nu_l \\ l^- \end{pmatrix}_L, \qquad \begin{pmatrix} q_u \\ q_d \end{pmatrix}_L; \qquad l_R^-, \qquad q_{uR}, \qquad q_{dR}$$

where L and R refer in practice to the sign of the projection of the spin vector onto the momentum vector: L negative, R positive, through the chirality operator γ_5, defined as the product of the four Dirac matrices γ^i, such as:

$$q_L = \frac{1}{2}(1-\gamma_5)q, \qquad q_R = \frac{1}{2}(1+\gamma_5)q \tag{3.1}$$

The three fermionic families appear to have identical properties in terms of gauge interactions, while they differ only by their mass and flavour quantum number. The main characteristics of the fermions are summarized in table 3.1.

The Standard Model includes the Glashow-Salam-Weinberg (GSW) theory of electroweak interaction and the Quantum Chromodynamics (QCD) which describes the strong interaction while does not consider gravitation: since this can be neglected

Quarks			Leptons		
name	charge	mass	name	charge	mass
up u	$+\frac{2}{3}$	$1.5 \div 3$ MeV	electronic neutrino ν_e	0	
down d	$-\frac{1}{3}$	$3 \div 7$ MeV	electron e^-	-1	0.51 MeV
charme c	$+\frac{2}{3}$	1.25 ± 0.09 MeV	muonic neutrino ν_μ	0	
strange s	$-\frac{1}{3}$	105 ± 25 MeV	muon μ^-	-1	105 MeV
top t	$+\frac{2}{3}$	175 GeV	tauonic neutrino ν_τ	0	
bottom b	$-\frac{1}{3}$	4.20 ± 0.07 MeV	tau τ^-	-1	1.777 GeV

Table 3.1: Standard Model Fermions: u, d, s masses are extracted from hadron masses and remains over active investigation; c and b masses are "running" masses evaluated in the Standard Model scheme. Top quark mass recent measurements are disscused in section 3.3.3.

in the processes occurring at energies reached with current accelerators, the Standard Model results to be an effective theory for particle physics up to the energy reached so far, but cannot be considered a final theory.

The strong interaction is characterized by the coupling g_s and the colour charge, postulated in order to satisfy Fermi-Dirac statistics, since assuming that baryons ar formed by three quarks and meson of quark-antiquark pairs, all the hadronic spectrum can be explained. The experimental lack of non colourless states, brings to the "confinement hypothesis" which leads in turn to the unobservability of free quarks.

The electroweak interaction is characterized by two gauge coupling constants: g and g', related by:

$$g \sin\theta_W = g' \cos\theta_W = e \tag{3.2}$$

where e is the electron electric charge and θ_W is the electroweak mixing angle.

Electromagnetic interactions are associated with the fermion electric charges, while the quark flavours are related to electroweak phenomena. The strong forces are flavour conserving and flavour independent. On the other side, the carriers of the electroweak interaction (γ, Z, $W^\pm$) do not couple to the quark colour. Thus it seems natural to take colour as the charge associated with the strong forces and try to build a quantum field theory based on it. The QCD Lagrangian contains not only the kinematic terms for the field involved, and the interaction among quarks and gluons, but also cubic and quartic gluon self interactions. The auto-interaction among gauge fields is at the basis of the asymptotic freedom, the weakness of the coupling constant at short distances, and confinement. Furthermore an important consequence from an experimental point of view is that quark can emit gluons, thus hadronic decay, such as of the $Z \rightarrow q\overline{q}$ can result in $Z \rightarrow q\overline{q}g$, e.g. with an additional hadronic shower related to the emission of a gluon.

The main characteristics of electroweak processes can be summarized as follows:

- 100% breaking of the parity symmetry P (left ↔ right) and charge conjugation symmetry C (particle ↔ antiparticle), while CP is still a good symmetry;
- the $W^{\pm}$ bosons couple to the fermionic doublets where the electric charge of the two fermion partners differ of one unit. All fermions couple with the same universal strength;
- the doublet partners of the up, charm and top are mixing of the three quarks with charge $-\frac{1}{3}$ through the CKM element matrix. The mixing of mass eigenstates to form electroweak eigenstates is observed also in the neutrino oscillations;
- neutral currents are flavour conserving: both γ and Z couple to a fermion and its own antifermion, while mixing have never been observed;
- neutral currents depend on the fermion electric charge Q_f so that fermions with the same Q_f couple with the same universal couplings. Neutrinos, which have null electric charge, have non-zero coupling with the Z;
- γ have same interactions with both fermion chiralities
- Z couplings are different for left-handed and right-handed fermions; the neutrino coupling involves for instance only left-handed chiralities;
- whilst the Lagrangian contains cubic and quartic self-interactions among the gauge bosons, at least a pair of charged W bosons are present, but there are not neutral vertices with only photons and Z bosons;

A Higgs model

The $SU(2)_L$ group results in a electroweak Lagrangian which, although well describing many of the above listed characteristics, well defined and renormalizable from the theoretical point of view, produces massless gauge bosons.

The mass generation involves a gauge symmetry breaking, which can occur keeping a symmetric form of the Lagrangian, to preserve renormalizability, but with a degenerate set of states of minimal energy, allowing non-symmetric results for the "vacuum". Thus the symmetry is broken by the ground state: the mechanism is called Spontaneous Symmetry Breaking (SSB) of the electroweak group to the electromagnetic subgroup:

$$SU(3)_C \otimes SU(2)_L \otimes U(1)_Y \rightarrow SU(3)_C \otimes U(1)_{QED}$$

The SSB mechanism generates the masses of the weak gauge bosons and give rise to the appearance of a physical scalar particle in the model, the Higgs boson. The fermion masses and mixing are also generated through the SSB.

As a result of the Goldstone theorem, there are massless excitations associated with a SSB: if a Lagrangian is invariant under a continuous symmetry group H, but the vacuum is invariant only under a subgroup $H \subset G$, then there exists as many massless spin-0 particles as the generators of G not belonging to H.

By choosing a doublet of complex scalar fields with a potential invariant under local $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ transformations, but with a set of infinite degenerate ground states such as:

$$|\langle 0|\phi^{(0)}|0\rangle| = \sqrt{\frac{-\mu^2}{2h}} \equiv \frac{v}{\sqrt{2}} \tag{3.3}$$

it is possible to give an opportune parameterization in terms of a v and a real field $H(x)$ plus a phase term given by three real $\theta^i(x)$ fields. By choosing the physical unitary gauge $\theta^i(x) = 0$ two quadratic terms arise for $W^\pm$ and Z boson, giving them a mass term such as:

$$M_W = M_Z \cos\theta_W = \frac{1}{2} vg \tag{3.4}$$

together with a massive scalar particle H: the Higgs boson. This mechanism is the minimal approach, providing a single Higgs boson, although it can be complicated to give more. The main advantage is a precise prediction for the gauge boson masses, which experimentally follows the above equation, once the M_Z, M_W and $\sin^2\theta$ are measured.

The Higgs mass value is not foreseen, although its form can be obtained as:

$$M_H = \sqrt{2h}v \tag{3.5}$$

So far there is no experimental evidence of the existence of one (or more) Higgs boson(s). Direct searches at LEP and Tevatron fixed a lower bound:

$$M_H = 114.4\,\mathrm{GeV}/c^2 (95\% C.L.) \tag{3.6}$$

Fermion masses

While a mass term for fermions is not allowed in the original Lagrangian, the assumption used for the Higgs mechanism brings to a Yukawa-type Lagrangian term of the form:

$$\mathscr{L}_Y = \frac{1}{2}(v + H)\{c_1\overline{d}d + c_2\overline{u}u + c_3\overline{e}e\} \tag{3.7}$$

generating the fermion masses:

$$m_d = c_1\frac{v}{\sqrt{2}}, m_u = c_2\frac{v}{\sqrt{2}}, m_e = c_3\frac{v}{\sqrt{2}}, \tag{3.8}$$

The fermion masses are arbitrary, since c_i are not provided by the theory, but their coupling is fixed in terms of mass and proportional to the Higgs mass:

$$\mathscr{L}_Y = -\left(1 + \frac{H}{v}\right)\{m_d\overline{d}d + m_u\overline{u}u + m_e\overline{e}e\} \tag{3.9}$$

QED and QCD corrections

High order electroweak contributions as well as well-known QED and QCD corrections have an high impact on the phenomenology.

In the QED, the photon propagator gets vacuum polarization corrections, induced by virtual fermion-antifermion pairs. This kind of QED loop correction can be taken into account through a redefinition of the QED coupling, which depends on the energy scale. The resulting QED running coupling $\alpha(s)$ decreases at large distances. This can be intuitively understood as the charge screening generated by the virtual fermion pairs. The physical QED vacuum behaves as a polarized dielectric medium.

The strong coupling also “runs”. However, the gluon self-interactions generate an antiscreening effect, through gluon-loop corrections to the gluon propagator, which spreads out the QCD charge. Since this correction is larger than the screening of the colour charge induced by virtual quark-antiquark pairs, the net result is that the strong coupling decreases at short distances. Thus QCD has the required property of asymptotic freedom: quarks behave as free particles when $Q^2 \to \infty$. QCD corrections increase the probability of the Z and the $W^{\pm}$ to decay into hadronic modes. Therefore, their leptonic branching fractions become smaller.

Quantum corrections offer the possibility to be sensitive to heavy particles, which cannot be kinematically accessed, through their virtual loop effects. In QED and QCD the vacuum polarization contribution of a heavy fermion pair is suppressed by inverse powers of the fermion mass. At low energies, the information on the heavy fermions is then lost. This "decoupling" of the heavy fields does not apply at higher energies, since Standard Model involves a broken chiral gauge symmetry. As a consequence the heavy top quark generates corrections to the $W^{\pm}$ and Z propagators, which increase quadratically with the top mass. This effect is originated in the strong breaking of weak isospin generated by the top and bottom quark masses, i.e., the effect is actually proportional to $m_t^2 - m_b^2$.

Because of the $SU(2)_C$ symmetry of the scalar sector (the so-called custodial symmetry), the virtual production of Higgs particles does not generate any quadratic dependence on the Higgs

Higher-order corrections to the different electroweak couplings are non-universal and usually smaller than the self-energy contributions, with one one interesting exception, the $Z \to \bar{b}b$ vertex, which is sensitive to the top quark mass. The peculiarity of the $Z\bar{b}b$ vertex leads to well defined contributions to new physics scenarios and is as well very sensitive to the SSB mechanism. Therefore, a precise measurement of the observables involved brings to a better understanding of the theoty.

More in generale, the precision measures of electroweak observables serve as an important tool for testing the theory, non only the Standard Model, since they provide an important consistency test for every model under consideration. By comparing precision data with the predictions it is in principle possible to derive indirect constraint on all the parameters of the model. The information so obtained is complementary to the information gained from direct production of these particles.

In order to drive precise theoretical predictions, two kinds of theoretical uncertainties have to be kept under control: the uncertainties from unknown higher-order corrections, as the predictions are derived only up to a finite order in perturbation theory and parametric uncertainties caused by the experimental errors on input parameters.

3.1.1 Beyond the Standard Model: the reasons for LHC

The Standard Model provides a beautiful theoretical framework which is able to accommodate all our present knowledge on electroweak and strong interactions. It is able to explain any single experimental fact and, in some cases, it has successfully passed very precise tests at the 0.1% to 1% level. In spite of this impressive phenomenological success, the Standard Model leaves too many unanswered questions to be considered as a complete description of the fundamental forces. We do not understand yet why fermions are replicated in three (and only three) nearly identical copies. Why the pattern of masses and mixings is what it is. Are the masses the only difference among the three families? What is the origin of the Standard Model flavour structure? Which dynamics is responsible for the observed CP violation?

In the gauge and scalar sectors, the Standard Model Lagrangian contains only four parameters: g, g', μ^2, and h. We can trade them for α, M_Z, G_F, and M_H; this has the advantage of using the three most precise experimental determinations to fix the interaction. In any case, one describes a lot of physics with only four inputs. In the fermionic flavour sector, however, the situation is very different. With three generations of fermions, we have 13 additional free parameters in the minimal Standard Model: 9 fermion masses, 3 quark mixing angles and 1 phase. Taking into account non-zero neutrino masses, we have three more mass parameters plus the leptonic mixings: three angles and one phase (three phases) for Dirac (or Majorana) neutrinos.

The source of this proliferation of parameters is the set of unknown Yukawa couplings. The origin of masses and mixings, together with the reason for the existing family replication, constitute at present the main open problem in electroweak physics. The problem of fermion mass generation is deeply related to the mechanism responsible for the electroweak SSB. Thus, the origin of these parameters lies in the most obscure part of the Standard Model Lagrangian: the scalar sector. The dynamics of flavour appears to be *terra incognita* which deserves a careful investigation.

The Standard Model incorporates a mechanism to generate CP violation, through the single phase naturally occurring in the CKM matrix. Although the present laboratory experiments are well described, this mechanism is unable to explain the matter-antimatter asymmetry of our Universe. A fundamental explanation of the origin of CP-violating phenomena is still lacking.

The first hints of new physics beyond the Standard Model have emerged recently, with convincing evidence of neutrino oscillations showing that $\nu_e \to \nu_{\mu,\tau}$ and $\nu_\mu \to \nu_\tau$ transitions do occur. The existence of lepton-flavour violation opens a very interesting window to unknown phenomena.

Still other problems arise from many fields. The lack of an unified theory and the fact that the Standard Model is not able to deal with gravity. In the Standard

Models predictions not even the QCD and electroweak couplings converge to the same value. The gauge hierarchy problem: if there is new physics at an energy scale which is much above the electroweak scale of the SM, the Higgs scalar of the Standard Model acquires a mass of the order of this new scale, but we want the Higgs mass not to leave the typical range of the order of the electroweak scale (between 100 and 1000 GeV) because the vacuum expectation value is related to quantities like the W mass which are of O(100 GeV).

The Higgs particle is the main missing block of the Standard Model framework. The successful tests of the Standard Model quantum corrections with precision electroweak data confirm the assumed pattern of SSB, but do not prove the validity of the minimal Higgs mechanism embedded in the Standard Model. The present experimental bounds put the Higgs hunting within the reach of the new generation of detectors. The LHC should find out whether such scalar field indeed exists, either confirming the Standard Model Higgs mechanism or discovering completely new phenomena. Many interesting experimental signals are expected to be seen in the near future. New experiments will probe the Standard Model to a much deeper level of sensitivity and will explore the frontier of its possible extensions. Large surprises may well be expected, probably establishing the existence of new physics beyond the Standard Model and offering clues to the problems of mass generation, fermion mixing, and family replication. Then baryogenesis, inflation, dark matter.

3.1.2 LHC physics program

Given the open questions of the Standard Model, LHC as described in chapter 1 is designed to reach the following main goals [24]:

- search for the Standard Model Higgs from the experimental limit up to the theoretical 1 TeV/c^2 upper bound;

- perform precise direct and indirect measurements of the Standard Model observables to check the consistency of the model and look for deviations as signal for new physics;

- more in general, search for physics beyond the Standard Model, such as Supersymmetry, technicolour, extra-dimensions, etc..

3.2 Top quark physics

The Standard Model top quark is a spin $\frac{1}{2}$ and charge $\frac{2}{3}$ fermion transforming as a colour triplet under $\mathrm{SU(3)_C}$ of strong interactions.

The top mass, about 35 times larger than the mass of the b quark, is very close to the scale of electroweak symmetry breaking, raising many questions about the actual role of the heavier quark in the symmetry breaking, in the Higgs mechanism. An accurate study of the top quark and eventual anomalies on its production and decays could enlighten on the presence of eventual lighter particles and non standard couplings with other particles, leading to an first evidence of new physics beyond the standard model.

Furthermore, the top quark mass enters in the electroweak precision observables as an input parameter via quantum effect, i.e. loop corrections. As a distinctive feature, the large numerical value of $\mathrm{M_{top}}$ gives rise to sizeable corrections that behave as powers of $\mathrm{M_{top}}$, in contrast to the corrections associated with all other particles of the Standard Model. The top mass enters into the prediction of the W mass via loop corrections containing virtual top quarks, giving rise to terms proportional to M_{top}^2/M_Z^2. Moreover a precise measurement of the top quark mass, together with W mass, provides a constraint on the Higgs mass.

Top decays proceed through the channel $t \to Wb$ with a Branching Ratio (BR) of 0.99, yielding energetic b-jets. The fact that the electroweak decay is faster than the hadronization time scale implies that the top quark exists only as a free quark, so that the effects from new physics should show up very clearly by comparing measurements with the precise Standard Model predictions. Some SUSY particles and heavy resonances have the top quark as decay product: as a consequence the Standard Model production of the top quark is the background to many new physics channels.

3.2.1 Decay Width

The on-shell decay width Γ_t of the top quark is known with a theoretical accuracy ($< 1\%$): although not impressive, it is better than any foreseasable measurement.

Being the $t \to bW$ decay the dominant one by far, any further consideration will be restricted to this case.

It is useful to quantify the decay width in units of the lowest order decay with m_W and m_b set to zero and $|V_{tb}|$ is set to one:

$$\Gamma_0 = \frac{G_F m_t^3}{8\pi\sqrt{2}} = 1.67\,\mathrm{GeV} \tag{3.10}$$

At the leading order, m_W has to be considered:

$$\Gamma_{LO}(t \to bW)|V_{tb}|^2 = \Gamma_0 \left(1 - \frac{m_W^2}{m_t^2}\right)^2 \left(1 + 2\frac{m_W^2}{m_t^2}\right) = 0.885\Gamma_0 = 1.56\,\mathrm{GeV} \tag{3.11}$$

Using radiative corrections, known to the second order in QCD and to the first order in the EW theory, we obtain:

$$\Gamma(t \to bW)|V_{tb}|^2 \approx 0.887\Gamma_0 = 1.42\,\mathrm{GeV} \tag{3.12}$$

3.3 Top quark physics at the LHC

Several properties have been already studied at the Tevatron, such as kinematical properties of top quarks production, cross section, mass measurement, reconstruction of the decay final states. Most of them are anyway limited by the small sample of top quarks collected.

At the LHC, top quarks will be produced copiously, due to the large center-of-mass energy as well as the high luminosity. These samples can be used not only for precision measurements of Standard Model parameters such as m_W and m_t, but also for detector commissioning, alignment and calibration. Furthermore, Standard Model processes involving $W^{\pm}$, Z^0 bosons and top quarks constitute the primary sources of background in many Higgs boson and new physics searches.

The large top quark mass ensures that top production is a short-distance process and that the perturbative expansions, given by a series of powers of the small parameter $\alpha_s(\mathrm{M_{top}})\tilde{1}$, converges rapidly. Because of the large statistics that can be collected in a relatively short time, at LHC, the measurements of top events will be dominated by experimental and theoretical systematic errors. Accurate studies of top production and decay mechanism will provide interesting tests of QCD. An accurate measurement of the cross section will provide an independent indirect determination of the top quark mass. Asymmetries in the rapidity distribution of top and antitop are sensitive to the parton distribution function of the proton. Anomalies in the $\mathrm{t\bar{t}}$ rate would indicate the presence of non-QCD production channels, leading to scenarios beyond the Standard Model, as well as parity violating asymmetries.

