

DATA-DRIVEN ESTIMATION OF Z^0
BACKGROUND CONTRIBUTIONS TO THE
HIGGS SEARCH IN THE $H \rightarrow \tau^+\tau^-$
CHANNEL WITH THE CMS EXPERIMENT
AT THE LHC

ARMIN BURGMEIER

DIPLOMA THESIS

AT THE PHYSICS DEPARTMENT OF
THE KARLSRUHE INSTITUTE OF TECHNOLOGY

*Referent: Prof. Dr. G. Quast
Institut für Experimentelle Kernphysik*

*Korreferent: Prof. Dr. W. de Boer
Institut für Experimentelle Kernphysik*

August 1, 2011

Deutsche Zusammenfassung

Schon seit jeher erforschen Menschen systematisch ihre Umgebung und versuchen sie zu ihrem Vorteil zu verändern. Diese Anstrengungen haben den heutigen Wohlstand erst ermöglicht. Dennoch ist die gezielte Forschung heute intensiver denn je - ein Zeichen dafür dass noch längst nicht alles verstanden ist was es zu verstehen gibt und dass eine Entdeckung oftmals viele neue Fragen aufwirft.

In der Grundlagenforschung wird die Erweiterung des Wissens der Menschheit angestrebt ohne gezielt Anwendungsfälle im Hinterkopf zu haben. Oft haben sich solche Entdeckungen später jedoch als äußerst nützlich erwiesen, darunter beispielsweise die Supraleitung. Die technischen Herausforderungen sind heutzutage so groß, dass alleine um die Experimente zu errichten neue Konzepte und Technologien entwickelt werden müssen. Das prominenteste Beispiel hierzu ist wohl das World Wide Web, das 1989 am Forschungszentrum CERN entwickelt wurde um Forschungsergebnisse innerhalb immer größer werdenden Teams austauschen zu können.

Im Bereich der Elementarteilchenphysik, in den diese Arbeit einzuordnen ist, führten Messungen an Teilchenbeschleunigern und kosmischer Strahlung zur Entwicklung des Standardmodells der Teilchenphysik. Das Standardmodell beschreibt die kleinsten bisher bekannten Bauteile der Materie, genannt Teilchen, und mit Ausnahme der Gravitation auch deren Wechselwirkungen. Es erlaubt Rückschlüsse auf die Entwicklung des frühen Universums zu ziehen als auch Vorhersagen für zukünftige Experimente zu treffen. Eine zentrale Vorhersage des Standardmodells ist die Existenz des Higgs-Bosons, ein bisher experimentell nicht nachgewiesenes Teilchen welches aber notwendig ist damit die Theorie konsistent ist und die experimentellen Befunde beschreiben kann. Im ersten Kapitel dieser Arbeit wird neben dem allgemeinen Aufbau des Standardmodells aufgezeigt, wie es zur Einführung des Higgs-Bosons kommt. Die Theorie ist jedoch nicht in der Lage, die Masse des Higgs-Bosons vorherzusagen; sie verbleibt ein freier Parameter.

Der Large Hadron Collider (LHC) am CERN ist das neue Flaggschiff der experimentellen Elementarteilchenphysik. Dabei handelt es sich um einen 27 km langen Ringbeschleuniger 100 m unter der Erde. Nach jahrzehntelanger Planung, Forschungs- und Entwicklungsarbeit und Konstruktion sowie diversen Verzögerungen konnte im Frühjahr 2010 das Forschungsprogramm aufgenommen werden. Im Ring werden zwei Protonenstrahlen auf Energien von 7 TeV (später 14 TeV) beschleunigt und gezielt an vorher festgelegten Punkten zur Kollision gebracht. Um diese Punkte herum wurden riesige Detektoranlagen errichtet, die die Kollisionsprodukte mit erstaunlicher Genauigkeit vermessen. Abbildung 1 zeigt den gewaltigen CMS-Detektor. Der LHC und insbesondere das CMS-Experiment werden im zweiten Kapitel dieser Arbeit ausführlich beschrieben.

Die Forscher am LHC verfolgen mehrere Ziele. Im Vordergrund steht jedoch der Nachweis des Higgs-Bosons, da er am Vorgängerexperiment LEP und am US-amerikanischen

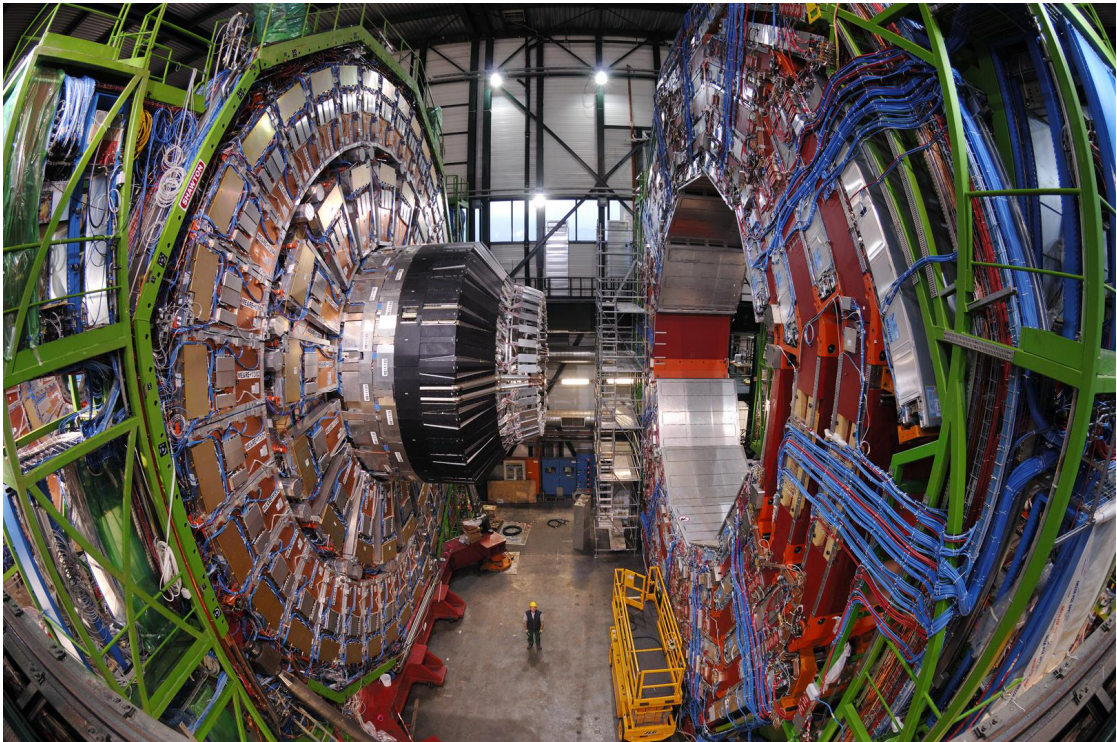


Abbildung 1: Der CMS-Detektor kurz vor Fertigstellung im November 2006.

Tevatron-Beschleuniger nicht gelungen ist. Der LHC wird nach wenigen Jahren Laufzeit in der Lage sein, das Higgs-Boson entweder nachzuweisen oder seine Existenz auszuschließen falls es nicht existiert. Desweiteren erhofft man sich, Zeichen von neuer, nicht vom Standardmodell beschriebener, Physik am LHC zu finden. Dies könnten neue Teilchen oder neue fundamentale Wechselwirkungen sein die erst bei sehr hohen Energien im TeV-Bereich eine Rolle spielen.

Unter Nominalbedingungen finden in jedem der LHC-Experimente 40 Millionen Kollisionen pro Sekunde statt von denen nur ein Bruchteil interessante Ereignisse liefert die für neue Entdeckungen relevant sind. Daher werden nur knapp 200 Kollisionen pro Sekunde aufgezeichnet während der Rest verworfen wird. Dennoch fallen so im Jahr Datenmengen in der Größenordnung von mehreren 10 Petabytes an die gespeichert und verarbeitet werden müssen: Aus den Rohdaten des Detektors müssen Teilchenspuren rekonstruiert und einzelne Teilchen identifiziert werden. Mit Hilfe von Monte Carlo-Simulationen werden die erwarteten Signale im Detektor modelliert. Im dritten Kapitel werden die dazu eingesetzten Software-Pakete näher erläutert.

Es kann jedoch nicht ein Rechenzentrum allein die enormen anfallenden Datenmengen bewältigen. Daher wurden viele Rechenzentren von Universitäten und Forschungseinrichtungen miteinander vernetzt und mit spezieller Software ausgestattet um sich das Speichern und Verarbeiten der LHC-Daten untereinander aufzuteilen. Dieser Verbund wird "Worldwide LHC Computing Grid" (WLCG) genannt. Das WLCG ist hierarchisch

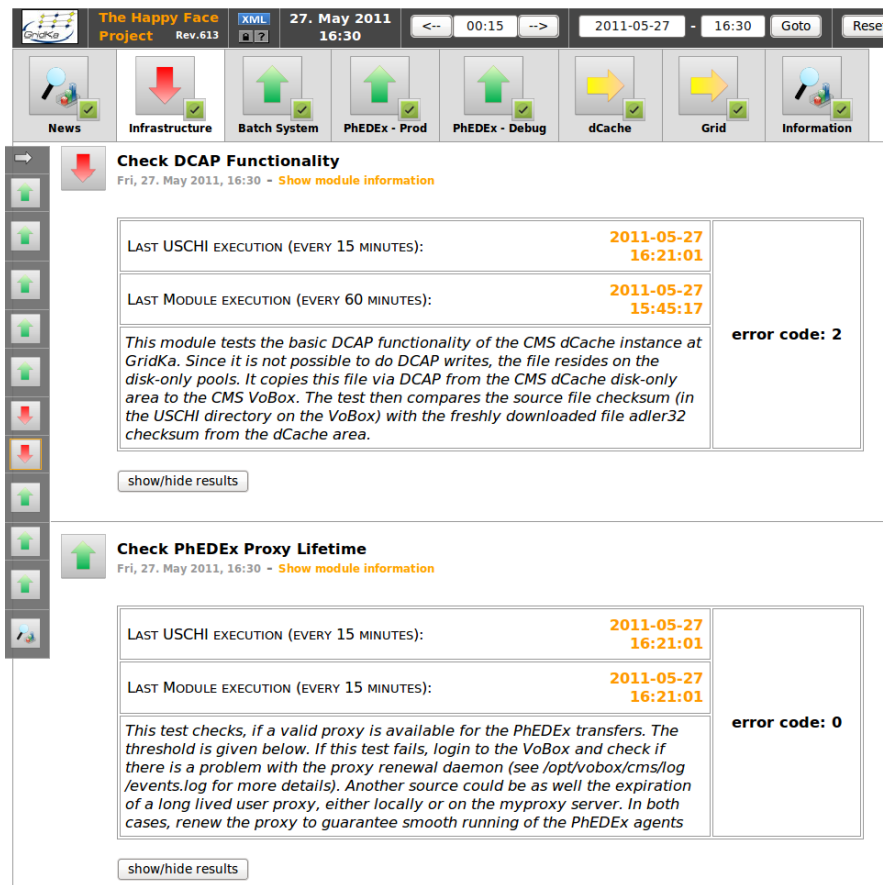


Abbildung 2: Bildschirmfoto der HAPPYFACE-Webseite. In der obersten Leiste kann der Zeitpunkt verändert werden zu dem der Status des Rechenzentrums angezeigt wird. Darunter können in einer weiteren Leiste zwischen verschiedenen Kategorien navigiert werden wobei die Pfeile gleich andeuten ob es in einer Ebene eventuell ein Problem gibt. Auf der Hauptseite werden dann einzelne Module angezeigt die eine bestimmte Funktionalität des Zentrums testen und je nach Ergebnis des Tests ein Problem (roter Pfeil) andeuten oder nicht (grüner Pfeil).

aufgebaut: Die oberste Ebene (Tier-0) befindet sich direkt am CERN und ist mit der Datenauslese der Experimente verbunden. Von dort werden die Daten an die nächste Ebene, die Tier-1-Zentren, verteilt. Diese halten eine Kopie der Daten vor und führen Rekonstruktionssoftware aus.

Das deutsche Tier-1-Zentrum (GridKa) steht in Karlsruhe am KIT. Zum Betrieb eines solchen Zentrums müssen viele verschiedene Systeme funktionieren: Nur beispielhaft genannt seien hier Speichersysteme die mehrere Petabyte an Daten vorhalten, die Verteilung der zur Verfügung stehenden Rechenleistung an einzelne Rechanforderungen,

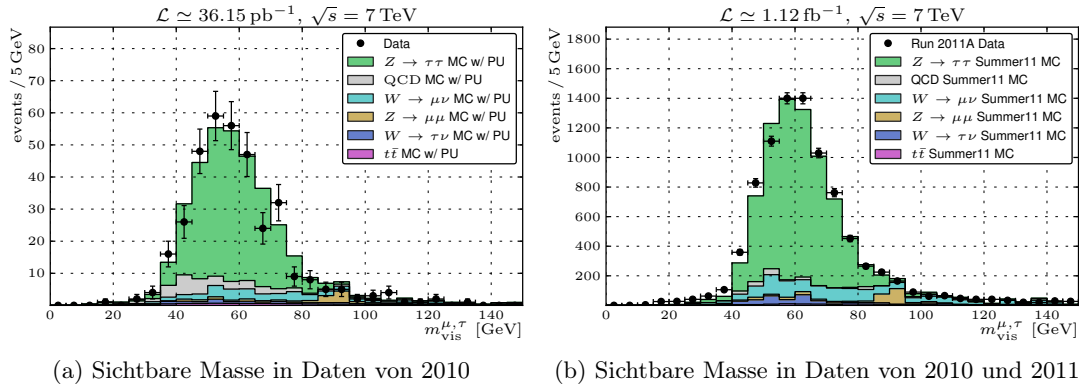
die Software auf den einzelnen Rechenknoten und die Netzwerkverbindungen innerhalb des Systems und zu den anderen Grid-Zentren. Diese Systeme müssen durchgehend überwacht werden damit das Tier-1-Zentrum seinen Anforderungen gerecht werden kann und sich die Wissenschaftler am LHC darauf verlassen können. Die meisten der Komponenten bieten in der Tat Systeme zur Überwachung an. Diese werden auch eingesetzt, allerdings ist es aufwendig und ermüdend, sämtliche Komponenten einzeln überprüfen zu müssen.

Daher wurde am Institut für Experimentelle Kernphysik des KIT eine Lösung entwickelt, die es sich zum Ziel gesetzt hat, die komplette Überwachung eines Rechenzentrums zu vereinheitlichen. Das HAPPYFACE PROJECT fragt die einzelnen, bereits existierenden, Überwachungskomponenten periodisch ab und stellt das Ergebnis kohärent auf einer einzigen Webseite dar, so dass sich auf einen Blick erkennen lässt, ob das Grid-Zentrum zuverlässig funktioniert oder ob Probleme bestehen. Da keine neuen Informationen generiert werden sondern lediglich bestehende Informationen gesammelt und ausgewertet werden spricht man in diesem Zusammenhang gerne von "Meta Monitoring". Weiterhin bietet HAPPYFACE die Möglichkeit, den Zustand des Zentrums zu einem früheren Zeitpunkt abzufragen sodass neue Probleme mit früheren in Beziehung gesetzt werden können. Abbildung 2 zeigt die von HAPPYFACE erzeugte Webseite.

Im Rahmen dieser Arbeit wurde HAPPYFACE signifikant weiterentwickelt: Neben der Implementation von neuen Modulen wurden im Kernbereich diverse Überarbeitungen durchgeführt um die Skalierbarkeit und Fehlerbehandlung zu verbessern. Im vierten Kapitel wird das Grid-Konzept präsentiert und anschließend detailliert auf HAPPYFACE eingegangen. Neben GridKa wird HAPPYFACE inzwischen auch an den Zentren in Aachen, Göttingen und Hamburg eingesetzt. Eine Einführung am CERN soll noch dieses Jahr erfolgen.

Der zweite Teil der Arbeit widmet sich der Suche nach dem Higgs-Boson. Das Higgs-Boson ist kein stabiles Teilchen sondern es zerfällt quasi instantan in zwei weitere Teilchen. Es kommt auf die (unbekannte) Masse des Higgs-Bosons an, welche Zerfallsmodi dabei dominant sind: Das Higgs-Boson koppelt bevorzugt an massereiche Teilchen, allerdings muss der Zerfall kinematisch erlaubt sein, d.h. der Energieerhaltungssatz darf nicht verletzt werden. Ein leichtes Higgs-Boson zerfällt also tendenziell in leichte Teilchen und ein schweres Higgs-Boson in schwerere. In dieser Arbeit wird der Zerfall in zwei τ -Leptonen thematisiert. Das τ -Lepton ist so etwas wie der schwere Bruder des Elektrons: Bis auf die Masse hat es die gleichen Eigenschaften wie ein Elektron oder auch ein Myon. Im Verhältnis zur Higgs-Masse ist das τ -Lepton jedoch immernoch sehr leicht, sodass dieser Zerfall des Higgs-Bosons nur für Massen unterhalb von etwa 150 GeV beobachtet werden wird. Dies ist allerdings auch der Bereich, in dem Präzisionsmessungen von elektroschwacher Physik die Higgs-Masse erwarten lassen, sodass es sich lohnt, diesen Zerfall zu studieren.

Es gibt noch ein weiteres Teilchen im Standardmodell welches in ein τ -Paar zerfällt: Das Z^0 -Boson mit einer Masse von etwa 91 GeV. Dieses ist sehr gut bekannt, sodass es sich anbietet die Suche nach τ -Paaren daran auszuprobieren. Da das τ -Lepton selbst wiederum nicht stabil ist sondern in Hadronen, ein Myon oder ein Elektron zerfällt ist die Rekonstruktion von Ereignissen in denen ein τ -Lepton vorkommt aufwendig. Das



(a) Sichtbare Masse in Daten von 2010

(b) Sichtbare Masse in Daten von 2010 und 2011

Abbildung 3: Die Verteilung der sichtbaren Masse von $\mu + \tau$ -jet-Ereigniskandidaten in gemessenen (schwarz) und simulierten (farbige Balken) Daten. Das linke Schaubild zeigt die Verteilung anhand von Daten die im Jahr 2010 genommen wurden während das rechte Bild alle Daten bis Anfang Juli 2011 beinhaltet. Die unterschiedlichen Farben markieren Beiträge von unterschiedlichen physikalischen Prozessen. In 2011 wurde ein Vielfaches der Datenmenge von 2010 aufgezeichnet und entsprechend sind die statistischen Fehler im rechten Schaubild deutlich kleiner.

τ -Lepton muss anhand der Zerfallsprodukte zunächst identifiziert und seine Eigenschaften anschließend rekonstruiert werden. Im fünften Kapitel dieser Arbeit wird mit den bisherigen Daten des CMS-Detektors vorgestellt, wie dies im einzelnen geschieht. Dabei werden insbesondere Endzustände betrachtet bei denen ein τ -Lepton in ein Myon zerfällt und das andere in Hadronen. Es wird eine Ereignisselektion vorgestellt die hilft, Untergrundbeiträge aus anderen physikalischen Prozessen die zu einer ähnliche Signatur im Detektor führen zu unterdrücken.

In Abbildung 3 wird die Verteilung der rekonstruierten Masse der Zerfallsprodukte zweier τ -Leptonen gezeigt. Der größte Beitrag kommt aus dem Zerfall von Z^0 -Bosonen. Die Verteilung der Masse ist deutlich gegen die Masse des Z^0 -Bosons verschoben, da beim τ -Zerfall auch Neutrinos entstehen die nicht nachgewiesen werden können. Diese tragen einen Teil des Impulses weg der dann nicht zur Massenrekonstruktion zur Verfügung steht. Die so rekonstruierte Masse wird daher auch "sichtbare Masse" genannt.

Obwohl der $Z^0 \rightarrow \tau^+ \tau^-$ -Zerfall gut geeignet ist um die Methoden zur Rekonstruktion von τ -Paaren auszuprobieren, stellt er einen gewissen Nachteil bei der Higgs-Suche dar. Aus einem gefundenen τ -Paar kann nämlich nicht geschlossen werden, ob es aus einem Z^0 - oder einem Higgs-Zerfall stammt. Der Nachweis ist daher statistisch zu erbringen: Beobachtet man mehr τ -Paare als von Z^0 -Zerfällen erwartet, so hat man möglicherweise ein Higgs-Signal gefunden. Es sind daher zwei Kriterien ausschlaggebend für die statistische Signifikanz eines solchen Signals: Zum einen die Qualität der Massenrekonstruktion, sodass Z^0 - und Higgs-Zerfallsprodukte aufgrund der unterschiedlichen Masse der beiden Teilchen getrennt werden können. Zum anderen ist die genaue Kenntnis des Z^0 -Beitrags

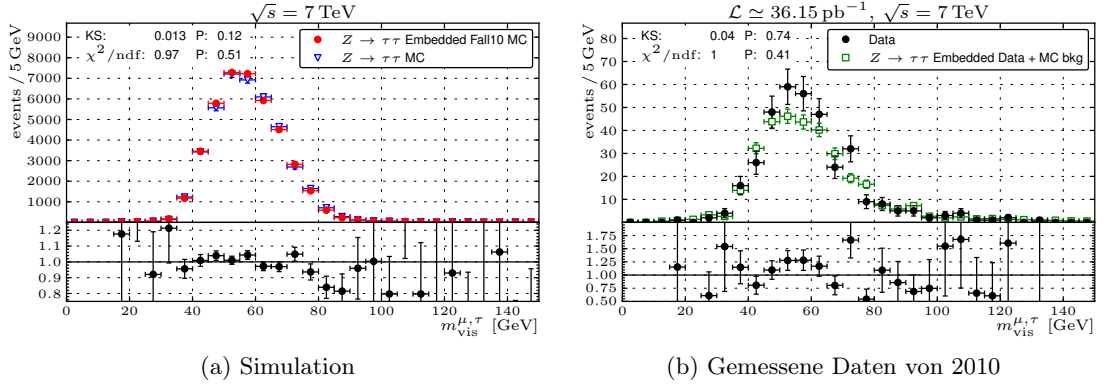


Abbildung 4: Verteilung der sichtbaren Masse von $Z \rightarrow \tau\tau$ -Ereignissen und $Z \rightarrow \mu\mu$ -Ereignissen wobei die Ersetzungsmethode angewandt wurde auf Simulation (links) und Daten des CMS-Experiments (rechts). Eine gute Übereinstimmung innerhalb der statistischen Ungenauigkeiten zeigt das Funktionieren der Methode.

wichtig, um bei einem beobachteten Überschuss nicht befürchten zu müssen, dass dieser nur aus Unkenntnis des Untergrunds resultiert.

Im sechsten und letzten Kapitel dieser Arbeit wird daher eine Methode vorgestellt, wie dieser Beitrag aus Daten abgeschätzt werden kann. Auf diese Weise muss man sich nicht ausschließlich auf Simulationen verlassen, die naturgemäß mit diversen systematischen Unsicherheiten behaftet sind. Um diese Abschätzung vorzunehmen nimmt man Zerfälle von Z^0 -Bosonen in zwei Myonen die man gerade mit dem CMS-Experiment sehr sauber selektieren kann. Anschließend werden in so einem Ereignis die rekonstruierten Myonen entfernt und durch simulierte τ -Leptonen ersetzt. Dadurch werden alle anderen Bestandteile des Ereignisses die von zusätzlichen Kollisionen oder den Wechselwirkungen der Protonüberreste entstehen direkt aus gemessenen Daten übernommen. Gerade diese Anteile sind es, die nur mit relativ großen Unsicherheiten simuliert werden können. Aufgrund der Lepton-Universalität zerfallen Z^0 -Bosonen genau gleich häufig in Myonen und τ -Leptonen. Daher eignet sich diese sogenannte Ersetzungsmethode um die Anzahl an $Z \rightarrow \tau\tau$ -Zerfällen präzise zu bestimmen.

Zum ersten Mal wurde die Ersetzungsmethode auf Daten des CMS-Experiments angewandt. In Abbildung 4 ist der Vergleich von solch künstlichen Ereignissen mit normalen $Z \rightarrow \tau\tau$ -Ereignissen in der sichtbaren Masse zu sehen. Die Übereinstimmung der Verteilung bestätigt das Funktionieren der Methode die im weiteren Verlauf der Higgs-Suche im Kanal $H \rightarrow \tau\tau$ ein mächtiges Werkzeug sein wird um die statistische Signifikanz sowohl bei einer eventuellen Entdeckung als auch beim Formulieren von Ausschlussgrenzen zu verbessern.

Contents

Introduction	1
1 The Standard Model of Particle Physics	3
1.1 Quantum Field Theory	4
1.2 Quantum Electrodynamics	6
1.3 Electroweak Unification	7
1.3.1 Weak Isospin and Parity Violation	8
1.3.2 The Glashow-Weinberg-Salam Model	8
1.4 The Higgs Mechanism	10
1.5 Quantum Chromodynamics	12
1.6 Experimental Verification	13
1.6.1 Higgs Boson Production	14
1.6.2 Higgs Boson Decay	15
1.6.3 Higgs Exclusion Limits	16
1.6.4 The $H \rightarrow \tau\tau$ Channel	17
1.7 Limitations of the Standard Model	17
2 The CMS Experiment at the LHC	19
2.1 The Large Hadron Collider	19
2.1.1 LHC Beam and Injection	22
2.1.2 Luminosity	23
2.2 The CMS Experiment	24
2.2.1 Coordinate System	25
2.2.2 Silicon Tracking Detector	26
2.2.3 Electromagnetic Calorimeter	27
2.2.4 Hadronic Calorimeter	28
2.2.5 Muon System	29
2.2.6 Particle Identification	30
2.2.7 Data Acquisition	31
3 High Energy Physics Software and Frameworks	35
3.1 Monte Carlo Event Generation	35
3.1.1 Pythia	36
3.1.2 Powheg	36
3.1.3 Tauola	36
3.2 ROOT	37

3.3	CMSSW	37
3.3.1	Event Data Model	38
3.3.2	Dataset Bookkeeping	38
3.3.3	Conditions Database	39
3.3.4	Detector Simulation	39
3.3.5	Event Reconstruction	40
3.3.6	Particle Flow	41
3.4	Analysis Workflow	42
4	The LHC Computing Grid	45
4.1	Grid Structure	45
4.2	Grid Architecture and Components	46
4.3	The HappyFace Project	48
4.3.1	Motivation	48
4.3.2	HappyFace Goals and Features	49
4.3.3	HappyFace Architecture	53
4.3.4	HappyFace Installation	55
4.3.5	HappyFace Core System	55
4.3.6	Available Modules	57
4.3.7	Conclusions and Future Work	61
5	Analysis of $\tau^+\tau^-$ Final States	63
5.1	τ Identification and Reconstruction	63
5.1.1	Leptonic τ Reconstruction	64
5.1.2	Hadronic τ Reconstruction	65
5.2	Mass Reconstruction	68
5.2.1	Visible Mass	68
5.2.2	Collinear Approximation Mass	69
5.2.3	SVfit Mass	71
5.3	The $\mu + \tau$ -jet Final State	72
5.3.1	Background Contributions	73
5.3.2	Selection Cuts	74
5.4	Update with 2011 Data	78
5.5	Possible Improvements	80
6	The Embedding Technique	83
6.1	Systematics Overview	84
6.2	$Z \rightarrow \mu\mu$ Selection	87
6.3	Particle Replacement	88
6.4	Direct Normalization	90
6.4.1	$\tau\tau \rightarrow \mu + \tau$ -jet Branching Ratio	91
6.4.2	$Z \rightarrow \mu\mu$ Selection Efficiency	91
6.4.3	Phase Space Restriction	92
6.4.4	Muon Isolation Efficiency	93

6.4.5	Trigger Efficiency on Data	95
6.5	Closure Test	96
6.6	Application on Data	97
6.7	Normalization With a Fit to the Mass Distribution	98
6.8	Conclusions and Outlook	100
Summary and Conclusions		101
A Additional HappyFace Modules		103
A.1	List of HappyFace Modules	103
A.2	Module Descriptions	104
A.2.1	CMSPhedexAgents	104
A.2.2	CMSPhedexErrorLog	105
A.2.3	DashboardDatasetUsage	106
A.2.4	dCacheDatasetRestoreLazy	107
A.2.5	dCacheInfoPool	108
A.2.6	dCacheTransfers	109
A.2.7	JobsDist	110
A.2.8	JobsEfficiencyPlot	111
A.2.9	SAM	112
B Datasets Used		113
B.1	Simulation	113
B.2	Data	114
C Additional Plots		115
C.1	$\tau^+\tau^-$ Final States in 2010 CMS Data	116
C.2	$\tau^+\tau^-$ Final States in 2011 CMS Data	117
C.3	Embedding Monte Carlo Closure Test	118
C.4	Embedding on 2010 CMS Data	119
List of Figures		119
List of Tables		122
Bibliography		125

Contents

Introduction

Today, being a researcher requires dedication to a narrow field of study since the vast existing knowledge must be learned before reaching a point where new discoveries can be made. Also, scientific breakthroughs often require teams of many scientists and years of planning and construction of complex machines. This is especially true for experimental high energy physics where hundreds or even thousands of people work together at the same experiment.

Most experimental input in high energy physics was obtained with collider experiments at SLAC, DESY Fermilab or CERN. Many fascinating and surprising results were found and many composite and fundamental particles have been discovered. This led to the formulation of the Standard Model of Particle Physics, a theory which describes all fundamental particles and interactions and which allows to predict the outcome of future experiments. Chapter 1 presents the basic ideas and predictions of the Standard Model. Special attention is given to the Higgs mechanism which postulates the existence of the Higgs boson, a particle that has not yet been observed experimentally.

The Large Hadron Collider (LHC) at CERN is the most powerful particle accelerator ever built. After tens of years of planning and construction, the physics program finally started in 2010. The LHC and its experiments are masterpieces of engineering pushing forward frontiers in technologies such as cryogenics, semiconductor sensors and superconducting magnets. One primary purpose of the experiments is to discover or exclude the Standard Model Higgs boson and to verify the Standard Model at energies at the TeV scale. Any deviations from the Standard Model in this energy region will be visible to the LHC experiments. Chapter 2 describes both the collider machine and the CMS experiment, one of the two general-purpose detectors at the LHC.