3.3.1 $t\bar{t}$ production at LHC

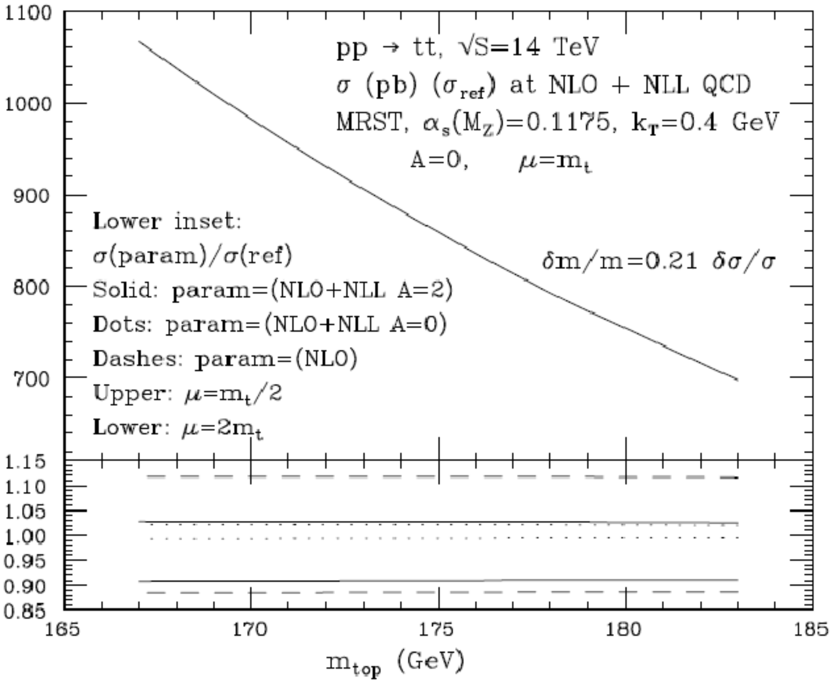

Figure 3.1: Production cross section calculated for top pair production at LHC [30].

The top pair production at LHC has been computed to be 488 pb at leading order and 833^{+52}_{-39} pb (±3.5% PDF error), at the next-to-leading order (Figure 3.1), about 100 times higher than the one at Tevatron. At low luminosity LHC will then produce 8×10^6 $t\bar{t}$/y (almost one top pair per second). The LHC will act as a top factory and will allow the top quark properties to be determined with significant precision by measuring observables in production and decay and exploiting all possible channels.

The dominant production mechanisms are gluon-gluon fusion (90%) and $q\bar{q}$ annihilation (10%). Within the Standard Model the top quark decays almost exclusively via $t\bar{t} \rightarrow WWb\bar{b}$.

The structure of the $t\bar{t}$ final state affects the direct determination of the mass. Initial (ISR) and final (FSR) state gluon radiation contribute to the amount of energy carried by the jets produced in the decay, and need to be taken into account when jets are combined to extract the top quark mass. The details of the structure of these jets, such as the fragmentation function and their shapes will influence the experimental determination of the jet energy scales as well as the determination of the efficiency with which b-jets are tagged.

3.3.2 $\mathrm{t\bar{t}}$ decay channels

The signature of the $\mathrm{t\bar{t}}$ system is classified according to the W^+W^- decay as dileptonic (10%), semi-leptonic (44%) or fully hadronic (46%).

The fully hadronic final state is characterized by the nominal six jets topology $\mathrm{t\bar{t} \rightarrow WWb\bar{b} \rightarrow qqqqb\bar{b}}$. It has the largest branching ratio, 46%, and kinematics can be fully reconstructed, but it is affected by a large background from QCD multi-jet production which makes the isolation of the signal rather challenging, and ISR and FSR. In addition the all-jet final state poses difficulties on the trigger so that an accurate study is needed to determine appropriate threshold on the six-jet topology.

Experiments at Fermilab, showed that is possible to isolate the signal from the background, just relying on selection cuts and a high b-tag efficiency.

The trigger menus examined so far by CMS consider multi-jet trigger thresholds for 1, 2 or 4 jets, for which a jet $\mathrm{E_T}$ threshold of 170, 80 and 55 GeV respectively is applied at low luminosity. Further studies are required to determine appropriate thresholds for a six-jet topology.

The lepton + jets channel is the golden channel for the measurement of top mass since it is easily triggered and has a BR of 29.6%, that is 2.5×10^6 events for a luminosity of $10\,\mathrm{fb^{-1}}$. The hadronically decaying top can be fully reconstructed by combining the two light quark jets into a W candidate (rescaled to the nominal W mass) and then adding one of the b-tagged jets. The leptonic decaying top can be partially reconstructed by imposing $\mathrm{E_T}(\nu) = \mathrm{E_T^{miss}}$ and $\mathrm{M_{l\nu} = M_W}$. The main background to this process arises from W+jets production and $\mathrm{t\bar{t}} \rightarrow \tau + \mathrm{X}$. The expected mass resolution is $1 \div 2\,\mathrm{GeV/c^2}$, where the main contributions to the overall uncertainty come from the b-jet energy scale and from the theoretical uncertainty on the FSR (Final State Radiation).

Another interesting analysis is based on the search for a J/ψ in the final state, which is easily reconstructed in the dimuon decay. The top mass depends on the invariant mass of the system lepton+J/ψ. This analysis is unrealistic at low luminosity, while it becomes promising at full luminosity with an expected sample of about 1000 events/y. The interesting feature of this analysis is that it's free from jet energy scale systematic uncertainty. The main limitation comes instead from the theoretical uncertainties on the fragmentation, limiting the expected precision to $1\,\mathrm{GeV/c^2}$.

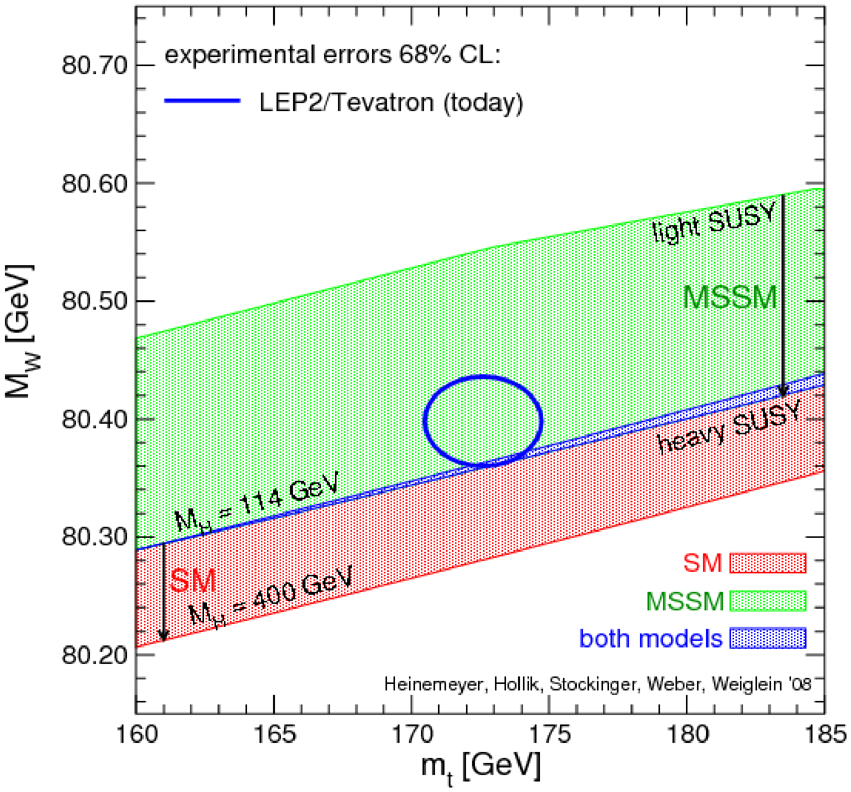

Figure 3.2: W and top quark masses measurements with prediction for different Higgs masses.

3.3.3 Top quark mass measurements

The mass of the top quark measured by CDF and D0 experiments at Tevatron from the direct observation of top events is [23]:

$$\mathrm{M_{top}} = 174.2 \pm 3.3 \mathrm{GeV}/c^2 \quad (3.13)$$

while the mass from the Standard Model electroweak fit is:

$$\mathrm{M_{top}} = 172.3^{+10.2}_{-7:6} \mathrm{GeV}/c^2 \quad (3.14)$$

The precision of the top quark mass measurement will be improved at LHC, being limited by systematic uncertainties. A combined precision (from different channels) on $\Delta\mathrm{M_{top}}$ less than 1 GeV/c^2 is achievable at the LHC. Within the precision of 15 MeV/c^2 on the W boson mass reaching at LHC, these measurements improve the error on $\ln(m_H)$ by a factor 2 compared to the current measurements.

Chapter 4

High Level Objects

This section will describe the ingredients needed for the study of the top quark and how they are reconstructed in CMS. Since the top quark decays have always hadron on the final states, jets are fundamental components for any analysis.

Although tracks reconstruction involves mainly the leptonic channels, they are important also in the fully hadronic channel, since they are used to tag jets arising from b quarks. Since b-tag will be an important instrument for the analysis, a brief description of the track reconstruction and the b-tag algorithm is given.

4.1 Jets

The huge QCD cross section ensures that jets will dominate high-p_{T} physics at the LHC. Jets will not only provide a benchmark for understanding the detector, but will also serve as an important tool in the search for physics beyond the Standard Model. Event signatures for SUSY, Higgs boson production, compositeness, and other new physics processes require accurate reconstruction and measurement of jets coming from high-p_{T} quarks and gluons. The problems with associating a jet measured in a calorimeter with a scattered parton is an old, persistent problem in hadron collisions. Jet energy resolution and response linearity are key factors in separating signal events from backgrounds. Missing transverse energy resolution, which historically has played an important role in the discoveries of the W-boson and the top quark and the search for new phenomena at hadron colliders, is closely related to the calorimeter jet energy response.

Readout cells in HCAL are arranged in a tower pattern in η, ϕ space, projective to the nominal interaction point. The cells in the barrel region have segmentation of $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$, becoming progressively larger in the endcap and forward regions. Since the ECAL granularity is much finer than HCAL, calorimeter towers (ECAL plus HCAL) are formed by addition of signals in η, ϕ bins corresponding to individual HCAL cells. In total there are 4176 such towers. The towers are used as input to several jet clustering algorithms. The energy associated with a tower is calculated as the sum of all contributing readout cells which pass the online zero-suppression threshold and any additional offline software thresholds. For the purpose of jet clustering, the towers are treated as massless particles, with the energy given by the tower energy, and the direction defined by the interaction point and the center of the tower.

Optimum performance of higher-level objects reconstructed from calorimeter towers requires careful selection of these inputs because calorimeter noise contributions can have significant impact on the reconstruction of low-$\mathrm{E_T}$ jets. For that reason $\mathrm{E_T}$ cuts ($\mathrm{E_T} > 0.5\,\mathrm{GeV}$) are used in jet reconstruction, eventually combined with energy cuts ($E > 0.8\,\mathrm{GeV}$) to eliminate more noise in central η region allowing a better reconstruction at low $\mathrm{E_T}$[40]. Further energy threshold at the individual cell level have been tested for every calorimeter region in order to refine the noise rejection. A 50% efficiency can be already reached for $\mathrm{E_T} \simeq 20\mathrm{GeV}$ while is near to 1 already at 30 GeV.

4.2 Jet Reconstruction

The first jet algorithms for hadron physics were based on simple cones [41]. Over the last two decades, clustering techniques have greatly improved in sophistication. Three principal jet reconstruction algorithms have been coded and studied for CMS: the iterative cone [42], the midpoint cone [43] and the inclusive k_T jet algorithm [44] [45]. The midpoint-cone and k_T algorithms are widely used in offline analysis in current hadron collider experiments, while the iterative cone algorithm is simpler and faster and commonly used for jet reconstruction in software-based trigger systems.

The jet algorithms may be used with one of two recombination schemes for adding the constituents. In the energy scheme, constituents are simply added as four-vectors. This produces massive jets. In the $\mathrm{E_T}$ scheme, massless jets are produced by equating the jet transverse momentum to the $\Sigma\mathrm{E_T}$ of the constituents and then fixing the direction of the jet in one of two ways: 1) $\sin\theta = \mathrm{E_T}/\mathrm{E}$ where E is the jet energy (usually used with cone algorithms), or 2) $\eta = \Sigma\mathrm{E_{Ti}}\eta_\mathrm{i}/\Sigma\mathrm{E_T}$ and

$\phi = \mathrm{E_{Ti}}\phi_\mathrm{i}/\Sigma\mathrm{E_T}$ (usually used with the k_T algorithm). In all cases the jet $\mathrm{E_T}$ is equal to $\hat{p}_\mathrm{T}$.

The inclusive k_T algorithm merges, in each iteration step, input objects into possible final jets and so the new jet quantities, the jet direction and energy, have to be calculated directly during the clustering. The cone jet algorithms, iterative and midpoint, group the input objects together as an intermediate stage and the final determination of the jet quantities (recombination) is done in one step at the end of the jet finding.

4.2.1 Iterative cone

In the iterative cone algorithm, an $\mathrm{E_T}$-ordered list of input objects (particles or calorimeter towers) is created. A cone of size R in η-ϕ space is cast around the input object having the largest transverse energy above a specified seed threshold. The objects inside the cone are used to calculate a proto-jet direction and energy using the $\mathrm{E_T}$ scheme. The computed direction is used as seed to a new proto-jet. The procedure is repeated until the energy of the proto-jet changes by less than 1% between iterations and the direction of the proto-jet changes by $\Delta\mathrm{R} < 0.01$. When a stable proto-jet is found, all objects in the proto-jet are removed from the list of input objects and the stable proto-jet is added to the list of jets. The whole procedure is repeated until the list contains no more objects with an $\mathrm{E_T}$ above the seed threshold. The cone size and the seed threshold are parameters of the algorithm. When the algorithm is terminated, a different recombination scheme may be applied to jet constituents to define the final jet kinematic properties.

4.2.2 Midpoint cone

The midpoint-cone algorithm was designed to facilitate the splitting and merging of jets. The midpoint-cone algorithm also uses an iterative procedure to find stable cones (proto-jets) starting from the cones around objects with an $\mathrm{E_T}$ above a seed threshold. In contrast to the iterative cone algorithm described above, no object is removed from the input list. This can result in overlapping proto-jets (a single input object may belong to several proto-jets). To ensure the collinear and infrared safety of the algorithm, a second iteration of the list of stable jets is done. For every pair of proto-jets that are closer than the cone diameter, a midpoint is calculated as the direction of the combined momentum. These midpoints are then used as additional seeds to find more proto-jets. When all proto-jets are found, the splitting and merging procedure is applied, starting with the highest $\mathrm{E_T}$ proto-jet. If the proto-jet does not share objects with other proto-jets, it is defined as a jet and removed from the proto-jet list. Otherwise, the transverse energy shared with the highest $\mathrm{E_T}$ neighbor proto-jet is compared to the total transverse energy of this neighbor proto-jet. If the fraction is greater than a value f (typically 50%) the proto-jets are merged, otherwise the shared objects are individually assigned to the proto-jet that is closest in η-ϕ. The procedure is repeated, again always starting with the highest-$\mathrm{E_T}$ proto-jet, until no proto-jets are left. This algorithm implements the energy scheme to calculate the proto-jet properties but a different recombination scheme may be used for the final jet. The parameters of the algorithm include a seed threshold, a cone radius, a threshold f on the shared energy fraction for jet merging, and also a maximum number of proto-jets that are used to calculate midpoints.

4.2.3 Inclusive k_T

The inclusive k_T jet algorithm is a cluster-based jet algorithm. The cluster procedure starts with a list of input objects, stable particles or calorimeter cells. For each object i and each pair (i, j) the following distances are calculated:

$$d_i = (\mathrm{E}_{\mathrm{T},i})^2 R^2, \tag{4.1}$$

$$d_{ij} = min\{\mathrm{E}_{\mathrm{T},i}^2, \mathrm{E}_{\mathrm{T},i}^2\} R_{ij}^2 \;\; with \;\; R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2 \tag{4.2}$$

where R^2 is a dimensionless parameter normally set to unity. The algorithm searches for the smallest d_i or d_{ij} . If a value of type d_{ij} is the smallest, the corresponding objects i and j are removed from the list of input objects. They are merged using one of the recombination schemes listed below and filled as one new object into the list of input objects. If a distance of type d_i is the smallest, then the corresponding object i is removed from the list of input objects and filled into the list of final jets. The procedure is repeated until all objects are included in jets. The algorithm successively merges objects which have a distance $\mathrm{R_{ij}} < \mathrm{R}$. It follows that $\mathrm{R_{ij}} > \mathrm{R}$ for all final jets i and j.

4.2.4 CMS algorithms

Four basic algorithms are implemented in CMS:

- Midpoint Cone, from CDF implementation, seeded with $\mathrm{E_T} > 1\,\mathrm{GeV}$, cone size 0.5 or 0.7 and overlap threshold f = 0.75

- Iterative Cone, seeded with $\mathrm{E_T} > 1\,\mathrm{GeV}$ and cone size 0.5 or 0.7
- k_T from external package (Butterworth, Cox, Waugh), with $\mathrm{D} = 1$
- k_T FastJet from external package (Salam, Cacciari), with $\mathrm{D} = 1$ or $\mathrm{D} = 0.6$

4.3 Jet Calibration

The jet energy resolution is influenced by a multitude of physics and detector effects: gluon radiation in both initial and final states, underlying events, pile-up, uncertainties in the jet fragmentation models and in general the complexity of the hadronization process due to the increasing of the strong coupling constant at smaller energies while the hadronic shower evolves. On the other hand, effects due to the detector layout such as out of cone showering, jet containment and separation, low resolution for low $\mathrm{E_T}$ and in general non-linearity of the response of calorimeters, different response of the HCAL from ECAL (jets have both hadronic and e/γ components), the electronic noise, dead materials and cracks and so on have to be taken into account. Other effects are related to the magnetic field: particles with $p_\mathrm{T} < 0.8\,\mathrm{GeV/c}$ loops in the barrel, while up to $p_\mathrm{T} < 1.6\,\mathrm{GeV/c}$ have large deflection (more than 0.5 radians).

The jet calibration is performed using physics analysis, with the major objective to parameterize calibration parameters as function of direction, energy and flavour of the jet.

The HCAL calibration system will be used to set the initial absolute energy scale, understand the detector response and uniformity and to monitor the time stability during data taking.

The initial calibration came from the quality control test performed with a collimated radiation source for the scintillating tiles quality test and the validation of the digital converters.

The energy scale is obtained combining test beam data taken with $e^\pm$, $\pi^\pm$ and muon beams with radiation source for a limited number of modules, then translated for the full system.

The initial test beam calibration is made without magnetic field which instead will influence the showering corrections mainly in the transition regions barrel-endcap and endcap-forward. Furthermore at high luminosity the endcaps response will be degraded by radiation damages.

For that reason re-calibration and updates of the scale constants will come from the analysis of physics events and a constant monitoring of each full channel and electronic response. In particular minimum bias events will provide high statistics test of the uniformity of the energy scale in ϕ; higher order moments of the energy distribution will help to detect effects of miscalibration. The statistics needed to calibrate each tower better than 2% is collected in $1 \div 2$ hours.

The barrel and part of the endcap, up to $|\eta| < 2.4$ are covered by the tracker and can be calibrated with isolated energetic particles from events such as from decays of the form $\tau \rightarrow \pi\nu$ from $W \rightarrow \tau\nu$ and $Z, \gamma^* \rightarrow \tau\tau$ or isolated energetic particles from QCD-jets. Both methods allow calibration better than 2% during one month from charged particles with transverse momentum from $15\,\mathrm{GeV/c^2}$ to $70\,\mathrm{GeV/c}$ using the E/p ratio. The $|\eta| > 2.4$ jets are calibrated using $\mathrm{E_T}$-balance in di-jet events or $\gamma/Z + jet$ events.