Operating the CMS experiment and analyzing its results requires dedicated software which is introduced in Chapter 3. This includes Monte Carlo techniques to simulate particle collisions and to model their signal in the detector as well as reconstruction and identification of individual particles from the raw detector output.

The LHC experiments produce several petabytes of data per year. The task of storing and processing this huge amount of data is shared by many computing centers all around the world which form the Worldwide LHC Computing Grid. Each Grid site consists of many different components such as network infrastructure, storage systems and batch systems which distribute computing jobs to individual worker nodes. Chapter 4 explains the components and the procedure of submitting Grid jobs. Furthermore, it presents HAPPYFACE, a tool to dramatically ease monitoring of the various components of a Grid center. Substantial contributions to HAPPYFACE have been carried out within the scope of this thesis.

Introduction

It should not be forgotten that all these technical challenges are undertaken in order to be able to analyze the data output of the CMS detector, such as the search for the Higgs boson. The Higgs boson, if it exists, will be observed via its decay products. Since its mass is not known a priori and in order to verify the properties of a potential discovery, a wide range of possible decay channels must be studied. If the Higgs boson is light ($m_H \lesssim 150$ GeV) it predominantly decays into a pair of τ leptons. An analysis of this final state was performed and is described in Chapter 5.

The Z^0 boson can also decay into two τ leptons. This allows the analysis of this final state to be commissioned, but it also makes it more difficult to observe a Higgs signal: when a slight excess in $\tau\tau$ final states is observed it could either be interpreted as a Higgs discovery or as a statistical fluctuation. In order to make this decision it is essential that the expected contribution of $Z \rightarrow \tau\tau$ events is known very precisely. This contribution can be estimated from simulated events, however this procedure results in large systematic uncertainties. It is therefore preferable to estimate this number from data. Chapter 6 discusses and presents results obtained with a new method in which the muons in measured $Z \rightarrow \mu\mu$ events are replaced by simulated τ leptons. This *Embedding* method was applied the first time to CMS data within the scope of this thesis.

1 The Standard Model of Particle Physics

In particle physics, physicists strive to understand the most basic building blocks of matter and the interactions between them. The gold foil experiment by Ernest Rutherford in 1909 [1] marks the start of modern particle physics. The experiment yielded insights into the substructure of the atom, namely that it consists of a dense, charged core and surrounding electrons.

Since then the experimental methods only changed marginally. All modern particle accelerators still perform scattering experiments of various particles, even if most machines collide two beams instead of smashing one beam at a fixed target. This way the composition of nuclei was probed, inelastic scattering experiments revealed the substructure of the proton and myriads of new composite and fundamental particles have been discovered.

The results obtained in such scattering experiments is used as an input to formulate theories which not only explain the results of past experiments but which are also able to predict the outcome of future experiments. New experiments eventually verify or falsify existing theories. *Quantum field theory* (QFT), the theory modern particle physics is based on, is especially remarkable in this regard as it has been verified to an unprecedented accuracy.

There are four different forces between particles known to date. Three of them can be successfully described by quantum field theory. Together they are referred to as the *Standard Model of Particle Physics*. In QFT, interactions are mediated via force carrier particles, so-called gauge bosons. They are summarized in Table 1.1. Gravity is the only force which could not yet be consistently formulated as a QFT. It is described by the theory of *General Relativity* which does not take quantum mechanical effects into account.

While forces are mediated via gauge bosons, all matter consists of fermions. These particles come in three “generations” where particles in different generations only differ in their masses but have exactly the same properties otherwise. Table 1.2 summarizes the fermions. For each fermion there is also an anti-fermion with the same properties as the fermion but inverse couplings to the Standard Model interactions.

In the following a brief introduction to the ideas (Section 1.1) and the essential results (Sections 1.2, 1.3, 1.4 and 1.5) of quantum field theory are given, especially in the electroweak sector of the Standard Model which is most relevant for the remainder of this thesis. Section 1.6 presents experimental results and Section 1.7 concludes by briefly discussing the shortcomings of the current theory. A full introduction into QFT can be found in [2].

Force	Carrier	Mass [GeV]	Range	Example
Strong	8 Gluons	0	10^{-15} m	Holding together nuclei
Weak	W^\pm boson	80.4	10^{-18} m	Radioactive β decay
	Z^0 boson	91.2		
Electromagnetic	Photon	0	∞	Radio communication
Gravitation	(Graviton)	0	∞	Motion of planets

Table 1.1: The four fundamental interactions. For each force also an example of an interaction based on the corresponding force is given. The graviton is not part of the Standard Model and has not yet been observed experimentally. However, if gravity can be described by a quantum field theory the corresponding force carrier would be called “Graviton”.

1.1 Quantum Field Theory

The basic idea of quantum field theory comes from classical field theory. In classical field theory a dynamical system minimizes the action S as it propagates from one state to another. S can be expressed as an integral of the Lagrangian L or the Lagrangian density \mathcal{L} :

$$S = \int \mathcal{L}(\phi, \partial_\mu \phi) d^4x. \quad (1.1)$$

The Lagrangian depends on one or more fields ϕ and their derivatives $\partial_\mu \phi$, usually composed of a kinetic term, a (rest) mass term and interaction terms. From the principle of least action the Euler-Lagrange equation can be derived:

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi)} \right) - \frac{\partial \mathcal{L}}{\partial \phi} = 0 \quad (1.2)$$

In classical field theory the fields are real or complex functions. In QFT a process called “second quantization” takes place which replaces the fields by operators obeying the same commutation relations as the classical variables. The states the operators can be applied on are specified by the number of particles with a certain momentum p (and spin in case of non-scalar fields). The ensemble of such states is called “Fock space”. As with the quantum mechanical harmonic oscillator, there exist ladder operators a_p and a_p^\dagger which create or destroy a particle with momentum p . The field $\phi(x)$ can be written in terms of a_p and a_p^\dagger as a Fourier integral. For example, for a scalar field it is given by

$$\phi(x) = \int \frac{d^3p}{(2\pi)^3} \frac{1}{\sqrt{2p^0}} \left(a_p e^{ip_\mu x^\mu} + a_p^\dagger e^{-ip_\mu x^\mu} \right) \quad (1.3)$$

with $p^0 = \sqrt{\vec{p}^2 + m^2}$.

	Generation			Charge	Weak Isospin	Color
	1	2	3			
Leptons	$\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}$	$\begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}$	$\begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}$	0 $-e$	$+\frac{1}{2}$ $-\frac{1}{2}$	– –
Quarks	$\begin{pmatrix} u \\ d \end{pmatrix}$	$\begin{pmatrix} c \\ s \end{pmatrix}$	$\begin{pmatrix} t \\ b \end{pmatrix}$	$+\frac{2}{3}e$ $-\frac{1}{3}e$	$+\frac{1}{2}$ $-\frac{1}{2}$	r, g, b r, g, b

Table 1.2: The various fermions are categorized into leptons (top) and quarks (bottom). The quarks interact strongly, electromagnetically and weakly. They form mesons and baryons such as pions, protons or neutrons. The leptons only interact electromagnetically (if charged) and weakly. The second and third generation fermions are so heavy that they eventually decay to the first generation ones, except for the neutrinos which are approximately massless. Charge, weak isospin and color denote the coupling strength to the electromagnetic interaction, the weak interaction or the strong interaction, respectively.

Physical Lagrange densities should yield the known differential equations for free particles, that is the Klein-Gordon equation for scalar fields, the Dirac equation for spin- $\frac{1}{2}$ fields and the Maxwell equations for massless spin-1 fields.

In order to calculate an observable quantity, such as a cross section or a decay width, the transition amplitude $\langle f|H_{\text{int}}|i\rangle$ must be computed where $|f\rangle$ and $|i\rangle$ denote the initial and final states, respectively, and H_{int} is the interaction Hamiltonian which can be derived from the Lagrangian density. Eventually, observables are proportional to the magnitude squared of the matrix element.

Since, for interacting processes, such matrix elements cannot be computed analytically, one resorts to perturbation theory. Coupling constants such as the electric charge e or the weak or strong coupling constants α_W or α_S for the weak or strong interactions, respectively, are used as the perturbation parameter. The procedure of deriving this perturbation series is highly nontrivial and at this point only a reference to [2] shall be given.

Feynman Diagrams. Every term in the perturbation series can be assigned a schematic drawing called a *Feynman diagram* [3]. A Feynman diagram consists of propagators (possibly virtual particles with a specific momentum and spin) and vertices (interactions between particles). Every element in the Feynman diagram corresponds to a term in the matrix element. These translations from diagram elements to mathematical terms are known as *Feynman rules*. Propagators not connected to a vertex represent external particles in the initial or final state.

Figure 1.1 shows example Feynman diagrams for the $e^+e^- \rightarrow \mu^+\mu^-$ process of quantum electrodynamics (QED). Every vertex contributes a factor of $\sqrt{\alpha}$ to the matrix element, and therefore diagrams with many vertices are higher order in perturbation theory. This

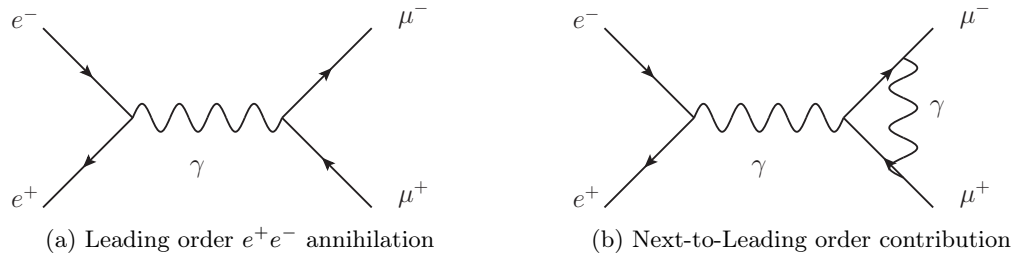


Figure 1.1: Example Feynman diagrams for the $e^+e^- \rightarrow \mu^+\mu^-$ process (e^+e^- annihilation). The first diagram shows the leading order contribution. The second diagram shows one of many second order contributions where one additional photon is exchanged between the final state particles. Other second order contributions include photon exchange of the initial state particles or pair production of the photon leading to a fermion loop.

explains why diagram 1.1a is the leading order diagram of the process and 1.1b is a higher order diagram with lower contribution to the matrix element.

For higher order effects it usually happens that loops occur in Feynman diagrams. In this case it must be integrated over all possible momenta of the particles within the loop, which leads to divergencies. In order to circumvent such divergencies and to obtain finite numbers for observables a procedure called renormalization must be applied. Renormalizability is a feature of a particular quantum field theory, and in fact the reason why no quantum field theory can be formulated for gravity is that such a theory would not be renormalizable. However, the mathematical concepts behind renormalizability are again beyond the scope for this thesis.

1.2 Quantum Electrodynamics

Quantum electrodynamics is the theory which describes all electromagnetic effects and interactions. It is the simplest of the three interactions of the Standard Model but its basic ideas are also applicable to the weak and strong interactions.

As a quantum field theory, QED is fully characterized by its Lagrangian density. The Lagrangian density for free fermions is given by

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi} (i\cancel{\partial} - m) \psi, \quad (1.4)$$

where ψ is a Dirac spinor field and $\bar{\psi} = \psi^\dagger \gamma^0$. The motivation for this form of the Lagrangian density is that plugging it into the Euler-Lagrange equations leads to the well-known Dirac equation for spin- $\frac{1}{2}$ fermions.

Gauge Transformations. Classical electrodynamics is a gauge theory, which means that the four-potential A_μ can be transformed as

$$A_\mu \rightarrow A'_\mu = A_\mu - \partial_\mu \Lambda(x) \quad (1.5)$$

with an arbitrary scalar field Λ . This transformation has no effect on the observable quantities \vec{E} and \vec{B} : they are invariant under *local* gauge transformations.

This property motivates a similar invariance in quantum electrodynamics. For the spinor field ψ a *global* phase transformation

$$\psi \rightarrow \psi' = e^{i\alpha} \psi \quad (1.6)$$

vanishes when applied in the Lagrangian density. However, in classical electrodynamics Λ may depend on space-time. If $\alpha = \alpha(x)$, i.e. it is a *local* phase transformation, then an additional term appears because of the derivative in the Lagrangian density. This additional term is nonzero and therefore breaks the gauge invariance.

In order to restore gauge invariance the derivative ∂_μ is substituted by the *covariant derivative*,

$$D_\mu = \partial_\mu + ieA_\mu(x), \quad (1.7)$$

where A_μ is a new vector field. Its transformation property under local gauge transformations can easily be derived by requiring the Lagrangian density to be invariant:

$$A_\mu \rightarrow A'_\mu = A_\mu - \frac{1}{e} \partial_\mu \alpha(x). \quad (1.8)$$

This transformation is astonishingly similar to the gauge transformation of classical electrodynamics, Equation 1.5. At this point it is easy to identify the field A_μ as the photon field. In other words, postulation of local gauge invariance introduces the photon into the theory. The phase transformation, Equation 1.6, is the symmetry transformation of the U(1) group. Therefore, the theory is said to be invariant under U(1) transformations.

With the covariant derivative in place an additional term in the Lagrangian density shows up, $e\bar{\psi}\gamma^\mu\psi A_\mu$, which describes the coupling of fermions to the field A_μ . However, the Lagrangian density needs to be extended further to fully account for the photon field:

$$\mathcal{L}_{\text{QED}} = \bar{\psi} (i\partial\!\!\!/ - m) \psi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - e\bar{\psi}\gamma^\mu\psi A_\mu, \quad (1.9)$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the electromagnetic *field strength tensor*. The $\frac{1}{4} F_{\mu\nu} F^{\mu\nu}$ term is the kinetic term for the photon field. A mass term along the lines of $\frac{1}{2} m^2 A_\mu A^\mu$ must not be added because it would spoil gauge invariance again. This is in perfect agreement with the observation that the photon is a massless particle.

Equation 1.9 is the full Lagrangian of QED. The Euler-Lagrange equations for A_μ lead to the inhomogeneous Maxwell equations, $\partial_\mu F^{\mu\nu} = e\bar{\psi}\gamma^\nu\psi = ej^\nu$.

1.3 Electroweak Unification

As QED, the theory of the weak interaction is a quantum field theory. In the first part of this section, the differences between the two interactions are discussed. In the second

part, gauge invariance is postulated in order to obtain the force carriers of the weak interaction, the W and Z bosons. It turns out that, in order to describe experimental results consistently, the electromagnetic and weak interactions need to be unified into a single electroweak interaction.

1.3.1 Weak Isospin and Parity Violation

The weak interaction can turn charged leptons to neutrinos or up-type quarks to down-type quarks and vice versa. This motivates a spin-like formalism where particles are arranged in doublets of Dirac fields,

$$\psi = \begin{pmatrix} \psi_\nu(x) \\ \psi_e(x) \end{pmatrix}_L. \quad (1.10)$$

Particles in such a doublet are said to have *weak isospin* $T = \frac{1}{2}$ where the third component of weak isospin is $T_3 = +\frac{1}{2}$ for neutrinos and up-type quarks and $T_3 = -\frac{1}{2}$ for charged leptons and down-type quarks. T_3 can be seen as the “charge” of weak interaction.

The index L in Equation 1.10 denotes a left-handed doublet. The *helicity* of a particle is defined as the projection of its spin to its momentum:

$$h = \frac{\vec{\sigma} \cdot \vec{p}}{|\vec{\sigma}| \cdot |\vec{p}|}. \quad (1.11)$$

For a spin- $\frac{1}{2}$ particle it can be either $+1$ (right-handed) or -1 (left-handed). The famous Wu experiment [4] has shown that the weak interaction only couples to left-handed fermions. In other words, right-handed fermions form a singlet with respect to the weak interaction with $T = T_3 = 0$. Therefore, right-handed charged leptons only interact electromagnetically and right-handed neutrinos are not observed at all. The different coupling of left-handed and right-handed fermions is known as *parity violation* of the weak interaction.

1.3.2 The Glashow-Weinberg-Salam Model

For the weak interaction similar ideas as for QED can be applied: the Lagrangian density for this fermion doublet, $\bar{\psi}(i\cancel{\partial} - m)\psi$, should be invariant under gauge transformations. In contrast to the QED case there are now three linearly independent ways to transform the phase of a doublet of complex fields, or more generally the $SU(2)$ symmetry group. They are given by the Pauli matrices $\vec{\sigma}$ which are therefore called the generators of the $SU(2)$ group. The transformation of the fields is given by

$$\psi \rightarrow \psi' = e^{\frac{i}{2}\vec{\sigma} \cdot \vec{\alpha}(x)}\psi. \quad (1.12)$$

Again, to restore invariance of the Lagrangian density under this transformation a covariant derivative is introduced:

$$D_\mu = \partial_\mu - \frac{i}{2}g\vec{\sigma} \cdot \vec{W}_\mu \quad (1.13)$$

where g is a coupling constant and \vec{W}_μ are three new vector fields. It is now tempting to identify these as the W^+ , the W^- and the Z^0 , the carriers of the weak force found experimentally. However, there is a catch here: it has been observed that the Z boson couples differently to neutrinos (or up-type quarks) than to charged leptons (or down-type quarks), for example by measuring branching fractions of the Z boson.

To solve this problem, both the electromagnetic and the weak interaction must be considered together, leading to electroweak unification or the Glashow-Weinberg-Salam (GWS) model [5]. In this case the covariant derivative becomes

$$D_\mu = \partial_\mu - \frac{1}{2}ig\vec{\sigma} \cdot \vec{W}_\mu - \frac{1}{2}ig'B_\mu, \quad (1.14)$$

where the last term comes from the U(1) symmetry of QED. It is the same as Equation 1.7 where e and A_μ have been renamed to $\frac{g'}{2}$ and B_μ for reasons that will become obvious soon. The covariant derivative is plugged into the kinetic Lagrangian density term for left-handed fermion fields to obtain the couplings of the fields to the gauge bosons:

$$i\bar{\psi}D_\mu\gamma^\mu\psi = \bar{\psi} \left(i\partial_\mu + \frac{1}{2}g\vec{\sigma} \cdot \vec{W}_\mu + \frac{1}{2}g'B_\mu \right) \gamma^\mu\psi \quad (1.15)$$

$$= \bar{\psi} \left(i\partial_\mu + \frac{1}{2} \begin{pmatrix} g'B_\mu + gW_\mu^3 & gW_\mu^1 - igW_\mu^2 \\ gW_\mu^1 + igW_\mu^2 & g'B_\mu - gW_\mu^3 \end{pmatrix} \right) \gamma^\mu\psi. \quad (1.16)$$

What can be learned from Equation 1.16 is that what actually couples to the fermion fields are not the B_μ and \vec{W}_μ fields, but linear combinations of them:

$$W_\mu^+ = \frac{W_\mu^1 - iW_\mu^2}{\sqrt{2}} \quad (1.17)$$

$$W_\mu^- = \frac{W_\mu^1 + iW_\mu^2}{\sqrt{2}} \quad (1.18)$$

$$A_\mu = \frac{g'B_\mu + gW_\mu^3}{\sqrt{g^2 + g'^2}} \quad (1.19)$$

$$Z_\mu^0 = \frac{-g'B_\mu + gW_\mu^3}{\sqrt{g^2 + g'^2}} \quad (1.20)$$

The mixing of the B and W^3 fields explains how the coupling of the Z boson also has an electromagnetic component, and therefore also depends on the electric charge. The mixing can also be parameterized by the electroweak mixing angle, called Weinberg angle:

$$\tan \vartheta_W = \frac{g'}{g}. \quad (1.21)$$

Experimentally, it can be determined by cross section measurements of elastic neutrino-nucleon scattering [6].

However, there is still one problem remaining. It has been verified experimentally that the W^\pm and the Z^0 particles are not massless [7, 8]. Introducing a mass term for the gauge fields would spoil the gauge invariance, though. This inconsistency can be explained theoretically by introducing the Higgs mechanism which is discussed in the following section.

1.4 The Higgs Mechanism

The idea is that instead of adding a mass term for the gauge bosons directly into the Lagrangian density, new fields are added. The mass terms arise from the interaction of the gauge boson fields with the new fields. This mechanism was first proposed by P. Higgs and others [9, 10, 11].

Since three masses need to be generated, the new field needs at least three degrees of freedom. The simplest way to do this is to introduce a doublet of complex scalar fields,

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}. \quad (1.22)$$

The Lagrangian density is extended by terms which are invariant under $SU(2) \otimes U(1)$ transformations:

$$\mathcal{L}_\Phi = (D_\mu \Phi)^2 - \underbrace{\mu^2 |\Phi^\dagger \Phi| - \lambda |\Phi^\dagger \Phi|^2}_{-V(\Phi)}, \quad (1.23)$$

where μ has the dimensions of a mass and λ is a dimensionless constant. The reason why there cannot be any Φ^6 or higher terms is that it would lead to a non-renormalizable theory, however showing this is beyond the scope of this thesis.

Figure 1.2 shows the form of the potential for $\mu^2 > 0$ and $\mu^2 < 0$. In the second case the minimum of the potential is not at the origin but at

$$|\langle \Phi_0 \rangle| = \frac{v}{\sqrt{2}} = \sqrt{\frac{-\mu^2}{2\lambda}}. \quad (1.24)$$

The newly introduced parameter v is called the *vacuum expectation value*. Since the Lagrangian density is invariant under $SU(2) \otimes U(1)$ transformations the vacuum expectation value will not change under such a transformation. Therefore, it is possible to find a transformation such that

$$\langle \Phi_0 \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \quad (1.25)$$

The system will spontaneously fall into one of the many possible ground states. In the ground state the system is located in a minimum of the potential where it is no longer invariant under $SU(2)$ transformations. Figure 1.2 visualizes this: the potential does not change when one rotates the coordinate system around the Y axis. However, if the

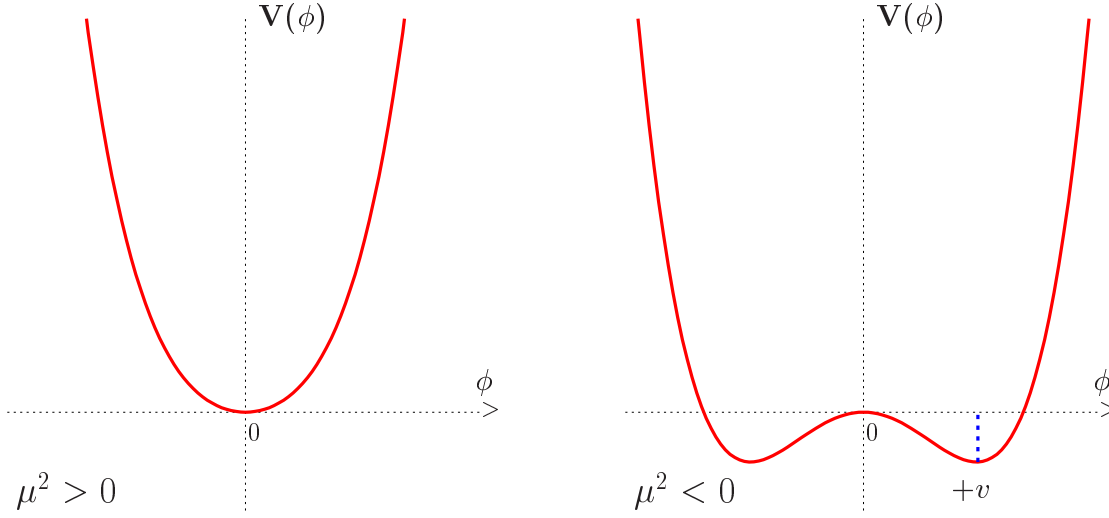


Figure 1.2: Potential of the field Φ for (a) $\mu^2 > 0$ and (b) $\mu^2 < 0$. In the second case there is more than one ground state and the system chooses one at random. From [12].

center of the rotation is located in the minimum of the potential, then in the $\mu^2 < 0$ case the curve does not stay invariant. This phenomenon is called *spontaneous symmetry breaking*. The theory remains unbroken under $U(1)$, however. This will be the reason why the photon remains massless in the following.

The gauge boson masses arise from the kinetic term in the Lagrangian density when it is evaluated at the vacuum:

$$(D_\mu \Phi)^2 = \Phi^\dagger \left(\partial_\mu + ig\vec{W}_\mu \cdot \frac{\vec{\sigma}}{2} + \frac{1}{2}ig'B_\mu \right) \left(\partial^\mu - ig\vec{W}^\mu \cdot \frac{\vec{\sigma}}{2} - \frac{1}{2}ig'B^\mu \right) \Phi \quad (1.26)$$

$$= \frac{1}{2} \cdot \frac{1}{4} v^2 \left(|gW_\mu^1 - igW_\mu^2|^2 + |g'B_\mu - gW_\mu^3|^2 \right) + \dots, \quad (1.27)$$

where

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \quad (1.28)$$

has been expanded around the vacuum. The real field $H(x)$ is called the *Higgs* field. It results as a remaining degree of freedom from the original complex doublet and, by construction, vanishes at the minimum of the potential.

In Equation 1.27 terms containing H or ∂_μ have been omitted. They lead to the kinetic term for the Higgs field and to interaction terms between gauge bosons and the Higgs field. What remains are mass terms for the W bosons and the Z boson (note that $|gW_\mu^1 - igW_\mu^2| = |gW_\mu^1 + igW_\mu^2|$, and therefore the mass term in Equation 1.27 accounts

1 The Standard Model of Particle Physics

for both the W^+ and the W^-). There is no mass term for the photon field, so it remains massless. The masses can be directly read from Equation 1.27:

$$m_W = g\frac{v}{2}, \quad m_Z = \sqrt{g^2 + g'^2}\frac{v}{2}. \quad (1.29)$$

This implicates that the W and Z masses are not independent but related by the Weinberg angle:

$$\frac{m_W}{m_Z} = \cos \vartheta_W. \quad (1.30)$$

Evaluating the potential terms in the original Lagrangian leads to the mass term for the Higgs field, $\mu^2 H^2$. The mass of the Higgs boson, $\sqrt{2}\mu$, is a free parameter of the theory and cannot be predicted. Even though the value of v is known to date by measurements of the W mass and the Weinberg angle, the values of μ and λ are unknown.

1.5 Quantum Chromodynamics

Quantum Chromodynamics (QCD) is the quantum field theory of strong interaction. The charge of strong interaction is called color, however this is just an analogy and has nothing to do with actual colors. Quarks are arranged in color-triplets,

$$\psi = \begin{pmatrix} \psi_r \\ \psi_g \\ \psi_b \end{pmatrix}, \quad (1.31)$$

so the Lagrangian density is postulated to be invariant under $SU(3)$ transformations. In a similar way as for the electromagnetic and weak interactions this leads to eight gauge bosons, called gluons. A gluon carries both color and anti-color so that, due to color conservation, quarks change their color when interacting with a gluon. The leptons do not carry color charge and therefore form a $SU(3)$ singlet.

The gluons are color-charged themselves, and therefore interact with each other. This is different from QED where photons do not carry electric charge and it leads to an important consequence. The effective potential of a color-charged particle is proportional to

$$V_c(r) \propto \alpha \frac{1}{r} + \beta r. \quad (1.32)$$

The first term is attributed to the color charge of quarks, and as with the electromagnetic interaction the potential diminishes at large distances. The second term, which originates from gluon self-coupling, however leads to much energy being stored in the color field for color charges which are far apart from each other. At distances of about 1 fm it is energetically favorable to create a new quark-antiquark pair out of the vacuum to shorten the distances between individual quarks. The conclusion of this is that no free quarks can be observed since they always arrange with other quarks or antiquarks to form color-neutral objects, called hadrons. This effect is called *color confinement* of QCD. Possible

arrangements include mesons (color and anticolor) and baryons (red, green and blue or anti-red, anti-green and anti-blue).

The masses of mesons or baryons are usually much higher than the masses of their quark constituents. For example, the proton, which consists of two up quarks and one down quark (so-called *valence quarks*), has a mass of 950 MeV whereas the quarks themselves have masses around 5 MeV. The remainder of the mass is contributed by the color field between the three quarks, i.e. carried by the gluons they exchange. The gluons can also, temporarily, generate additional quark-antiquark pairs in accordance with Heisenberg's uncertainty principle. Therefore, the probability of, for instance, finding a strange or a charm quark within the proton is not zero. Such temporary quarks are called *sea quarks*.

Parton Density Functions. The total momentum of a hadron is split among its constituents (called partons), the valence quarks, sea quarks and gluons. The Bjorken scale variable,

$$x = \frac{p_P}{p_H}, \quad (1.33)$$

denotes the fraction of the full hadron momentum that a parton carries. x is not deterministic but everytime a parton performs an interaction its momentum can be different. One can think of the gluons constantly exchanging momentum between the quarks. This behavior can be described with *parton density functions* (PDFs). Let $f_d(x)$ be the PDF for down quarks within a proton. Then $f_d(x) dx$ equals the probability of finding a down quark with momentum between x and $x + dx$ inside the proton. The actual parton density functions depend on the energy of the hadron. They can be measured via inelastic proton scattering.

Formation of Jets. Due to confinement, quarks or gluons cannot exist alone. When a high-energetic quark or gluon is created in particle collisions, their energy is high enough to create not only one quark-antiquark pair, but many of them. This leads to the formation of a whole bunch of hadrons all of which move into approximately the same direction. Such a bunch is called a *jet*. Since some hadrons are unstable, also leptons and photons created when they decay can be part of a jet.

1.6 Experimental Verification

The Standard Model has been verified by hundreds of experiments. An example is the prediction of the top quark: After the bottom quark was discovered and it was clear that there exists a third generation of quarks and the search for the top quark started. Eventually, it was discovered by the CDF collaboration at Tevatron [14] in 1995. It was heavier than originally expected but in full agreement with the Standard Model. In 2000 the τ neutrino was experimentally observed [15].

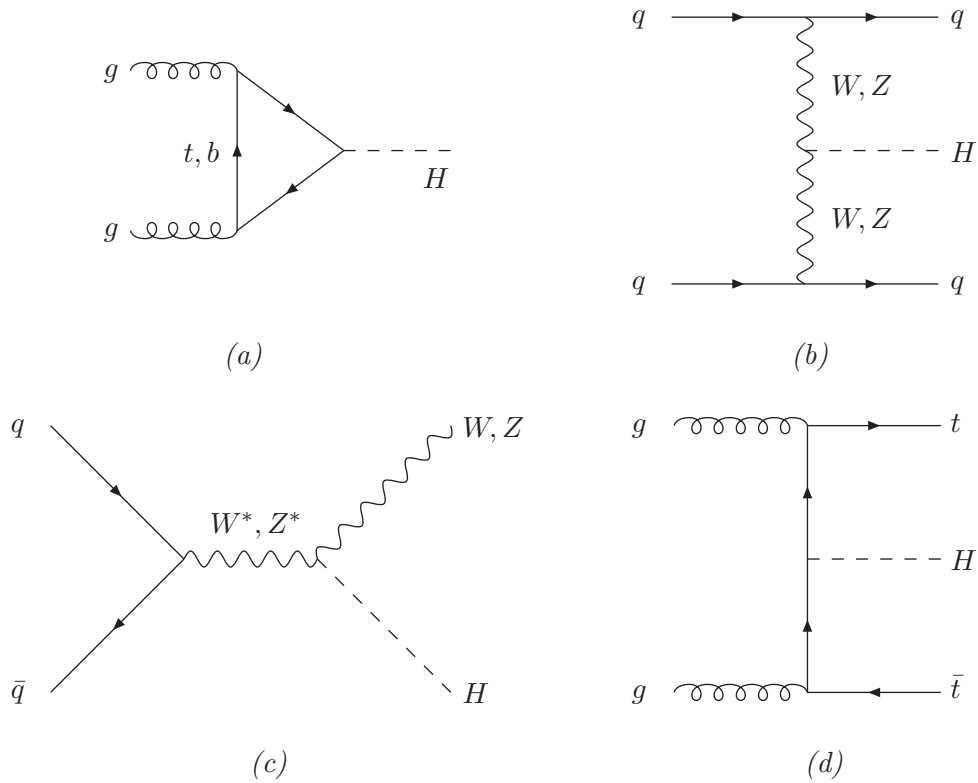


Figure 1.3: Leading order Feynman diagrams for Higgs production at a hadron collider. The processes are called (a) gluon fusion, (b) vector boson fusion, (c) Higgs strahlung and (d) quark associated production. From [13].

The Higgs boson mass is the last parameter of the Standard Model remaining to be measured, and an actual measurement would confirm the Higgs mechanism as described in Section 1.4 being realized in nature. Experimentalists at various particle accelerators have tried to perform such a measurement, or in other words, to discover the Higgs boson.

1.6.1 Higgs Boson Production

At colliders the Higgs boson can be produced via different mechanisms. The most important ones at hadron colliders are shown in Figure 1.3. At LEP, the Large Electron-Positron collider, and also at Tevatron which collides protons and anti-protons, the Higgs strahlung process is the most dominant production process (at LEP with electrons instead of quarks in the initial state). The LHC (see chapter 2 for details) collides two proton beams where the probability of finding an antiquark is much lower. Therefore, the gluon fusion process is the dominant production process at the LHC.

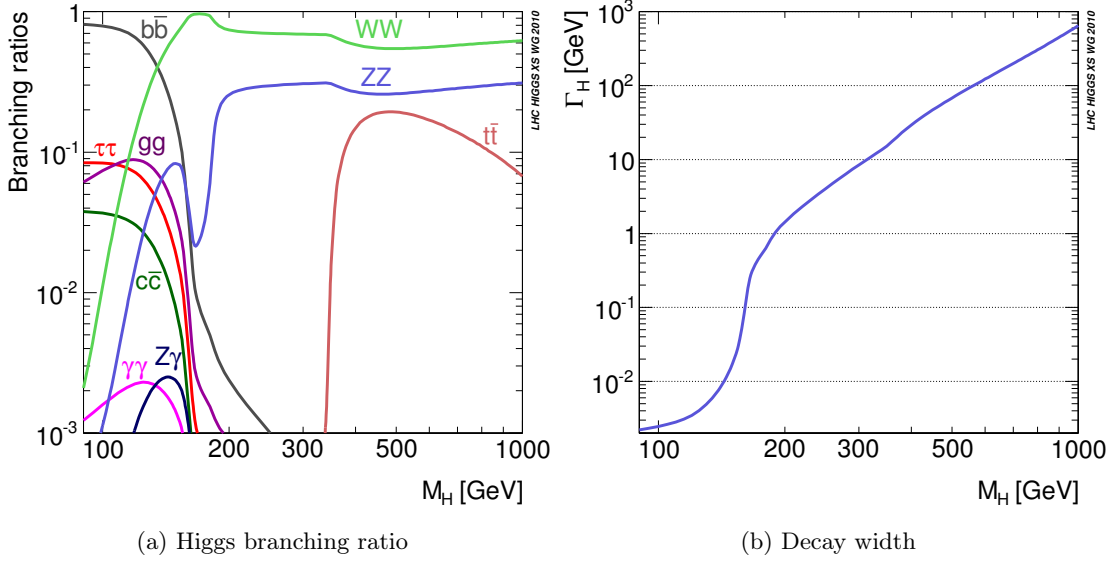


Figure 1.4: The branching ratio (a) and the decay width (b) of the Standard Model Higgs boson as a function of its mass, from [16].

The vector boson fusion process is also very interesting at the LHC. Its production cross section is about one order of magnitude lower than for gluon fusion, however it has a very clean event topology where there are two jets in opposite hemispheres of the detector and very low activity between these two jets (known as the rapidity gap). This allows for a very clean separation against background processes because only few other non-Higgs processes have this feature.

The associated quark production process only plays a minor role because due to the very heavy particles in the final state the available phase space limits the production cross section. Replacing the top quarks by bottom quarks leads to a higher phase space but lower coupling to the Higgs boson.

1.6.2 Higgs Boson Decay

The coupling of fermions or vector bosons to the Higgs boson is proportional to their mass. Therefore, the Higgs boson dominantly decays to particles with high masses as long as the decay is kinematically allowed. The branching ratio depends on the mass of the Higgs boson and can be seen in Figure 1.4a.

For masses below 135 GeV the most prominent decays are $H \rightarrow b\bar{b}$ and $H \rightarrow \tau^+\tau^-$. Also, the decay into two photons is possible but suppressed because it requires a top quark or W boson loop since the Higgs boson does not couple to the photon directly. This channel is promising for low Higgs boson masses because it can be clearly distinguished from other signatures and so suffers only from low background contributions.

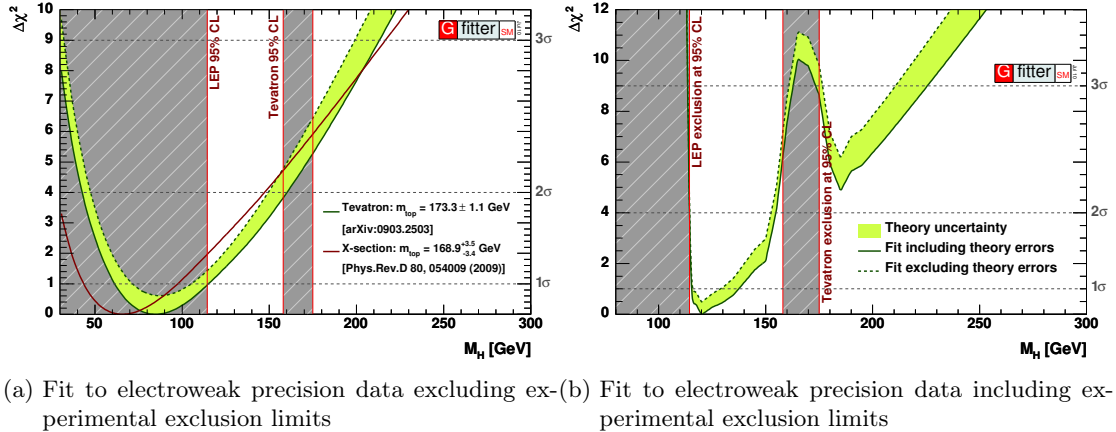


Figure 1.5: Results of a fit of the Higgs boson mass to electroweak precision data. From [20].

For high Higgs masses above 135 GeV the decay to vector boson pairs ($H \rightarrow W^+W^-$ and $H \rightarrow ZZ$) becomes available and quickly dominates all other channels.

Once a Higgs signal shows up in experimental data, its properties must be confirmed in order to verify that the signature found is indeed the Standard Model Higgs boson and not another new particle. Apart from spin and CP properties [17] this also includes couplings to fermions and gauge bosons. Therefore, all the decay channels in Figure 1.4a need to be studied carefully.

1.6.3 Higgs Exclusion Limits

Eventually, the results from all available channels in the relevant mass ranges are combined in order to increase overall statistical significance of a potential discovery. Also, results from different experiments at LEP [18] and Tevatron [19] were combined.

Neither one experiment alone nor the full combination did result in the discovery of the Higgs boson yet. However, since a Higgs boson signal would have been observed if the Higgs mass were within certain mass ranges, these masses can be excluded with a high level of confidence (C.L.), usually 95 %. The LEP experiments were able to exclude a Higgs boson mass below 114 GeV this way. The experiments at Tevatron excluded a Higgs boson mass between 158 GeV and 173 GeV with the highest contribution from the $H \rightarrow W^+W^-$ channel.

The Higgs boson mass depends weakly on the masses of the top quark and the W boson. Therefore, precise knowledge of these parameters from LEP and SLC [21] can be used to constrain the Higgs boson mass from a theoretical point of view. This is implemented by a fit of the Higgs boson mass to such electroweak precision data [20]. The best estimate obtained this way leads to a mass of $M_H = 84^{+30}_{-23}$ GeV as can be seen in Figure 1.5a. This is below the LEP exclusion limit but the allowed region is still in

1.7 Limitations of the Standard Model

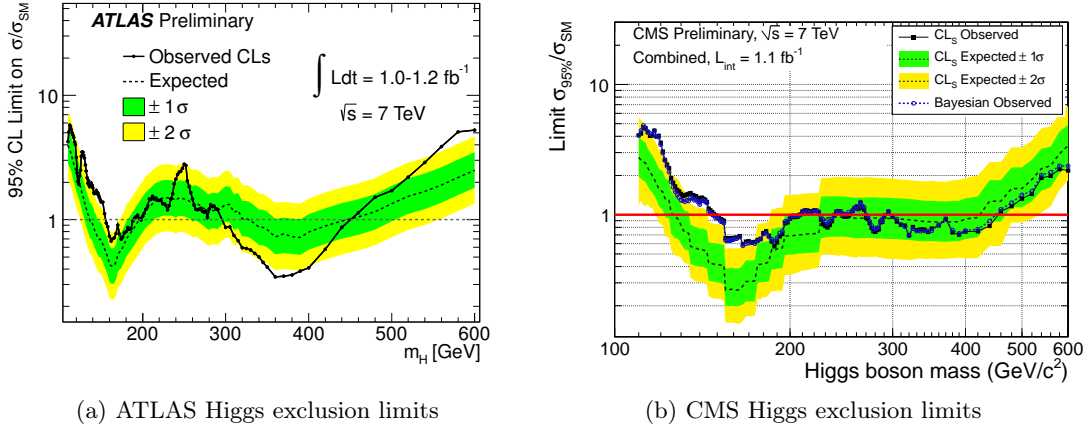


Figure 1.6: Higgs exclusion limits of the LHC experiments ATLAS and CMS. For a given Higgs mass, the production cross section of the Higgs boson is lower than the solid line with a probability of 95 %. The Y axis is normalized to the Standard Model cross section. The dashed line represents the expected exclusion limit given the hypothesis that the Higgs boson does not exist. From [22, 23].

the 2σ ($\Delta\chi^2 < 4$) range. Figure 1.5b shows the result of the fit taking into account the LEP and Tevatron limits which leads to an estimate of the Higgs boson mass of $m_H = 121_{-6}^{+17}$ GeV.

The experiments at the LHC, ATLAS and CMS, recently published exclusion limits with 1 fb^{-1} of data [22, 23] in a combination of all channels. The plots are reproduced in Figure 1.6. Both experiments can nearly exclude all masses between 160 GeV and 450 GeV.

1.6.4 The $H \rightarrow \tau\tau$ Channel

As shown in Figure 1.5b electroweak precision measurements suggest that the Standard Model Higgs boson has a mass below 150 GeV [20]. Even though the branching ratio for the $b\bar{b}$ channel is significantly higher than the $\tau^+\tau^-$ one it is more experimentally challenging to separate the $b\bar{b}$ signal from background processes. The reason is that this requires *b-tagging*, a technique for identification of jets originating from a b-quark and not from another quark or a gluon [24]. Such b-tagging is possible but leads to high systematic errors. Therefore, the $\tau^+\tau^-$ channel is the most promising one for discovery of the Higgs boson, along with $H \rightarrow \gamma\gamma$. This is the primary reason for studying $\tau^+\tau^-$ final states later in this thesis (Chapter 5).

1.7 Limitations of the Standard Model

It is known already that the Standard Model cannot be the most basic theory of particle physics. For once it does not include a theory of gravity, but there are also other problems

1 *The Standard Model of Particle Physics*

that cannot be solved within the Standard Model.

The first direct observation of physics beyond the Standard Model is the discovery of neutrino oscillation at Super-Kamiokande [25] and SNO [26]. In the Standard Model, neutrinos are massless, however neutrino oscillation can only be described if the difference of neutrino masses squared is nonzero.

Another evidence for physics beyond the Standard Model comes from cosmology: the velocity of distant galaxies rotating around the galaxy center is higher than what is predicted with Newton's laws based on the mass present within the galaxy. This suggests that there is additional mass which cannot be seen, that is it does neither interact strongly nor electromagnetically, because otherwise it could be observed with telescopes. This property gives it the name "dark matter". The Standard Model does not have a particle which can describe dark matter. The only stable particle which interacts only weakly is the neutrino which is too light to account for the observed rotation curves.

The theory of supersymmetry solves this problem by introducing new stable, massive particles. Supersymmetry postulates that, for every Standard Model particle, there exists a new particle (called its superpartner) with different mass and whose spin differs by $\frac{1}{2}$. Dark matter could consist of such a superpartner. Furthermore, there are at least 5 Higgs bosons in supersymmetric theories.

Other shortcomings of the Standard Model include its high number of free parameters and the inability to explain why exactly there are 3 generations of fermions with a strict mass hierarchy.

Apart from discovery or exclusion of the Higgs boson in the full mass range, observing any signal of "New Physics" is the primary goal of the Large Hadron collider which is discussed in the next chapter.

2 The CMS Experiment at the LHC

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is the world's most powerful particle accelerator to date [27]. It is a proton-proton ring collider installed at CERN¹ near Geneva at the border between Switzerland and France. The LHC is located in the same tunnel that previously hosted the Large Electron Positron Collider (LEP) [28] which was shut down in 2000 in order to allow the construction of the LHC to begin. Having successfully achieved collisions with a center-of-mass energy of 7 TeV the LHC surpassed the Tevatron [29] at Fermilab near Chicago as the world's most energetic particle collider. However, its design energy of 14 TeV is anticipated to be reached only after a technical stop in 2014.

In contrast to LEP, the LHC collides two proton beams instead of beams of electrons and positrons. In contrast to protons, electrons are very light particles so they suffer much more from synchrotron radiation which scales as $(E/m)^4$. The reason to use two proton beams instead of a proton and an anti-proton beam as does the Tevatron is that anti-proton production is the limiting factor in achieving high collision rates. The downside of hadron colliders is that since protons are composite particles what actually collides are the components of the protons (partons). However, the proton remnants or their interactions are also visible in the detector (“underlying event”) and need to be separated from the hard process. Another difference to lepton colliders is that the energy that goes into the collision is not fixed because the partons carry a variable fraction of the full proton momentum (3.5 TeV).

The LHC gives access to physics studies at energy scales up to about 1 TeV. One of the major goals of the LHC is to discover or to exclude the existence of the Standard Model Higgs boson. Previous accelerators, such as LEP and the Tevatron, were only able to set exclusion limits for limited mass ranges [18, 19].

As discussed in Section 1.7, there is strong evidence for physics beyond the Standard Model. It is expected that such new physics manifests itself at the TeV scale. It is therefore essential to verify the Standard Model in all aspects at high energies. Supersymmetric particles, if they exist, are expected to have masses between hundreds of GeV and several TeV, a region that can be probed by the LHC [30].

There are four major experiments installed around the LHC ring. The LHC brings the particle beams to collide at these so-called *interaction points*.

- **ALICE**² [31] is a detector designed for studying heavy ion collisions. For about one month every year the LHC is colliding lead ions instead of protons. In such

¹Conseil Européen pour la Recherche Nucléaire (engl.: European Organization for Nuclear Research)

²A Large Ion Collider Experiment

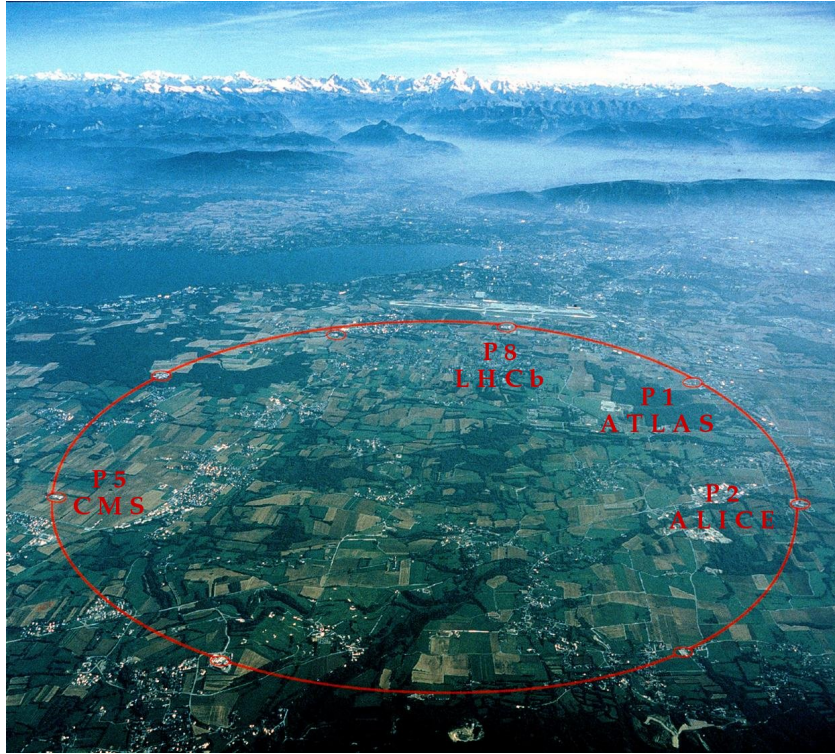


Figure 2.1: Aerial view of the area where the LHC is located. There are eight locations where the tunnel can be accessed, numbered P1 to P8. Experiments have been installed at four of them. In the background the airport of Geneva, Lake Geneva and the Alps can be seen.

collisions, due to very high temperature, a new state of matter where quarks and gluons can propagate freely is created for very short times, called a quark-gluon plasma. ALICE is studying such quark-gluon plasmas which resemble the state of the universe only microseconds after the big bang. ALICE is located at interaction point 2 (P2) on the LHC ring.

- **ATLAS** [32] is a general-purpose particle detector. It is designed to identify various particles created in the collisions to be able to find any new and yet unknown particles or physics processes, especially the Higgs boson. ATLAS is cylindrical in shape, 45 m long and has a diameter of 22 m, as high as a four floor tower building. It weighs 7,000 t and is installed at Point 1 (P1).
- **CMS**³ [33, 13] is again a general-purpose particle detector with complementing detector technologies compared to ATLAS. It is only roughly half the size of ATLAS, however it weighs 12,500 t. As its name implies, CMS especially focuses on excellent muon reconstruction. ATLAS and CMS are supposed to cross-check

³Compact Muon Solenoid

Parameter	LHC		Tevatron	Unit
	Design	Achieved		
Circumference	26.7	26.7	6.28	km
Beam Energy	7000	3500	980	GeV
Number of Particles per Bunch	1.15	1.1	p : 2.9	10^{11}
			\bar{p} : 1.0	10^{11}
Number of Bunches	2808	1390	103	
Bunch Spacing	25	50	132	ns
Crossing Angle at IP	285	100	136	μrad
Luminosity	100	18	4	$10^{32} \text{ cm}^{-2} \text{ s}^{-1}$
β^* at IP	0.55	1.5	0.28	m

Table 2.1: Important Collider parameters of the LHC and the Tevatron. For the LHC both the original design value and the best value achieved so far are reported. The design values are taken from [27] whereas the achieved values are not yet published and are also subject to change. They are taken from the current monitoring displays. The Tevatron parameters are taken from [35, 36]. Luminosity and β^* are explained in Section 2.1.2.

each other's results and eventually combine them for higher significance. CMS is situated at P5, opposite to ATLAS on the ring.

- **LHCb** [34] is designed to study B mesons with high precision. There is special interest in the decay of such mesons to learn about the asymmetry between matter and antimatter (known as CP violation). LHCb is located at P2.

There exist also three minor experiments for very specific purposes, namely TOTEM [37] (total cross section measurement), LHCf [38] (forward physics) and MoEDAL [39] (search for magnetic monopoles). These experiments share the cavern with one of the four major experiments listed above.

LHC operation started on September 10, 2008 when for the first time beams circulated in the machine in both clockwise and counterclockwise directions. However, on September 19, 2008 a quench of a superconducting magnet occurred, leading to about fifty magnets damaged and a leak of about 6 tonnes of liquid helium [40]. The necessary repairs and enhancements to the Quench Protection System (QPS) to avoid similar incidents in the future took more than a year [41]. On November 30, 2009 the LHC restarted operation and finally delivered first collisions at the (reduced) nominal center-of-mass energy of 7 TeV to the four experiments on March 30, 2010.

An aerial view of the area where the collider is located underground is shown in Figure 2.1. Table 2.1 shows important design parameters of the LHC and the values achieved so far. For comparison, the parameters are also given for the Tevatron.

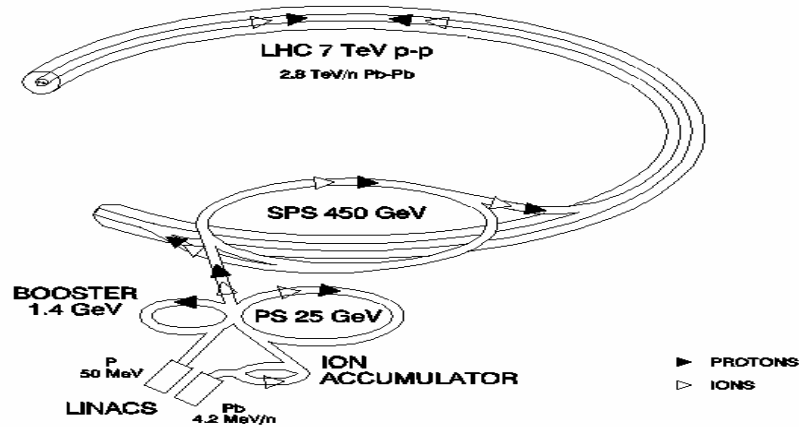


Figure 2.2: LHC injector complex. Protons are first accelerated by the LINAC 2 to an energy of 50 MeV. Then they are subsequently injected into the Proton Synchrotron Booster (PSB, 1.4 GeV), the Proton Synchrotron (PS, 26 GeV) and the Super Proton Synchrotron (SPS, 450 GeV). In heavy ion mode the particles start at another LINAC but then traverse the same injection chain. From [27].

2.1.1 LHC Beam and Injection

Before the beam is injected into the main LHC ring the particles are pre-accelerated by a sequence of smaller particle accelerators (injectors). First, protons are accelerated by a linear accelerator, LINAC 2, to an energy of 50 MeV. Then they run through a bunch of ring accelerators, the Proton Synchrotron Booster (PSB, 1.4 GeV), the Proton Synchrotron (PS, 26 GeV) and the Super Proton Synchrotron (SPS, 450 GeV). At the energy of 450 GeV the protons are injected into the LHC where they are accelerated to their final energy of (currently) 3.5 TeV and then brought to collision at the various interaction points. Figure 2.2 visualizes the LHC injection chain.

Once in the LHC, the beams are steered by 1232 superconducting dipole magnets around the ring. The magnetic field is at 0.5 T at injection energy and then ramped up synchronously with the beam energy to 8.3 T at 7 TeV. Quadrupole and sextupole magnets are used to focus the beam in the transverse directions. Acceleration of the particles is performed by 400 MHz radio frequency (RF) cavities which also correct longitudinal injection errors. This acceleration principle is the reason why the protons in the LHC are separated into bunches (of $1.1 \cdot 10^{11}$ protons each) instead of a continuous stream of particles. The bunches are injected in so-called bunch trains where the spacing between two bunches is currently 50 ns. Between bunch trains, there is an *abort gap*, a region free of particles of about 3 μ s. The abort gap is required to have enough time to increase the current in the kicker magnets in case the beams need to be removed (“dumped”) from the machine.

2.1.2 Luminosity

The rate for a certain process to occur at a particle collider can be separated into two terms, the *cross section* σ and the (instantaneous) *luminosity* L :

$$\frac{dN}{dt} = L \cdot \sigma \quad (2.1)$$

The cross section is a quantity which solely depends on the physics process in question, such as Higgs Boson production. It can be calculated by theoretical means using quantum field theory as sketched in Section 1.1. This takes also into account the incoming particles, their energies and parton density functions. The unit of the cross section is m^2 , however it is often given in barn with $1 \text{ b} = 10^{-28} \text{ m}^2$. The cross section can be thought of as a quantum mechanical analogy to the geometric cross section of two objects hitting each other.

The luminosity, on the other hand, is determined by the parameters of the collider. The higher the luminosity the more collisions between individual particles occur and therefore the more likely it is for a certain process to happen. It depends on quantities such as the number of protons in a bunch, the number of bunch crossings per second and the quality of beam focusing at the intersection point. For Gaussian beam profiles the luminosity is given by

$$L = \frac{N_b^2 n_b f_{\text{rev}} \gamma_r}{4\pi \epsilon_n \beta^*} F, \quad (2.2)$$

where N_b is the number of particles per bunch, n_b the number of bunches per beam, f_{rev} the revolution frequency and γ_r the relativistic gamma factor. The quantity ϵ_n is called the normalized transverse beam emittance and β^* is the beta function at the collision point. These two factors are a measure of beam focusing and overlap at the collision point. The geometric luminosity reduction factor F originates from the fact that the two bunches do not cross each other head-on but the crossing angle is finite (see Table 2.1). It is given by

$$F = \frac{1}{\sqrt{1 + \left(\frac{\theta_c \sigma_z}{2\sigma^*}\right)^2}}, \quad (2.3)$$

where θ_c is said crossing angle, σ_z the root-mean square of the bunch length and σ^* the width of the beam size distribution.

During operation in 2011, the LHC has reached the highest luminosity ever obtained at a hadron collider. However, it is still far below its design value and in future operation the luminosity will be increased further by enhancing the number of bunches in the machine, by focusing the beams better or even by increasing the particles per bunch above the design value.

Another important quantity related to the instantaneous luminosity is the integrated luminosity, defined by

$$\mathcal{L} = \int L dt, \quad (2.4)$$

where the integral goes over the whole running time of the experiment. The integrated luminosity is a measure of the amount of data an experiment has acquired. It is expressed in inverse barn. For example, considering a physics process with a cross section of σ and a collider having acquired an integrated luminosity of \mathcal{L} , one expects on average $N = \sigma \cdot \mathcal{L}$ collisions.

Therefore, for an experiment to predict how many times a process occurred, or, vice-versa, to measure the cross section of a certain process by counting how many times it occurred, it is essential that the integrated luminosity is known. The luminosity can be measured to the level of a few percent by a special LHC operation mode, called a *van-der-Meer scan* [42]. In this mode interaction rates are measured while the two beams are swiped through each other.

During a fill, the instantaneous luminosity is constantly decreasing because the collisions cause the beam intensities to decrease and the beams to defocus. To obtain a good online luminosity measurement, relative measurements can easily be performed during normal operation. In CMS, the forward hadronic calorimeter is exploited [43] because, in the forward region, the number of calorimeter cells containing hits is correlated to the number of collisions in a bunch crossing and therefore the luminosity. About two or three times a year a van-der-Meer scan is performed to calibrate this relative measurement. This takes about two hours for each of the four experiments.

2.2 The CMS Experiment

The CMS experiment is one of the largest particle detectors ever built. It measures 22 m in length and 16 m in diameter. The detector consists of various subdetectors which are arranged in layers around the interaction point. CMS consists of a central barrel which is arranged symmetrically around the interaction point and so-called endcaps on both sides of the barrel. In the endcaps detector components are aligned perpendicular to the beampipe in order to improve spatial resolution in the forward regions.

A schematic view of the detector and its components is depicted in Figure 2.3. The different detector components of CMS are, from innermost to outermost, the silicon tracker, the electromagnetic calorimeter, the hadronic calorimeter and the muon system. These systems are explained in more detail in the following subsections.

Between the hadronic calorimeter and the muon system, there is a superconducting solenoid which creates a homogeneous field of up to 3.8 T inside and 1.5 T in the outer region. Its purpose is to bend the trajectories of charged particles in order to deduce their transverse momentum from the curvature of reconstructed tracks in the silicon tracker and the muon chambers.

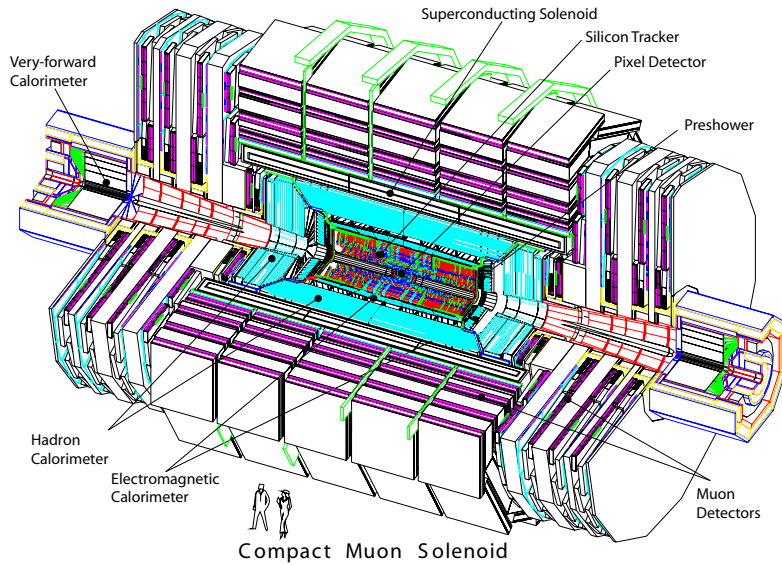


Figure 2.3: Schematic overview of the CMS detector and its various components. The silicon pixel tracking detector is located in the innermost region right next to the interaction point. Going radially outside, the silicon strips tracking detector, the electromagnetic and hadronic calorimeters, the superconducting solenoid and the muon chambers with the iron return yoke inbetween follow. From [33].