4.3.1 Jet Corrections

Current jet corrections for Monte Carlo productions are derived from studies over the jet response which were made with fully simulated QCD di-jets events (without pile-up) over the range $0 < p_\mathrm{T} < 4000\,\mathrm{GeV/c}$ reconstructed with all the three techniques using the $\mathrm{E_T}$-scheme.

Comparisons between Monte Carlo simulation particle-level and reconstructed jets were made by applying the same jet algorithm to stable particles (excluding neutrinos and muons) and calorimeter cells, respectively. A matching criterion, based on the distance $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ was used to associate Monte Carlo particle-level and reconstructed jets. The data were divided into η bins where the ratio of reconstructed jet transverse energy ($\mathrm{E}_\mathrm{T}^{rec}$) to the Monte Carlo particle-level jet transverse energy ($\mathrm{E}_\mathrm{T}^{MC}$), as a function of $\mathrm{E}_\mathrm{T}^{MC}$ was fit using an iterative procedure.

Corrections are provided in the framework implementation as final object collections and correction values to be applied on the fly, scaling the jets Lorentz vectors such as the final average correction is 1.

Those correction are blindly applied to the analysis.

On the other hand, an analysis of individual jet flavors (uds, c, b, gluon) shows that different corrections are needed. Light quarks, b and gluon generated jets have different cone size and fragmentation properties and so the original parton can be guessed from the study of the jet. Flavor specific correction can be in that way evaluated and used to improve event selection and specific studies in analysis involving jets, in particular when the identification of a b-jet can improve the event topology reconstruction.

The flavor level corrections can be applied once the nature of the jet is established

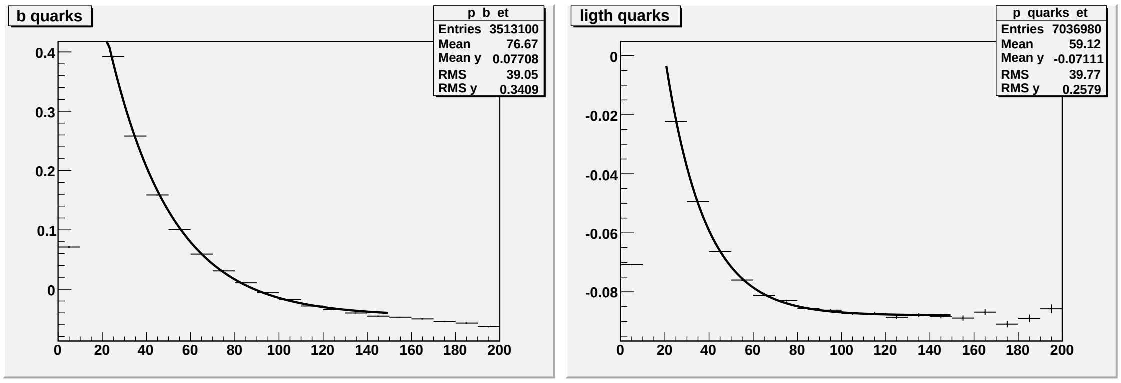

Figure 4.1: $\mathrm{E_T}$ deviation of b-jets and light quarks jets from the generated partons $\mathrm{E_T}$.

or guessed from the jet characteristics or the analysis path.

Flavor corrections are currently under implementation in the framework, and some versions are deployed as deviation from the average QCD jets. For that reason they do not guarantee that the reconstructed mass peak is perfectly superimposed to the one obtained with the generated partons.

This analysis anyway does not make use of those corrections, due to the preliminary stage of their implementation until very advanced stages of the work. For that purpose, flavor level corrections were obtained with similar techniques from the analysis itself. We run the kinematic fitter over the whole $\mathrm{t\bar{t}}$ sample, using also parton level information. We provided a geometric association using $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ with jets passing the basic trigger selection. For those jets, we extracted the $\mathrm{E_T}$ deviation from the associated partons, separately for light quarks and b. The deviation is defined as:

$$\frac{\mathrm{E_T}^{parton} - \mathrm{E_T}^{jet}}{\mathrm{E_T}^{jet}} \tag{4.3}$$

An exponential fit of the form:

$$f(\mathrm{E_T}) = e^{a+b\mathrm{E_T}} + c \tag{4.4}$$

is found, whose parameters are shown in table 4.1

Parameter	b	light quarks
a	2.45 $\times 10^{-2} \pm 5.2 \times 10^{-3}$	-1.32 $\times 10^{-2} \pm 2.2 \times 10^{-2}$
b	-3.51 $\times 10^{-2} \pm 1.5 \times 10^{-4}$	-5.57 $\times 10^{-2} \pm 7.4 \times 10^{-4}$
c	-4.54 $\times 10^{-2} \pm 3.8 \times 10^{-4}$	-8.80 $\times 10^{-2} \pm 2.0 \times 10^{-4}$

Table 4.1: Parameters for the flavor-dependent corrections.

4.4 Tracks

The study of all top quark decay channels rely on tracks reconstruction, both directly as for semileptonic channels, or through the usage of b-tagging algorithms. Even if we will not use directly tracks information, they play a relevant role in the top quark physics so we are going to give a brief description of tracks reconstruction and lepton identification in the following sections.

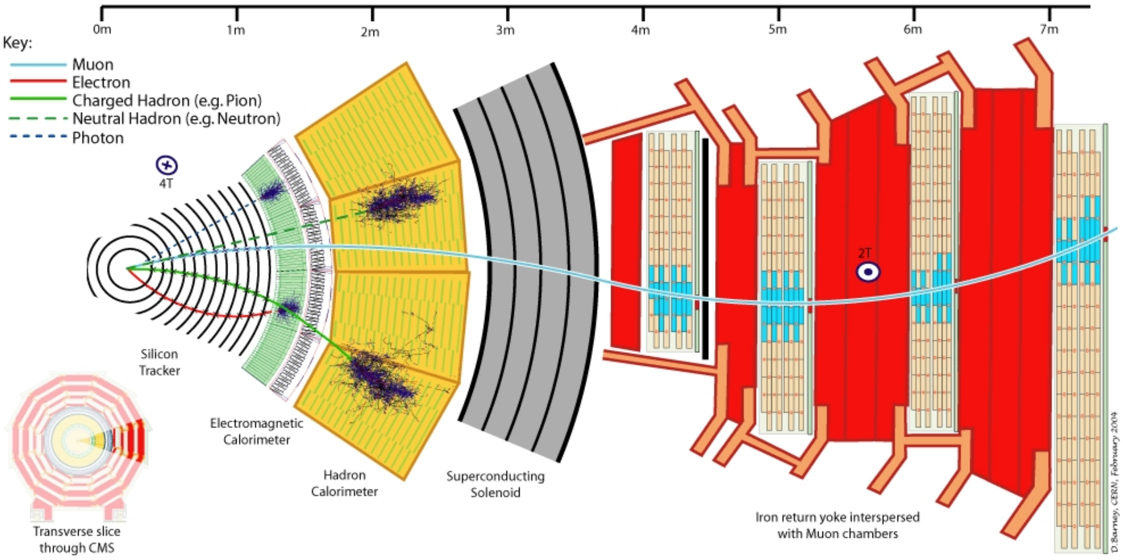
0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D Barney, CERN, February 2004

Figure 4.2: Particle tracks in the CMS slice.

4.4.1 Tracks reconstruction

The CMS Tracker plays a central role for the tracks reconstruction. The characteristics of being practically hermetic for a particle originating from the centre of the detector, having a magnetic field almost constant in a large part of the tracker volume and concentrating most of the support structure in the tracking layers, close to the sensors, allow an efficient search for hits during the pattern recognition stage, and a fast propagation of trajectory candidates.

The reconstruction accounts for a typical step length, for propagation of track parameters, of the order of the distance between 2 layers and an attribution of material to layer instead of the detailed distribution of passive material, in order to simplify the estimation of energy loss and multiple scattering, which can be done at the position of the sensitive elements without requiring additional propagation steps. The track reconstruction is decomposed into 5 logical parts:

- Hit reconstruction, which in turn consists of clustering of strips or pixels and estimating a position and its uncertainty.
- Seed generation, providing initial trajectory candidates for the full track reconstruction. They can be obtained externally to the Tracker, using inputs from other detectors, with a poor initial trajectory parameters precision, or internally. In this case each seed is composed from the set of reconstructed hits that are supposed to come from 1 charged particle track. Five parameters are needed to start trajectory building, e.g. 3 hits, or 2 hits and a beam constraint which will be anyway removed during the final fit.
- Pattern recognition, or trajectory building, based on a combinatorial Kalman filter method. The filter proceeds iteratively from the seed layer, starting from a coarse estimate of the track parameters provided by the seed, and including the information of the successive detection layers one by one. On each layer, the track parameters are known with a better precision, up to the last point, where they include the full tracker information. At each step multiple trajectory candidates are created, since several hits on the new layer may be compatible with the predicted trajectory, plus an additional one, in which no measured hit is used, to account for the possibility that the track did not leave any hit on that particular layer. Trajectory candidates are grown in parallel in order not to bias the result, and limited in number by criteria based on their normalized χ^2 and number of valid and invalid hits.
- Ambiguity resolution, based on the fraction of hits that are shared between 2 trajectories; ambiguities in track finding arise because a given track may be reconstructed starting from different seeds, or because a given seed may result in more than 1 trajectory candidate. These ambiguities, or mutually exclusive track candidates, must be resolved in order to avoid double counting of tracks.

- Final track fit using collected hits and track parameters. The trajectory is refitted to remove eventual constraint using a least-squares approach, implemented as a combination of a standard Kalman filter complemented with a smoother. The Kalman filter is initialized at the location of the innermost hit with an estimate obtained during seeding, then proceeds in an iterative way through the list of hits. For each valid hit the position estimate is re-evaluated using the current values of the track parameter and the trajectory is propagated to the surface associated with the next hit. The smoothing stage is made of a second filter, initialized with the result of the first one and running backward toward the beam line. This procedure yields optimal estimates of the parameters at the surface associated with each hit and, specifically, at the first and the last hit of the trajectory. Estimates on other surfaces are then derived by extrapolation from the closest hit.

4.4.2 Muon reconstruction and identification

The muon reconstruction is performed using both the muon system and the silicon tracker, using the concept of regional reconstruction in order to allow its use in both the offline reconstruction and the High-Level Trigger (online event selection).

Tracks reconstruction is not performed ever the entire tracker, but only in that part which can possibly be involved in the reconstruction of a charged particle track compatible with the hits in the muon chambers. The method depends strongly on the identification of a good "seed", providing initial values of the 5 trajectory parameters and their errors, that can start the reconstruction with high efficiency and reliability.

For offline reconstruction a seed-generation algorithm has been developed, which performs local reconstruction in the entire muon system and uses patterns of segments reconstructed in the CSC and/or DT chambers as initial seeds.

Muon reconstruction is performed in 3 stages:

- local reconstruction (local-pattern recognition) from a seed identification in a specific chamber, leading to a track segment.

- standalone reconstruction, using only information from the muon system. In the barrel DT chambers, reconstructed track segments are used as measurements in the Kalman-filter procedure. In the endcap CSC chambers, where the magnetic field is inhomogeneous, the individual reconstructed constituents (three-dimensional hits) of the segments are used instead. Reconstructed hits from the RPC chambers are also included. A backward Kalman filter is then applied, working from outside in, and the track parameters are defined at the innermost muon station. Finally, the track is extrapolated to the nominal interaction point and a vertex-constrained fit to the track parameters is performed.

- global reconstruction, extending the muon trajectories to include hits in the silicon tracker. Starting from a stand-alone reconstructed muon, the muon trajectory is extrapolated from the innermost muon station to the outer tracker surface. Silicon layers compatible with the muon trajectory are then determined, and a region of interest within them is defined in which to perform regional track reconstruction. The determination of the region of interest is based on the track parameters and their corresponding uncertainties of the extrapolated muon trajectory, obtained with the assumption that the muon originates from the interaction point. Inside the region of interest, initial candidates for the muon trajectory (regional seeds) are built from pairs of reconstructed hits. The 2 hits forming a seed must come from 2 different tracker layers, both from pixel or silicon strip layers. In addition, a relaxed beam-spot constraint is applied to track candidates above a given transverse momentum threshold to obtain initial trajectory parameters, use as seed for the track-reconstruction inside the selected region of interest.

4.4.3 Electrons

The electrons reconstruction relies both on the tracker and on the energy deposit in several crystals in the ECAL. Approximately 94% of the incident energy of a single electron or photon is contained in 33 crystals, and 97% in 55 crystals. Summing the energy measured in such fixed arrays gives the best performance for unconverted photons, or for electrons. The presence in CMS of material in front of the calorimeter results in bremsstrahlung and photon conversions. Furthermore, because of the strong magnetic field the energy reaching the calorimeter is spread in ϕ . The spread energy is clustered by building a cluster of clusters, called “supercluster”, which is extended in ϕ.

The cluster identification is made through two algorithms: “hybrid” and “Island”. The hybrid algorithm was designed to reconstruct relatively high energy electrons in the barrel, then was tuned to allow efficient reconstruction of electron showers down to $p_{\mathrm{T}} = 5\,\mathrm{GeV/c}$. It uses the $\eta - \phi$ geometry of the barrel crystals to exploit the knowledge of the lateral shower shape in the η direction (taking a fixed bar of 3 or 5 crystals in η), while searching dynamically for separated energy in the ϕ-direction.

The Island algorithm is more appropriated when looking for small deposits of energy in individual clusters, for example when making a calorimetric isolation cut. It starts by a search for crystals with an energy above a certain threshold. Using them as seed position, adjacent crystals are examined, scanning first in η and then in ϕ. Along each scan line, crystals are added to the cluster until a rise in energy or crystal that has not been read out is encountered. In the same way as energy is clustered at the level of calorimeter cells or crystals, non-overlapping Island clusters can be clustered into superclusters. The procedure is seeded by searching for the most energetic cluster and then collecting all the other nearby clusters in a very narrow η-window, and much wider ϕ-window.

4.5 b-tagging

The top quark decays almost exclusively into a W-boson and b-quark. The capability to identificate the jet produced by the b-quark improves the isolation and identification of top decays as well as many other physics channels, enabling the rejection of lower energetics events.

The inclusive tagging of b-jets relies upon relatively distinct properties of b-hadrons such as large proper lifetime ($\tau \approx 1.5\,\mathrm{ps}$, $c\tau \approx 450\,\mu\mathrm{m}$), large mass, decays to final states with high charged track multiplicities (on average 5) relatively large semileptonic branching ratio (about 19%) and a hard fragmentation function.

In CMS, algorithms for b-tagging can be applied not only offline, but also in the High Level Trigger, and for that reason they should be efficient also in terms of computing resources and may not rely on fully reconstructed objects. Three techniques are used:

1. take advantage of the relatively large semileptonic branching ratio, identifying electrons and muons arising from jets cone.
2. use secondary vertex identification combined with 3)
3. use of tracks multiplicity through impact parameter measurement [46]

4.5.1 Track counting impact parameter based b-tagging

The main advantage of the method is simplicity, since it relies upon the selection of good quality tracks and cut on impact parameters significance, without the need of further steps such as reconstruction of secondary vertex. For that reason it can be used for on-line b-tag based selection at the trigger level.

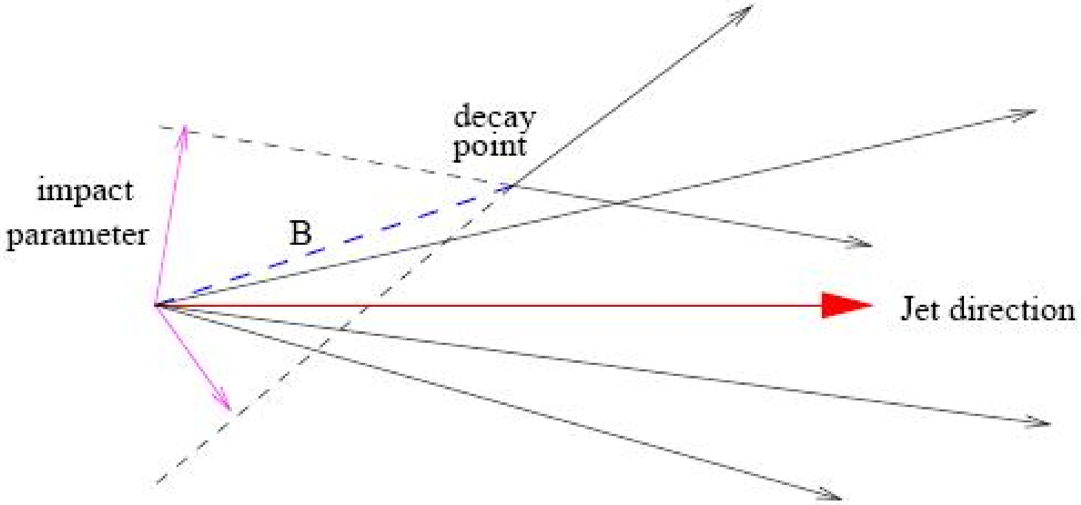

Figure 4.3: Representation of an hadronic jet originating from a b-quark.

Tracks originated from B decays have large impact parameter since they come from a displaced vertex, while the impact parameter of tracks coming from the primary vertex are compatible with 0, within the track reconstruction resolution.

The B direction is reconstructed using the axis of the jet cone, improved with the sum of the momenta of tracks associated to the jet with high quality. The primary vertex identification information are then needed.

To take into account the experimental resolution, the track impact parameter significance is used, defined as the ratio between the track impact parameter and its uncertainty.

The track counting method was originally based on the simple requirement of a minimum number of good quality tracks with an impact parameter significance exceeding a given threshold. Then it has been integrated with the usage of a continuous tagging variable, a *discriminator*, defined as the significance of the n^{th} track, once tracks are ordered by decreasing impact parameter.

The algorithm performances are limited by inefficiencies in track reconstruction, resolution of track parameters, efficiency in reconstructing the primary vertex. Mistagging can be produced by secondary interactions and decay of long-lived particles as well as superimposing of pile-up events tracks.

Chapter 5

Event Samples and High Level Trigger

The goal of the analysis is to reconstruct the top quark mass from the fully hadronic channel, which has the advantages of the largest branching ratio fraction, 46% and that kinematics can be fully reconstructed. It is anyway affected by a large background from QCD multi-jet production which makes the isolation of the signal rather challenging. For that reason a multi-jet trigger is required, allowing a good rejection of the background, in order to reduce the final rate, at least to fit with the High Level Trigger request.

In this section signal and background samples are described and a set of thresholds are evaluated for the multi-jet trigger for the online selection. The trigger thresholds given here have the effect to give a realistic starting point, even if they can be better evaluated in future developments and integrated with an online b-tagging to fit with more strict requests.

5.1 Monte Carlo and Simulation tools

The CMS offline software framework provides tools both for the analysis and the full Monte Carlo production chain. Physics objects are generated through Monte Carlo generators such as PYTHIA [37], then propagated into the CMS detector, whose response is based on GEANT [39]; the digitization then will include the response of the detector and electronics in the channel "hits". Digitized samples are then used for the reconstruction of high level objects such as tracks and jets used for this analysis.

5.2 Signal and background Monte Carlo Samples

For the analysis, $\mathrm{t\bar{t}}$ events are used from the official CMS production, generated at the leading order with cross section $\sigma = 0.488\,\mathrm{nb}$. In our calculation, however, the next-to-leading order production cross section will be $\sigma = 0.833\,\mathrm{nb}$. The event generator is TopRex 4.11 [38] for pythia 6.227.

TopRex provides the simulation of several important processes in pp and $\mathrm{p\bar{p}}$ collisions, not implemented in pythia. Some of these processes include top quarks whose spin polarizations are taken into account in the subsequent decay of the top quarks. Several non-SM top quark decay channels are included, too. All calculated subprocesses can be accessed from pythia as external processes. In addition, TopRex can be used as stand-alone event generator, providing partonic final states before showering. In this mode the control of the event generation is taken by TopRex itself.

The QCD sample used are generated using pythia 6.409 from the official CMS production, divided in 21 samples with different ranges of the hard scatter $\hat{p}_{\mathrm{T}}$.

Table 5.1 shows the main characteristics of the samples used for the official production, i.e. $\hat{p}_{\mathrm{T}}$ range, events generated and generation cross section.

5.2.1 Mass Scan Samples

The official production configuration was also used to build samples with top quark mass different than $175\,\mathrm{GeV/c^2}$, needed for the procedure used for the top mass measurement. The 21 generated samples consist of 100000 events with mass ranging from $165\,\mathrm{GeV/c^2}$ to $185\,\mathrm{GeV/c^2}$ in steps of $1\,\mathrm{GeV/c^2}$.

5.3 Multi-Jet trigger

For the sake of the fully hadronic decay channel study, a multi-jet trigger needs to be set up, defining a set of cut-off values for physics observables that can be used both to discard not interesting events or applied offline for the final analysis.

A set of cut-off values have been evaluated using QCD and top signal rates at $\mathcal{L} = 2 \times 10^{33}\ \mathrm{cm^{-2}\,s^{-1}}$ to favor the fully hadronic top decay candidates events,

Sample	Generated	Cross section (nb)
$\mathrm{t\bar{t}}$ ($m_t = 175$ GeV/c^2)	2767326	0.833
QCD		
$\hat{p}_{\mathrm{T}} < 15$ GeV/c	601338	5.52×10^{6}
$15 < \hat{p}_{\mathrm{T}} < 20$ GeV/c	1261976	1.46×10^{6}
$20 < \hat{p}_{\mathrm{T}} < 30$ GeV/c	2147944	6.32×10^{5}
$30 < \hat{p}_{\mathrm{T}} < 50$ GeV/c	1152979	1.63×10^{5}
$50 < \hat{p}_{\mathrm{T}} < 80$ GeV/c	893240	2.16×10^{4}
$80 < \hat{p}_{\mathrm{T}} < 120$ GeV/c	1243257	3.08×10^{3}
$120 < \hat{p}_{\mathrm{T}} < 170$ GeV/c	1260951	4.94×10^{2}
$170 < \hat{p}_{\mathrm{T}} < 230$ GeV/c	934870	1.01×10^{2}
$230 < \hat{p}_{\mathrm{T}} < 300$ GeV/c	800840	2.45×10^{1}
$300 < \hat{p}_{\mathrm{T}} < 380$ GeV/c	1272037	6.24
$380 < \hat{p}_{\mathrm{T}} < 470$ GeV/c	781003	1.78
$470 < \hat{p}_{\mathrm{T}} < 600$ GeV/c	1317613	6.83×10^{-1}
$600 < \hat{p}_{\mathrm{T}} < 800$ GeV/c	592580	2.04×10^{-1}
$800 < \hat{p}_{\mathrm{T}} < 1000$ GeV/c	718458	3.51×10^{-2}
$1000 < \hat{p}_{\mathrm{T}} < 1400$ GeV/c	615085	1.09×10^{-2}
$1400 < \hat{p}_{\mathrm{T}} < 1800$ GeV/c	298782	1.06×10^{-3}
$1800 < \hat{p}_{\mathrm{T}} < 2200$ GeV/c	314815	1.45×10^{-4}
$2200 < \hat{p}_{\mathrm{T}} < 2600$ GeV/c	764396	2.38×10^{-5}
$2600 < \hat{p}_{\mathrm{T}} < 3000$ GeV/c	752036	4.29×10^{-6}
$3000 < \hat{p}_{\mathrm{T}} < 3500$ GeV/c	512868	8.44×10^{-7}
$\hat{p}_{\mathrm{T}} > 3500$ GeV/c	556968	1.08×10^{-7}
TOTAL QCD	18794036	5.748×10^{7}

Table 5.1: Signal and Background Monte Carlo samples.

so that the effective signal loss is negligible, while a significant QCD background is discarded.

The physics observable is mainly the jet $\mathrm{E_T}$ estimated for the four leading jets. Thresholds are evaluated starting from best statistical significance cuts (see Section 6.2) and increasing their value to reach the desired rate.

The design HLT rate is below 100 Hz, reachable with most of the algorithms used offline. Thresholds over jet transverse energy, total event transverse energy, jet multiplicity and b-tag algorithms can be used with online object, which are evaluated using a coarse detectors segmentation and less refined reconstruction

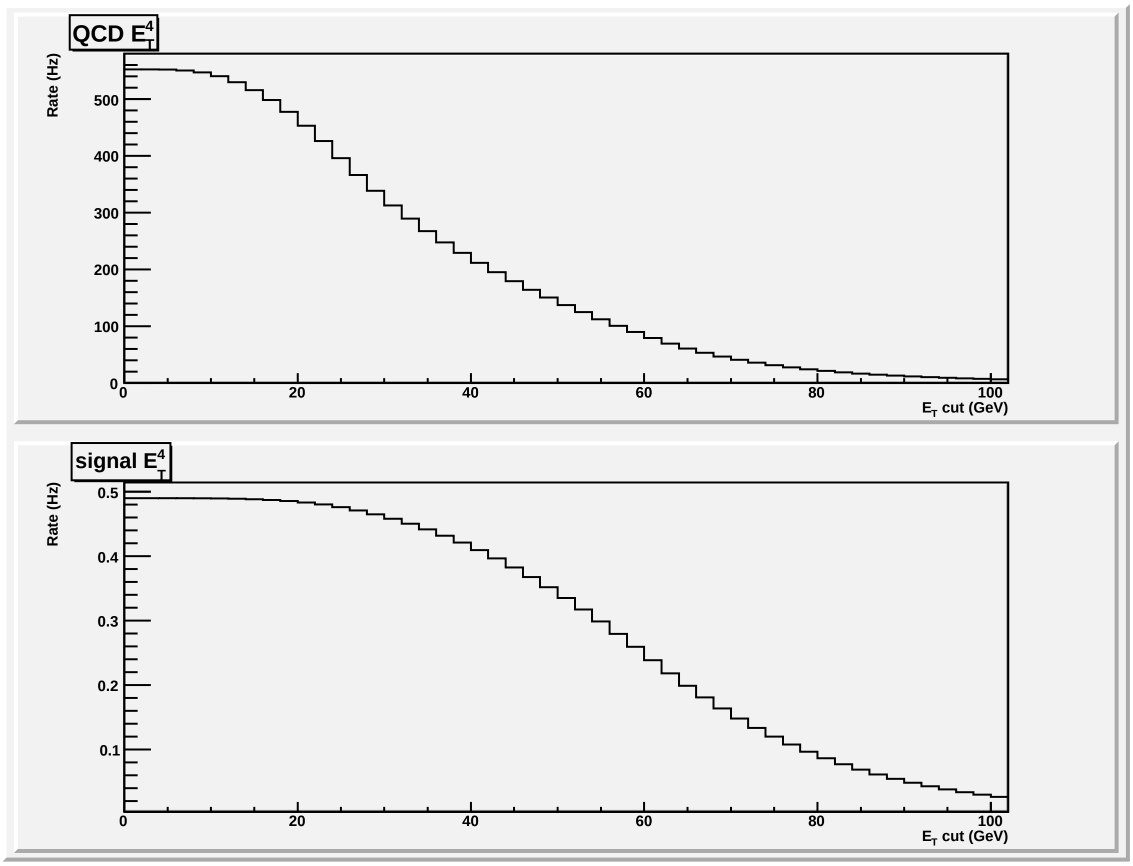

Figure 5.1: QCD rate (top) and $t\bar{t}$ efficiency (bottom) as a function of the threshold on the fourth jet.

algorithms. For the sake of the analysis we have available offline reconstructed objects, so we have to emulate the effect of the trigger by using offline cuts to reduce the total rate. As a first step aim to reach a rate of the order of 100 Hz. In order to reach a valid starting point, we need to cut on the leading jets with the

Parameter	threshold
E_T^1	115 GeV
E_T^2	70 GeV
E_T^3	60 GeV
E_T^4	40 GeV

Table 5.2: Trigger thresholds.

thresholds shown in table 5.2. We are aware that any b-tag request will cut off the most part of QCD background while keeping high the $\mathrm{t\bar{t}}$ efficiency, but we perform offline this selection, since we have to deal with the pour statistic of the mass scan samples which can affect our results. We performed test to ensure that all the work made, included the kinematic selection, keeps its validity with stronger trigger cuts.

5.3.1 QCD Rates

The effect of the multi-jet trigger can be evaluated starting from the QCD production rates at a given luminosity.

We use the effective cross section $\hat{\sigma}$, defined as:

$$\hat{\sigma} = \sigma \times \epsilon$$

where ϵ is the efficiency of the selection. The production rate R can be then expressed as:

$$R = \hat{\sigma} \times \mathcal{L} = \sigma \times \epsilon \times \mathcal{L}$$

If we focus on the initial LHC luminosity $\mathcal{L} = 2 \times 10^{33}\ \mathrm{cm^{-2}\,s^{-1}}$, the resulting rate would be:

$$R(Hz) = 2 \times \hat{\sigma}(nb) = 2 \times \epsilon \times \sigma(nb)$$

The resulting QCD rates for the different p_{T} ranges, as well as the total QCD rate are shown in table 5.3.

The expected QCD rate is still high but we know that a reduction up to a factor ~ 10 can be easily achieved by requiring a b-tag.

5.3.2 $\mathrm{t\bar{t}}$ efficiency

The $\mathrm{t\bar{t}}$ sample consists of 2767326 events, whose 46% is known to have a fully hadronic decay from Monte Carlo simulation, according to the expected branching ratio. After the preliminary event selection based on the multi-jet trigger described previously, the sample is reduced to the 25% of the total (54% of the fully hadronic set).

This results in an effective cross section of about $\hat{\sigma} = 205\,\mathrm{pb}$, with a signal to background ratio (S/B) of about S/B = 1/517.

The effect of the trigger on the $\mathrm{t\bar{t}}$ signal, the QCD background and their ratio is summarized in table 5.4.

p_{T} range	Trigger Rates (Hz)
$\hat{p}_{\mathrm{T}} < 15$ GeV/c	0
$5 < \hat{p}_{\mathrm{T}} < 20$ GeV/c	0
$20 < \hat{p}_{\mathrm{T}} < 30$ GeV/c	0
$30 < \hat{p}_{\mathrm{T}} < 50$ GeV/c	0
$50 < \hat{p}_{\mathrm{T}} < 80$ GeV/c	10
$80 < \hat{p}_{\mathrm{T}} < 120$ GeV/c	66
$120 < \hat{p}_{\mathrm{T}} < 170$ GeV/c	75
$170 < \hat{p}_{\mathrm{T}} < 230$ GeV/c	39
$230 < \hat{p}_{\mathrm{T}} < 300$ GeV/c	14
$300 < \hat{p}_{\mathrm{T}} < 380$ GeV/c	4.5
$380 < \hat{p}_{\mathrm{T}} < 470$ GeV/c	1.5
$470 < \hat{p}_{\mathrm{T}} < 600$ GeV/c	0.62
$600 < \hat{p}_{\mathrm{T}} < 800$ GeV/c	0.20
$800 < \hat{p}_{\mathrm{T}} < 1000$ GeV/c	0.037
$1000 < \hat{p}_{\mathrm{T}} < 1400$ GeV/c	0.012
$1400 < \hat{p}_{\mathrm{T}} < 1800$ GeV/c	0.0012
$1800 < \hat{p}_{\mathrm{T}} < 2200$ GeV/c	0.00016
$2200 < \hat{p}_{\mathrm{T}} < 2600$ GeV/c	2.6×10^{-5}
$2600 < \hat{p}_{\mathrm{T}} < 3000$ GeV/c	4.6×10^{-6}
$3000 < \hat{p}_{\mathrm{T}} < 3500$ GeV/c	8.7×10^{-7}
$< \hat{p}_{\mathrm{T}} > 3500$ GeV/c	1.1×10^{-7}
TOTAL QCD	211

Table 5.3: QCD rates after the multi-jet trigger.

	Trigger
$\hat{\sigma}_{QCD}(pb)$	106×10^3
$\hat{\sigma}_{\mathrm{t\bar{t}}}(pb)$	205
efficiency	25%
S/B	1/517

Table 5.4: Effect of the multi-jet trigger on the efficiency and S/B.

Chapter 6

Kinematical Selection and b-tagging

After the specific multi-jet trigger selection, the effective cross section of all $\mathrm{t\bar{t}}$events amounts to 205 pb, corresponding to an efficiency of about 25%, corresponding to the 54% of the fully hadronic set.

The QCD events effective cross section amounts to 106 nb, which is still about three order of magnitude above the signal.

A further selection is then needed in order to achieve a reasonably good signal to background ratio, allowing a good top mass measurement.

The selection is made before the usage of the b-tagging which will give later an additional contribution to reach a valid starting point for the actual analysis. We stress that part of this selection can be used online to reduce the initial multi-jet trigger rate.

6.1 Jet multiplicity selection

A preliminary selection is made taking into account the minimal topology of a top fully hadronic decay.

The process $\mathrm{t\bar{t}} \rightarrow (jj)(jj)\mathrm{b\bar{b}}$ requires 6 jets in the final state. Anyway, jets can blend resulting as a single reconstructed jet, or either get lost along the beam axis. On the other hand, additional jets can be produced in the initial (ISR) or the final (IFR) state from gluon radiation. The resulting final topology is not well defined.

Jets produced from gluons, light quarks and b quarks differ, and a flavor dependent selection is possible. On the other hand, the request of at least 6 jets allows a first background reduction.

For that reasons, the selection of $\mathrm{t\bar{t}}$ hadronic events is made requiring offline at least 6 jets and no more than 8 jets. Moreover, only jets with $\mathrm{E_T} > 25\,\mathrm{GeV/c}$ and $|\eta| < 2.4$ are considered.

Combined to the trigger selection this request leads to an effective cross section of about 92 pb, corresponding to an efficiency of about 11% (about 24% of the fully hadronic set), while the QCD events effective cross section amounts to 14 nb, increasing the signal to background ratio up to 1/151.

	Trigger	$6 \leq \mathrm{N_{jets}} \leq 8$
$\hat{\sigma}_{QCD}(pb)$	$106{\times}10^3$	$14{\times}10^3$
$\hat{\sigma}_{\mathrm{t\bar{t}}}(pb)$	205	92
efficiency	25%	11%
S/B	1/517	1/151

Table 6.1: Effect of the $\mathrm{N_{jets}}$ selection on the efficiency and S/B.

6.2 Event Variables

As described above, jets reconstructed with an iterative cone algorithm with a size of 0.5 are used. A multi-jet trigger, whose thresholds are evaluated from the Monte Carlo generated event sample, is applied for a first rejection of the background. Furthermore the top fully hadronic decay foresees six hadronic jets with a specific event topology. For the purpose of the analysis, i.e. the top quark mass measurement, we need all of them and so we discard events with less than six jets, and keep events with no more than 8 jets. The top decay topology has further well defined traits which can help the improvement of the signal to background ratio up to values allowing a good mass measurement. They are related to event $\mathrm{E_T}$ collected in the calorimeter towers, the natural centrality and co-planarity of the event and other compound variables.

The selection cuts to be chosen are based on the statistical significance:

$$\frac{S}{\sqrt{S+B}} \tag{6.1}$$

corresponding to $1\,\mathrm{fb}^{-1}$ of integrated luminosity.

An optimal kinematic selection would require the best statistical significance achievable. Anyway in the specific case, although a significant background reduction is obtained, the final S/B ratio is still not satisfactory. We choose instead stronger cuts which are still within 95% of the maximum significance but provide a better S/B ratio. At each step, the cut giving the best performance is evaluated and applied for the next step.

The set of variable evaluated are the following:

- Σp_{x}, Σp_{y} and Σp_{z}, sums of the jet momentum components;
- $\mathrm{E_{T6}}$, the transverse energy of the sixth jet;
- $\Sigma\mathrm{E}$, the sum of the jet energies;
- $\Sigma\mathrm{E_T}$, the sum of the jet transverse energies;
- $\sum_{k=3}^{\mathrm{N_{jets}}} \mathrm{E_T}$, e.g. sub-leading jet total transverse energy, obtained removing the two most energetic jets;
- *centrality* $= \sum \frac{\mathrm{E_T}}{\sqrt{\hat{s}}}$, with: $\hat{s} = (\sum E)^2 - (\sum P_z)^2$ where all the sums run over all the reconstructed jets. The *centrality* represents the fraction of the total available energy going in the transverse plane;
- *aplanarity* $= \frac{3}{2}Q_1$ where Q_1 is the smallest of the three normalized eigenvalues of the sphericity tensor $M_{ab} = \sum_j P_{ja}P_{jb}$;
- $\cos\theta_1^*$and $\cos\theta_2^*$, cosines of the angle of the two leading jet with respect to the beam axis as computed in the jet system center of mass frame;
- $\mathrm{E_T}{}_1^*$ and $\mathrm{E_T}{}_2^*$, where $\mathrm{E_T}{}_i^* = \mathrm{E_T}\sin^2\theta_i^*$;
- $\prod_{i=3}^{\mathrm{N_{jets}}} \mathrm{E_T^*}^{\mathrm{N_{jets}}-2}$geometrical average of N-2 sub-leading jets $\mathrm{E_T}$;
- $M_{2j}^{min}$the minimum di-jet invariant mass;
- $M_{3j}^{min}$the minimum tri-jet invariant mass;
- $M_{2j}^{max}$the maximum di-jet invariant mass;
- $M_{3j}^{max}$the maximum tri-jet invariant mass.

The cut significance, gain and related S/B ratio are shown in table 6.2, ordered in decreasing statistical significance value. Already at the very first stage, a subset of few useful variables can be identified: *centrality*, $\prod_{i=3}^{\mathrm{N_{jets}}} \mathrm{E}_{\mathrm{T}}^{*\,\mathrm{N_{jets}}-2}$, *aplanarity*, $\sum_{k=3}^{\mathrm{N_{jets}}} \mathrm{E_T}$. Even if the quantities reported evolve after every step, the remaining variables have low effects on the purpose of improving the S/B ratio, so we concentrate on these four. Their distributions is shown in figures 6.1, 6.2, 6.3, 6.4.

variable	$\frac{S}{\sqrt{S+B}}$	cut	S/B gain	S/B
centrality	26.6	>0.81	3.25	1/47
$\prod_{i=3}^{\mathrm{N_{jets}}} \mathrm{E}_{\mathrm{T}}^{*\,\mathrm{N_{jets}}-2}$	25.3	>28	2.35	1/65
aplanarity	24.3	>0.065	1.77	1/86
$\mathrm{E_{T}}_2^*$	24.1	>63	1.42	1/107
$\sum_{k=3}^{\mathrm{N_{jets}}} \mathrm{E_T}$	23.9	>225	1.39	1/109
$\cos\theta_1^*$	23.6	>0.24	1.43	1/133
$\mathrm{E_{T}}_1^*$	23.6	>88	1.25	1/121
$\cos\theta_2^*$	23.5	>0.21	1.15	1/132
$\Sigma\mathrm{E_T}$	23.3	>455	1.11	1/136
$\mathrm{E_{T6}}$	23.2	>33	1.55	1/98
M_{2j}^{min}	23.2	>35	1.07	1/142
Σp_{z}	23.2	>855	0.99	1/153
M_{2j}^{max}	23.2	>240	0.96	1/158
Σp_{y}	23.1	>87.5	0.97	1/156
$\Sigma\mathrm{E}$	23.1	>600	0.97	1/158
M_{3j}^{min}	23.0	>100	1.08	1/140
Σp_{x}	23.0	>90	0.97	1/156
M_{3j}^{max}	23.0	>340	0.95	1/159

Table 6.2: Cuts evaluated at 95% of statistic significance, S/B achievable and gain over the current S/B.