2.2.1 Coordinate System

CMS uses a cylindrical coordinate system with the nominal interaction point as origin. The Z axis points in direction of the beampipe toward the Jura mountains. The X and Y axes span the plane perpendicular to the beampipe so that the Y axis points upwards to the surface and the three axes form a right-handed coordinate system.

The azimuthal angle ϕ is defined in the X - Y plane, counting from positive X direction to positive Y direction. The polar angle θ is defined between the Z axis and the X - Y plane. It equals 0° for directions parallel to the beampipe and it has a value of 90° for directions perpendicular to it. In radial direction, the transverse momentum of a physics object is given as $p_T = \sqrt{p_x^2 + p_y^2}$.

Instead of the polar angle θ , often the *pseudorapidity* η is used. It is defined as

$$\eta = -\log \left(\tan \left(\frac{\theta}{2} \right) \right). \quad (2.5)$$

The pseudorapidity is 0 for directions perpendicular to the beampipe and goes to infinity as the polar angle approaches 0° . The reason to use the pseudorapidity instead of the polar angle is that the differential cross section $\frac{d\sigma}{d\eta}$ and the difference of two pseudorapidities $\eta_1 - \eta_2$ is approximately invariant under Lorentz boosts along the Z direction.

2 The CMS Experiment at the LHC

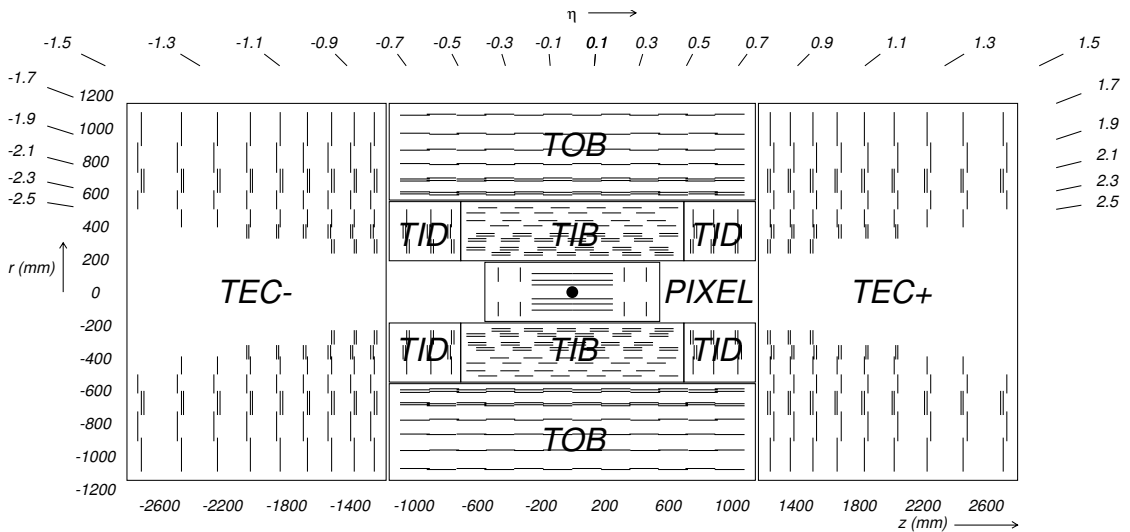


Figure 2.4: Schematic view of the inner tracking system. In the innermost region, there are three pixel layers in the barrel region and two layers in the disk region on each side. The silicon strips tracker is composed of multiple components: the Tracker Inner Barrel (TIB) region with 4 layers, the Tracker Inner Disk (TID) region with 3 layers, the Tracker Outer Barrel (TOB) region with 6 layers and the Tracker Endcap (TEC) regions with 9 layers. From [44].

This is perfectly true for massless particles and in the high energy limit $E \rightarrow \infty$. It is also possible to define the *rapidity* y for which the two relations are always fulfilled. However, this quantity depends on the energy of the particle in question and is therefore inconvenient to deal with.

The geometrical distance between two objects is expressed as ΔR , given by

$$\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}. \quad (2.6)$$

2.2.2 Silicon Tracking Detector

The purpose of the inner tracking detector is to measure the trajectories of charged particles. The region nearest to the interaction point is covered by the pixel tracker. It features three layers in the barrel region and two layers in the disk regions where the layers consist of individual modules sized $100 \mu\text{m} \times 150 \mu\text{m}$ [44]. The pixel sensors provide full three-dimensional information for each hit in the detector. In total, there are 66 million pixel channels.

The region starting at a distance of 20 cm from the beampipe is covered by 9 or 10 layers of silicon strip sensors, depending on the η region. A single strip can cover a long range in one direction and, therefore, fewer individual channels are required for full coverage. However, it only provides two-dimensional hit information. In order to compensate for

this, the strips in different layers are tilted toward each other by $110\ \mu\text{rad}$ even though ambiguities remain for high fluxes.

Figure 2.4 shows a schematic view of the layout of the tracking detector. The detector covers the complete region up to $|\eta| = 2.5$. With about $200\ \text{m}^2$ of total area the silicon tracker of CMS is the largest silicon detector ever built.

A silicon detector basically consists of a p-n-junction with high voltage applied so that the region between the p-doped and the n-doped side is fully depleted of free charge carriers through recombination of electrons and holes. If a charged particle crosses the depleted region it ionizes atoms so that electrons and holes are created, producing a measurable current due to the voltage applied.

When exposed to radiation, the quality of the sensors decreases due to radiation-induced imperfections in the semiconductor material. This requires applying higher voltages for full depletion of the sensor. When the depletion voltage exceeds the maximum voltage that can be applied, the signal to noise ratio increases until the signal cannot reliably be read out anymore. Since the inner tracker, and especially the pixel detector, is nearest to the interaction point, it will be exposed to a very high flux of particles. This employs special requirements for radiation hardness on the modules that were taken into account in the design phase of the experiment [45]. Still the lifetime of the tracker is lower than the expected operation time of the LHC so that it needs to be replaced around 2020 after about 10 years of running [46].

2.2.3 Electromagnetic Calorimeter

The purpose of the homogeneous electromagnetic calorimeter is to stop and measure the energy of electromagnetically interacting particles, i.e. photons and charged particles. The calorimeter consists of about 70,000 crystals of lead-tungstate, PbWO_4 . The choice of the material was driven by the requirement of radiation hardness (especially in the forward regions) and short scintillation decay times (80 % of the light is emitted within the LHC bunch spacing of 25 ns). Its high density of $8.28\ \text{g}/\text{m}^3$ and its short radiation length of 0.89 cm allow the construction of a very compact calorimeter that can be placed inside the solenoid [44].

Within the crystals, charged particles emit photons via bremsstrahlung. Photons in turn create electron-positron pairs which then create bremsstrahlung photons again. This process repeats until the energy of the particles is below the pair production threshold. Finally, photons excite the scintillating material which then re-emits the absorbed energy in the form of light. As the crystals are transparent for light it can fully traverse the crystal and is detected by avalanche photodiodes (APDs) in the barrel and vacuum photodiodes (VPTs) in the endcaps. Since the abortion of the initial particle cascade depends on the energy of the incoming charged particle or photon, the number of scintillator photons registered in the APDs or VPTs is a direct measure for the energy of the incoming particle. As this is a counting process the intrinsic energy resolution of the calorimeter scales as \sqrt{E} .

A schematic view of the electromagnetic calorimeter can be seen in Figure 2.5. The pre-shower detector in front of the endcaps is a lead-based sampling calorimeter. Its

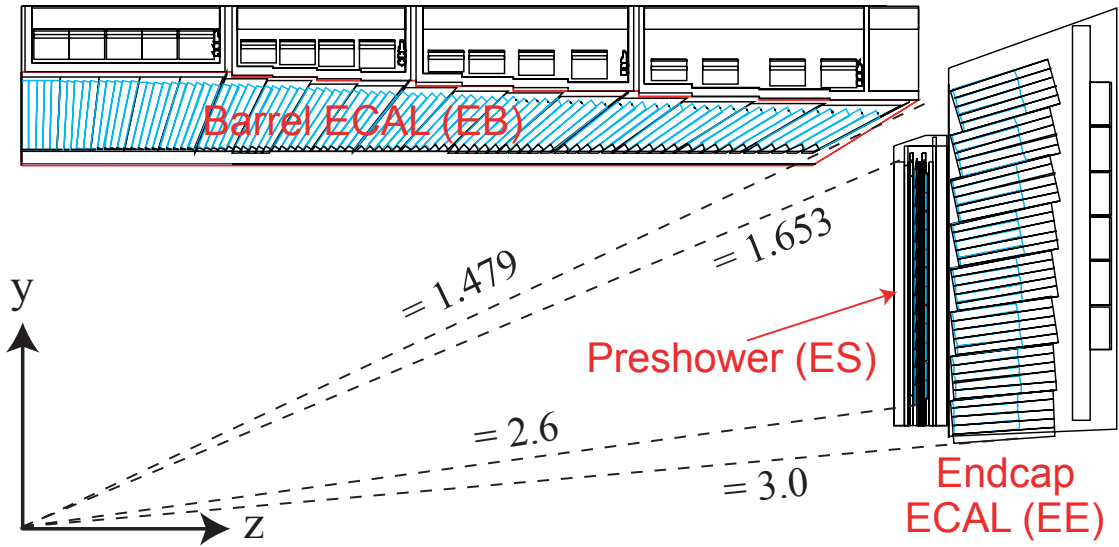


Figure 2.5: Schematic view of the electromagnetic calorimeter of CMS. It covers a pseudorapidity range up to 3.0. In the endcap, the indicated η regions (dashed lines) are not covered by the pre-shower detectors which in front of the endcaps aid in identifying neutral pions. From [33].

purpose is to identify π^0 candidates and to improve the spatial resolution of electrons and photons in the forward region.

2.2.4 Hadronic Calorimeter

The hadronic calorimeter of CMS is a sampling calorimeter which consists of steel and brass absorbers and plastic scintillator material. Again, the materials were chosen with respect to high density and good radiation hardness. In the absorber, strongly interacting particles (hadrons) interact with the matter, producing cascades of low energetic particles. Most energy is deposited in the absorber material, however some particles reach the scintillator material where photons are emitted and registered by photomultiplier tubes. Since only a fraction of the energy is actually detected, the calorimeter needs to be accurately calibrated to account for the energy loss in the absorber material. This explains why the relative energy resolution of the hadronic calorimeter is worse than the one of the electromagnetic calorimeter.

The total absorber width is about 5 interaction lengths in the barrel region and increases with higher pseudorapidity. In the barrel region, this does not allow hadronic showers to be fully stopped. Therefore, an additional calorimeter component is installed outside of the superconducting solenoid which uses the magnet coil material as an additional absorber. Figure 2.6 shows the layout of the various parts of the hadronic calorimeter.

The forward calorimeter is installed at a distance of about 10 m from the interaction

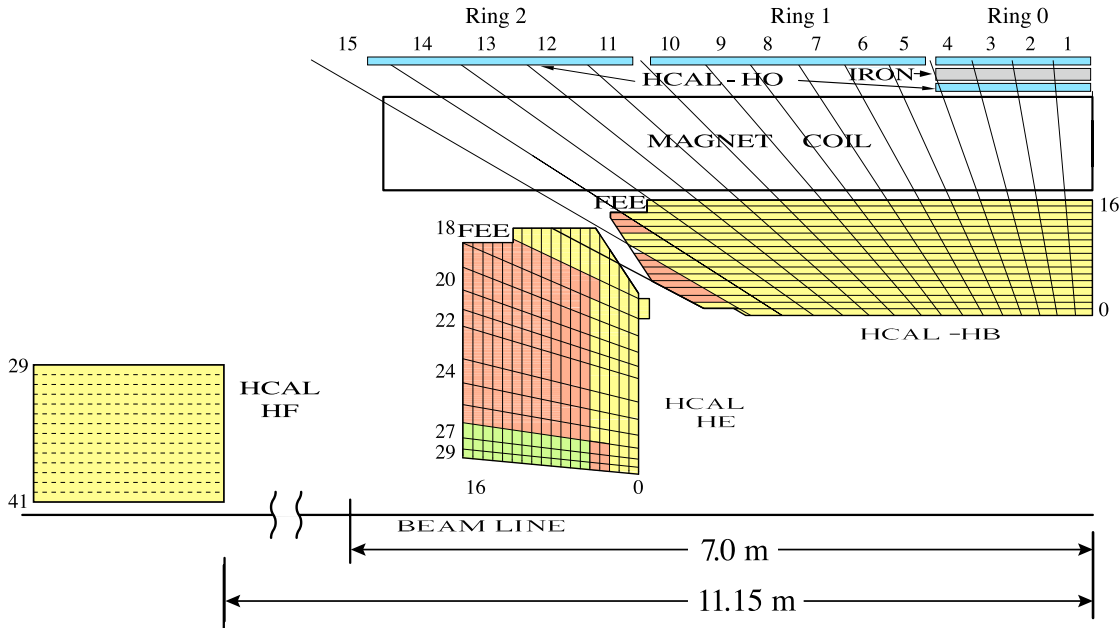


Figure 2.6: Schematic view of the hadronic calorimeter of CMS. The system consists of four parts: The barrel region (HB), the endcap region (HE), the forward region (HF) and a region outside of the magnet coil (HO). From [47].

point. It extends the pseudorapidity coverage from $|\eta| < 3.0$ to $|\eta| < 5.0$ and is also used for relative luminosity measurement (see Section 2.1.2).

2.2.5 Muon System

Behind the hadronic calorimeter only minimum ionizing particles have not yet been stopped. Apart from neutrinos, which cannot be detected by CMS, the only stable particles that can reach this point are muons. Therefore, outside of the hadronic calorimeter and also outside of the magnet coil, additional detector components have been installed. They are commonly referred to as the “muon system” and they are dedicated to identifying muons and increasing precision of the measurement of muon tracks. The muon system covers a pseudorapidity range of $|\eta| < 2.4$.

Figure 2.7 shows the layout of the muon system. Its components are installed in layers between the iron return yoke. In the barrel region drift tubes are used as detector technology and in the endcap cathode strip chambers are used. In both cases the detection principle is based on a gas chamber containing a conducting wire with a high voltage applied. When a muon crosses the chamber it ionizes the gas, causing charged particles to be accelerated toward the wire. On their way they collide with other gas particles, ionizing them as well. This produces a cascade of charged particles reaching the wire which results in a measurable signal.

The third component of the muon system are resistive plate chambers which are in-

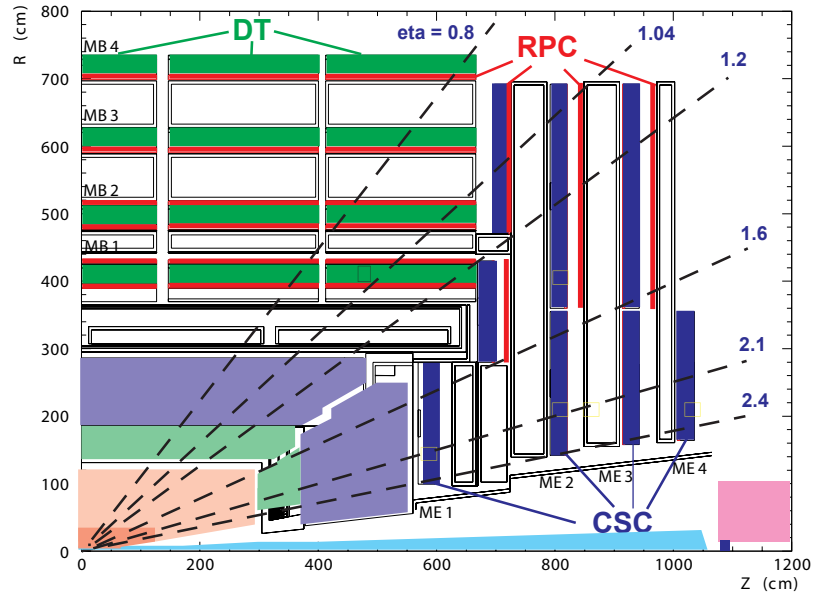


Figure 2.7: View of the CMS detector with a highlight on the muon system. It consists of three different components: The drift tubes (DT) in the barrel region, the cathode strip chambers (CSC) in the endcap region and the resistive plate chambers (RPC) in both regions. There are four layers of muon chambers. The iron return yoke is shown in white between the muon system components. From [33].

stalled both in the barrel and in the endcap regions. They have a very good time resolution in the order of 3 ns. Therefore, they are mainly used as input for the Level-1 trigger (see Section 2.2.7) and to assign a measured muon to the correct bunch crossing.

The identification of muons is crucial for the experiment because many important physics objects such as Z and W bosons, Higgs bosons or τ leptons can have muons in their final state. The identification efficiency of the muon system of CMS is better than 98 % and the uncertainty of the muon momentum is in the order of 1 % below 100 GeV [33].

2.2.6 Particle Identification

CMS is capable of distinguishing between photons, electrons, muons, neutral hadrons (e.g. neutrons) and charged hadrons (e.g. pions, protons or kaons) as all of them lead to different signatures in the detector:

- **Photons**, since they are not charged, do not produce any hits in the silicon tracker. They deposit all their energy in the electromagnetic calorimeter.
- **Electrons** produce hits in the silicon tracker and are stopped in the electromagnetic calorimeter.

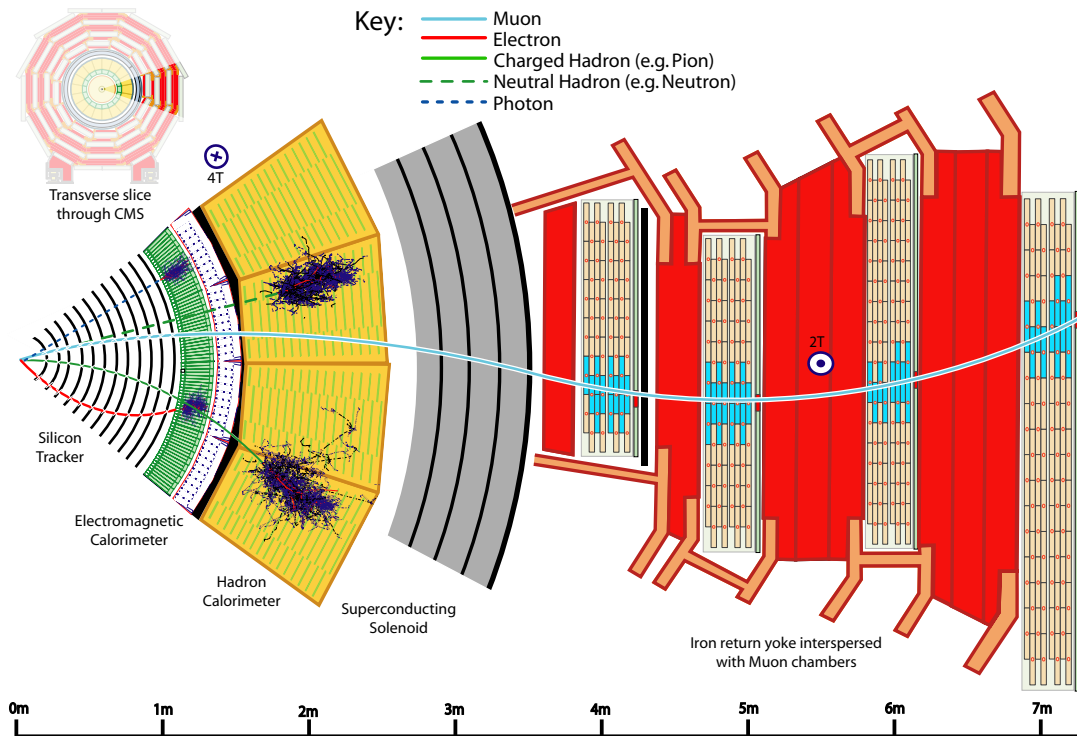


Figure 2.8: Slice through the CMS detector. It is shown how different particles lead to different signatures in the various subdetectors of CMS. The dashed lines do not produce hits in the silicon tracker. From [48].

- **Muons** also produce hits in the silicon tracker, however they only deposit very little energy in the calorimeters. Muons are the only particles causing a signal in the muon chambers.
- **Charged Hadrons** lead to hits in the tracking detector and little energy deposits in the electromagnetic calorimeter. They are stopped in the hadronic calorimeter.
- **Neutral Hadrons** will neither produce tracker hits nor electromagnetic calorimeter hits. They are fully stopped in the hadronic calorimeter.

Figure 2.8 visualizes the five different cases.

2.2.7 Data Acquisition

When running under design conditions the LHC will produce about 40 million bunch crossings every second in CMS. The data recorded by the detector after a bunch crossing is called an *event*. The average size of an event is about 1.0 MB. If all events recorded in CMS were stored this would correspond to a rate of several TB/s.

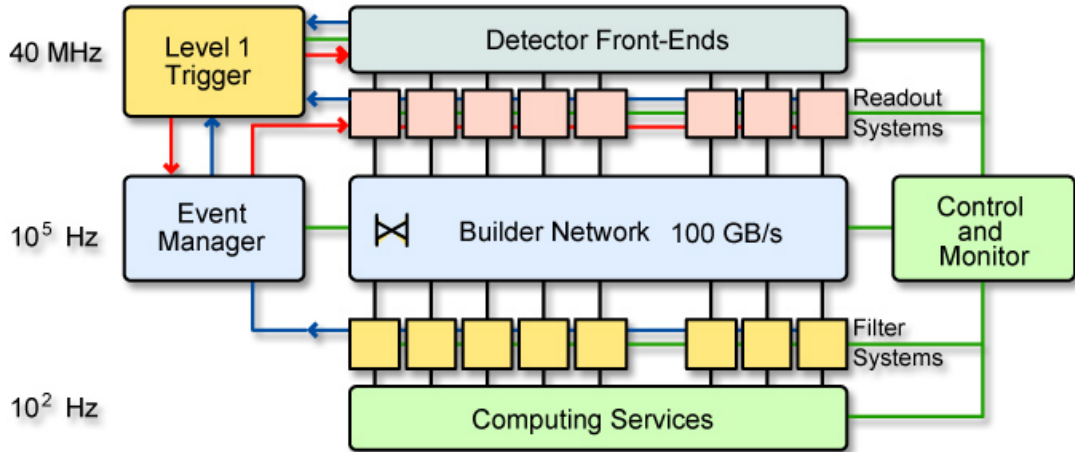


Figure 2.9: There are 40 million bunch crossings every second. This rate is first reduced by the Level 1 Trigger to around 100,000 events per second. In a second step a computer farm (“Filter Systems”) reduce the rate to about 100 events that are archived in and made available for data analysis via Computing Services. From [33].

Since this rate could never be archived, a sophisticated trigger system is required to reduce the event rate to a manageable amount of about 200 events per second. In many collisions only low energetic interactions occur that do not reveal any insights into rare physics processes whose study the collider was built for. Such events do not need to be stored and can be safely filtered out. The purpose of the trigger system is to quickly decide whether an event is worth storing or not. Its decision is based on energy, multiplicity and combination of measured particles and missing transverse momentum. There exist various different triggers for different purposes. For example, a single muon trigger requires at least one muon whose transverse momentum exceeds a certain threshold to exist in the event. Other triggers are activated by energetic jets, electrons or a combination of other criteria (so-called cross-triggers). It is sufficient for a single trigger to fire for the event to be archived. The parameters of the triggers are chosen so that the total output rate does not exceed what can be processed and permanently stored. With changing machine conditions and increasing luminosity the trigger parameters are updated accordingly.

The triggering process is twofold. The first step is called the *Level 1 Trigger* (L1) which reduces the rate to 100 kHz. The Level 1 Trigger is implemented in Hardware for efficiency reasons. The event data is stored in hardware buffer while the trigger makes its decision. The buffer size limits the runtime of the trigger algorithm. At maximum it has 3 μ s to make a decision, including latency for data retrieval [49]. The Level 1 Trigger does not have access to the full detector data but only the calorimeters and the muon chambers.

The second step is the *High Level Trigger* (HLT). It is implemented in software and runs

on a computer farm based on commercially available CPUs. This allows for flexibility in upgrading the trigger algorithms. Also, one can profit from further advancements in computer technology, allowing the final event rate to be increased in future. The event rate is reduced from 100 kHz to about 200 Hz which already exceeds the design rate of 100 Hz. A 100 GB/s link distributes the event data to the computers running the trigger algorithm. Again due to buffering constraints the average execution time of the HLT is 40 ms per event even though for single events the decision might take up to a maximum of 1 s.

Figure 2.9 visualizes the full trigger procedure.

Event Labeling. Every event recorded by CMS is uniquely identified by a three-tuple of run number, luminosity section and event number. The run number is a sequential number which increases every time a so-called *run* of the CMS detector is started. This is performed manually every time when the detector is supposed to start taking data. Configuration changes, for example adding or updating trigger algorithms, requires to start a new run. For each run the luminosity section number starts at 1 and is incremented about every 23 s. It is assumed that the instantaneous luminosity is constant during a luminosity section. Finally the event number uniquely identifies the event within a run and luminosity section.

2 *The CMS Experiment at the LHC*

3 High Energy Physics Software and Frameworks

In High Energy Physics there are many daunting and repetitive tasks that need to be carried out by computers. This includes analysis of huge amounts of data, simulation of physics processes, computation of Feynman diagrams, detector simulation or reconstruction of particles from the detector response. The High Energy Physics community has developed a rich set of software tools and frameworks for their specific needs. These efforts are necessary because many problems that need to be solved are not present in other disciplines and therefore commercial solutions are not available.

In this chapter a brief overview of the common software packages that have been used to obtain the results presented in this thesis is given.

3.1 Monte Carlo Event Generation

In order to interpret the measured High Energy Physics data it needs to be compared to the expectation from the Standard Model. This is achieved by simulating the physics processes in question with Monte Carlo generator programs. As the name implies they use Monte Carlo techniques to produce single events for a certain process. Usually many events are generated this way so that the distributions of quantities such as the transverse momentum, the pseudorapidity or the azimuthal angle can be compared to the ones obtained from real detector output.

However, the bare physics process (“Matrix-element level”) is not what is measured by the detector. In a proton-proton collision what enters into the hard physics process are not the protons as a whole but one of the quarks and gluons contained in the proton. The proton remnants are also part of the measured event. Their interactions result in additional low-energetic activity in the event, called the *underlying event*. Both Quarks and gluons both from the hard process and from the underlying event can radiate gluons. Such radiations occur frequently and tend to be soft. This is called the *parton shower*. Single quarks or gluons in the final state cannot exist alone but form jets as discussed in Section 1.5. This is known as the *hadronization* process.

These three processes, underlying event, parton shower and hadronization, involve mostly processes with low momentum transfer. Due to the divergence of the strong coupling constant α_S in QCD, perturbation theory cannot be applied in this regime. Therefore, heuristic models are needed to describe these phenomena. The parameters of such a model are tuned so that the model correctly describes both data from previous experiments and new LHC data.

The final step before simulated data can be compared to the experiment consists of detector simulation. This accounts for effects such as electrical noise, finite detector resolution or detector inefficiencies. Section ?? describes this in more detail for the CMS experiment.

3.1.1 Pythia

PYTHIA is a general purpose Monte Carlo event generator [50]. Its current version, PYTHIA 6.4, is written in FORTRAN77. It exists already since a 1978 and therefore it is well tested and widely accepted in the community. There is also a C++-based version, Pythia 8 [51], but it is not yet used extensively.

PYTHIA takes the beam particles and energies as input parameters as well as the process to model, such as $Z \rightarrow \mu^+\mu^-$. PYTHIA only takes the leading order (LO) in perturbation theory into account. For the hadronization process, PYTHIA uses the Lund string model [52]. Its output are event descriptions in HEPMC format, a standardized format for particle interaction events.

There exist many tunes for PYTHIA each of which describes different distributions in different energy regions better than others. The two most prominent ones are D6T and z2. D6T was developed with data from the CDF experiment at the Tevatron and then was extrapolated to LHC energies. z2 was commissioned directly with early CMS data and so far was found to describe most distributions better than D6T [53]. Therefore, all Monte Carlo samples used in this thesis use the z2 tune.

3.1.2 Powheg

POWHEG¹ [54] is a tool for next-to-leading order matrix element calculations. What makes it different from other tools such as MC@NLO [55] is that, as its name implies, it makes sure that no negative event weights occur. In order to avoid weighted events for the subsequent analysis an “unweighting” procedure can be applied. An example for such a procedure is generating more events than necessary and then discarding an event with probability $1 - w$ when w is the event’s weight.

POWHEG only performs the matrix element step. It supports a limited number of physics processes, including several $2 \rightarrow 2$ processes such as W and Z boson production [56, 57], top quark production [58, 59] and Higgs production [60, 61]. The subsequent parton showering and hadronization processes are not covered by POWHEG, but they can be performed by PYTHIA.

3.1.3 Tauola

TAUOLA [62] is a package for simulating τ lepton decays. Its main feature is proper simulation of spin correlation in τ decays, depending on the spin of the mother particle and its production process. This results in a non-isotropic angular distribution of the

¹POsitive Weight Hardest Emission Generator

decay products. Such polarization effects are not described by PYTHIA itself, however, external libraries such as TAUOLA can be plugged into PYTHIA for this purpose.

3.2 ROOT

ROOT [63, 64] is a data analysis framework developed at CERN. It is written in C++ and pretty much replaced the FORTRAN-based PAW [65] for this task by now.