6.3 Event selection

The kinematical selection which provides the best performance, as discussed above, requires:

- *centrality* >0.8

- $\sum_{k=3}^{\mathrm{N_{jets}}} \mathrm{E_T} > 215$ GeV
- *aplanarity* > 0.065
- $\prod_{i=3}^{\mathrm{N_{jets}}} \mathrm{E_T^*}^{\mathrm{N_{jets}}-2} > 30$ GeV

The performance of each cut is presented in table 6.3. Such a selection has an efficiency of 1.36% on $\mathrm{t\bar{t}}$ events ($m_t = 175\,\mathrm{GeV/c^2}$) corresponding to an effective cross section of about $\hat{\sigma} = 11\,\mathrm{pb}$, and S/B = 1/23

	$\mathrm{t\bar{t}}$ efficiency (%)	$\hat{\sigma}_{\mathrm{t\bar{t}}}$ (pb)	$\hat{\sigma}_{QCD}$ (pb)	S/B
start	11%	92	14×10^3	1/151
centrality > 0.8	4.06%	34	2×10^3	1/47
$\sum_{k=3}^{\mathrm{N_{jets}}} \mathrm{E_T} > 215$	2.94%	25	838	1/34
aplanarity > 0.065	1.80%	15	383	1/26
$\prod_{i=3}^{\mathrm{N_{jets}}} \mathrm{E_T^*}^{\mathrm{N_{jets}}-2} > 30$	1.36%	11	258	1/23

Table 6.3: Cuts chosen for the kinematic selection and related S/B ratio.

The kinematical selection used here corresponds to the application of subsequent 1-dimensional cuts. Such a naive selection manages to reach a reasonably good S/B even without fully considering all correlations among variables.

A more refined selection fully accounting for all correlations could be obtained by recurring to Multi-Variated Analysis Techniques (like neural networks for instance), but this approach is beyond the scope of this analysis.

6.4 b-tag based selection

By recurring to b-tagging we can improve the S/B ratio.

The algorithm used for the analysis is the track counting one, since it is the first implemented and so deeply studied for the Monte Carlo samples. In some way it has determined also the choice of the iterative cone algorithm, since it was the first with b-tag information available.

The CMS implementation foresees two discriminators, related to the second (TC2) track for high efficiency selections and the third (TC3) track for high purity selections. A set of reference cuts (table 6.4) and plots (figure 6.5) are provided for the Monte Carlo samples [47].

For the sake of the analysis, both the cuts referred as “Medium” and “Tight” were used, since they allows a good efficiency selection. Requiring at least one

tagging efficiency (%)	udsg-jets	c-jets	b-jets
Loose TC2 >2.3	10.3 ± 0.1	32.8 ± 0.2	71.5 ± 0.2
Medium TC2 >5.3	0.991 ± 0.008	10.9 ± 0.1	51.0 ± 0.2
Tight TC3 >4.8	0.108 ± 0.003	2.97 ± 0.06	32.4 ± 0.2

Table 6.4: Tagging efficiency in the QCD 80-120 Monte Carlo sample. These three operating points are defined in order to achieve a fraction of respectively 10%, 1% and 0.1% of udsg-jets.

Medium b-tag, the resulting efficiency is slightly reduced to 1.13%, corresponding to an effective cross section $\hat{\sigma} = 9.4\,\text{pb}$, while the signal to background ratio is much better S/B = 1/6.9. If we require at least two jets to be tagged as b-jets, the efficiency is reduced to 0.53%, corresponding to an effective cross section $\hat{\sigma} = 4.4\,\text{pb}$, and we reach S/B = 1/2.8.

These values of S/B can be improved respectively by 1/4.5 and 1/1.5 if we require one or two Tight b-tags.

N b-tags	$\mathrm{t\bar{t}}$ efficiency (%)	$\hat{\sigma}_{\mathrm{t\bar{t}}}$ (pb)	$\hat{\sigma}_{QCD}$ (pb)	S/B
before tagging	1.36%	11.3	258	1/23
≥ 1 b-tag	1.13%	9.4	65	1/6.9
≥ 2 b-tag	0.53%	4.4	12	1/2.8

Table 6.5: Efficiency for the kinematical selection cuts before b-tagging and requiring at least one or two b-tag using the Medium cut.

N b-tags	$\mathrm{t\bar{t}}$ efficiency (%)	$\hat{\sigma}_{\mathrm{t\bar{t}}}$ (pb)	$\hat{\sigma}_{QCD}$ (pb)	S/B
before tagging	1.36%	11.3	258	1/23
$ge1$ b-tag	0.80%	6.7	30	1/4.5
$ge2$ b-tag	0.19%	1.5	2.34	1/1.5

Table 6.6: Efficiency for the kinematical selection cuts before b-tagging and requiring at least one or two b-tag using the Tight cut.

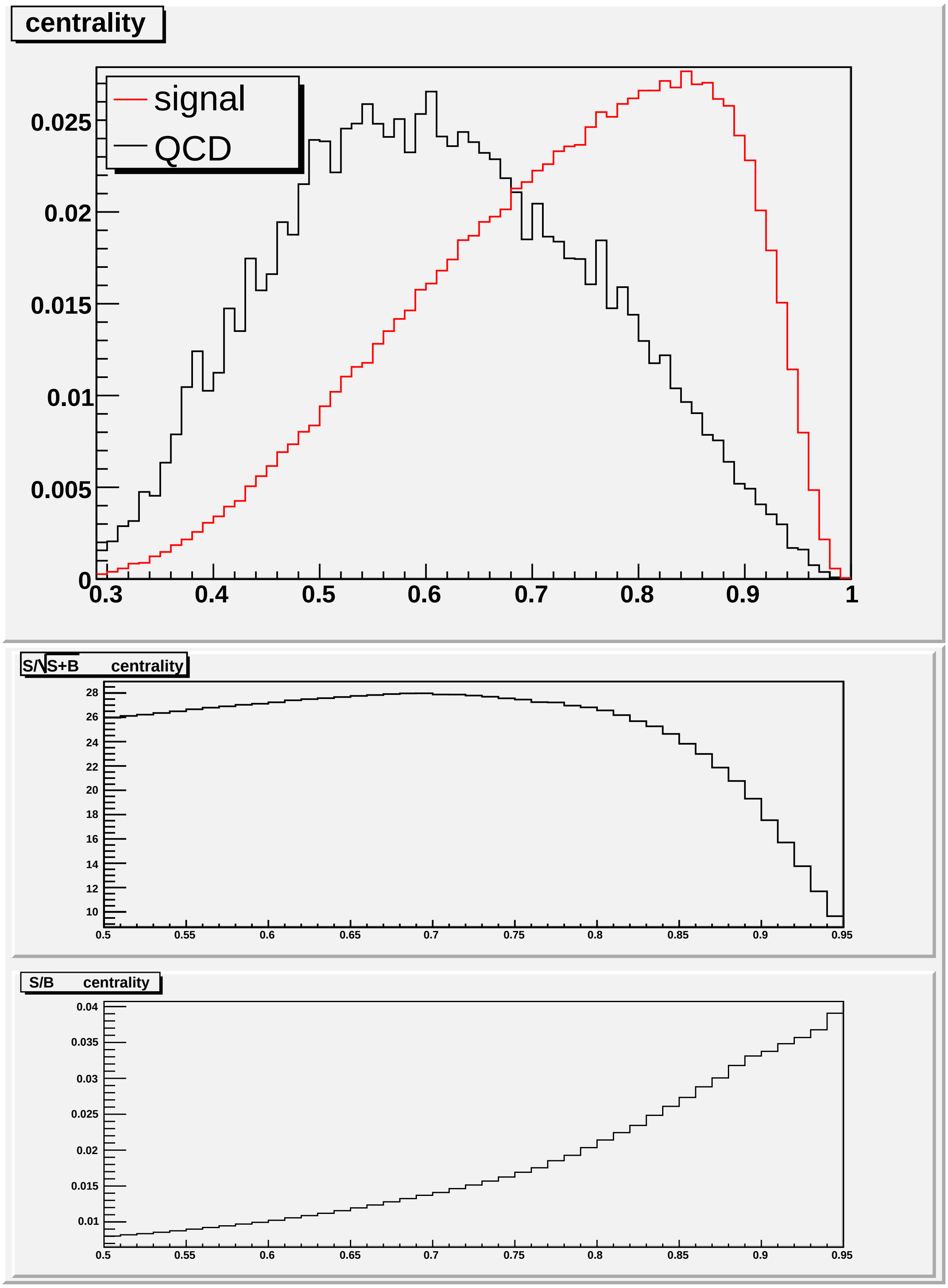

Figure 6.1: Distributions of *centrality* for $t\bar{t}$ (top) and QCD (bottom) events and related S/B and $\frac{S}{\sqrt{S+B}}$.

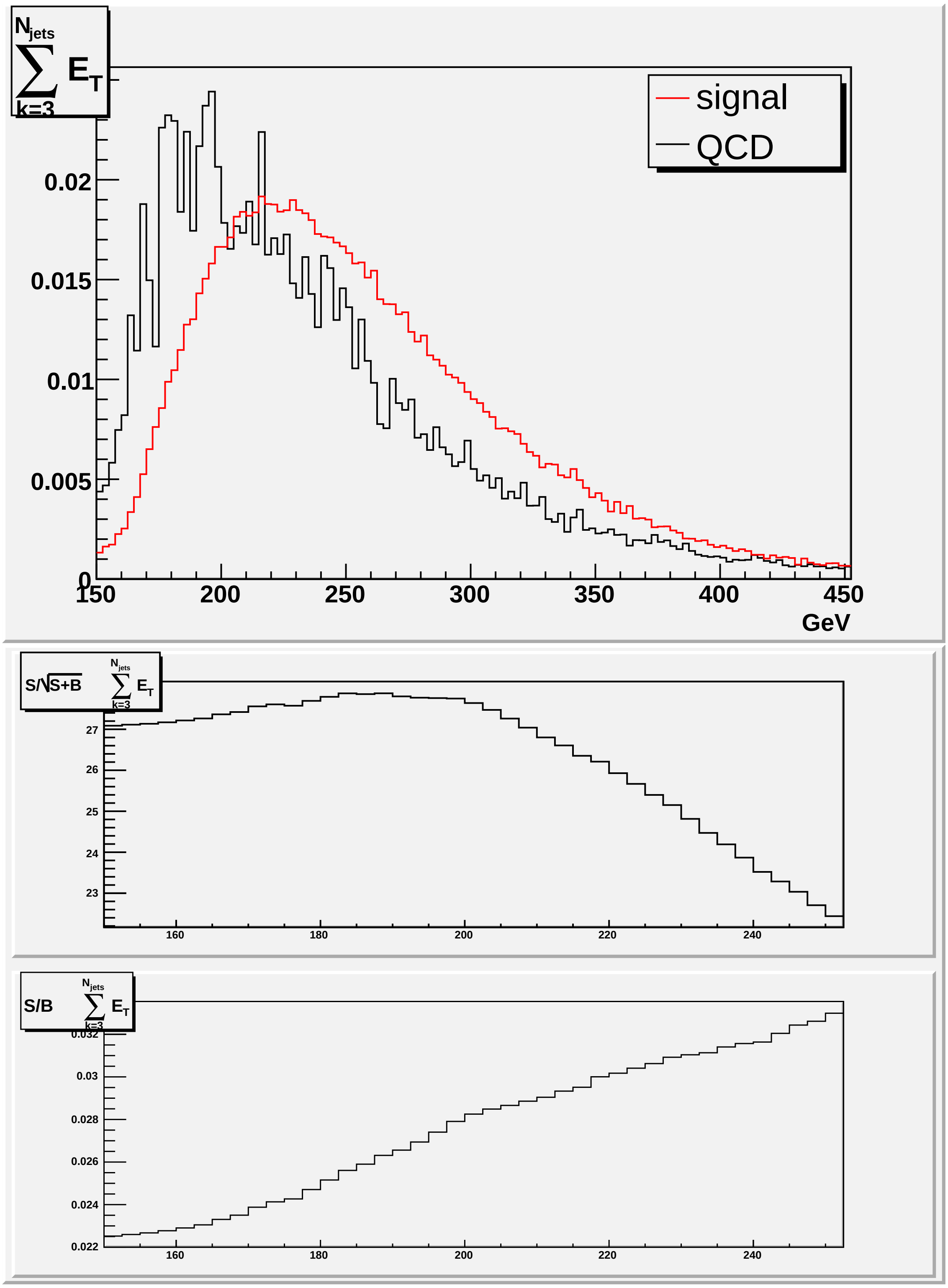

Figure 6.2: Distributions of $\sum_{k=3}^{\mathrm{N_{jets}}} \mathrm{E_T}$ for $t\bar{t}$ (top) and QCD (bottom) events and related S/B and $\frac{S}{\sqrt{S+B}}$.

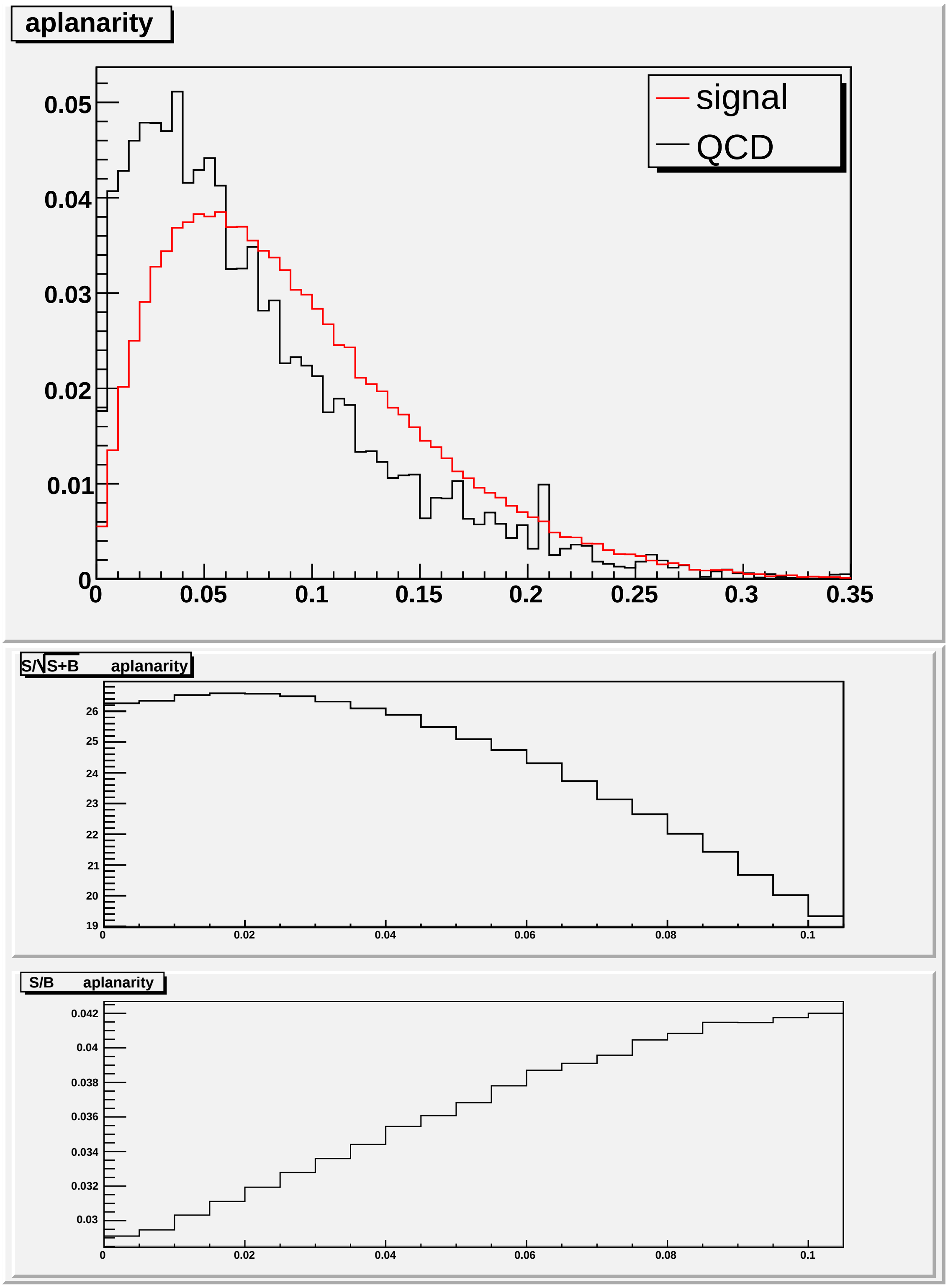

Figure 6.3: Distributions of *aplanarity* for $t\bar{t}$ (top) and QCD (bottom) events and related S/B and $\frac{S}{\sqrt{S+B}}$.

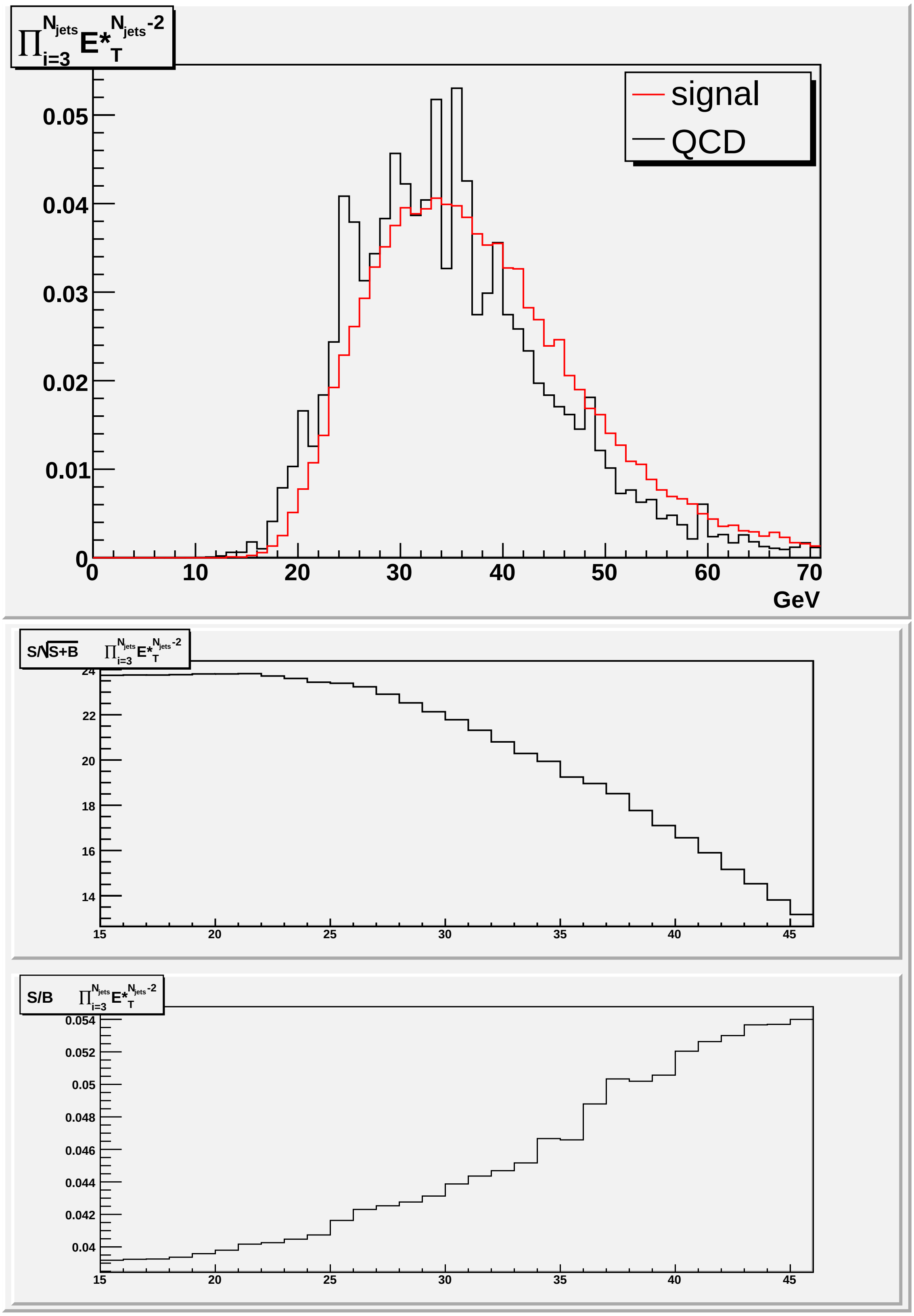

Figure 6.4: Distributions of $\prod_{i=3}^{\mathrm{N_{jets}}} \mathrm{E}_{\mathrm{T}}^{*\,\mathrm{N_{jets}}-2}$ for $\mathrm{t\bar{t}}$ (top) and QCD (bottom) events and related S/B and $\frac{S}{\sqrt{S+B}}$.