Its primary purpose is statistical analysis of huge amounts of data. For this use-case it provides an object-oriented approach where various statistical objects such as graphs, histograms or n-tuples² are represented as classes. These classes provide methods for common operations such as computing the mean, root-mean-square or the integral. Also, all classes allow instances to be serialized into a compressed binary format in `.root` files.

ROOT is especially efficient in handling n-tuples, or a more general data structure which ROOT calls *trees*. Individual elements in a tree (“branch”) can again contain many other elements. The type of a branch can be freely chosen: apart from integral and floating point numbers also strings, structs, arrays or STL containers can be used. In fact all data from the CMS detector is stored as trees in `.root` files.

ROOT can also be used interactively. For this purpose it includes a C/C++ interpreter, called CINT [66]. It can be used to try out complicated commands such as a fitting procedure on the command line. Additionally ROOT includes a graphical browser, that can be used to browse `.root` files and visualize graphs, histograms or trees. For the visualization many options are provided to be able to tweak the resulting plots to ones liking. The plots can not only be generated via the browser but also via the ROOT API to be displayed on the screen or saved into a file.

3.3 CMSSW

CMSSW is the software framework of the CMS collaboration. It processes collision events for all purposes within the CMS experiment: it is used in the High Level Trigger, in event reconstruction, in Monte Carlo event generation and in data analysis. The main CMSSW code is written in C++.

CMSSW makes heavy use of a modular architecture. CMSSW modules are arranged in “paths” where all modules in a path are executed sequentially. The output of a given module is available as input for all following modules. The modules can be configured with configuration files written in PYTHON. This allows for a flexible and powerful configuration so that in many cases when parameters need to be changed recompilation of the C++ module code can be avoided. Also, since PYTHON is a well known programming language, many collaborators do not need to learn a special configuration language. The entry point for a CMSSW program is a configuration file which can be called with the

²An n-tuple is a list of n numbers which belong to the same object, such as the four-vector, reconstruction quality and vertex coordinates of a particle

`cmsRun` command. This entry point configuration specifies what other modules to load and which parameters to use for them.

There exist four different types of modules:

- **Source.** This kind of module serves as a data source for a CMSSW workflow. The first module in such a workflow is always a source module. Possible sources include `.root` files, Monte Carlo event generators or the data acquisition system which delivers collision data from the CMS detector.
- **Producer.** A producer module generates new data from existing event content. Tasks such as running the detector simulation on generated Monte Carlo events, reconstruction of physics objects such as muons or jets from the raw detector output or computation of new variables.
- **Filter.** Filter modules can be used to prevent certain events from being processed further. When the filter function returns `false` for a given event then it is discarded and not processed by the following modules. This is typically used in event selection tasks where only events originating from a certain process shall be filtered from all data events.
- **Analyzer.** An analyzer module is used to create final output such as histograms or n-tuples used for further analysis and plot generation outside the CMSSW framework.

Many external programs such as ROOT or PYTHIA are available within the CMSSW framework. This allows them to be configured using CMSSW configuration files instead of having to resort to their respective format. Also, the output of such external programs can directly be used by other CMSSW modules.

3.3.1 Event Data Model

All recorded data in CMS is stored in events. An event contains the detector response from a single triggered bunch crossing. This event-based organization of the detector data is known as the *Event Data Model* (EDM). The output of CMSSW modules such as reconstruction algorithms can simply be added to the event content where special metadata information keeps a history of the processing of the event. This allows central distribution of datasets on which the common algorithms have already been run.

Different events are regarded as independent from each other. This implies that the processing of many events can be trivially parallelized since different events can simply be processed by different computers.

3.3.2 Dataset Bookkeeping

As explained before already all data used by the CMS experiment, both detector output and Monte Carlo simulation, are stored in `.root` files. In order to quickly retrieve the

data that is interesting for a particular analysis the *Dataset Bookkeeping System*, or DBS, was deployed [67]. DBS organizes many `.root` files into so-called datasets. A dataset is characterized by its name, a file list and various metadata such as the CMSSW configuration file used to generate the data. All `.root` files in a dataset have the same origin, for example they all contain data from the same CMS run period or they all contain Monte Carlo data where a certain physics process was simulated.

Monte Carlo events for many common processes are produced centrally by CMS and made available in DBS. When an analysis requires Monte Carlo data for exotic processes or simply more events than officially produced then the group can perform a private production and also inject the resulting dataset to DBS so that the data can be accessed by others.

3.3.3 Conditions Database

For consistent data analysis the conditions of the CMS detector during data taking is archived. This includes parameters with which the individual subdetectors were operated, such as calibration factors for the calorimeters or alignment constants. These settings are saved in a central database at CERN, the *Conditions Database*. The reason these conditions are not saved directly with the event content is that it can happen that it needs to be corrected in hindsight.

When a CMSSW module needs information from the Conditions Database it makes a connection to the database. A so-called *Global Tag* is used to identify the record in the database to query the conditions. The same procedure is also performed for Monte Carlo datasets (with a different Global Tag) in which case the conditions of the detector simulation are stored.

Not all data taken by CMS can be used for analyses. When during data taking a subdetector component has a problem and does not deliver proper data then the collision events taken during that time need to be discarded. For this purpose all subdetector experts certify valid luminosity sections in all recorded data by making sure the whole detector was fully operable when the data was taken. They regularly publish a whitelist of valid runs and luminosity sections in JSON format [68] that can be used by analyzers to filter bad collision events.

3.3.4 Detector Simulation

??

In order to obtain the detector response for a simulated physics process a full simulation of the CMS detector needs to be performed. The GEANT4 [69] framework is used for this purpose. GEANT4 is written in C++ and simulates the traversal of particles through matter. An accurate geometric description of the CMS detector is available for GEANT4.

The interactions of particles with detector material is converted into hits of the various detector components which are then read by the data acquisition system, a process called digitization. In this step, electronic noise which occurs in the real detector is included in the simulation. Afterwards, the format of a simulated event is compatible to one

measured with the CMS detector so that all upcoming steps such as High Level Trigger decision and reconstruction of physics objects can be performed with the same software.

The simulation of a single event with the full GEANT4 detector description takes a relatively long time in the order of tens of seconds. Therefore, another detector simulation which is less precise is also available, called *FastSim*. All official CMS Monte Carlo event productions, and therefore all Monte Carlo samples used in this thesis, use the full detector simulation, however.

3.3.5 Event Reconstruction

Based on the individual hits in the detector components various higher-level physics objects are reconstructed through the following procedures.

- **Track Fitting.** Hits in the silicon pixel and strip detectors are combined to tracks of charged particles. The fit starts with three subsequent hits in the innermost layers of the detector. Then, the track is extrapolated outwards, searching for compatible hits in the next layers. The search stops when the end of the tracker is reached or no more hits are found. Eventually, a curved track is fitted to the whole collection of hits in order to obtain the track parameters from which particle properties such as its four-momentum can be deduced.

The algorithm described above is called the *Combinatorial Track Finder* (CTF) and is used for track reconstruction in CMS by default [70]. There exist other algorithms such as the *Gaussian Sum Filter* (GSF) [71], the *Deterministic Annealing Filter* (DAF) or the *Multi Track Finder* (MTF) which can be used for special purposes such as electron reconstruction or track reconstruction in jets [72].

- **Vertex Reconstruction.** A *primary interaction vertex* is the point in space where the collision of two protons occurred. The points of subsequent decays of collision products are called *secondary vertices*. Primary vertices can be reconstructed by finding a common set of tracks which can be extrapolated to the same position within the beampipe. The vertex position is fitted based on candidate tracks with an adaptive vertex fit [73].
- **Jet Algorithms.** When high-energetic quarks or gluons are produced in a collision then, due to confinement in QCD, jets of particles are created as described in Section 1.5. The sum of the four-momenta of the jet constituents is a good approximation to the four-momentum of the original quark or gluon which is usually of interest in analyses.

Jets can be constructed from tracks or also from deposits in the electronic and/or hadronic calorimeter. There exist various algorithms to compose jets from such objects which can be categorized into either cone-based algorithms or clustering algorithms. Cone-based algorithms, such as *IterativeCone*, start at a high-energetic entry point (“seed”) in η - ϕ -space and combine all objects in a fixed cone around

that seed to a jet. Clustering algorithms start with two single objects with the lowest difference in four-momentum and then iteratively add as many objects to the jet as long as the difference in four-momentum between the jet and the candidate is low enough. How exactly that difference is measured and what thresholds are used depends on the individual algorithms. Examples for these types of algorithms used in CMSSW include the k_T algorithm [74] and the Cambridge-Aachen algorithm [75]. In contrast to the cone-based algorithms the shape of jets constructed with a clustering algorithm is not constrained to be circular in η - ϕ -space.

There are two important properties which modern jet algorithms should provide: infrared safety and collinear safety. An algorithm is said to be *infrared safe* if additional low-energetic objects in the event do not change the output of the algorithm. The most common problem infrared unsafe algorithms have is that two separate jets could be merged into a single jet when there is a soft particle between them. An algorithm is *collinear safe* if its output is independent from whether there is a single object with all the energy or when there are two or more collinear objects each of them carrying a part of the energy of the real physical particle. The latter case can easily happen for example when calorimeter deposits are split between adjacent calorimeter cells or when quark or gluon radiation leads to a parallel track. Cone-based algorithms usually require special precautions in contrast to clustering algorithms which are intrinsically both infrared and collinear safe. An example for an infrared safe cone algorithm is the *SISCone* algorithm [76].

- **Muon Reconstruction.** Muons produce both tracks in the silicon tracker and the muon chambers. The track reconstruction in the muon system works the same way as it does in the silicon tracker except for the fact that the energy loss of the muon in the iron material between the individual chambers and the inhomogeneous magnetic field in the outer region are taken into account in the parametrization of the muon track.

In addition to these so-called *stand-alone muons* which are reconstructed using either the silicon tracker or the muon system alone it is also possible to combine the information from the two components. The tracks from the muon chambers are extrapolated to the silicon tracker region in order to find a matching track there. When such a match can be found a global fit of both the tracker and muon chamber hits is performed. A muon reconstructed this way is called a *global muon*.

3.3.6 Particle Flow

The PARTICLE FLOW algorithm [77] attempts to reconstruct stable particles in the CMS detector individually. This includes photons, electrons, muons, charged pions, protons and neutrons. First, every subdetector is considered individually and a clustering of calorimeter cells is performed. At this point, individual particles cannot be identified yet since for example a track in the silicon detector can belong to any charged particle. Therefore, in the next step geometric links between tracks, electromagnetic and hadronic

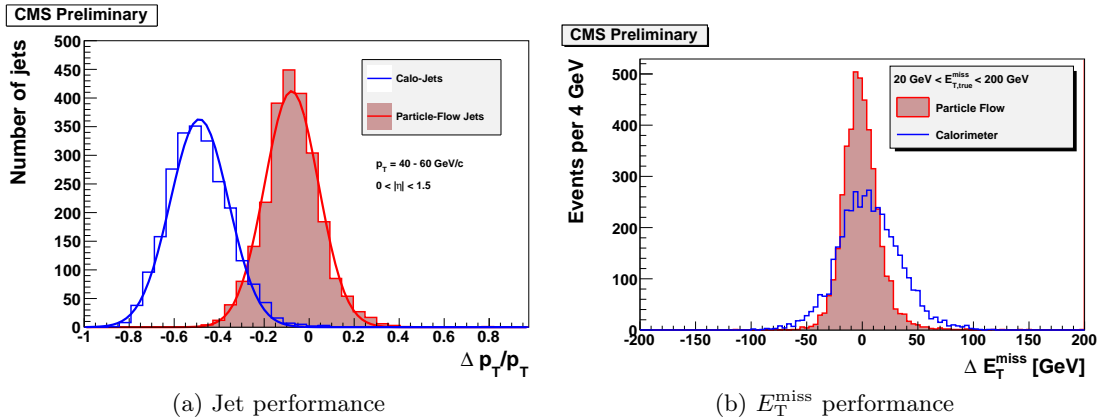


Figure 3.1: Plot (a) shows the relative resolution, $(p_T^{\text{reco}} - p_T^{\text{gen}})/p_T^{\text{gen}}$, for calorimeter jets and PARTICLE FLOW jets. Only jets in the central detector region and with transverse momentum between 40 GeV and 60 GeV were considered. The plot was generated from a Monte Carlo sample of QCD events. Plot (b) shows the absolute resolution of the missing transverse momentum in simulated $t\bar{t}$ events. From [77].

calorimeter clusters are established. This combines information from the three systems where possible, to prevent double counting and to unambiguously identify the particles in question.

PARTICE FLOWS combines all subdetectors in an ideal way to also reconstruct the four-momenta of the identified particles more precisely than with a single subdetector alone. This is an improvement especially for jets, missing transverse energy (E_T^{miss}) and hadronically decaying τ leptons. PARTICLE FLOW jets are generated by clustering particles identified by PARTICLE FLOW instead of tracks or calorimeter deposits. E_T^{miss} can simply be computed as the negative sum of the transverse momenta of all identified particles. Figure 3.1 shows examples of the improvement in jet and E_T^{miss} reconstruction when compared to calorimeter-only information. Not only are the distributions narrower but in the jets case it can also be observed that it is much more central around zero.

3.4 Analysis Workflow

The development of an analysis is an iterative process where the actual analysis code needs to run many times over the datasets. Therefore, a typical analysis is divided into multiple steps so that time-consuming steps do not need to be repeated in every iteration. First, the required datasets are looked up in DBS for both data and Monte Carlo events. Before the analysis itself is involved, the datasets are preprocessed so that only the information required by the analysis is contained in the result. This possibly includes omitting events that are not of interest. This process is called *skimming*. For example, if an analysis does not require electrons the output will not contain any reconstructed

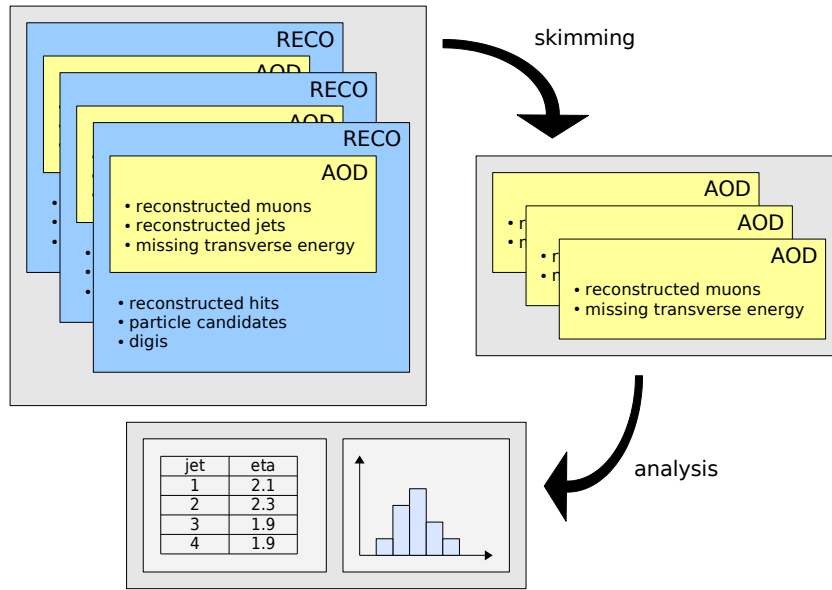


Figure 3.2: Typical analysis workflow: the reconstructed datasets are first skimmed to remove irrelevant information to be able to run over the skimmed data quickly in the subsequent analysis process. The result of the analysis consists of histograms and n-tuples. From [78].

PARTICLE FLOW electrons. This way the size of the datasets can be reduced drastically so that the actual analysis code can quickly run over them and also so that they can be copied to the local institute cluster or even a personal computer.

The skimming procedure itself can again be performed in multiple steps. After the reconstruction algorithms have been run on a dataset it is stored in *RECO* format. This still includes many low-level information that are not used in many analyses such as individual tracker hits. Therefore, a first pass skimming is already performed centrally by CMS leading to data in *AOD*³ format which is often skimmed again according to the needs of the particular analysis. Eventually the analysis code is run, producing histograms or n-tuples as visualized in Figure 3.2.

³Analysis Object Data

3 High Energy Physics Software and Frameworks

4 The LHC Computing Grid

The data rate of the LHC experiments is in the order of 10 PB per year [79, 80]. Due to this large amount of data the traditional High Energy Physics approach to computing where all data are stored and processed at one large computing facility is not viable anymore. The expected number of CPU cores required to run reprocessing and skimming jobs is about $O(100,000)$ which is again about one order of magnitude above what today's largest scientific computing centers can provide.

Therefore, the Worldwide LHC Computing Grid (WLCG) was funded and designed. In addition, due to international funding constraints and for redundancy reasons a distributed approach was chosen. The idea behind the WLCG is to distribute the required resources between many computing centers, also called *Grid sites*. All Grid sites deploy software which adheres to standardized interfaces to ensure interoperability between them. This way the computing power of many centers around the world can be combined to meet the computing requirements of the LHC experiments.

4.1 Grid Structure

The WLCG is structured hierarchically into four layers, also known as *Tiers* [79, 82]. Figure 4.1 visualizes its structure for the CMS experiment.

There is only one root layer, or Tier-0, which is located at CERN with direct connections to the LHC experiments. The raw experiment data are archived at the Tier-0 and a first pass reconstruction is performed. Both raw and reconstructed data are replicated to the next layer in the Grid, the Tier-1 centers, in such a way that all Tier-1s together have a complete copy of all LHC data available. This provides redundant storage so that all data are still available in case of data loss at one center. During LHC shutdown and maintenance periods when there is no new data being acquired the Tier-0 is also used for data reprocessing.

The Tier-1s are fairly large computing centers with high storage capabilities and a direct dedicated broadband connection to the Tier-0 at CERN and in many cases direct connections to other Tier-1s for redundancy purposes. Data reprocessing jobs are run and the results are stored at the Tier-1 centers. Also, reconstructed data are transferred to the Tier-2s for processing by users and Monte-Carlo datasets produced by the Tier-2s are transferred to a Tier-1 for permanent storage and further distribution. There are 11 Tier-1s all of which have agreed to provide a very high availability of 99 % service quality toward the experiments [83].

The Tier-2s are connected to at least one Tier-1 center but there are usually no direct links between the Tier-2s. Apart from official Monte Carlo production the main mission

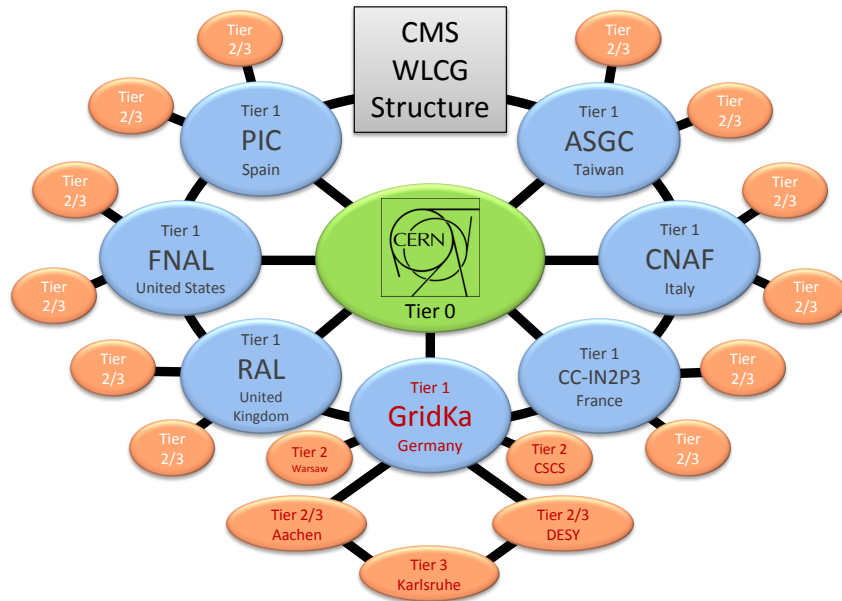


Figure 4.1: The multi-layered structure of the CMS part of the LHC Computing Grid. From [81].

of the Tier-2s is to execute user jobs such as custom analysis or skimming jobs.

Local institute clusters and individual desktop computers that are connected to a Tier-2 center form the Tier-3s. Officially they are not part of the WLCG and this is also why there is no dedicated usage pattern for these sites. However, they are used for end-user analyses and final visualization of analysis results and therefore are a central part in the chain of typical High Energy Physics analyses.

4.2 Grid Architecture and Components

In this section, the required steps to run computing jobs in the Grid and what components are involved in job submission, execution and retrieving the results are discussed.

The WLCG is organized into multiple so-called *Virtual Organizations*, or VOs. For example, there is one VO representing each LHC experiment. In order to access resources on the Grid a user needs to be registered with one or more VOs. The first step is obtaining a *Grid certificate* which authenticates the user in the Grid via the user's institute which is then signed by a VO representative. The certificate allows cryptographically secure authentication and communication between all Grid components using the known and well-established *Public Key Infrastructure* principle.

Once the certificate is set up and registered with a VO, computing jobs can be submitted to the Grid. This can be done by logging into a machine which provides a User Interface (UI) to access the Grid. Often, these User Interfaces are available on machines

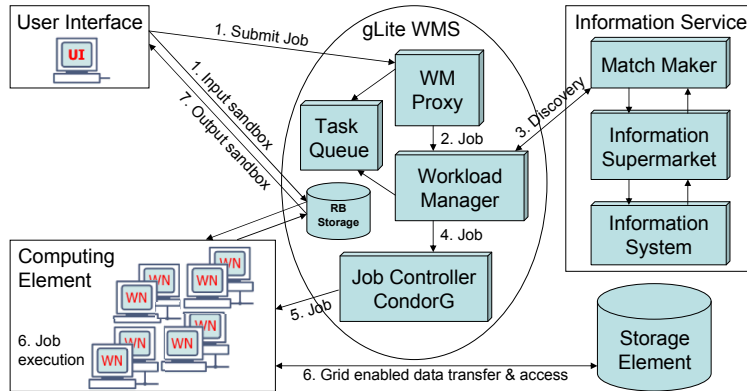


Figure 4.2: Typical work flow of a Grid job: a job is submitted by the user from a UI to the WMS. Depending on the demands of the job (for example in case it requires input data to analyze) and the current load of the Grid the WMS decides to which computing center the job is sent for execution. When the job has finished the Output Sandbox is sent back to the user via the WMS. From [84].

at the local institute or at the national Tier-2s or Tier-3s. Such a UI basically consists of tools and commands for working with the Grid, including mechanisms for authentication with the VO's *Virtual Organization Membership Service* (VOMS) using the Grid certificate. Once authenticated a so-called "VOMS proxy" is created which can subsequently be used for authentication with the *Workload Management Service* (WMS) to which jobs are eventually submitted. For security reasons the lifetime of the proxy is limited (max. 200 h) but it can be renewed in a matter of seconds with the Grid certificate at hand.

To submit a computing job, a so-called *Input Sandbox* needs to be prepared. This is an archive containing all files that the job requires to run, such as executables and shared libraries required by the executables. It also contains a job description file formatted according to the *Job Description Language*, or JDL [85]. The JDL file not only specifies which executable in the Input Sandbox to run on the *Worker Node* (the computer which eventually executes the job) with what input parameters, but also the requirements for the job, such as time, memory or data availability. Based on the requirements, the available resources, the current load and special user privileges (for example, German users might be granted additional resources at the German Grid centers), the WMS decides to which Grid site the job is sent. The JDL also specifies what files to send back to the user in an *Output Sandbox* when the job has finished running.

Once it was decided by the WMS to which center to send a job it submits the Input Sandbox to that center's *Computing Element*, or CE. The CE acts as an interface between the WLCG and a local batch system which queues the job for execution on a Worker Node.

When a job either requires or produces large amounts of data then these data are not

transferred via the Input or Output Sandboxes, respectively. Instead, they should be copied to a Storage Element (SE) for performance reasons. In the WLCG a SE is a mass storage system which is responsible for bookkeeping of all available data on disk or tape storage. This way large data do not need to be copied more than once for many similar jobs. The WMS takes care to only send jobs to sites whose SEs have the data required by the corresponding job. Additionally, sites can be blacklisted or whitelisted in the JDL. This mechanism is used especially when analyzing LHC data (both simulated and “real” detector output) since a typical dataset is usually at least several 100 GB in size and potentially available already at the SEs of multiple Tier-2 centers.

In summary, from a user point of view the steps to submit a job to the Grid are: obtaining a certificate (only once), creating a VOMS proxy (once every week or so), writing a JDL file and submitting it to the WMS. The WMS returns a unique identifier for each job which can then be used to query the status of the job or to fetch its result when it has finished.

As has been illustrated in this section a Grid site consists of many different components and subsystems such as the the CEs, SEs, the worker nodes, the authentication infrastructure, data transfers to and from other Grid centers or the storage and networking systems themselves. All of these components need to work together correctly for the Grid site to be fully operative, so a problem with one of them can render the whole center non-functional. Therefore an efficient monitoring for the various subsystems must be available to be able to quickly react in case problems arise. The HAPPYFACE PROJECT, presented in the next section, is such an efficient monitoring solution.

4.3 The HappyFace Project

Scope. The work on the HAPPYFACE PROJECT was a technical contribution within the scope of this thesis. Several improvements to its core have been made and various modules (described in Section 4.3.6) were improved or developed.

4.3.1 Motivation

Most of the available components for Grid sites provide their own monitoring, such as DCACHE [86] or PHEDEX [87]. Existing monitoring software like DASHBOARD [88] provide extensive site-spanning information for a single aspect only, such as job monitoring. Usually such monitoring software is web-based, so that it can be accessed from anywhere using a web browser.

However, to get a quick overview about the global health of the entire system, this is inconvenient since many different websites of monitoring tools need to be checked periodically. This results in many browser windows or tabs open which makes it cumbersome to systematically check the site’s status. Often, multiple websites also provide partially redundant information so that it takes longer to only browse the relevant information. Another inconvenience of this approach is that many monitoring websites require specific parameters to be provided every time when accessing the content for it to show the

information one is actually interested in. These parameters could be, for example, the Grid site to be shown in case the monitoring system handles multiple sites, or the time range for which to show information. Another issue is that due to the complicated nature of these systems the way the monitoring information is presented is often complicated, varies from system to system and is not understandable to non-experts.

Many of these monitoring websites are fairly complex so that it takes them a considerable amount of time to gather the information to show, for example if they need to query other services which are under load themselves. This results in long loading times for the user and thus reduces efficiency in finding potential problems. This effect is amplified if the website does not show all information at once but requires the page to be reloaded to access more information.

Often, problems of services at a computer center are correlated. For example, if the batch system monitoring shows many failed jobs in the past hour and at the same time a *Transfer Agent* reports failed data transfers to the site's SE then it is likely that the jobs fail *due to* a problem with data transfer when they try to write their results to the SE. These kind of correlations can be difficult to notice when only looking at each component's monitoring individually.

4.3.2 HappyFace Goals and Features

These issues lead to the development of the HAPPYFACE PROJECT (or HAPPYFACE in short) [89]. The main idea of HAPPYFACE is that it does not generate any new information but instead aggregates available information from existing monitoring sources and visualizes it at a single place: HAPPYFACE is a *Meta Monitoring Framework*.

The development of HAPPYFACE started in Karlsruhe in 2007 as a part of the diploma thesis of V. Mauch [90]. During operation of the Tier-1 center GridKa in Karlsruhe it became obvious that new monitoring techniques were required to reliably fulfill the service quality all Tier-1 centers in the WLCG have to provide.

The major design goals of HAPPYFACE are the following:

- **Simplify monitoring.** There should be only one single entry point for accessing all monitoring information, providing detailed information about potential problems at the Grid site. The consequences of this are that such a tool, when widely deployed, reduces the manpower needed for administration and maintenance shifts and that, if there are instructions available what to do when a certain problem occurs, it also allows shifts to be performed even by non-experts.
- **Quick access.** Another design criterion of HAPPYFACE was its performance: it should not take longer than a few seconds to load the website. This implies that the monitoring sources are not queried in real-time because many of them do not meet this requirement. Instead all information has to be pre-fetched and cached locally.

HAPPYFACE runs in a regular interval such as every 15 minutes. In each run it queries all external sources and stores the results locally for display on its website. This approach also minimizes load on the monitoring sources.

- **Easy to deploy.** It should be easy to set up a HAPPYFACE instance on a GNU/Linux machine. Also reconfiguring and updating the software should be easily possible.
- **Multiple information sources.** Information from many different sources in various formats should be accepted. Ideally, a whole center's status can be solely monitored in HAPPYFACE only. If there is a problem then a link to the original monitoring source should be available which can be used to access further information.
- **Modular architecture.** With the experiences made with the first version of HAPPYFACE it went through a major redesign in 2009. It proved useful to build upon a modular architecture which allows to easily add additional information sources to the monitoring, or also to remove any during operation without interfering with the rest of the system. Also, new modules can be developed and deployed very easily by adhering to a simple interface.

The actual HAPPYFACE functionality is therefore divided into many pluggable modules. A module can either be rated (also called *test*) or unrated (*plot*). Tests usually download data from an external monitoring source and then analyze it to see whether the service is in a good state or not. The status is represented as a floating point number ranging from 0 to 1. So for each module also a “warning” state (0.5) can be defined. Plots on the other hand simply download binary information, usually a graph, and show it on the website.

Individual modules are grouped into different categories. For example, there can be a category for modules related to the storage system, a category for the batch system and a category for Grid infrastructure tests. Categories are represented by tabs on the HAPPYFACE website and which provide access to the modules which belong to it when clicked at. Based on the results of test modules the category is also assigned a rating. Different algorithms are available, such as averaging over all tests or taking the lowest rating value as category status.

- **History functionality.** There should be a way to check the center's status at every chosen time in the past. This can be useful to find out at working hours what happened during the weekend, for example whether two problems occurred at the same time which is a strong hint for them to be correlated. Comparing similar states at different times allows as well to track down correlation of problems easily.

HappyFace Website. Figure 4.3 shows a screenshot of the website generated by HAPPYFACE. On the very top of the page there is a header bar which, amongst showing an icon of the site and the current HAPPYFACE version, allows to navigate back and forth in time. On startup the most recent data are shown.