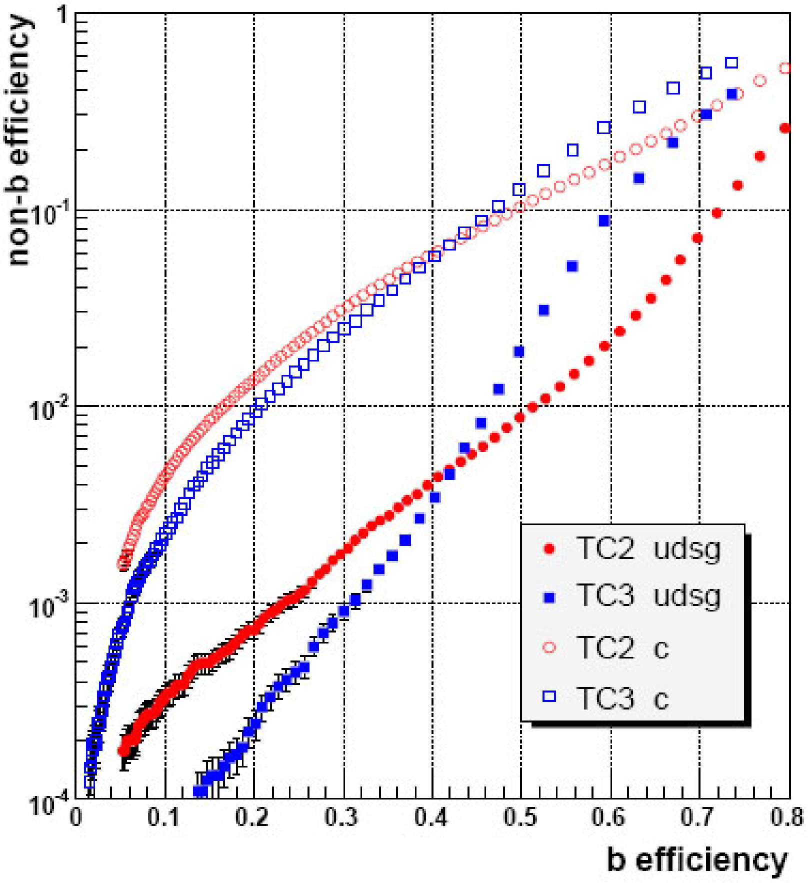

Figure 6.5: Performance of the Track Counting in the QCD 80-120 Monte Carlo sample requiring two (TC2) or three (TC3) tracks with a given impact parameter significance.

Chapter 7

Measurements of the top quark mass

The goal of this analysis is the measurement of the top quark mass in the all-hadronic channel, i.e. where both W's decay into a $q\overline{q}$ pair.

The method used relies on comparison of event samples with the expected signal and background probability density functions (*p.d.f.*) through a likelihood maximization. These *p.d.f.*, also referred as "templates" are derived from Monte Carlo simulated events, both for signal and background, The final result we present here is not properly the measure of the top quark mass, but an evaluation of the expected uncertainty of the method to understand its applicability with the actual data expected for the first year of data taking.

7.1 Mass Reconstruction

The first step consists on the reconstruction of the top quark mass associated to events surviving the cuts. The reconstructed mass is not to be considered the event-per-event mass measurement, instead it is the mass dependent quantity which will be used to extract the top mass measurement through the likelihood maximization.

The jet four vectors are the main ingredients of the event topology reconstruction. As discussed above, calorimetry non-compensation, response non-linearity, energy losses and instrumentation cracks require jet corrections. The iterative cone reconstructed jets are, for that reason, corrected using jet correction provided in the software framework.

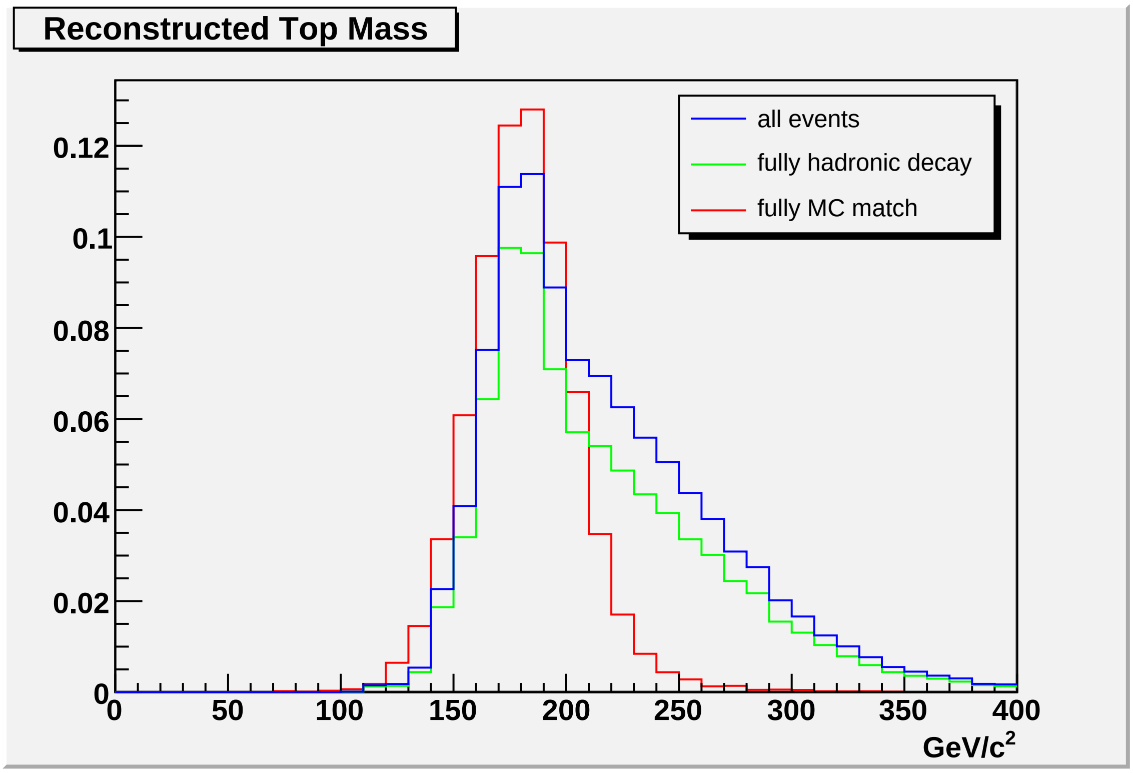

Figure 7.1: Reconstructed top mass for the full sample compared with the subset of fully hadronics events (from Monte Carlo truth) and the subset of events where jets are fully matched with the generated partons. All distribution are normalized to the same area.

Moreover, jets arising from the W decay have different fragmentation and decay properties from the b-jets, and so, flavor specific corrections have to be applied. As discussed in the previous chapter, a flavor-dependent correction function has been extracted from the Monte Carlo generated particle matched to jets. An equivalent feature is being integrated in the framework, while a jet flavor tagging based on the jet properties is needed once real data will be analyzed. These corrections can be applied only after choosing an hypothesis on the jet nature and will be applied dynamically from the kinematic fitter described below, together with an imposed mass of 5 GeV/c^2 for b-jets and 0.5 GeV/c^2 for light-quark jets.

Considering only the six leading jets and requiring 0, 1 or 2 b-tags we have multiple combinations of jets to reconstruct the decay $t\bar{t} \rightarrow WbWb \rightarrow (jj)b(jj)\bar{b}$. Namely, out of the 90 possible combinations when no b-tag is required, we are left

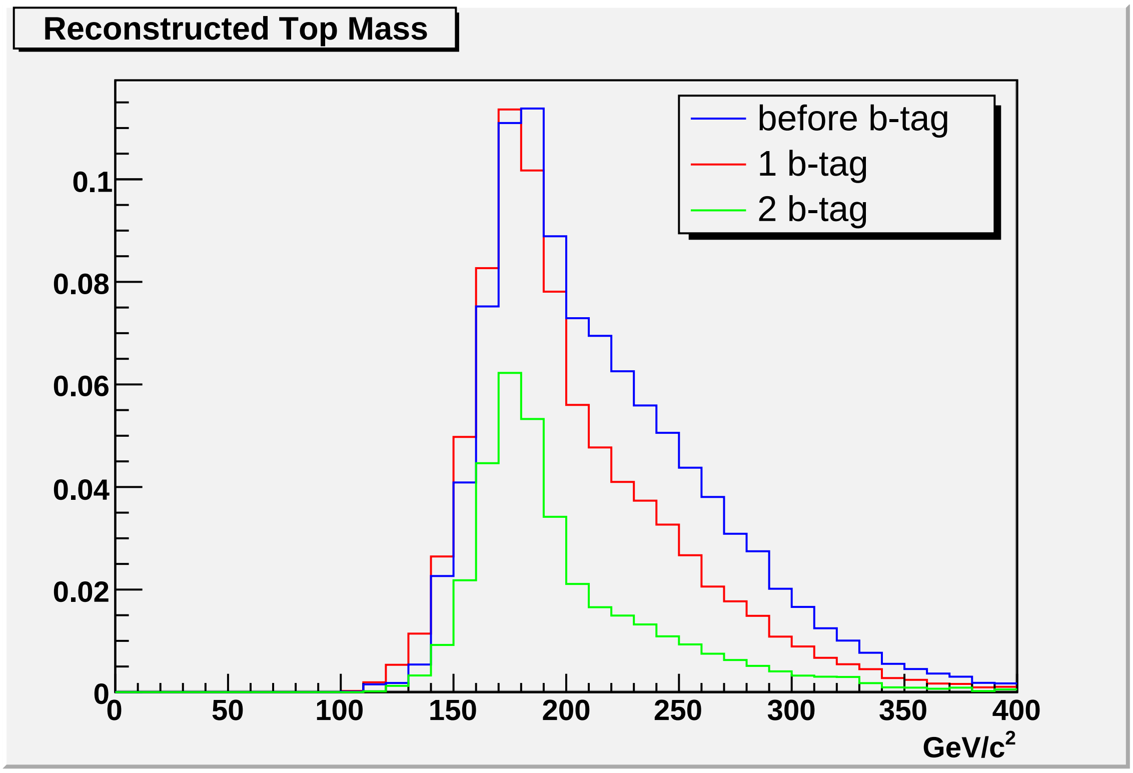

Figure 7.2: Reconstructed top mass without requiring b-tag, requiring at least 1 b-tag and at least 2 b-tag. Both distribution are normalized to the same area.

with 30 combinations for events with a single tag and 6 for events with 2 tags. For each combination we evaluate a χ^2 as follows:

- two jets are considered as b and, a nominal mass of 5 GeV/c^2 is applied to them, together with flavor specific corrections;
- the remaining jets are corrected with a nominal mass of 0.5 GeV/c^2;
- the four candidate light quarks are combined together in two doublets m_{jj^1}, m_{jj^2}, to reconstruct the W masses, constrained to the nominal value $M_W = 80.4\,\mathrm{GeV/c^2}$ with a natural width $\Gamma_W = 2.1\,\mathrm{GeV/c^2}$;
- the two doublets are the associated to the b-jets to form two triplets m_{jjj^1}, m_{jjj^2}

A χ^2 of the form

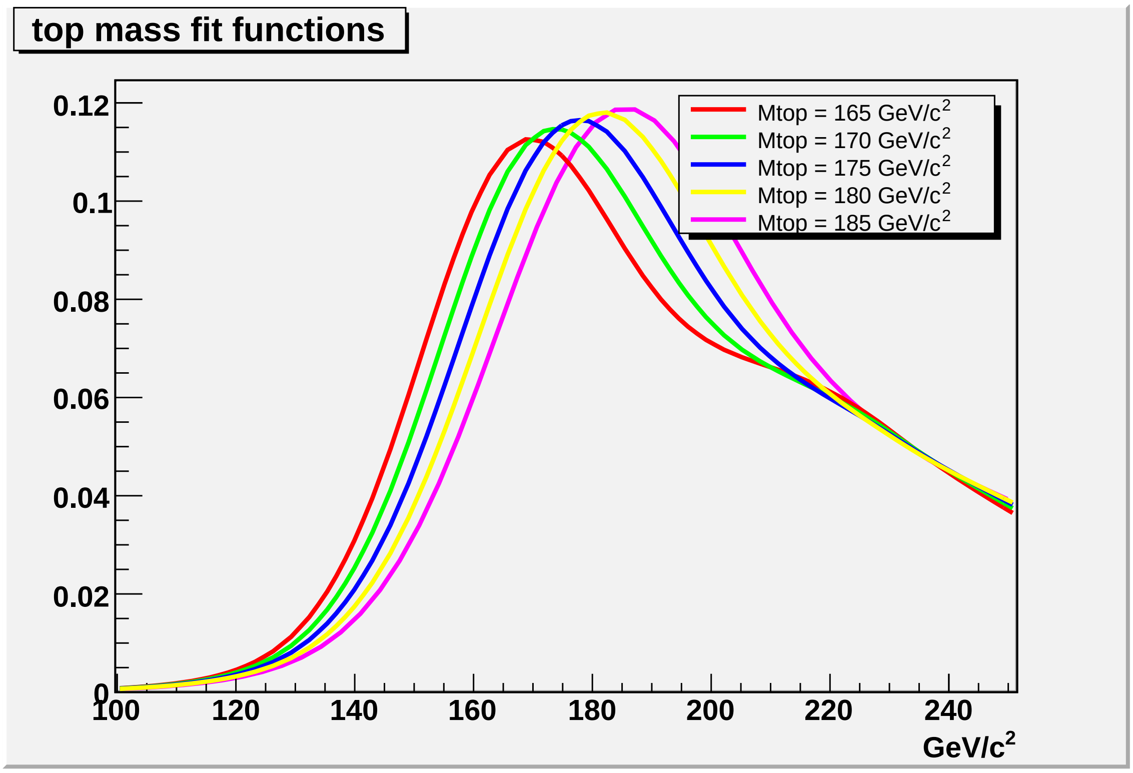

Figure 7.3: Evolution of the fit function for some of the reconstructed masses for events with at least 1 b-tag.

$$\begin{aligned}\chi^2 \;=\; & \frac{(m_{jj^1}-m_W)^2}{\Gamma_W^2} + \frac{(m_{jj^2}-m_W)^2}{\Gamma_W^2} + \frac{(m_{jjj^1}-m_t)^2}{\Gamma_t^2} + \frac{(m_{jjj^2}-m_t)^2}{\Gamma_t^2} + \\ & + \sum_{i=1}^{N} \frac{({p_T}_i^{fit} - {p_T}_i^{data})^2}{\sigma_1^2} \qquad (7.1)\end{aligned}$$

is evaluated, where the only free parameters are the reconstructed top mass m_t and ${p_T}_i^{data}$ which are the transverse momenta of the jets.

The χ^2 expression is minimized so that we obtain an invariant top quark mass for each combination. The invariant mass chosen for each event is then the one corresponding to the combination with the lowest χ^2; this is the reconstructed mass, $\mathrm{m}_\mathrm{t}^\mathrm{rec}$.

The resulting distribution is not expected to be a faithful description of the top quark mass, since jet reconstructed with poor resolution, as well as lost jets and hard

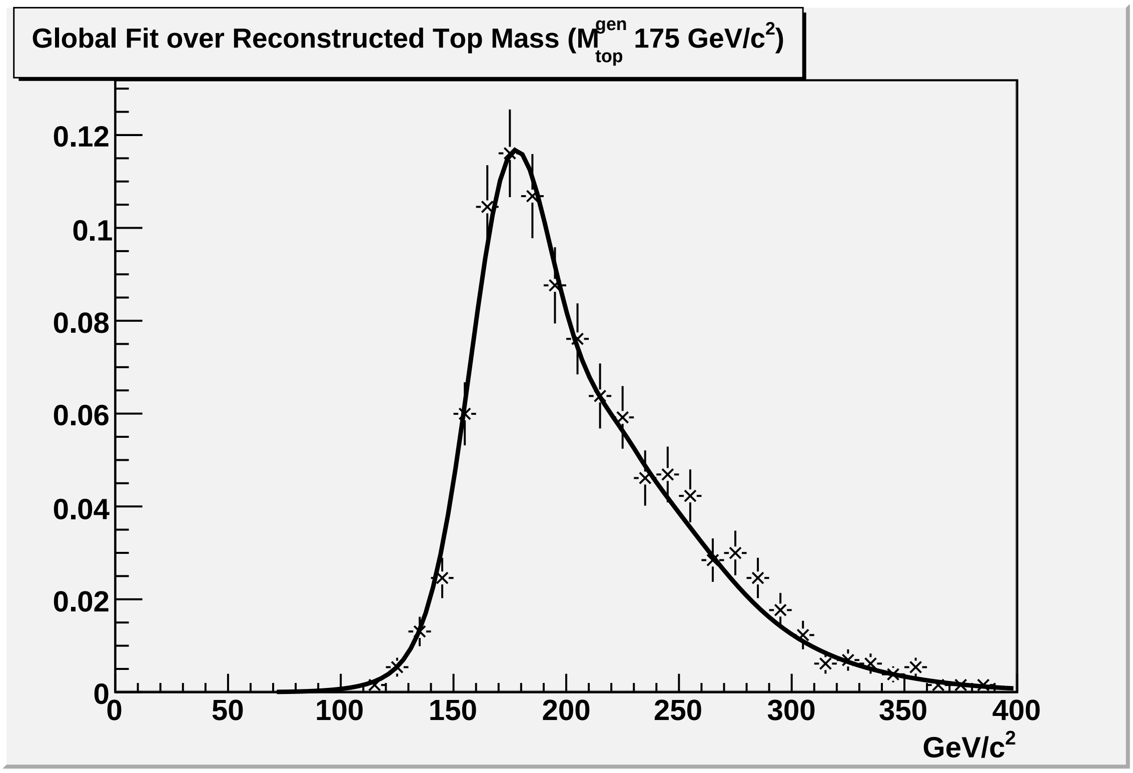

Figure 7.4: Reconstructed top mass for signal samples generated with mass from 175 GeV/c^2 with global fit superimposed. Events with at least 1 b-tag.

radiating gluons jets bring to a consistent amount of incorrect assignment. Although it is a good observable candidate for the measurement itself: in the Monte Carlo generated event for instance the distributions are clearly dependent of the generated top mass, as we will see later.

As shown in figure 7.2, the request of at least one b-tag, not only increases the S/B ratio, but improves also the overall shape of the reconstructed top mass, shrinking the distribution. This is even more the case when we require two b-tags.

7.2 Measured Mass

In this analysis 21 $t\bar{t}$ samples are used, generated with masses ranging from 165 GeV/c^2 to 185 GeV/c^2 in steps of 1 GeV/c^2. With this method is crucial to know with a certain precision the shape expected for the samples generated at different top quark mass.

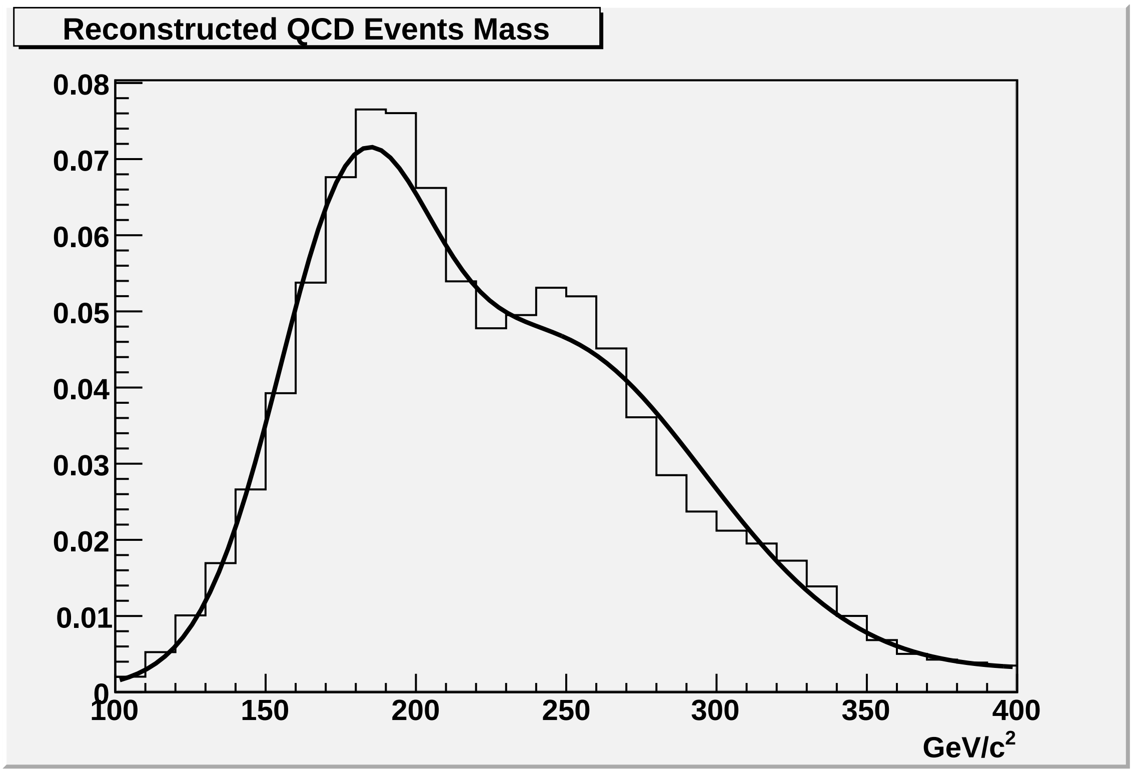

Figure 7.5: Background reconstructed with the kinematical fitter with fit superimposed. Events with at least 1 b-tag.

Since we have only a discrete knowledge of the signal templates and reduced size of the samples, we need to reach a better *p.d.f.* evaluation. To reach this goal, the mass scan samples are interpolated with a continuous arbitrary curve, function only of the generated mass. The signal is so described as a sum of two Gaussian distributions describing the core of the distribution, plus a gamma distribution describing the combinatoric component. The resulting probability density function is normalized to 1. Each parameter of this function has a linear dependence on the generated top quark mass. The analytical expression is given by:

$$
\begin{aligned}
P_{sig}(m|\mathrm{M_{top}}) \quad = \quad & \delta_7 \, \frac{\delta_2^{1+\delta_1}}{\Gamma(1+\delta_1)} \, (m-\delta_0)^{\delta_1} \, e^{-\delta_2(m-\delta_0)} + \\
& +\delta_8 \, \frac{1}{\sqrt{2\pi}\delta_4} \, e^{\frac{-(m-\delta_3)^2}{2\delta_4^2}} + \\
& +(1-\delta_7-\delta_8) \, \frac{1}{\sqrt{2\pi}\delta_6} \, e^{\frac{-(m-\delta_5)^2}{2\delta_6^2}}
\end{aligned} \tag{7.2}
$$

where:

$$\delta_1 = \alpha_i + \beta_i \, (\mathrm{M_{top}} - 175) \tag{7.3}$$

$m = \mathrm{m_t^{rec}}$ is the reconstructed top mass and $\mathrm{M_{top}}$ is the top quark mass used in the $\mathrm{t\bar{t}}$ sample generation. The dependence of the parameters on the top mass allows to interpolate between the discrete set of Monte Carlo mass templates. The in table 7.1. The fit function evolution is shown in figure 7.3, while 7.4 shows how actually fit with the signal sample generated at $\mathrm{M_{top}} = 175\,\mathrm{GeV/c^2}$.

A very similar function gives a good interpolation for the background, of course without the top mass dependency. So the fit function is still a sum of two Gaussian and one Gamma functions an in 7.2, where δ_i are now the constant values in table 7.2.

i	α_i	β_i
0	230	2.5
1	0.047	174
2	17	212
3	37	0.86
4	0.079	4.1×10^{-8}
5	0.013	4.8×10^{-5}
6	0.79	0.15
7	0.38	0.12
8	3.9×10^{-4}	3.5×10^{-4}

Table 7.1: Parameters for the signal sample fit.

	value
δ_0	4.0 $\times 10^{-11}$
δ_1	14
δ_2	0.032
δ_3	178
δ_4	25
δ_5	243
δ_6	52
δ_7	0.91
δ_8	0.049

Table 7.2: Parameters for the QCD sample.

7.3 Likelihood Fit

Once the templates are calculated for both the signal and the QCD events surviving the cuts, our goal is to compute the probability of the masses of the candidate events surviving the kinematic cuts to come from the expected *p.d.f.*'s.

The function is divided in two parts: the first part accounts for the signal-background discrimination and includes, as a parameter to be minimized, the number of signal tags, n_s, the number of background tags n_b, and the probability for the top mass given the i^{th} template; the second term is a constraint on the amount of background tags N_b, which we obtain from the background sample surviving the kinematic cuts. The analytic expression of the likelihood is:

$$L = e^{-\frac{(n_s+n_b-N)^2}{2\sigma_N^2}} \prod_{i=1}^{N} \frac{n_s\, P_{sig}(m_i|\mathrm{M_{top}}) + n_b\, P_{bkg}(m_i)}{n_s+n_b}\, e^{-\frac{(n_b-N_b)^2}{2\sigma_{N_b}^2}} \tag{7.4}$$

where N is the total number of tag with error $\sigma_N = \sqrt{N}$ and $\sigma_{N_b} = \sqrt{N_b}$ is the error over the Monte Carlo generated tags surviving cuts.

Since the background estimate is affected by large uncertainties we prefer for the time being to drop the constraint on n_b and reformulate the likelihood as:

$$\prod_{i=1}^{N} L = \mathscr{P}_s\, P_{sig}(m_i|\mathrm{M_{top}}) + (1-\mathscr{P}_s)\, P_{bkg}(m_i) \tag{7.5}$$

where the only free parameter is the purity $\mathscr{P}_s$, of the signal, defined as the ratio between the number of signal events and the number of all candidates.

The computation is obtained minimizing the negative logarithm of the likelihood with respect to $\mathscr{P}_s$. The likelihood can be interpolated with a third degree polynomial: the minimum is taken to be the mass measurement.

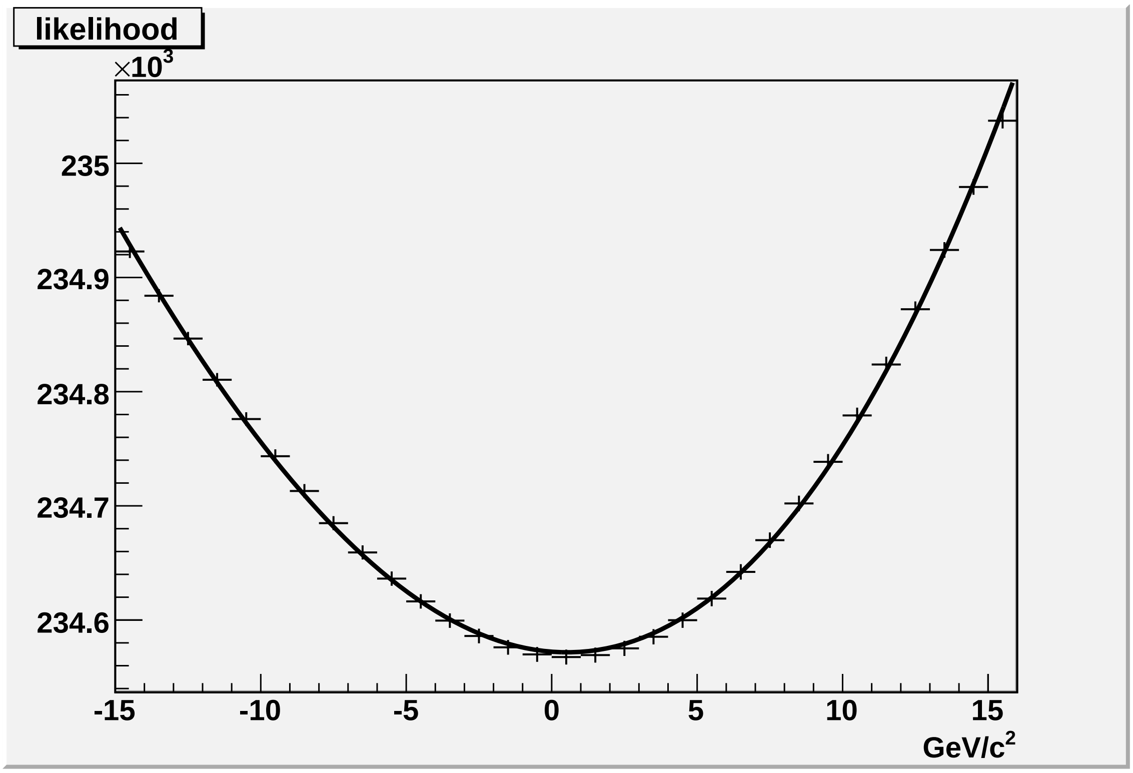

Figure 7.6: Likelihood evaluated for signal sample generated with mass 175 GeV/c^2. Events with at least 1 b-tag.

7.4 Pseudo Experiments

Waiting for the data from the experiment, we study the performance of the method by using pseudo-experiments, i.e. sets of N events where N_s events are generated according to the $\mathrm{m}_{\mathrm{t}}^{\mathrm{rec}}$ distribution characteristic of $\mathrm{t\bar{t}}$ events generated with mass M_{top} and N_b events generated according to the $\mathrm{m}_{\mathrm{t}}^{\mathrm{rec}}$ distribution characteristic of background events.

To check for possible biases introduced by the method used as well as its statistical power we produced a set of pseudo-experiments extracting pseudo data with the predicted amount of pseudo-events according to the expected S/B ratio.

We fix the total amount to be the one foreseen for 1 fb^{-1} of integrated luminosity for each of the 21 input top masses. Then we do the same for the QCD background, with a total amount of pseudo-experiments such as the S/B ratio obtained with the kinematic cuts is respected.

We perform the measurement over an ensemble of 100 different sets of pseudo-events for each mass, measuring the result top mass and taking the mean value. The resulting masses are plotted with respect to the related generated mass. The plot can be interpolated with a line: the resulting function is used to re-scale the measured mass of a real sample and give an estimation of the expected uncertainties.

7.5 Top Quark Mass Measurement

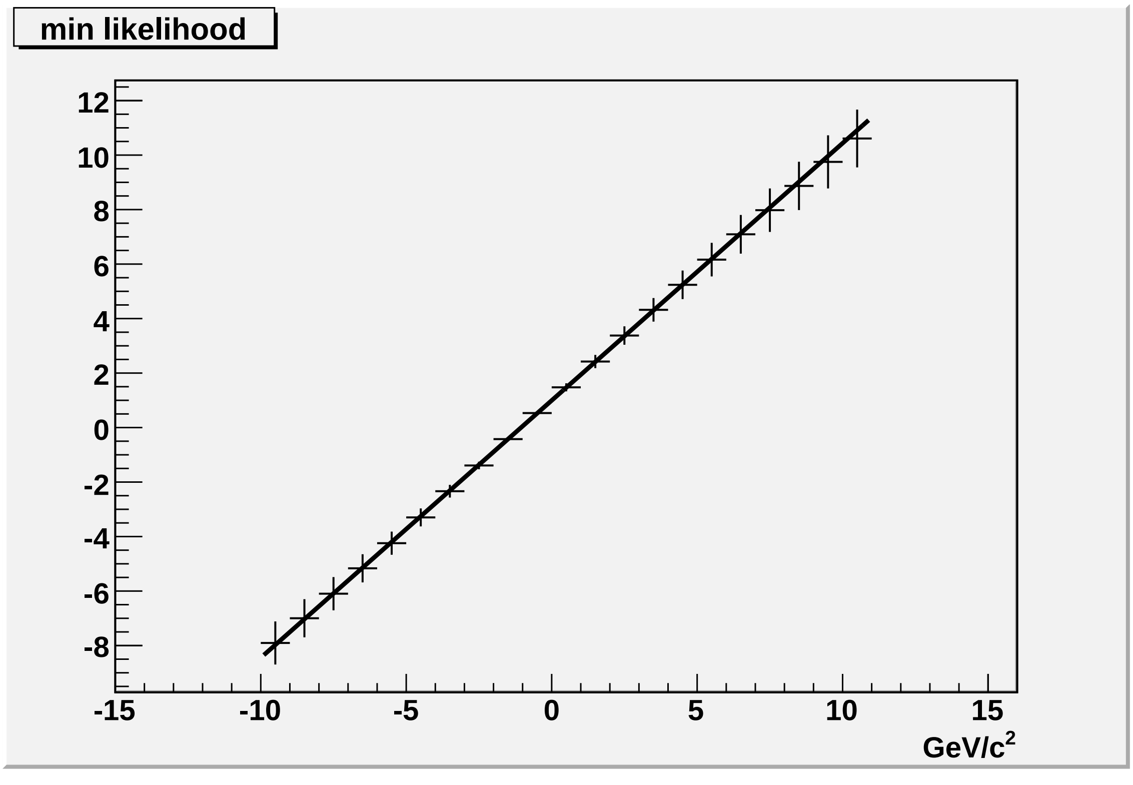

Figure 7.7: Measured vs. generated top quark mass assuming no background. A linear fit is superimposed.

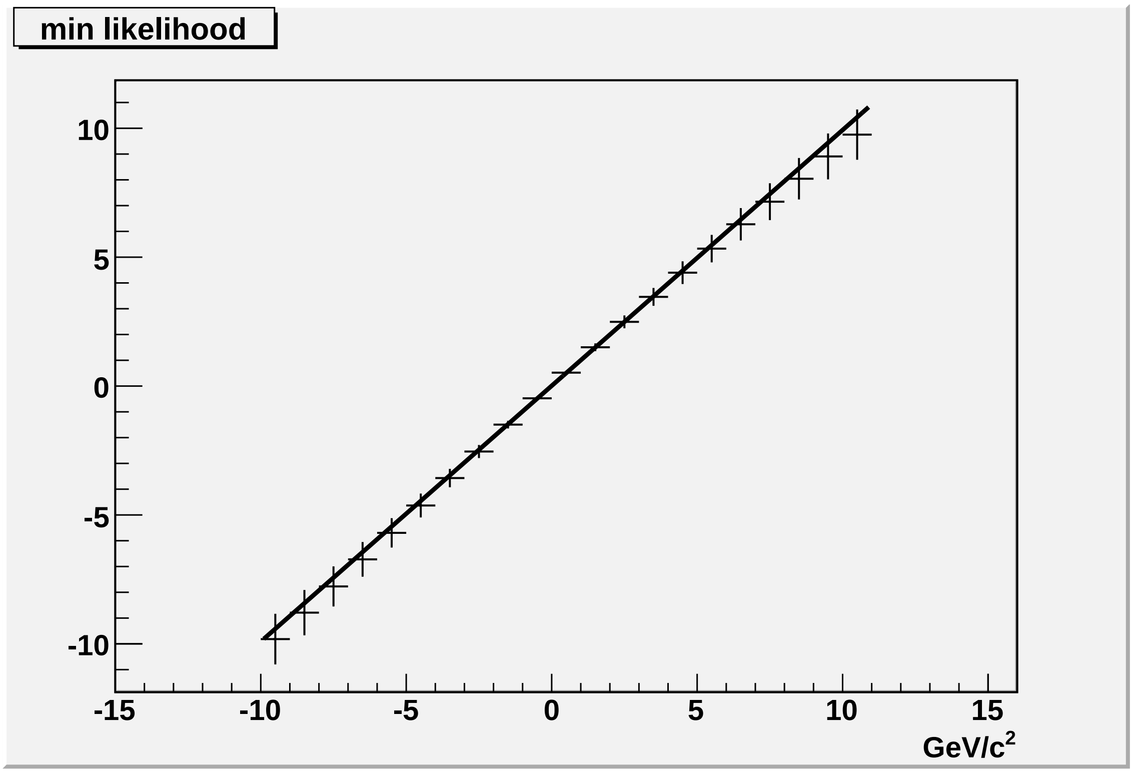

Figure 7.8: Measured vs. generated top quark mass using the background according to the expected S/B ratio. A linear fit is superimposed.

7.5.1 Statistical uncertainties on the top mass measurement

Since we have, for the time being, no data available for a top quark mass measurement, we are interested here on the expected performance of our method in terms of expected uncertainties.

The statistical uncertainty depends on the amount of signal, its purity and the sharpness of the signal templates. We can estimate this uncertainty using pseudo-experiments from the likelihood fit, extracting the values M^+ and M^- corresponding to half unit of increment with respect to the minimum M^{min} :

$$-\ln(M^{\pm}) = -\ln(M^{min}) + 1/2 \tag{7.6}$$

We consider as statistical uncertainty the average value:

$$\Delta M_{stat} = \frac{|M^{+} - M^{min}| + |M^{-} - M^{min}|}{2} \tag{7.7}$$

The distribution on such value is shown in figure 7.9. The Average value is then:

$$\Delta M_{stat} = 0.5\,\mathrm{GeV}/c^{2} \tag{7.8}$$

and does not depend much on the generated top mass.