The screenshot shows the HAPPYFACE website interface. At the top, there is a header bar with the project name 'The Happy Face Project', version 'Rev. 613', and a date/time display '27. May 2011 16:30'. Below the header is a row of navigation tabs: 'News', 'Infrastructure', 'Batch System', 'PhEDEx - Prod', 'PhEDEx - Debug', 'dCache', 'Grid', and 'Information'. Each tab has a status indicator: a green arrow pointing up for 'News', 'Batch System', 'PhEDEx - Prod', and 'PhEDEx - Debug'; a yellow arrow pointing right for 'dCache' and 'Grid'; and a red arrow pointing down for 'Infrastructure'. A 'Goto' button and a 'Reset' button are also present.

The main content area displays two module status sections. The first section is titled 'Check DCAP Functionality' and shows the following details:

LAST USCHI EXECUTION (EVERY 15 MINUTES):	2011-05-27 16:21:01	error code: 2
LAST MODULE EXECUTION (EVERY 60 MINUTES):	2011-05-27 15:45:17	
<p><i>This module tests the basic DCAP functionality of the CMS dCache instance at GridKa. Since it is not possible to do DCAP writes, the file resides on the disk-only pools. It copies this file via DCAP from the CMS dCache disk-only area to the CMS VoBox. The test then compares the source file checksum (in the USCHI directory on the VoBox) with the freshly downloaded file Adler32 checksum from the dCache area.</i></p>		

Below this table is a 'show/hide results' button.

The second section is titled 'Check PhEDEx Proxy Lifetime' and shows the following details:

LAST USCHI EXECUTION (EVERY 15 MINUTES):	2011-05-27 16:21:01	error code: 0
LAST MODULE EXECUTION (EVERY 15 MINUTES):	2011-05-27 16:21:01	
<p><i>This test checks, if a valid proxy is available for the PhEDEx transfers. The threshold is given below. If this test fails, login to the VoBox and check if there is a problem with the proxy renewal daemon (see /opt/vobox/cms/log/events.log for more details). Another source could be as well the expiration of a long lived user proxy, either locally or on the myproxy server. In both cases, renew the proxy to guarantee smooth running of the PhEDEx agents</i></p>		

Below this table is a 'show/hide results' button.

Figure 4.3: Screenshot of the HAPPYFACE website. The bar at the top allows navigation in time, the tabs below switch between categories. For each category the modules belonging to it are shown in the main area, along with a quick module navigation bar on the left.

Below the header bar there is a row of tabs representing the available categories. Clicking on one tab shows all modules that belong to the corresponding category. The content of all tabs is loaded in advance from the webserver so that switching between them does not cause any delay due to reloading. The arrows on the tabs indicate the status of the category: a green arrow pointing upwards represents a status value of 1.0 (everything OK), a yellow arrow pointing to the right is shown for status 0.5 (warning) and a red arrow pointing downwards for 0.0 (critical). It is also possible to choose a different visualization theme for the three states, for instance there is also a version which shows happy or sad faces¹. The symbol showing a chart instead of an arrow indicates that this is an unrated category or module (a “plot”, not a “test”).

The minor symbol in the lower right corner of the main category icon indicates whether

¹In fact this visualization theme was the first one available, giving the project its name.

Module: dcache_datastore_lazy

Start: 2011-05-24 16:47 Stop: 2011-05-26 16:47 Variable: **total**
 Interval: 48 hours Stop: now Legend: bottom inside right
 Interval: 48 hours Stop: 2011-05-26 16:47

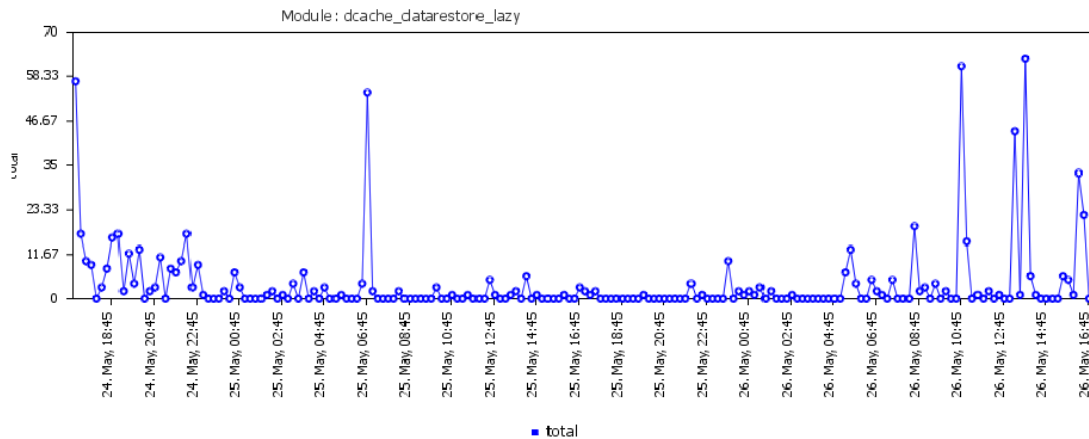
Variable(s): total

Figure 4.4: Number of staging requests at GridKa from May 24 to May 26, 2011. There are various other quantities that can be plotted, depending on the module chosen. The time range can be freely selected.

all modules within the category executed successfully or not. A module can fail for example if its download of external data are not successful or if a downloaded file is not formatted as expected (for instance, invalid XML). If a module could not be executed it is given the special status value -1 . It can also happen that the user is not authorized to see a certain module, for example because his or her certificate does not allow viewing the module or the user did not authenticate at all. In this case the module status is -2 .

The navigation bar on the left allows to quickly browse between the modules of the currently selected category. This way the user can quickly jump to a module which indicates a problem.

The main view shows the output of the modules in the selected category. Each module begins with a small header containing an icon which indicates the module status, its title and execution time. Below, extended module information is available. It is initially hidden but can be made visible by clicking on the “Show Module Information“ link. The module information includes the name of the PYTHON script of the module, configuration parameters which affect the module’s rating (“Definition”), a link to the original monitoring information (“Source”) and instructions for shifters what it means and/or what to do if the module indicates a problem. The module information box also contains a simple plot generator: any numeric variable that a module stores in its database table

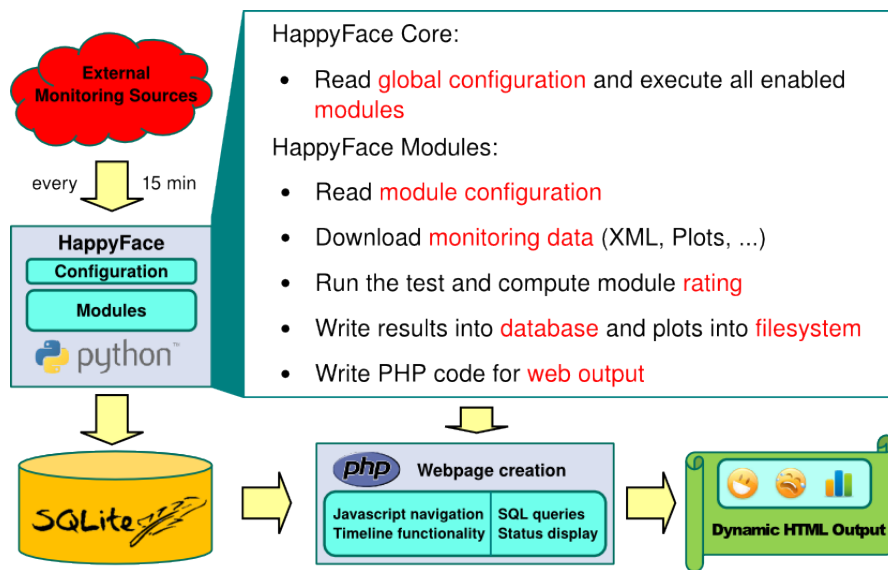


Figure 4.5: Work flow in the HAPPYFACE framework: external monitoring information is queried in a regular interval and the results of the rating is stored into a SQLITE database. Also, a PHP code fragment is generated which visualizes the data stored in the database on a website.

can be plotted against time. Examples for such variables include the module’s status, the amount of free or used space for a module monitoring a DCACHE pool or the number of failed PHEDEX transfers in a corresponding module. As an example Figure 4.4 shows the total number of requests for staging a file from tape to disk at GridKa for a time period of 48 hours.

Below the module information box there is the main output which is specific to every module. Selected modules are discussed in Section 4.3.6.

4.3.3 HappyFace Architecture

Internally HAPPYFACE basically consists of two parts: The HAPPYFACE core and the modules. The core is responsible for providing common functionality to all modules, for executing all enabled modules and for creating the website structure. Modules query an information source, rate it, write the results into a database and generate code to read it from the database and display it on the website generated by the core.

Figure 4.5 visualizes the general data flow in HAPPYFACE. The main HAPPYFACE code is written in the PYTHON programming language [91]. It is supposed to be run periodically by a “cronjob”, but it can also be executed manually for development and testing purposes. For each run, the modules query their external monitoring sources, and, in the case of tests, evaluate the output and eventually compute the module status which is then written to a SQLITE database [92]. Also, all other relevant information that is used for displaying the module on the website, such as the current configuration

settings of the module or a link to the monitoring source is written to the database as well. Plot modules store the image they have received on the filesystem and put a link to its location into the database. In addition, each module generates PHP code [93] which is executed via the webserver when the website is displayed in a browser. It fetches information from the database and generates the HTML code for the browser.

It is important to note that this two-stage process is essential for the history functionality to work properly: the PYTHON part only cares about the latest data currently being retrieved and stores it. It does not encode any information about the latest run in the generated PHP code itself but it writes everything that needs to be shown on the website into the database. This is because the PHP code is not only used to display the most recent data but it can also show a previous dataset for another point in time by simply querying another row in the database.

However, this procedure also requires taking care when developing an update for a module. When a new field is added to a database table then that new field will be empty for records which were created before the module update. The generated PHP code therefore needs to be prepared for that case when the user navigates back in time. Similar considerations need to be taken into account when removing a database field.

Technical choices. Of course there are many programming languages and database solutions available that could have been chosen for HAPPYFACE. For the project to be easy to install and operate only common and widely spread technologies were considered.

- PYTHON was chosen as the project's primary language because it is well documented, has an exhaustive standard library and is already heavily used and acknowledged by the High Energy Physics community. Since most of the execution time of HAPPYFACE is spent on waiting for data from remote servers, performance considerations of the PYTHON code itself do not need to be taken into account.
- PHP was chosen for the web output because, unlike PYTHON, it is available on virtually all webservers.
- SQLITE was chosen because it is a very lightweight database solution. The whole database is stored in a single file on the filesystem. Nevertheless, continuous operation at KIT for more than two years has proven that it scales well for sizes up to O(10 GiB) and there are no indications that it will run into scaling problems in the future. The database being stored in a single file allows to easily move the whole HAPPYFACE instance coherently to another location. However if the filesystem is being backed up periodically then copying a file this large the backup run can turn out to be problematic, especially if copying the file takes longer than the interval between two HAPPYFACE runs.

So if, for this or some other reason, migration to a different database solution becomes necessary at one point then not much code needs to be changed due to the use of wrapper libraries for database access, namely SQLOBJECT [94] on the PYTHON side and PDO [95] on the PHP side. These wrapper libraries provide a

uniform interface to a wide range of database systems such as SQLITE, MYSQL [96] or POSTGRESQL [97]. For a few performance critical operations SQL is used directly, bypassing the wrapper, however this is only standard SQL code that is also available with other database solutions.

4.3.4 HappyFace Installation

HAPPYFACE can run on any modern Linux computer. All that is required is a webserver (such as APACHE [98]), PYTHON, PHP, SQLITE and SUBVERSION [99] for code management.

In a directory that can be accessed via the webserver, the following command needs to be run to acquire the latest version of the code:

```
svn co https://ekptrac.physik.uni-karlsruhe.de/public/HappyFace/
trunk myHFInstance
```

Next, a cronjob is set up to run the main script `run.py` periodically. 15 minutes has been proven to be a good interval. This can be done by adding the following line to `crontab`, for example by running `crontab -e` on the command line:

```
*/15 * * * * cd /path/to/myHFInstance/HappyFace && ./run.py >/dev/
null 2>&1
```

This sets up a basic HAPPYFACE instance with a few example modules activated. The next step is to customize HAPPYFACE to the site's requirements by editing the configuration file, `run.cfg`. The configuration file is rather self-explanatory, however it is suggested to copy it to `local/cfg/run.local` before making site-specific changes. The configuration options in `local/cfg/run.local` override the options given in `run.cfg`. This way updating the software does not conflict with local modifications to the configuration.

The generated website is stored in `webpage/index.php` by default. The SQLITE database will also be created at this location as well as the so-called archive directory where downloaded binary data like plots are stored.

More documentation on installing and running a HAPPYFACE instance is available in the official HAPPYFACE documentation [100].

4.3.5 HappyFace Core System

The HappyFace core consists of the main script (`run.py`), some components which provide common functionality for all modules, and the web output routines. `run.py` reads the main HAPPYFACE configuration, `run.cfg`, which contains configurable parameters in an `.ini`-like format. Examples for such parameters include the title of the website to generate (`index.php`) and the output directory where to store it. It also specifies what modules to run and their organization into categories. Then, it instantiates the module classes, downloads all data required by the modules, calculates the module ratings and finally generates the website.

The different tasks are handled by the following components, which are implemented as PYTHON classes.

- **ConfigService.** Besides the global HAPPYFACE configuration file there is one extra configuration file for each module, containing module-specific configuration such as the URL of the monitoring source(s) or instructions for shifters. Apart from that, there is also an optional local configuration file for each module. The idea is again to keep the basic default configuration in a different file than any additional settings which are specific to a particular site. The default configuration is shipped with HAPPYFACE and sites can choose to alter some settings in the local configuration, so that when updating HAPPYFACE they are not lost.

The *ConfigService* class takes care of loading the configuration files in the correct order and merging the settings. It not only loads the configuration file of the module in question but also the ones of all the modules it derives from, directly or indirectly. For a module called `MyModule`, the global configuration file is called `MyModule.cfg` and it must be located at the same place where the PYTHON class `MyModule.py` is. The local configuration file is expected to be at `local/cfg/MyModule.local` relative to `run.py`.

- **DownloadService.** The *DownloadService* class handles the download of input data for the various modules. In the module configuration each module can register files to download, or its PYTHON code can call the *DownloadService* directly. These configuration parameters are located in the `[downloadservice]` section of the configuration file and for each file to download there should be one line formatted according to the following:

```
[downloadservice]
plot=command|type|options|url
```

command specifies the program that is used for the download, such as `wget` or `curl`. *type* specifies the type of the file to download, for example `png`, `xml` or `html`. *options* specifies any additional options to pass to the download command and *url* specifies the URL to download. The tag before the equal sign (“plot” in this example) can be used in the module code to refer to the downloaded file.

Once all modules have registered their downloads the *DownloadService* takes care of downloading them, making sure not to download the same file twice even if requested by multiple modules. It runs all the download commands in parallel. When all files have been downloaded (or if a configurable timeout expires) HAPPYFACE execution proceeds to the next step, which consists of actually running all the modules.

- **ModuleBase.** Each module is represented by a PYTHON class which derives from the *ModuleBase* class which in turn is provided by the HAPPYFACE core system. Modules need to implement three instance methods that are called by the core at appropriate times: `__init__`, `process` and `output`.

`__init__(self, module_options)` is the constructor of the class. The `module_options` argument is a dictionary with construction options for the module. It contains for

example the timestamp of the current run, the directory into which to store downloaded images and the category of the module as specified in the global configuration. The reason why they are passed as a dictionary instead of separate arguments is that this way another parameter can be added easily in the future without having to adapt all the modules. In the constructor a module can read module-specific configuration options from the `ConfigService`, it can obtain additional download tags and it can register the database table fields it needs to visualize the result of its test.

`process(self)` runs the test and sets `self.status` to the status of the module, which is a floating point number between 0.0 and 1.0 where 1 means that the tested service is good and 0 indicates a critical status. This method is also used to assign values to the database fields declared in the `__init__` function.

`output(self)` creates the PHP code fragment to visualize the module output. Basically it creates a large string containing PHP code and calls the `PHPoutput` function on it, which is provided by `ModuleBase`. `PHPoutput` simply adds code which is common for all modules to the output, such as an icon indicating the module's status and the module title in the headline.

- **WebCreator.** Once all modules have been executed (which is also done in parallel, since modules do not interfere with each other, except for database access which is protected by a lock) the *WebCreator* is called. It creates the output page `index.php`, adds code for the HAPPYFACE header and footer and navigation to it and then fills it with each module's output.
- **CssService.** Apart from a configuration file and the main PYTHON code every module can also have a CSS associated with it. The CSS file defines styles that can be referred to in the HTML code that the generated PHP script produces. The idea behind CSS is to separate the content and the visual appearance of HTML elements.

The *CssService* class copies each module's CSS file to an appropriate location and references them in the main `index.php` file so that they are properly loaded by the browser.

4.3.6 Available Modules

This section describes some available HAPPYFACE modules in detail, especially ones that were developed or considerably improved in the scope of this thesis. However there are many more modules available, see Appendix A for a more exhaustive list.

JobsStatistics. The *JobsStatistics* module shows information about running and queued batch jobs at a center. If the total number of jobs exceeds a certain threshold then the module indicates a warning or a problem if a given fraction of the jobs are inefficient. An *inefficient* job has a low ratio of wall time to CPU time, in the case of the JobsStatistics module a job is considered inefficient if said ratio is below 10 %. The exact numbers for

4 The LHC Computing Grid

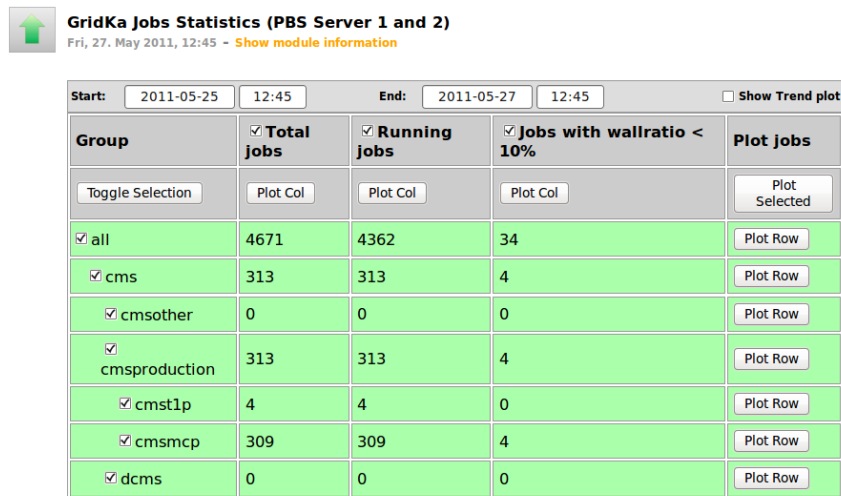


Figure 4.6: The “JobsStatistics” module in HAPPYFACE: the screenshot shows running and queued computing jobs at GridKa. Jobs are assigned to different groups which again are ordered hierarchically: All jobs that are contained in a certain group are also contained in its parent group. In this example, “cmsmcp” is a child group of “cmsproduction” which in turn is a child group of “cms”. Non-CMS jobs are contained in the “all” group but are not listed explicitly. The various “Plot” buttons can be used to generate a plot of running, total and/or inefficient jobs as a function of time for all the groups or a combination of them.

the threshold and the number of inefficient jobs for a warning or an error to be reported can be specified in the module configuration file since sensible values for these numbers depend on the size and the setup of each individual center.

The *wall time* of a job is simply the time since it started running (the name stems from the fact that this is equivalent to the time that a regular clock hanging on the wall would be showing). The *CPU time* is the time that the Worker Node’s processor dedicated to computing the job. On a single core machine this is always less than the wall time. It can be less if the job cannot proceed with computing because it is waiting for Input/Output operations (I/O) to complete, such as reading from or writing to disk, or receiving data from the network. Inefficient jobs therefore indicate problems with the storage system or the network.

Figure 4.6 shows screenshot of the module running in the HAPPYFACE instance at GridKa. The jobs are categorized into different groups, for example to differentiate between data reprocessing and Monte Carlo production jobs, or between jobs from different LHC experiments.

The module works by downloading an XML file containing the information about running jobs. The XML file is generated by a “producer” script which has access to the batch system and then makes the file available to HAPPYFACE by putting it on a

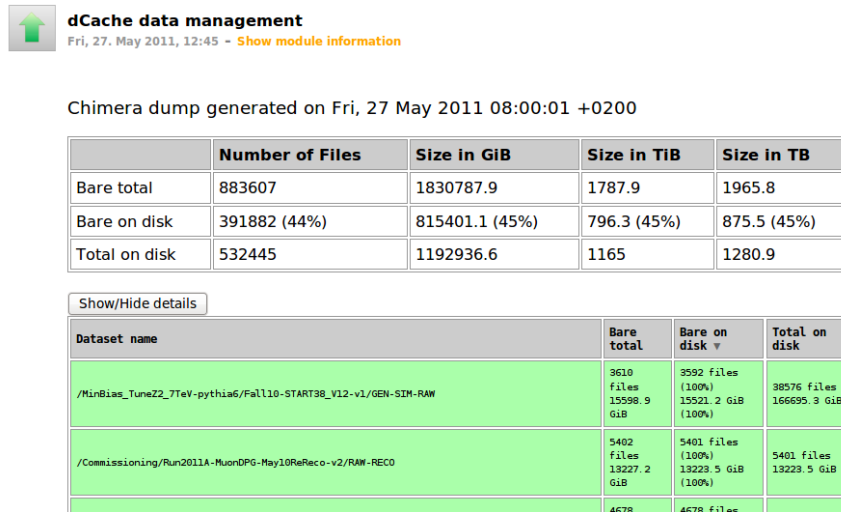


Figure 4.7: The “dCacheDataManagement” module in HAPPYFACE: in the main table the total number of files and the total file size of all datasets available in the DCACHE system at GridKa is shown (“Bare total”). Additionally, the fraction of all datasets which is available on disk (both in number of files and in file size) is given as “Bare on disk”. The total space occupied on disk (“Bare on disk” plus replicas) is shown as “Total on disk”. Clicking on the “Show/Hide Details” button brings up a list of all available datasets. A dataset is highlighted in green when at least 95 % of it is available on disk and thus ready to use for end-user analyses.

webserver. HAPPYFACE contains a producer for the PBS batch system as an external script in the `externals/` directory and additional producers for Conder, LSF and PBS are developed by CMS [101].

A possible future extension to this module is to not only show the number of running and queued jobs but also the number of jobs finished recently, such as during the last two hours. This would allow to also indicate a problem when many of these jobs did not finish successfully. If in addition the exit codes of the finished jobs is known then conclusions can be drawn on the reason most of the jobs failed. However, this information is not provided by the batch system, therefore other means need to be found to make this information available to HAPPYFACE.

dCacheDataManagement. The DCACHE software manages a large amount of stored data ($O(\text{PB})$ at GridKa). It manages multiple pools of disk space and takes care of balancing load between them. It can also transfer files that have not been accessed for a while to tape storage and recover them again if needed. Another feature is that DCACHE can copy the same file to multiple pools if it is requested very often (so-called *replicas*). This way the rate with which a file can be read from a single pool does not limit reading access to it. Furthermore, multiple files that have the same origin are grouped into so-

4 The LHC Computing Grid

called *datasets*. For example, all files produced by a Monte-Carlo simulation of a certain physics process are contained in the same dataset, or all files containing CMS data from the same LHC run period.

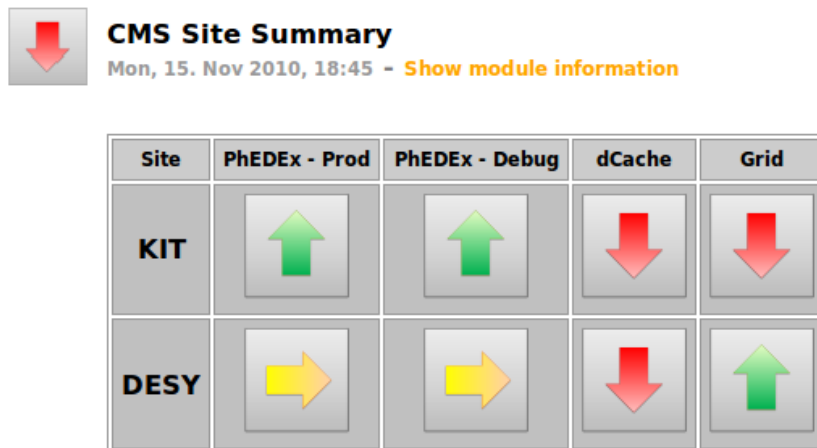
The CHIMERA tool [102] provides a way to obtain a list of all files available in a DCACHE system. There is a helper script in HAPPYFACE (`externals/cms/chimera.py`) which, given such a CHIMERA dump, counts the number of files that are available on disk, individually for each dataset, writes the results into an XML file and again copies it to a webserver. This information is particularly interesting to user analyses since a dataset which has all files available on disk can quickly be analyzed. However if many files of a dataset are only available on tape storage then iterating over it will take much longer because the files need to be read from tape first.

The *dCacheDataManagement* module now visualizes that information in HAPPYFACE. For each dataset it shows the number of files which belong to the dataset, the number of files which are available on disk, with and without duplicates. This way popular datasets can be identified (by looking for datasets with many replicas) and it can be checked whether a sensible fraction of the overall capacity is used for replicas (Typical values range from 5% to 10%). By comparing the local files with central CMS databases, files that do not belong to any dataset can be identified and are assigned to a special “Unassigned” dataset. The `chimera.py` script can also generate a list of these unassigned files so it can be investigated further where those files come from and why they are not cleaned up properly. Figure 4.7 shows a screenshot of the *dCacheDataManagement* module.

RSSFeed. The *RSSFeed* module allows to embed an RSS feed into the HAPPYFACE website. The feed content is read in each HAPPYFACE run and stored in the database as with the other modules. The actual feed parsing work is done by UNIVERSAL FEED PARSER [103]. It handles the RSS 0.9x, RSS 1.0, RSS 2.0, CDF, Atom 0.3, and Atom 1.0 formats. This module is a “plot” module, so no rating is performed.

The idea behind this module is to be able to show any additional (often relatively short-lived) information that could be useful when viewing the HAPPYFACE output. For example, if there is a known problem which cannot be solved immediately an entry can be added to the feed which is then picked up by the RSSFeed module. This way shifters can conveniently be advised to ignore the corresponding HAPPYFACE module since the problem is known and being worked on.

Summary. The current (or also some previous) state of all HAPPYFACE modules can be exported as an XML document. This allows third-party tools to query information from HAPPYFACE. For example there is a Firefox plug-in which shows the current state of a HAPPYFACE instance in the Firefox status bar. Another usage of the XML export feature is the *Summary* module. This module shows the status of selected categories of multiple HAPPYFACE instances in a table. This is achieved by downloading the XML file of all instances, reading the status value of each category that is enabled in the module’s configuration and showing an arrow for each of it. The total module status is given by



Site	PhEDEx - Prod	PhEDEx - Debug	dCache	Grid
KIT	↑	↑	↓	↓
DESY	→	→	↓	↑

Figure 4.8: The “Summary” module in HAPPYFACE: the matrix shows the status of four categories of two different HAPPYFACE instances, one at KIT and one at DESY. The category naming needs to be negotiated between sites so that they can be attributed to each other by the module. Each arrow is a link to the corresponding category of the remote HAPPYFACE instance.

the worst category status of all sites. Figure 4.8 shows a screenshot of the module with four common categories and two instances.

This module drives the idea of HAPPYFACE to the extreme in the way that it allows many computing centers to be supervised starting from a single place. If the Summary module indicates a problem with a site then it can be clicked at to go to that site’s HAPPYFACE instance where the problem can be investigated further. Ideally, this module could allow for nationwide shifts: for instance instead of the German CMS centers at Aachen, Hamburg and Karlsruhe doing their own shifts individually they could do only one shift which is responsible for all three sites, reducing the overall manpower needed for such computing shifts to one third. The Summary module allows the shifter to conveniently check the status of all centers and to take action if a problem occurs by providing a link to the problematic modules where further instructions are available.

4.3.7 Conclusions and Future Work

The HAPPYFACE PROJECT simplifies monitoring of a complex system by aggregating relevant information and presenting them in a consistent way at a single website. It automatically runs tests to rate the status of the system and it allows via the history functionality to check its status at a previous point in time. This is a considerable improvement to the situation before where many websites with partly redundant information needed to be checked by the shift crew. The whole monitoring process could be even more automated if critical modules caused mail to be sent to the experts. Such

4 The LHC Computing Grid

functionality is planned for the future either within the HAPPYFACE core or as a separate tool which reads the HAPPYFACE database.

Since its early development the HAPPYFACE PROJECT has been deployed not only at GridKa but also at many other Tier-2 sites in Germany, such as Aachen, Hamburg or Göttingen. They not only installed the software for their centers but they also developed custom modules which met their requirements and which were added to the main HAPPYFACE code eventually. This is where the modular design of HAPPYFACE paid off since it makes it easy to extend the software by anyone and to adapt it to ones needs.

At the time of this writing several other Grid sites plan to set up a HAPPYFACE instance in the near future. Apart from various ATLAS sites in Germany it is planned to set up a central instance at CERN which monitors all CMS production and reprocessing jobs at the Tier-1 sites. This can be done by installing the corresponding XML producers on each of the Tier-1s and providing the output to the central HAPPYFACE instance. This way HAPPYFACE is not used as a site-specific monitoring solution but instead one specific component is monitored at many sites at a single place. However it fits this purpose equally well.

Other recent developments include support for user authentication which allows certain modules or categories to be visible for authorized users only. The actual authentication is performed using certificates, for example by allowing access to all users with a valid Grid certificate. It is also possible to make use of a more fine-grained control such as on the level of VO memberships, roles within a VO or even on a personal level. All necessary information to do so can be obtained from the certificate. An example for this functionality is the *User Space Monitoring* module which shows the space occupied by each user at the Aachen and DESY Tier-2 sites. For regular users only the space occupied by that user is shown, and whether or not it is over quota. Administrators however can see the occupied space of all users.