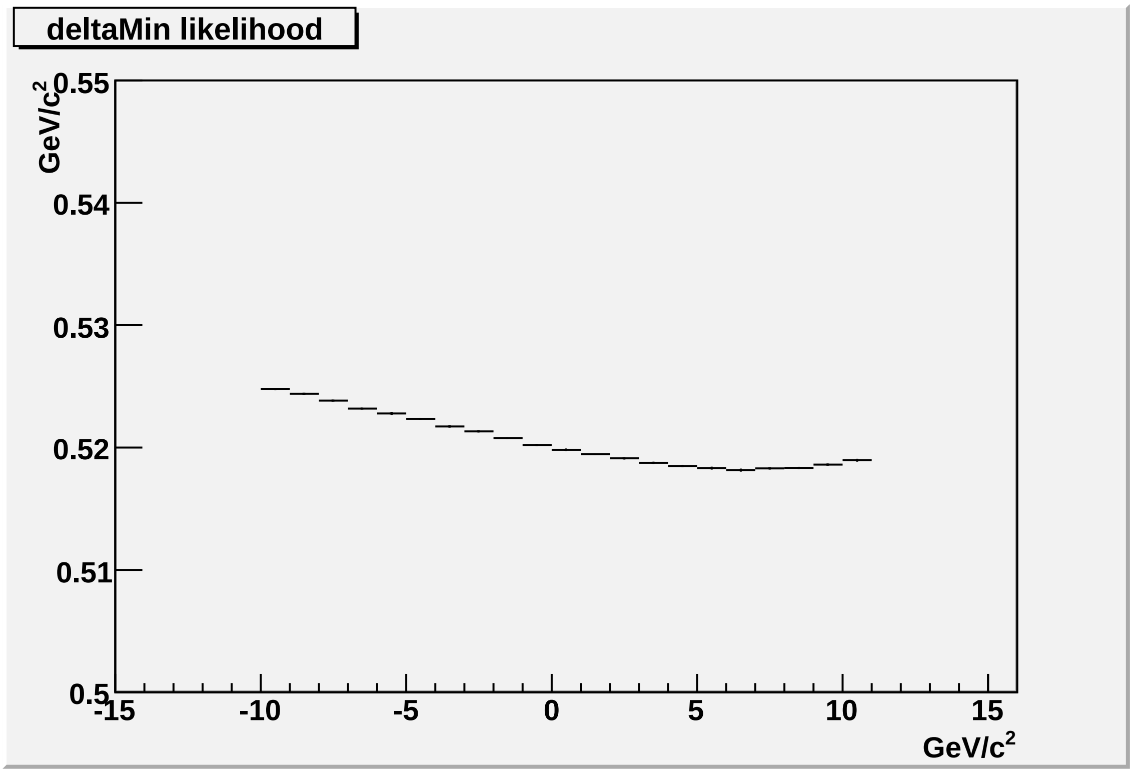

Figure 7.9: Statistical uncertainty as a function of $(\mathrm{M_{top}} - 175)\,\mathrm{GeV/c^2}$, calculated as $1/2(M^{+} - M^{-})$.

7.5.2 Systematic uncertainties on the top mass measurement

The evaluation of the systematic uncertainty on the mass measurement is a complicated issue for which we do not have all the tools available. We do not have, for

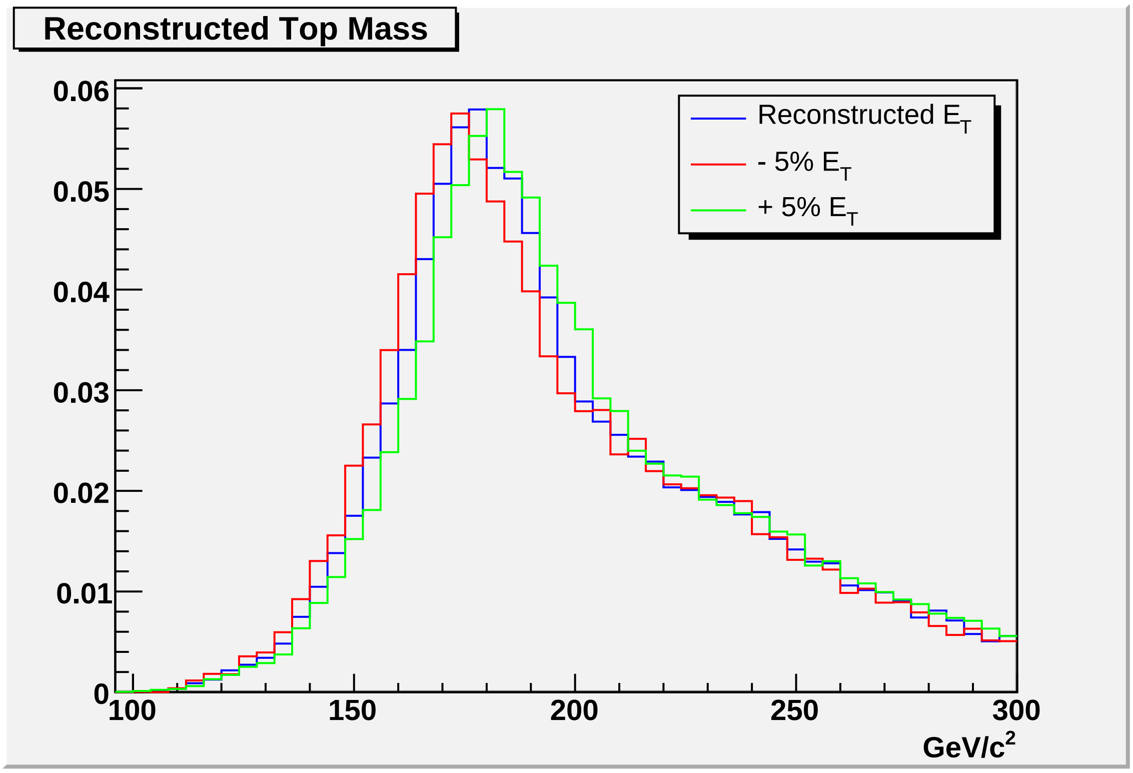

Figure 7.10: Reconstructed top mass for signal samples by increasing and decreasing each jet energy by 5% for sample generated with mass 175 GeV/c^2. Events with at least 1 b-tag.

instance, Monte Carlo generated with different assumption on Initial State Radiation (ISR) or Final State Radiation (FSR) or *p.d.f.*.

We expect, however, the systematic uncertainty, due to the lack of knowledge on the jet energy scale to be the dominating one.

In this case we assume a typical uncertainty of 5% on the jet energy scale and prepare a set of pseudo-experiments where the signal and background *p.d.f.* we draw the events from, are obtained increasing/decreasing each jet energy by 5%. Figure 7.10 shows for instance the distribution of the reconstructed mass for signal events for a +5% and -5% energy modification compared with the original one.

With such pseudo-experiments we would obtain a mass at 171.3 GeV/c^2 by increasing jet energy by 5% and 180.6 GeV/c^2 by decreasing jet energy by 5% for a generated mass of 175 GeV/c^2. We then estimate an expected systematic uncertainty of:

$$\Delta M_{syst} = \frac{180.9 - 171.3}{2} = 4.8\,\mathrm{GeV}/c^2 \tag{7.9}$$

This is clearly a large uncertainty. A big improvements to be expected from a two dimensional fit where we let the jet energy scale to vary and try to fit simultaneously the W masses. This is left for future developments.

A furter improvement can arise from a better evaluation of the background, since we have for the moment just discrete sample at different $\hat{p}_{\mathrm{T}}$: a data-driven background for instance, such as the whole QCD multi-jet sample from online streams, will give a more realistic characterization of the background.

7.5.3 Improving results with stronger b-tag cuts

Tables 6.5 and 6.6 show that the request for two b-tags can improve the S/B ratio. Moreover, Figure 7.2 shows that the same request shrink the overall shape of the reconstructed top mass. Both condition lead to better results on the likelihood comparison and the final. Unfortunately they result also in a lower efficiency thus reducing the overall statistic. Given the current reduced size of the Monte Carlo samples the events amount surviving is not enough to perform this step for the moment, but will result useful for measurement with real data.

Conclusion

In this work the $t\bar{t}$ events decay in the all-hadronic channel in the CMS experiment has been analyzed.

The analysis uses Monte Carlo simulated events both from official and private production and take advantages of the tools realized for the experiment community to deal with the large distributed CMS computing system.

The study is based on the LHC low luminosity $\mathcal{L} = 2 \times 10^{33}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ phase, foreseen for the first year of run, since the measurement can be done with the first fb^{-1} of data.

The data analyzed are the QCD multi-jet samples and $t\bar{t}$ samples from the official production, plus a private set of $t\bar{t}$ samples generated at different masses of the top quark.

A multi-jet trigger has been provided to fix a reasonable starting point. The integration of online available algorithms such as the request for a b-tag can provide a further reduction to reach the requested High Level Trigger rate

An offline selection has been provided to reduce the S/B ratio, but realized with tools available online such as track counting b-tag and jet multiplicity, so that they can be applied already at the High Level Trigger to reduce the initial rate. Parameters for the selection are based on the nominal event topology $t\bar{t} \to WWb\bar{b} \to qqqqb\bar{b}$ for the fully hadronic channel. The selection is based on the best statistical significance achievable and yields an expected $S/B \approx 1/23$ for $1\,\mathrm{fb}^{-1}$ of data.

The b-tag is applied in this analysis to provide a further S/B improvement, achieving $S/B \approx 1/7 \div 1/1.5$ depending on the used criteria.

The selection is applied to the background sample and to the samples generated at different top quark masses. The top quark mass candidate is reconstructed for all those samples using a kinematic fitter.

The kinematic fitter tries to find the best χ^2 association among jet so that, once associated a b quark to a b-tagged jet, two pairs of jet can reconstruct the nominal

W masses.

The resulting distributions are used to build *p.d.f.*'s, interpolating them with a continuous arbitrary curve. These curves are used to perform the top mass measurement through a likelihood comparison.

Pseudo-experiments are generated using these *p.d.f.*'s, populated with the total amount of events foreseen for $1\,\mathrm{fb}^{-1}$ of integrated luminosity for each of the 21 input top masses. The same procedure is applied for the QCD background, with a total amount of pseudo-experiments corresponding to the expected S/B ratio.

We perform the measurement over an ensemble of 100 different sets of pseudo-events for each mass, measuring the resulting top mass and estimating the statistical uncertainty. The resulting masses are plotted with respect to the related generated mass and interpolated with a line: the resulting function is used to re-scale the measured mass.

Since we have, for the time being, no data available for a top quark mass measurement, we are interested here on the expected performance of our method in terms of expected uncertainties. We evaluate the statistical uncertainty from the method described to be $0.5\,\mathrm{GeV/c^2}$.

We expect the systematic uncertainty due to the lack of knowledge on the jet energy scale to be the dominant one. We evaluate the systematic uncertainty assuming an typical uncertainty of 5% on the jet energy scale and we perform the mass reconstruction and pseudo-experiment increasing/decreasing each jet energy by 5%. Using them as input for the mass measurement, we then estimate an expected systematic uncertainty of $4.8\,\mathrm{GeV/c^2}$.

The current results suffer the high systematic uncertainty, which will be reduced with the online calibration of the calorimeters. They suffer also the limited size of the mass scan sample: providing mass scan samples populated with inclusive events expected at $1\,\mathrm{fb}^{-1}$ of integrated luminosity we expect significant improvements. A better evaluation of the background, for instance data-driven, e.g. using the whole QCD multi-jet sample from online streams, will provide a further reduction of systematic uncertainties and will provide a better S/B evaluation.

The work done has not the purpose to give realistic measurement of the top mass, but instead to build a measurement system for the top mass and check its validity. The method proves to be a powerful instrument to measure the top mass using events collected in the first year of run since it relies just on the knowledge of jets, which can be reached at very early stage of the detector operations, and the b-tag algorithms.

Bibliography

[1] LHC Homepage, http://cern.ch/lhc-new-homepage/

[2] The LHC Study Group: *The LHC Conceptual Design*, CERN/AC/95-05, 1995.

[3] CMS collaboration: *The Compact Muon Solenoid Technical Proposal*, CERN/LHCC 94-30.

[4] CMS collaboration: *CMS Physics Technical Design Report, Volume 1: Detector Performance and Software*, CERN/LHCC 2006-001, 2 February 2006.

[5] CMS collaboration: *CMS Physics Technical Design Report, Volume 2: Physics Performance*, CERN-LHCC-2006-021, 26 June 2006 and Journal of Physics G, Nuclear and Particle Physics, Volume 34, Number 6, June 2007.

[6] CMS collaboration: *CMS, The Trigger and Data Acquisition Project, Volume I: The Level-1 trigger, Technical Design Report*, CERN-LHCC-2000-038, 15 December 2000.

[7] CMS collaboration: *CMS, The Trigger and Data Acquisition Project, Volume I: Data Acquisition and High Level Trigger, Technical Design Report*, CERN-LHCC-2002-026, 15 December 2002.

[8] W. Adam et al. *The CMS High Level Trigger*, Eur. Phys. J. C 46 604 [hep-ex/0512077] 2005.

[9] CMS collaboration: *CMS, The Computing Project - Technical Design Report*, CERN/LHCC 2005-023, (2005)

[10] I. Foster, C. Kesselmann: *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Publishers ISBN 1-55860-475-8 (1999).

[11] I. Foster,C. Kesselmann, S. Tuecke: *The Anatomy of the Grid: Enabling Scalable Virtual Organizations* Lecture Notes in Computer Science (2001)

[12] LHC Computing Grid (LCG), Web Page, http://lcg.web.cern.ch/LCG/ and *LCG Computing Grid - Technical Design Report*, LCG-TDR-001 CERN/LHCC 2005-024, (2005)

[13] EGEE Homepage, http://www.cern.egee

[14] OSG Web Page, http:/opensciencegrid.org/

[15] PhEDEx - Physics Experiment Data Export - Project Homepage, http://cms-project-phedex.web.cern.ch/cms-project-phedex/

[16] ROOT, an object oriented data analysis framework, http://root.cern.ch/ and R. Brun, F. Rademakers *ROOT - An object Oriented Data Analysis Framework*, Nucl. Instrum. Meth. A 389 81, 1997.

[17] BOSS, Batch Object Submission system - Project Homepage, http:/boss.bo.infn.it/ and C. Grandi, A. Renzi, *Object Based system for Batch Job Submission and Monitoring (BOSS)* CMS NOTE 2003-005 (2003) and G. Codispoti et al., *BOSS: the CMS interface for job summission, monitoring and bookkeeping*, Workload Management and Workflows at the EGEE User Forum, CERN March 1st-3rd 2006

[18] D. Evans et al.: *CMS MC Production System Development & Design*, CHEP'07 International Conference on Computing in High Energy and Nuclear Physics , Victoria BC , Canada, 2-7 Sept 2007

[19] CRAB homepage, http://cmsdoc.cern.ch/cms/ccs/wm/www/Crab/ and D. Spiga et al., *The CMS Remote Analysis Builder (CRAB)* 14th Int. Conf. on High Performance Computing (HiPC 2007). Goa, India. Dec 18-21 2007. (vol. 4873, pp. 580-586). ISBN/ISSN: 978-3-540-77219-4.

[20] MySQL Home Page, http://www.mysql.com/

[21] SQLite Home Page, http://www.sqlite.org/

[22] gLite, Lightweigth Middleware for Grid Computing, http://glite.web.cern.ch/glite/

[23] Particle Data Group: *Review of Particle Physics* , Journal of Physics G33,1 , 2006.

[24] F. Gianotti: *Collider Physics at LHC*, CERN Yellow Reports 2000-007 (2000).

[25] F. Halzen, A.D. Martin: *Quarks and Leptons: an Introductory Course in Modern Particle Physics*, Wiley & Sons (1984).

[26] M. Guidry: *Gauge Field Theories: an Introduction with Applications*, Wiley & Sons (1991).

[27] I.J.R. Aitchinson, A.J.G. Hey: *Gauge Theories in Particle Physics*, Adam Hilger (1989).

[28] F. Mandl, G. Shaw: *Quantum Field Theory*, Wiley & Sons (1993).

[29] H.E. Haber, M. Schmitt: *Supersymmetry*, Particle Data Group (2002).

[30] M. Beneke at al.: *Top Quark Physics*, CERN-TH/2000-100, arXiv:hep-ph/0003033v1 (2000)

[31] C. Campagnari, M. Franklin, : *The discovery of the top quark*, Rev.Mod.Phys. 69 (1997) 137-212, UCSB-HEP-96-01 and HUTP-96/A023, arXiv:hep-ex/9608003v1 (1997)

[32] J. Hellis: *Beyond The Standard Model for Hillwalkers*, hep-ph/9812235 CERN Yellow Reports 1998-04 (1998).

[33] G.F. Giudice: *Physics Beyond the Standard Model*, hep-ph/9605390 CERN Yellow Reports 1996-04 (1996).

[34] A. Giammanco: *Top quark studies and perspectives with CMS*, CMS CR 2005/026, 20 October, 2005

[35] S.P. Mehdiabadi: *Top production and Search for SUSY at LHC*, CMS CR 2006/010, February 13, 2006

[36] F.P. Schilling: *Early Electroweak and Top Quark Physics with CSM*, CMS CR 2007/034, 20 June, 2007

[37] T. Sjstrand, P. Edn, C. Friberg, L. Lnnblad, G. Miu, S. Mrenna and E. Norrbin, Computer Physics Commun. 135 (2001) 238

[38] Comput.Phys.Commun. 148 (2002) 87-102 arXiv:hep-ph/0201292v1

[39] GEANT4 Collaboration, S. Agostinelli et al., *GEANT4: A simulation toolkit* Nucl. Instrum. and Methods A506, 2003.

[40] R. Demina et al., Calorimeter Cell Energy Thresholds for Jet Reconstruction in CMS, CMS Note 2006/020, 2006.

[41] UA1 Collaboration, G. Arnison et al., Hadronic Jet Production at the CERN Proton - Anti-Proton Collider, Phys. Lett. B132 (1983) 214. doi:10.1016/0370-2693(83)90254-X.

[42] S. V. Chekanov, Jet algorithms: A mini review, arXiv:hep-ph/0211298.

[43] J.W. Rohlf, Physics with jets at the LHC, Acta Phys. Polon. B36 (2005) 469-479. Available at http://th-www.if.uj.edu.pl/acta/vol36/pdf/v36p0469.pdf.

[44] J. M. Butterworth, J. P. Couchman, B. E. Cox, and B. M.Waugh, Kt-Jet: A C++ implementation of the K(T) clustering algorithm, Comput. Phys. Commun. 153 (2003) 8596, arXiv:hep-ph/0210022. doi:10.1016/S0010-4655(03)00156-5.

[45] S. D. Ellis and D. E. Soper, Successive combination jet algorithm for hadron collisions, Phys. Rev. D48 (1993) 31603166, arXiv:hep-ph/9305266. doi:10.1103/PhysRevD.48.3160.

[46] A. Rizzi, F. Palla, G. Segneri, Track impact parameter based b-tagging with CMS , CMS NOTE-2006/019

[47] J. Andrea, D. Bloch, D. Gel, P. Juillot, V. E. Bazterra, C. E. Gerber, F. Yumiceva, Evaluation of udsg Mistags for b-tagging using Negative Tags, CMS AN-2007/048

[48] A. Heister et al., Measurement of Jets with the CMS Detector at the LHC, CMS Note 2006/036 (2006).

[49] O. Kodolova, Jet Energy Measurements in CMS, CMS CR 2005-019 (2005). Presented at HCP 2005, Hadron Collider Physics Symposium, Les Diablerets, Switzerland, 46 July