It is further planned to introduce a configuration option to choose the database backend to use. With SQLITE the database file can grow to several tens of gigabytes of size which can be a problem in certain scenarios such as regular filesystem backup. Especially supporting POSTGRESQL in addition to SQLITE is envisioned.

5 Analysis of $\tau^+\tau^-$ Final States

One of the primary goals of the CMS detector is the discovery of the Higgs boson. In accordance with the energy-time uncertainty principle $\Delta t \cdot \Delta E \geq \hbar/2$ its decay width as shown in Figure 1.4b leads to a very short lifetime of the order of $O(10^{-25} \text{ s})$ in the whole mass range. Therefore, it cannot be detected directly but only its decay products can be seen in the detector. For low Higgs boson masses it predominantly decays into a $b\bar{b}$ pair or a $\tau^+\tau^-$ pair. For higher masses other decay channels such as W^\pm pair production or Z pair production become dominant.

As discussed in Section ??, the $\tau^+\tau^-$ final state is a promising candidate for discovery of the Standard Model Higgs boson. However, $\tau^+\tau^-$ final states are also interesting in searches for “New Physics”, i.e. observation of new particles or processes that are not described by the current Standard Model. Supersymmetry, as introduced in Section 1.7, is only the most prominent example. Many supersymmetric event topologies include τ leptons in the final state, for example in decays of supersymmetric Higgs bosons or stau leptons [104], the superpartner of τ leptons. τ leptons also play a role in more exotic models such as heavy vector bosons (Z' , W') [105], in heavy ion collisions [106] or in rare decays, such as $B \rightarrow \tau\nu$ [107].

In the following the identification and reconstruction of τ leptons by their decay products is described in section 5.1. Then, di-tau mass reconstruction is discussed in section 5.2 and a selection of $\mu + \tau$ -jet events from data of the CMS detector is presented in section 5.3. Finally, section 5.5 briefly discusses how the selection could be improved and what else can be done to reduce systematic errors.

5.1 τ Identification and Reconstruction

The τ lepton is the heavy sibling of the other charged leptons, the electron and the muon. Its properties are summarized in table 5.1. It is important to note that the given lifetime is not fixed but it is the mean of an exponential distribution.

In contrast to the electron or the muon the τ lepton can decay hadronically due to its higher mass. In almost all hadronic cases one or three charged particles (π^\pm or K^\pm , also called “prongs”) and several neutral particles (π^0 , K^0 or photons) end up in the final state. Because of lepton family number conservation, each decaying τ lepton results in a τ neutrino to be emitted. The τ lepton can also decay leptonically into either a muon or an electron in which case two neutrinos are emitted. Figure 5.1 shows example Feynman diagrams for a leptonic and a hadronic τ lepton decay.

In the detector the τ leptons can only be observed via their decay products. The decay products will lead to energy deposits in the calorimeters and charged tracks in the

Property	e^-	μ^-	τ^-
Mass	0.511 MeV	105.65837 ± 0.00004 MeV	1.77682 ± 0.00016 GeV
Lifetime	$> 4.6 \cdot 10^{26}$ y	$(2.197034 \pm 0.000021) \cdot 10^{-6}$ s	$(2.91 \pm 0.01) \cdot 10^{-13}$ s
Charge	$-e$	$-e$	$-e$
Spin	1/2	1/2	1/2
Weak isospin T_Z	-1/2	-1/2	-1/2

Table 5.1: Properties of the τ^- lepton in comparison with its lighter siblings, the electron and the muon. The properties of the e^+ , μ^+ and τ^+ are the same except for their charge which is $+e$. The measured values are taken from [30].

tracker, but so will quark- or gluon-induced jets as well. Therefore, a way to determine whether a particle or a jet of particles originate from a τ lepton or not, is required. There are several criteria that help make this decision, described in the following.

Since it can only decay weakly the τ lepton has a relatively long lifetime in the order of 10^{-13} seconds. This allows it to travel several hundred micrometers before it decays. This property can be exploited by looking for a secondary vertex displaced from the production vertex. For hadronic decays with three charged particles such a secondary vertex can be reconstructed with the CMS detector, given a very good understanding of the alignment of the detector and a good track reconstruction efficiency.

Another quantity which can be used for τ reconstruction is the missing transverse momentum, E_T^{miss} . Due to momentum conservation the total transverse momentum of the final state in an event must be zero since it was zero already before the collision. In longitudinal direction the same is valid, however particles escaping in the forward or backward regions into the beampipe can not be measured.

Therefore, E_T^{miss} is a measure for the direction and momentum of all particles that have not been observed in the detector, such as neutrinos or long-lived supersymmetric particles. Since there are two neutrinos produced in leptonic tau decays and one neutrino in hadronic decays a considerable amount of E_T^{miss} is also a hint for a τ decay.

5.1.1 Leptonic τ Reconstruction

For leptonic τ reconstruction an *isolation* requirement can be exploited: when a lepton has been reconstructed other particles in a cone in η - ϕ around the lepton track are considered. If the total transverse momentum of all particles within the cone does not exceed a certain threshold, the lepton is said to be isolated. It is expected that leptons from τ decays are isolated because there are no other visible particles coming from the τ decay. However, leptons from QCD processes are often surrounded by other particles which are part of a quark or gluon jet.

Lepton isolation can only discriminate against QCD processes though. Electrons or muons from electroweak processes such as W decays have the same isolation properties. To further identify leptons from τ decays the whole process needs to be taken into

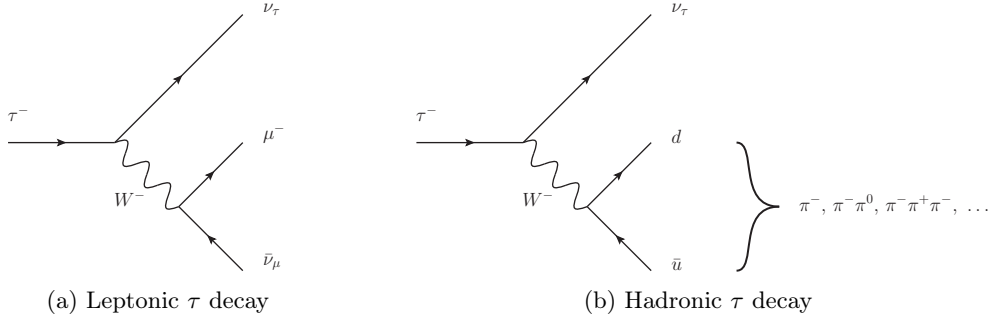


Figure 5.1: Example Feynman diagrams for leptonic (a) and hadronic (b) τ^- lepton decays. The same diagrams with charges conjugated are also allowed. There are more diagrams for τ decays possible, for example the muon can be replaced by an electron in the leptonic case or the down quark can be replaced by a strange quark in the hadronic case (however, such a decay is suppressed because it involves a transition between different quark generations). The two quarks in this example directly make up a π^- , however in the hadronization process additional gluon or W radiation can result in multiple mesons.

account. For example in $Z \rightarrow \tau\tau \rightarrow \mu\mu\nu\nu\nu\nu$ certain kinematic properties of the di-muon system can be exploited to discriminate against $Z \rightarrow \mu\mu$ [108].

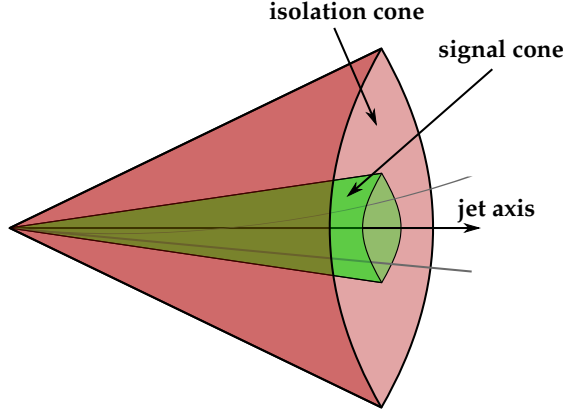
5.1.2 Hadronic τ Reconstruction

For hadronic τ reconstruction the isolation property can be exploited as well. However, since the τ -jet can consist of more than one particle a single isolation cone around the jet axis cannot be used. Instead two cones are used as depicted in Figure 5.2 [109].

Because of the high mass difference between a Z boson or a Higgs boson and a τ pair most rest energy of the boson will convert into kinetic energy of the τ leptons. This results in a high total momentum; the leptons are said to be heavily *boosted*. This implies that the system of decay products will be boosted as well and leads to collimation of τ -jets. Therefore, its constituents are expected to be within a small cone around the jet axis, the so-called signal cone. In a much wider cone around the signal cone an isolation criterion is required, in a similar way as discussed before for leptonic τ decay reconstruction: if there are particles with a transverse momentum greater than a certain threshold within the isolation cone then the jet in question is not considered a τ -jet.

A slight variation of the above “Fixed Cone” algorithm is to choose the cone size of the signal cone to be dependent on the energy of the jet: the more energetic the jet the more collimated it is, therefore the cone size is chosen to be inversely proportional to the jet energy. This is called the “Shrinking Cone” algorithm.

Typical cone sizes for the signal cone are $\Delta R_{\text{sig}} = 0.07$ for the “Fixed Cone” algorithm, where $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ is the Euclidean distance in the η - ϕ -plane. For the “Shrinking Cone”

Figure 5.2: Illustration of the signal and the isolation cone in a hadronic τ lepton decay.

Decay Mode	Resonance	Branching ratio
$\tau^- \rightarrow h^- \nu_\tau$		11.6 %
$\tau^- \rightarrow h^- \pi^0 \nu_\tau$	$\rho(770)$	26.0 %
$\tau^- \rightarrow h^- \pi^0 \pi^0 \nu_\tau$	$a_1(1260)$	10.8 %
$\tau^- \rightarrow h^- h^+ h^- \nu_\tau$	$a_1(1260)$	9.8 %
$\tau^- \rightarrow h^- h^+ h^- \pi^0 \nu_\tau$		4.8 %
Total		63.1 %
Other hadronic decay modes		1.7 %

Table 5.2: Prominent hadronic τ decay modes. h^\pm denotes a charged meson, in most cases a pion (but also kaons are possible). The decay modes for the positively charged tau lepton are the same with all charges in the final state conjugated. From [110].

Cone” algorithm $\Delta R_{\text{sig}} = 5 \text{ GeV}/E_T$, where E_T is the energy of the τ -jet, is a common choice. The minimum and maximum cone sizes are usually clamped to $0.07 \leq \Delta R_{\text{sig}} \leq 0.15$. For the isolation cone typical sizes vary between $\Delta R_{\text{iso}} = 0.3$ and $\Delta R_{\text{iso}} = 0.5$ [110].

Decay Mode Determination. The identification of hadronic τ decays can be improved further by reconstructing the decay mode. Table 5.2 shows the most prominent decay modes. The idea of this approach is to require the constituents of the τ -jet to match one of them. In some of these modes the τ decays to an intermediate strong resonance such as a $\rho(770)$ which then further decays to the final state. This fact can be exploited by requiring that the invariant mass of the final state is within the mass window of the resonance.

In CMS there are two algorithms for decay mode determination: the “Hadron Plus Strips” algorithm (HPS) and the “Tau Neural Classifier” (TaNC) [110, 111].

As a π^0 decays to two photons in virtually all cases the HPS algorithm attempts to reconstruct photons from “strips” in azimuthal direction in the electromagnetic calorimeter. This accounts for photon conversion effects: if electron-positron pairs are created they are heavily bended in the magnetic field of CMS, emitting synchrotron radiation. The algorithm then applies cuts on particle multiplicity and invariant mass to find out whether the jet is compatible to one of the hadronic τ decay modes or not. It can also determine the isolation of the τ -jet more accurately by accounting all objects that are not used for τ reconstruction to the isolation variable, even if it resides within the signal cone.

The TaNC algorithm uses PARTICLE FLOW photon candidates and combines them so that their invariant mass best matches the π^0 mass. It also considers unpaired photons to account for high energetic photons which could not be separated by the PARTICLE FLOW algorithm. With the π^0 s reconstructed the TaNC algorithm feeds five different neural networks, one for each of the decay modes given in table 5.2. An artificial neural network solves problems in pattern recognition and classification by replicating the structure of a human or animal brain in software [112]. A cut on the network output determines whether the τ -jet passes the TaNC discriminator or not.

Lepton rejection. Single leptons (electrons or muons) can be reconstructed as a jet and therefore be considered as a τ -jet candidate. To discriminate against such lepton fakes the leading track of the τ -jet is required not to have been reconstructed as a muon, and for electrons the PARTICLE FLOW multivariate electron discriminator $\text{PF}_{\text{mva}}^{e/\gamma}$ is required to be smaller than 0.6 [113, 114].

Fake rates. Despite all these methods it is still possible for a quark or gluon jet to be misidentified as a τ -jet. The probability of this happening is known as “fake rate”. The fake rate is a quantity used for comparing the performance of different τ identification algorithms. Also, it is important to be known for analyses so that it can be estimated how much background contamination to expect. Typically, in algorithms such as TaNC and HPS, a higher identification efficiency can be traded for a higher fake-rate by tuning parameters of the algorithm. This allows analyses with different requirements concerning signal efficiency and background contributions to choose a suiting working point.

Fake rates can easily be determined on data since it is easy to select a pure sample of QCD jets. The fake rates and also identification efficiencies of HPS and TaNC have been determined in [111]. Figure 5.3a shows the fake rate determined on different data samples compared to corresponding Monte Carlo samples for the HPS loose working point as a function of p_{T} . For $Z \rightarrow \tau^+\tau^-$ and also for light Higgs boson decays the low p_{T} region with $p_{\text{T}} < 60$ GeV is most relevant. In Figure 5.3b the fake rate is plotted against the identification efficiency which was determined using Monte Carlo truth information. Both HPS and TaNC are shown for different data samples. The trade-off between fake rate and efficiency of the various working points can be seen. For Higgs searches a loose working point is used in order to preserve a good efficiency to not suppress the already small Higgs signal even further.

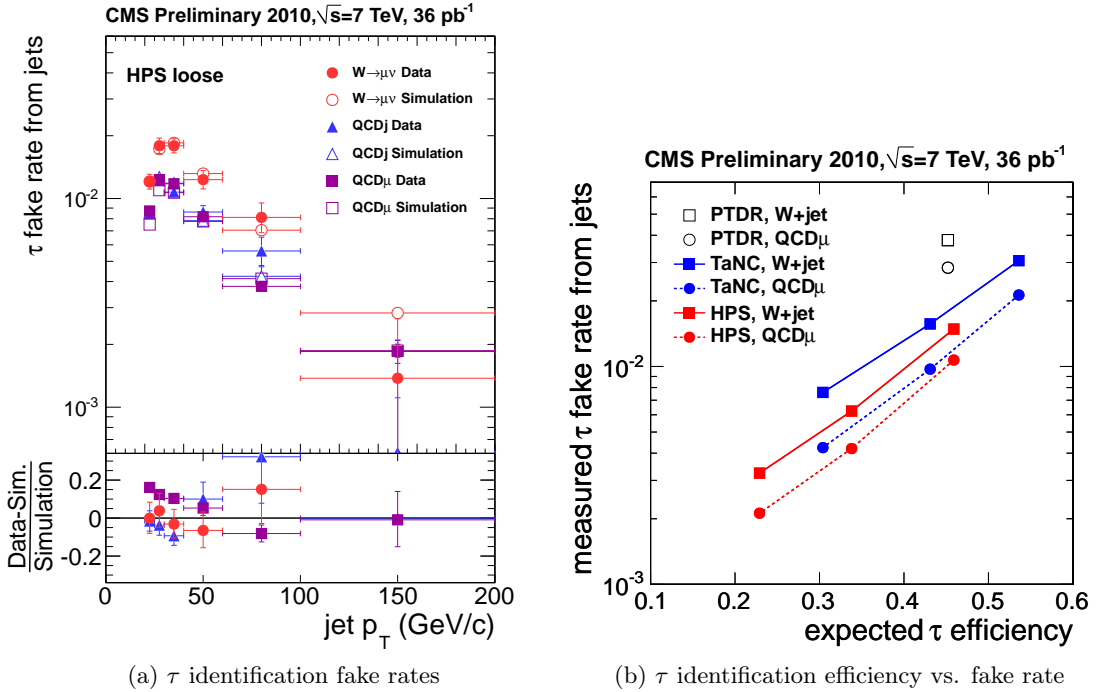


Figure 5.3: Fake rate of tau identification algorithms. Plot (a) shows the fake rate vs. jet p_T of the loose working point of the HPS algorithm measured on different data selections. The fake rates are compared to corresponding Monte Carlo samples. Plot (b) shows fake rate vs. efficiency for all working points of TaNC and HPS. Looser working points have higher efficiency and higher fake rates. From [111].

5.2 Mass Reconstruction

When two τ candidates have been identified it is desirable to estimate the mass of their mother particle, for example to test whether that mass is compatible with the Z mass or the Higgs mass. For $Z/H \rightarrow \mu\mu$ events this is straightforward since the invariant mass of the di-muon system equals the mass of the mother particle. However, in $Z/H \rightarrow \tau\tau$ this is different since the sum of the four-vectors of the τ candidates does not sum up to the four-vector of the mother particle as some momentum is carried away by the neutrinos. Therefore, multiple attempts to obtain a mass hypothesis have been developed in CMS. These methods are discussed in the remainder of this section.

5.2.1 Visible Mass

The *visible mass* is defined as the invariant mass of the two visible decay products (electron, muon or τ -jet). Given their four-momenta as $p^{\text{vis}1}$ and $p^{\text{vis}2}$ it can be written as

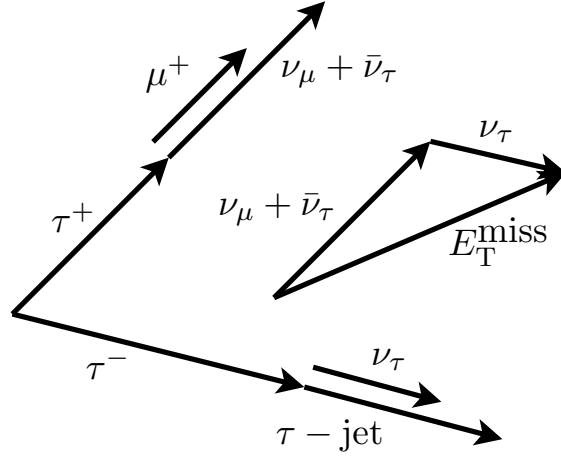


Figure 5.4: Principle of the collinear approximation: The direction of the neutrinos is assumed to be the same as the one of the visible decay products (“ μ^+ ” or “ τ -jet” in this figure). Making use of the missing transverse energy the four-vectors of the original τ leptons can be reconstructed under this assumption. From [12].

$$M_{\text{vis}}^2 = (p^{\text{vis}_1} + p^{\text{vis}_2})^2. \quad (5.1)$$

Since this mass definition does not take the neutrinos into account it does not peak at the mass of the parent resonance. For $Z \rightarrow \tau\tau \rightarrow \mu + \tau$ -jet the visible mass peak is at about 50 GeV. The blue curve in Figure 5.5 shows the visible mass distribution in $Z \rightarrow \tau\tau$ decays.

5.2.2 Collinear Approximation Mass

The *collinear approximation* is a method to reconstruct the four-vector of the parent resonance by making the following two assumptions, visualized in Figure 5.4:

- The neutrino(s) generated in a τ decay are collinear to the visible decay product (electron, muon or τ -jet), i.e. the direction of their momenta is the same.
- The $E_{\text{T}}^{\text{miss}}$ contribution in the event is only due to the neutrinos from the τ decays.

Formally, when $p_{\text{T}}^{\tau_1}$ and $p_{\text{T}}^{\tau_2}$ denote the real transverse momenta of the τ leptons and $p_{\text{T}}^{\text{vis}_1}$ and $p_{\text{T}}^{\text{vis}_2}$ denote the ones of the visible τ decay products, the first assumption can be written as

$$\vec{p}_{\text{T}}^{\tau_1} = x_1 \cdot \vec{p}_{\text{T}}^{\text{vis}_1}, \quad \vec{p}_{\text{T}}^{\tau_2} = x_2 \cdot \vec{p}_{\text{T}}^{\text{vis}_2} \quad (5.2)$$

where the x_i denote the fraction of the τ momenta carried away by the neutrinos.

The second assumption can be written as

5 Analysis of $\tau^+\tau^-$ Final States

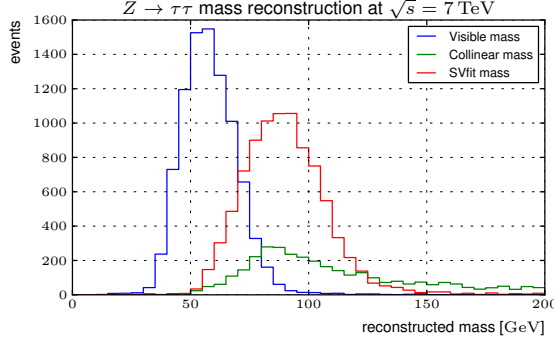


Figure 5.5: Comparison of the three mass reconstruction algorithms. The events shown in the histograms are from a $Z \rightarrow \tau\tau$ Monte Carlo sample. Only $Z \rightarrow \tau\tau \rightarrow \mu + \tau$ -jet events passing the selection described in Section 5.3.2 are considered.

$$\vec{p}_T^{\tau_1} + \vec{p}_T^{\tau_2} = \vec{p}_T^{\text{vis}_1} + \vec{p}_T^{\text{vis}_2} + \vec{E}_T^{\text{miss}} \quad (5.3)$$

Plugging 5.2 into 5.3 yields a system of equations for x_1 and x_2 . The solution is given by

$$x_1 = \frac{p_x^{\text{vis}_1} p_y^{\text{vis}_2} - p_y^{\text{vis}_1} p_x^{\text{vis}_2}}{p_y^{\text{vis}_2} p_x^{\text{miss}} - p_x^{\text{vis}_2} p_y^{\text{miss}} + p_x^{\text{vis}_1} p_y^{\text{vis}_2} - p_y^{\text{vis}_1} p_x^{\text{vis}_2}} \quad (5.4)$$

$$x_2 = \frac{p_x^{\text{vis}_1} p_y^{\text{vis}_2} - p_y^{\text{vis}_1} p_x^{\text{vis}_2}}{p_x^{\text{vis}_1} p_y^{\text{miss}} - p_y^{\text{vis}_1} p_x^{\text{miss}} + p_x^{\text{vis}_1} p_y^{\text{vis}_2} - p_y^{\text{vis}_1} p_x^{\text{vis}_2}}. \quad (5.5)$$

This allows the invariant di-tau mass to be calculated according to

$$M_{\text{coll}}^2 = (p^{\tau_1} + p^{\tau_2})^2 = \left(\frac{p^{\text{vis}_1}}{x_1} + \frac{p^{\text{vis}_2}}{x_2} \right)^2. \quad (5.6)$$

The collinear approximation only yields a physical solution for $0 < x_1 < 1$ and $0 < x_2 < 1$. If this is not the case either one of the two assumptions was spoiled or the directions of the two visible decay products are opposite to each other (back-to-back topology). In the latter case the collinear approximation is not applicable because no solution for both x_1 and x_2 can be found. For $Z \rightarrow \tau\tau$ events the collinear approximation yields a physical solution for about every second event only. Apart from reduced statistical precision this method also suffers from long non-Gaussian tails. Such long tails make it hard to separate the Higgs mass peak from the Z mass peak for a light Higgs boson. However, the collinear approximation mass peaks near the nominal Z mass. The green curve in Figure 5.5 shows a typical collinear mass distribution.

5.2.3 SVfit Mass

The *Secondary Vertex Fit* method (or “SVfit” in short) is a novel technique for di-tau mass reconstruction developed by CMS [115]. The method constructs a likelihood function which depends on several input quantities and then maximizes this likelihood with respect to a di-tau mass hypothesis.

Maximum Likelihood Method. Given a model with parameters $\vec{\vartheta}$ a likelihood function $\mathcal{L}(\vec{x}|\vec{\vartheta})$ describes the likelihood of measuring the values \vec{x} when the model is parametrized by $\vec{\vartheta}$. After a measurement \vec{x} has been performed the model parameters $\vec{\vartheta}$ are chosen so that the likelihood of obtaining \vec{x} is maximized. This is called the *Maximum Likelihood* method.

A likelihood function with multiple variables can be composed of multiple likelihood functions with a single variable if the variables are uncorrelated. In practice, the method also gives good results for only weak correlations [116] and can also be applied this way when the full multi-dimensional likelihood function is unknown. In the multi-dimensional case, given variables x_1 to x_N and likelihood functions $\mathcal{L}_1(x_1|\vec{\vartheta})$ to $\mathcal{L}_N(x_N|\vec{\vartheta})$, the combined likelihood function can be written as

$$\mathcal{L}(\vec{x}|\vec{\vartheta}) = \prod_{i=1}^N \mathcal{L}_i(x_i|\vec{\vartheta}) \quad (5.7)$$

If variables are correlated then the full multi-dimensional likelihood function must be known. However, in practice equation 5.7 still gives good results if the variables are only weakly correlated.

Secondary Vertex Fit. In the SVfit method, $\vec{\vartheta}$ basically consists of the four-vectors of the two τ leptons (which corresponds to six free parameters since the τ mass is known).

The following likelihood terms are used in the fit:

- **τ decay kinematics.** In a hadronic τ decay there are only two decay products, therefore the likelihood for the τ decaying depends only on the angle between the decay products. Leptonic decays however are three-body decays so that the τ momentum also depends on the invariant mass of the di-neutrino system which in this case is fitted as well.
- **Missing transverse momentum.** E_T^{miss} resolution was studied with $Z \rightarrow \mu^+ \mu^-$ events in data. A two-dimensional Gaussian with the measured resolution is matched against the E_T^{miss} in the event.
- **p_T balance.** An additional p_T balance term is added to account for the fact that applying a cut on p_T (see Section 5.3.2) introduces a bias in the mass distribution.
- **Secondary vertex information.** An additional parameter r is introduced in the fit which represents the flight distance of the τ lepton before it decays. The

Decay Channel	Branching ratio
$\tau\tau \rightarrow \mu + \tau_{\text{had}} + \nu\nu\nu$	22.5 %
$\tau\tau \rightarrow e + \tau_{\text{had}} + \nu\nu\nu$	23.1 %
$\tau\tau \rightarrow e + \mu + \nu\nu\nu\nu$	6.2 %
$\tau\tau \rightarrow \mu + \mu + \nu\nu\nu\nu$	3.0 %

Table 5.3: $\tau\tau$ decay channels studied in CMS, and their branching ratio. τ_{had} means a hadronically decaying τ lepton. The $\tau\tau \rightarrow \tau_{\text{had}} + \tau_{\text{had}} + \nu\nu$ and $\tau\tau \rightarrow e + e + \nu\nu\nu\nu$ decay channels exist as well, but because of high background contributions from QCD or $Z \rightarrow e^+e^-$, respectively, they are not analyzed for $H \rightarrow \tau\tau$ searches in CMS.

mean lifetime $c\tau = 87 \mu\text{m}$ is large enough for the CMS detector to resolve, both for three-prong and one-prong τ decays. The probability for a τ lepton to decay after the distance r (an exponential distribution) can be used to further constrain the secondary vertex position. Once it is known further information on the τ four-momenta can be inferred from the relative position of the secondary vertex with respect to the primary one.

However, since the alignment of the CMS tracking detector is not yet fully understood using this information would result in a large systematic uncertainty. For this reason secondary vertex information is not yet used for mass reconstruction. It could be added in the future though, once detector alignment and calibration are better understood.

The exact likelihood terms are given in [115]. In fact the name of the method is somewhat misleading since not only secondary vertex information is used but also other kinematic properties of τ decays.

Unlike the collinear approximation, the secondary vertex fit method, is able to give a mass hypothesis for every event, thus preserving full statistical precision. Also the width of the mass distribution is much narrower than in the collinear approximation and it is more symmetric. Unlike the visible mass distribution its peak is at the nominal mass of the mother particle. Figure 5.5 compares the SVfit mass distribution (red curve) with the other two mass definitions.

5.3 The $\mu + \tau$ -jet Final State

As discussed in the previous section a τ lepton can either decay into an electron, a muon or it can decay hadronically. For a τ pair this allows for six different combinations four of which are incorporated into Higgs searches in CMS. The four channels are summarized in Table 5.3 [117].

The fully hadronic channel $\tau\tau \rightarrow \tau_{\text{had}}\tau_{\text{had}}$ has the highest branching ratio but suffers from enormous QCD background (quark or gluon jets), despite the methods discussed

in the previous section. The $\tau\tau \rightarrow \mu + \tau$ -jet still has a relatively high branching ratio and is therefore studied in detail in the remainder of this chapter. In this channel the requirement of a muon being present in the event reduces the QCD background contamination to a manageable amount. Also, the muon is a very well-understood object: it can be measured precisely and it can be used as a reliable trigger for $\mu + \tau$ -jet events.

5.3.1 Background Contributions

There are several processes which have a jet and a muon in the final state and thus have the same signature as a $\mu + \tau$ -jet event. Many such processes do not involve τ leptons at all but lead to jets in the event which can be misidentified as a τ -jet (see Section 5.1.2). Jets can be created by gluon or W radiation in the initial or the final state, or by event activity from another proton collision (“pile-up”). The background processes considered in this analysis are the following:

- **QCD.** A quark or gluon jet can be misinterpreted as a τ -jet. A muon can appear in such events when a quark radiates a W boson which decays to a muon or a τ . Since at a hadron collider the production cross section for this background contribution is relatively large.
- **$W + \text{jets}$.** When a W boson is created it can decay to a muon or a τ . If there is another jet in the event that is misidentified as a τ -jet then such a $W + \text{jet}$ event can fake a $\mu + \tau$ -jet event.
- **$Z \rightarrow \mu\mu + \text{jets}$.** A Z boson decaying to two muons and either another jet in the event or one of the muons being misidentified as a τ -jet also appears the same way as a $\mu + \tau$ -jet event.
- **$t\bar{t} + \text{jets}$.** $t\bar{t}$ pair production processes also contribute to background for $\mu + \tau$ -jet. Since the top quark decays virtually always to a bottom quark under emission of a W boson a $\mu + \tau$ -jet event can be faked by one W boson decaying into a hadronically decaying τ and the other one to a muon, or by one W decaying to a muon and one of the b -jets or a hadronic decay of the other W boson being misidentified as a τ -jet. Due to the relatively low $t\bar{t}$ production cross section the contribution of this channel is small however.

There are other processes which contribute to the background (such as vector boson pair production), however their contribution is very small and therefore neglected in this analysis. For the signal and background processes Monte Carlo data samples from central CMS FALL10 production have been used to predict the number of data events to expect. The events are produced with the POWHEG generator where the parton shower was simulated with PYTHIA. The samples include simulation of additional proton interactions in the same bunch crossing (“pile-up”) with approximately the same number of additional interactions as observed in the 2010 data taking. Table 5.4 shows the Monte Carlo samples that were included in this analysis.

5 Analysis of $\tau^+\tau^-$ Final States

Process	σ_{prod} [pb]	$N_{\text{generated}}$	L_{equiv} [pb^{-1}]
$Z \rightarrow \tau\tau$	1,666	1,994,719	1,197
$Z \rightarrow \mu\mu$	1,666	1,998,931	1,200
$W \rightarrow \mu\nu$	10,438	3,993,866	383
$W \rightarrow \tau\nu$	10,438	3,990,741	382
$t\bar{t}$	65.83	996,022	15,130
μ -enriched QCD	86,679	28,315,088	334

Table 5.4: Monte Carlo samples used in this analysis. In this table σ_{prod} means the production cross section of the process, $N_{\text{generated}}$ the number of Monte Carlo events generated and L_{equiv} the equivalent integrated luminosity. This quantity equals the integrated luminosity required to produce on average $N_{\text{generated}}$ events for the process in question.

Trigger	p_{T}^{μ} threshold	Run Range
Run 2010A		
HLT_Mu9	9 GeV	132440 - 146239
Run 2010B		
HLT_Mu9	9 GeV	146240 - 147116
HLT_Mu11	11 GeV	147117 - 148068
HLT_Mu15_v1	15 GeV	148069 - 149442

Table 5.5: High Level Triggers used for the $\mu + \tau$ -jet selection. Due to the luminosity increasing over the year triggers with higher p_{T} threshold had to be used in later runs.

5.3.2 Selection Cuts

To separate genuine $\mu + \tau$ -jet events from the background processes as described in the previous section a set of selection cuts is applied to all events that include at least a reconstructed muon and a jet, called a $\mu + \tau$ -jet candidate. The selection is commissioned using $Z \rightarrow \tau\tau$ events since the Z boson is very well known and also decays into a τ pair. In the following $Z \rightarrow \tau\tau$ will be denoted as the *signal*. The event selection used in the $\mu + \tau$ -jet channel closely follows the official procedure described in [117].

First the event has to be selected by a single muon High Level Trigger. This trigger algorithm requires a muon with a transverse momentum above a certain threshold. This prevents selection of Monte Carlo events that would not have been recorded by the detector. Table 5.5 shows the triggers used for different run ranges. Next, the muon in the $\mu + \tau$ -jet candidate is required to be the one that activated the trigger (“Trigger matching”). This is relevant if there is more than one muon in the event. The efficiency of the trigger was determined with the so-called Tag and Probe method to be 0.9203 ± 0.0019

on data [118]. The efficiency of the trigger simulation is slightly higher so when comparing data events to Monte Carlo events on the percent level this effect needs to be accounted for.

For the muon so-called quality criteria concerning the reconstruction must be fulfilled to prevent selection of fakes or cosmic muons: the muon is required to be a *global* muon, i.e. it needs to have hits both in the inner silicon tracker and in the outer muon system which are compatible to each other. This is important for the purity of the sample and makes sure only muons with well-measured four-momentum are considered. Furthermore, there need to be more than ten hits in the tracker to make sure the momentum measurement is accurate since it is determined by the curvature of the muon trajectory in the tracker. Also, the χ^2 per degree of freedom of the track fit must be less than 10 [119].

For both the muon and the τ -jet a set of kinematic cuts is applied: the muon transverse momentum must be greater than 15 GeV and the jet transverse momentum must exceed 20 GeV. These cuts ensure that low-energy background activity is suppressed. This is essential at a hadron collider where there are high underlying event contributions in the low-energy region. Additionally, the pseudorapidity is constrained to $|\eta^\mu| < 2.1$ to be in full acceptance of the muon system and for the jet $|\eta^{\tau\text{-jet}}| < 2.3$ so that not only the leading track but also all constituents of the τ -jet are within the acceptance of the silicon tracker.

Next, the jet must have been identified as a τ -jet by the HPS algorithm. The methods for discriminating the τ -jet against a muon or an electron as described in Section 5.1.2 are applied to the τ -jet.

The distance between the muon and the τ -jet in η - ϕ space has to be larger than 0.5 to prevent selecting a muon which is part of a misidentified τ -jet. Also, since the τ pair originates from a neutral particle (either a Z or a Higgs boson) the sum of the charges of the muon and the τ -jet must equal 0.

At this point a solid set of events containing a high-quality muon and a τ -jet has been selected for which it is expected that Monte Carlo simulation to match data taken with the CMS detector. The events passing the cuts up to this point are said to pass the *preselection*.

Not only the shape of various distributions but also the total number of events can be predicted from simulation if the cross sections of all involved processes are known. The data used in this analysis is the full dataset that was taken in 2010 by CMS. It corresponds to an integrated luminosity of $\mathcal{L} = 36 \text{ pb}^{-1} \pm 4\%$. The technical details about the datasets used, both data and simulation, are available in Appendix B.

Figure 5.6a shows the distribution of the invariant mass of the muon and the τ -jet (visible mass). As can be seen the data agree well with the Monte Carlo simulation which is a good hint that all significant background contributions have been accounted for. However the $Z \rightarrow \tau\tau$ signal is still dominated by the background.

Background suppression. As can be seen in Figure 5.6a there are three major backgrounds to the $Z \rightarrow \tau\tau$ signal: the main backgrounds are quark and gluon QCD (gray),

5 Analysis of $\tau^+\tau^-$ Final States

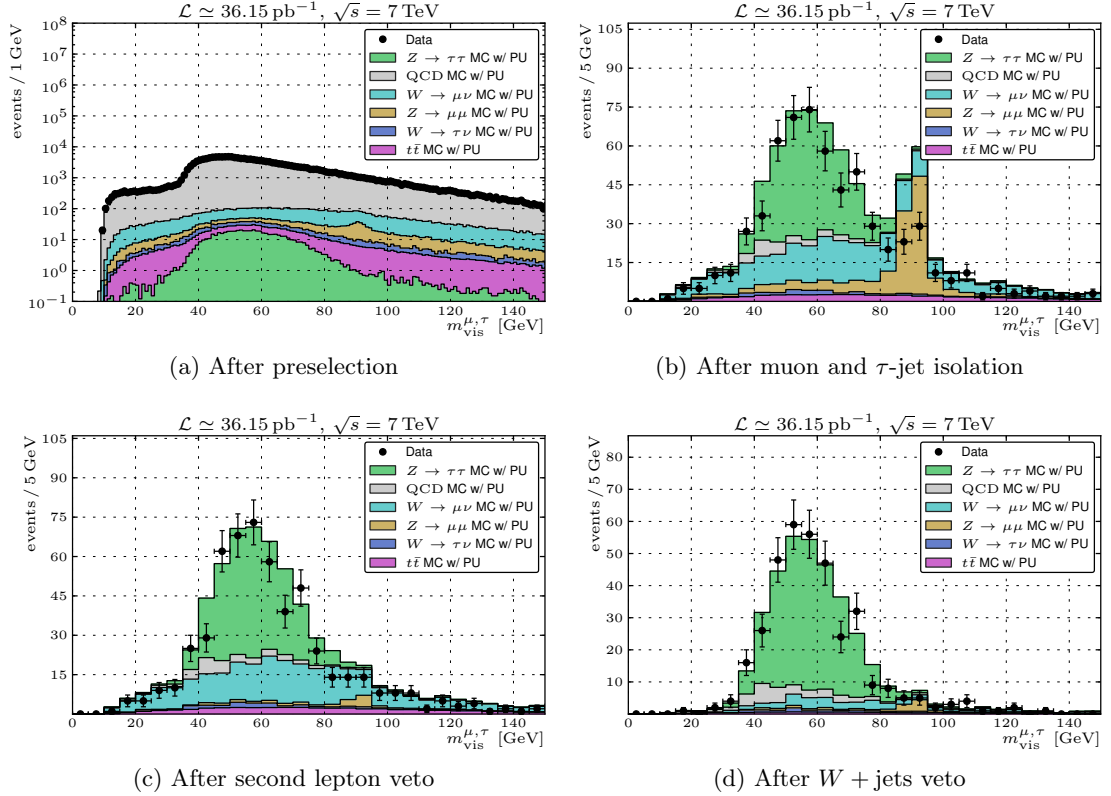


Figure 5.6: The visible mass distribution of $\mu + \tau$ -jet candidates in 2010 CMS data (black circles) and a POWHEG Monte Carlo sample (colored bars). The upper left plot shows the distribution after preselection as defined in this section. Application of the isolation criteria reduces contributions from QCD processes (upper right plot). Events containing a second muon were rejected in the lower left plot to discriminate against $Z \rightarrow \mu\mu$ background. Finally the lower right plot presents the final selection of $\mu + \tau$ -jet events with W + jets background reduced by a transverse mass cut.

W + jets (blue) and $Z \rightarrow \mu\mu$ (yellow). The three backgrounds are reduced individually one after the other:

- **Isolation.** The QCD background can be reduced by using an isolation criterion. This is done for both the muon and the τ -jet as described in sections 5.1.1 and 5.1.2. The isolation criteria for the muon and the τ -jet are slightly different however.

For the muon relative PARTICLE FLOW isolation is used. The quantity is given by

$$I_{\text{rel}}^{\text{PF}} = \frac{\sum p_{\text{T}}^{\text{charged}} + \sum E_{\text{T}}^{\text{neutral}} + \sum E_{\text{T}}^{\text{gamma}}}{p_{\text{T}}^{\mu}} \quad (5.8)$$

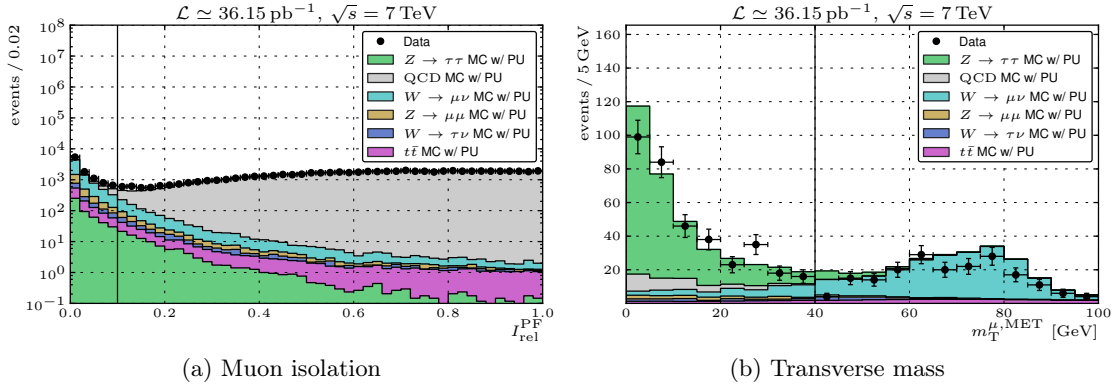


Figure 5.7: Plot (a) shows the muon isolation variable $I_{\text{rel}}^{\text{PF}}$ after the preselection. It is used to separate the signal from most of the QCD background. Plot (b) shows the transverse mass of the muon and E_T^{miss} after muon isolation and the second lepton veto. The $W + \text{jets}$ background peaks near the W mass whereas the signal tends to low values of M_T . The black lines on both plots indicate the selection cuts.

where the sums iterate over all PARTICLE FLOW charged hadron, neutral hadron or photon candidates, respectively, within a cone of $\Delta R = 0.4$ around the muon axis and with $p_T > 1 \text{ GeV}$. The $I_{\text{rel}}^{\text{PF}}$ distribution can be seen in Figure 5.7a. The cut is at $I_{\text{rel}}^{\text{PF}} > 0.1$ so that most QCD background is suppressed.

For hadronically decaying τ leptons the HPS loose isolation criteria as specified in [110] is applied. This requires no PARTICLE FLOW charged hadrons with $p_T > 1.0 \text{ GeV}$ and no PARTICLE FLOW photons with $E_T > 1.5 \text{ GeV}$ within an isolation cone of size $\Delta R = 0.5$.

Figure 5.6b shows the distribution of the visible mass after the isolation cuts, in linear scale. The $Z \rightarrow \tau\tau$ signal is now well visible already. The discrepancy between data and simulation around the Z boson mass indicates that the muon rejection for hadronically decaying τ leptons is more efficient on data than on simulation. However, this is not investigated further because the discrepancy vanishes when the $Z \rightarrow \mu\mu$ contamination is reduced in the next step.

- **Second lepton veto.** The significant contribution from events with $Z \rightarrow \mu\mu$ decays can be sufficiently suppressed by rejecting all events containing a second isolated muon with $p_T > 10 \text{ GeV}$ and opposite charge. The isolation requirement for the second muon is $I_{\text{rel}}^{\text{PF}} < 0.26$.

Figure 5.6c shows the distribution after this second lepton veto. The $Z \rightarrow \mu\mu$ contribution is removed without any significant impact on the signal.

- **Transverse mass cut.** The remaining background is $W + \text{jets}$ where the W decays into a muon. To suppress this background a new variable, the transverse mass M_T ,

5 Analysis of $\tau^+\tau^-$ Final States

Selection step	$Z \rightarrow \tau\tau$	$Z \rightarrow \mu\mu$	$W \rightarrow \mu\nu$	$W \rightarrow \tau\nu$	$t\bar{t}$	QCD	Data
Preselection	585	1247	5281	493	722	193658	199293
Isolation	276	139	235	17	53	35	616
2 nd lepton veto	275	19	235	17	51	35	552
Final selection	262	13	34	9	11	34	359

Table 5.6: Event yield after the various selection steps. It can be seen how each cut dramatically removes one particular background contribution as outlined in the text.

is introduced:

$$M_T = \sqrt{2p_T^\mu E_T^{\text{miss}} \cdot (1 - \cos \Delta\phi)} \quad (5.9)$$

where $\Delta\phi$ is the angle between the missing energy vector and the muon transverse momentum vector. The transverse mass can be regarded as the invariant mass of the muon and E_T^{miss} with the longitudinal component of the muon momentum ignored. Since in $W + \text{jets}$ events E_T^{miss} is an indicator for the neutrino momentum which results from the W decay M_T peaks near the W boson mass. This is illustrated in Figure 5.7b where it can also be seen that for $Z \rightarrow \tau\tau$ events M_T tends to lower values.

The cut is at $M_T < 40$ GeV. Figure 5.6d shows the invariant mass distribution after this cut. The $Z \rightarrow \tau\tau$ contribution dominates the distribution and all backgrounds are significantly reduced.

Distributions of more variables after the full selection are available in Appendix C.1. As shown in Table tab:analysis-event-yields, after the final selection there are 262 $Z \rightarrow \tau\tau$ events and 101 background events expected. 359 data events passing the selection are observed. These numbers could now be used for a $Z \rightarrow \tau\tau$ cross section measurement in proton-proton collisions at $\sqrt{s} = 7$ TeV. More work would be required to study selection efficiencies and to quantify systematic errors. This is for example performed in [117] but is out of scope for this thesis.

5.4 Update with 2011 Data

Until July 2011 CMS took about 1.1 fb^{-1} of integrated luminosity. This section presents a brief update of the analysis presented in the previous section on the full dataset. The same background processes as in the previous section are considered but the Monte Carlo samples used are from the SUMMER11 CMS Monte Carlo production. The primary difference to the previous Monte Carlo samples is different pile-up conditions as discussed below. The exact dataset names are given in Appendix B.

Trigger	Run Range	Int. Lumi. [pb ⁻¹]
Run 2011A		
HLT_IsoMu12_v1	160403 - 163261	44.4
HLT_IsoMu12_LooseIsoPFTau10_v4	163269 - 164236	156.9
HLT_IsoMu15_LooseIsoPFTau15_v*	165088 - 167913	887.2

Table 5.7: High Level Triggers used for the $\mu + \tau$ -jet selection in 2011. The increased luminosity with respect to the 2010 data taking requires more sophisticated triggers.

The selection steps of the analysis were slightly adapted to changed data taking conditions. In detail, the following

- **High Level Trigger.** Due to increased luminosity in 2011, a single muon trigger would lead to a too high data rate. Therefore, more sophisticated triggers are used in the analysis for 2011 data, summarized in Table 5.7. In the beginning of 2011 data taking, an isolated muon trigger was used. It is similar to the single muon trigger used in 2010 but it requires the muon to be isolated. For later runs, a cross trigger was used which requires both an isolated muon and an isolated PARTICLE FLOW τ -jet.
- **Pile-up Reweighting.** The higher luminosity comes with more collisions per bunch crossing. This leads to additional activity in an event which is commonly known as *pile-up*. Since the average number of collisions in an event was unknown when generating the Monte Carlo samples, a flat distribution of additional interactions up to 10 interactions was used for simulation. Above 10 interactions a Poissonian tail was applied. In order to take into account pile-up correctly, every Monte Carlo event is assigned a weight so that the distribution of reconstructed vertices matches between data and Monte Carlo simulation. Figure 5.8a shows the reweighted vertex distribution.
- **Pile-up Subtraction.** Pile-up activity contributes to the muon isolation variable if it happens to be around the muon in η - ϕ space. This additional activity is attempted to be subtracted from the isolation. The FastJet algorithm [120] allows to determine the area of a jet in η - ϕ space [53]. Summing up the areas of all jets in the event gives a measure of how much additional activity there is in the event. The density ρ , which is defined as the jet area of all jets divided by the total area, can be used to subtract the pile-up contribution to the isolation variable, assuming that the additional activity is evenly distributed over the whole event:

$$I_{\text{rel}}^{\text{PF}} = \frac{\sum p_{\text{T}}^{\text{charged}} + \sum E_{\text{T}}^{\text{neutral}} + \sum E_{\text{T}}^{\text{gamma}} - \rho \cdot \pi \cdot \Delta R^2}{p_{\text{T}}^{\mu}} \quad (5.10)$$

5 Analysis of $\tau^+\tau^-$ Final States

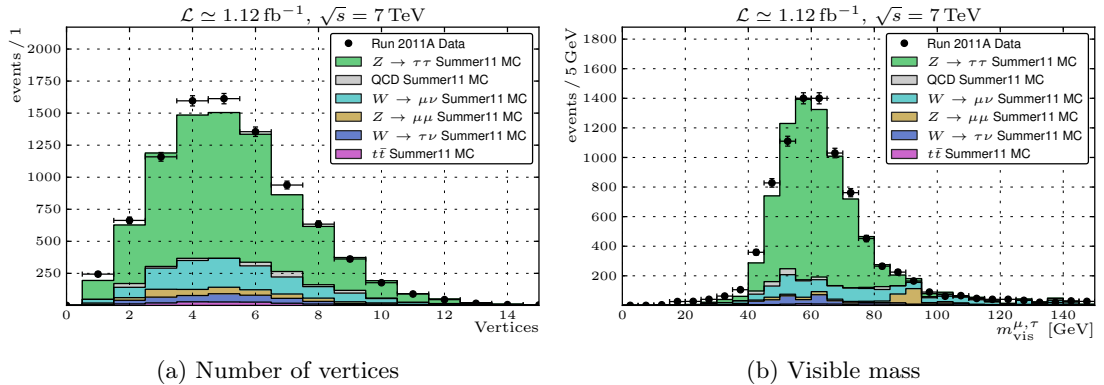


Figure 5.8: Number of reconstructed vertices (a) and visible mass (b) distributions after the $\mu + \tau$ -jet event selection on both 2010 and 2011 CMS data. The Monte Carlo events have been weighted so that the vertex distribution matches.

where ΔR is the radius of the isolation cone. This method is known as ρ -based pile-up subtraction.

- **Muon p_T cut.** The low-energetic region of $p_T^\mu < 20$ GeV is described poorly by simulation. Therefore, the cut was tightened correspondingly. This removes about 30 % of all events, however there are still enough events left in the 2011 dataset to obtain high statistical precision. Further study about this effect is in order at this point, however due to time constraints turned out to be impossible in the scope of this thesis.

Figure 5.8b shows the visible mass distribution for the full data samples with the changes mentioned above applied. The agreement between the Monte Carlo simulation and data is good, but it can be seen that systematic effects start to dominate the statistical uncertainties. Distributions of more variables are available in Appendix C.2.

5.5 Possible Improvements

Apart from additional systematic studies there is also other room for improvement. The analysis presented here is a simple cut-based procedure. The separation of signal and background can probably be improved by using multivariate analysis techniques such as likelihood ratios, boosted decision trees [121] or neural networks [112].

Another improvement that can be done to reduce systematic uncertainties is to rely less on Monte Carlo simulation for estimation of the number of background events. The detector simulation might not take effects into account that the real detector suffers from. These effects include detector (mis)alignment, calibration of energy measurement for various physics objects (electrons, missing transversum momentum, jets, etc.), underlying event and pile-up contributions.

A possibility how background contributions can be estimated from data is by inverting selection cuts. For example, requiring a non-isolated muon instead of an isolated muon selects nearly only QCD background. If the isolation variable is uncorrelated to another quantity the inverted selection gives the shape of the distribution of this quantity contributed by QCD processes. This can be used in a template fit to data to estimate the total number of QCD background events. The other background shapes in such a fit can be obtained similarly or they can be taken from Monte Carlo samples. Another way to estimate the total number of a certain background contribution is by choosing another variable which is uncorrelated to the first and invert the selection cut as well. For QCD background the opposite charge requirement is commonly used. This results in four regions: Events with opposite charge and an isolated muon (A), same charge and an isolated muon (B), opposite charge and a non-isolated muon (C) and same charge and a non-isolated muon (D). The QCD contribution to the signal region (A) can then be determined as

$$N_A = \frac{N_B \cdot N_C}{N_D}. \quad (5.11)$$

This method is known as the “ABCD” method.

Another method for background estimation from data for $Z \rightarrow \tau\tau$ background, the *embedding* technique, is presented in the next chapter.

5 Analysis of $\tau^+\tau^-$ Final States

6 The Embedding Technique

As mentioned in the beginning of the previous chapter one major motivation for studying $\tau^+\tau^-$ events is the search for the Higgs boson. The most prominent background for a $H \rightarrow \tau\tau$ signal is the $Z \rightarrow \tau\tau$ process. However, this background is mostly *irreducible* since it is the same decay. The only discriminating variable for a Higgs search is the mass of the resonance. There are other subtle differences, however they can only be exploited with high statistical precision far above the discovery threshold:

- Due to the different spin of the Higgs boson and the Z boson the angular distribution of the decay products will be different.
- Z bosons are usually boosted in a proton-proton-collider because they are produced by quark-antiquark annihilation with the quark carrying a much higher momentum fraction than the anti-quark on average. In contrast, most Higgs bosons will be observed in the central region since they are mainly produced via gluon fusion with the gluons having the same momentum distribution.

These properties will be verified later when sufficient statistical precision is available. In a discovery scenario, there are two criteria which directly affect the significance of the signal. The first is the quality of the mass reconstruction: The sharper the Higgs peak in the reconstructed mass spectrum, the easier it is to distinguish it from the $Z \rightarrow \tau\tau$ background. The SVfit algorithm presented in Section 5.2.3 is a promising method for mass reconstruction which allows to separate a possible Higgs signal from the Z peak.

The second criterion for signal significance is the understanding of the background. The lower the systematic (and statistical) uncertainties on the number of background events in the Higgs mass region the less likely it is that the signal originates from a fluctuation

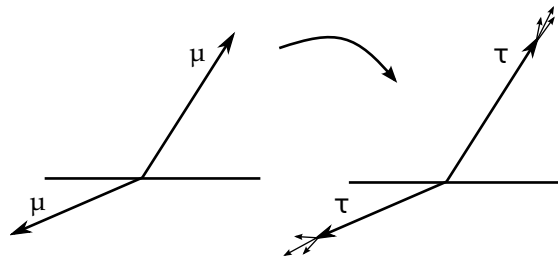


Figure 6.1: Principle of the embedding technique: In a measured $Z \rightarrow \mu\mu$ event the muons are removed and replaced by simulated τ leptons. The τ leptons are decayed by TAUOLA.

of the background and the more significant the signal is [122]. However, the $Z \rightarrow \tau\tau$ background can only be estimated from Monte Carlo simulation since a pure $Z \rightarrow \tau\tau$ sample cannot be obtained from data without significant contributions of background processes or Higgs boson decays. Such a Monte Carlo study includes uncertainties on the integrated luminosity, description of time-dependent pile-up and detector calibration. In the next section, a detailed discussion on the systematic error sources is given.

The *embedding* technique provides a data-driven way to estimate the number of $Z \rightarrow \tau^+\tau^-$ events. The idea is to select $Z \rightarrow \mu\mu$ events with little to no background contamination and then remove the muons from the event and replace them by simulated τ leptons. The procedure is illustrated in Figure 6.1. The hybrid events created this way have the exact same kinematic properties as regular $Z \rightarrow \tau\tau$ events because the coupling of the Z boson to muons is the same as its coupling to τ leptons. The only difference is the slightly smaller phase space in the $Z \rightarrow \tau\tau$ decay with respect to $Z \rightarrow \mu\mu$ because of the higher mass of the τ lepton. This effect accounts for 0.23 % more $Z \rightarrow \mu\mu$ decays than $Z \rightarrow \tau\tau$ ones. Since this number is precisely known from theory it can easily be corrected for, though, by weighting all embedded events accordingly.

6.1 Systematics Overview

The goal pursued with the embedding method is to reduce the systematic uncertainties on the estimated number of $Z \rightarrow \tau\tau$ events. This requires a good understanding of the systematic error sources involved. The following list briefly describes major systematic uncertainties and whether Monte Carlo samples, embedding samples or both are affected by them.

- **Luminosity.** The uncertainty on the integrated luminosity is estimated to be 4 % for 2010 [123] and 6 % for 2011. Every Monte Carlo sample needs to be normalized to the integrated luminosity of the data, and therefore the uncertainty on the integrated luminosity enters for all Monte Carlo studies. When comparing data to an embedded sample however there is no luminosity uncertainty involved because the $Z \rightarrow \mu\mu$ events are selected from the exact same data sample as the data events. The integrated luminosity is still not known exactly but it is guaranteed that the values for the embedded sample and the data sample are the same.

It should also be noted that when doing a cross section measurement the embedding method can be used to estimate the number N_{bkg} of background events with no uncertainty on the integrated luminosity. However, in the cross section equation,

$$\sigma = \frac{N_{\text{sig}} - N_{\text{bkg}}}{\epsilon \cdot \alpha \cdot \mathcal{L}}, \quad (6.1)$$

the integrated luminosity enters independently from that. Therefore, the method cannot be used to avoid the luminosity uncertainty in a cross section measurement.

- **Pile-up.** Even though pile-up effects, i.e. additional collisions in the same bunch crossing, are included in Monte Carlo simulations it is unclear to what extent the

simulation matches reality. It is known that by design it cannot be totally correct since pile-up is a time-dependent effect: at the beginning of a run when the beam intensities are high there will be more additional collisions than near the end of a run when the beam intensities have decreased. Pile-up has an influence on the isolation of muons and τ -jets when additional particles end up in the isolation cone. It also affects the missing transverse momentum. However, it is very hard to quantify the influence of this effect on the number of events passing the $\mu + \tau$ -jet selection.

Methods to deal with these difficulties exist. For instance, the number of reconstructed vertices in an event is a measure for pile-up activity. Therefore, a common approach is to reweight all Monte Carlo events so that the distribution of the number of reconstructed vertices in an event matches the distribution from data. A different approach uses the instantaneous luminosity in a bunch crossing and the total proton-proton inelastic cross section to estimate the number of pile-up events. Thus, all distributions are reweighted so that the number of pile-up events matches in data and simulation. However, it is still debatable whether all effects caused by pile-up are described correctly by such a reweighting procedure since this only corrects for pile-up on average and not on an event-by-event basis. Especially out-of-time pile-up, that is additional effects from previous bunch crossings (including but not limited to slow cooldown of calorimeter cells), is difficult to model correctly in Monte Carlo simulation and at the time of this study commissioning of out-of-time pile-up simulation is just about to start.

Since with the embedding technique all event content except the $Z \rightarrow \tau\tau$ decay is taken directly from data, pile-up effects are inherently both included and described correctly in an embedded sample. This is one of the major strengths of the embedding method.

- **Jet energy scale.** The measured energy of jets needs to be corrected for the mean of the jet energy distribution to be identical to the true jet energy. However, there are systematic uncertainties on these corrections which can lead to the mean being shifted against the true jet energy. The total uncertainty on the jet energy scale is 3% or lower for most of the transverse momentum and pseudorapidity regions covered by the $\mu + \tau$ -jet analysis [124].

Since non- τ jets are not directly involved in the $Z \rightarrow \tau\tau$ analysis, a slightly incorrect description of the jet energy scale is expected to be relatively small. However, it can affect the rate of QCD jets misidentified as τ jets (“fake rate”) and it has an influence on missing transverse momentum which is used for $W + \text{jets}$ rejection and for the more sophisticated mass reconstruction algorithms.

With embedded events, the jet spectrum is described correctly by design since as with pile-up all non- τ jets are taken directly from data. This means that if the measured jet energy differs from the real jet energy by a few percent and this leads to additional events to be dropped or kept during selection then it will be the exact same way in embedded events as in data events.

6 The Embedding Technique

- **τ -jet energy scale.** For jets from hadronic τ decays there is the same energy scale issue as with non- τ jets. However, since the τ leptons are simulated with Monte Carlo techniques in embedded events, this affects both Monte Carlo events and embedded events. The impact on the number of $Z \rightarrow \tau\tau$ events passing the $\mu + \tau$ -jet selection has been studied in [125] by varying the energy of τ - jets by 3 % and comparing how many events still pass the $\mu + \tau$ -jet selection. The event number differs at most by 4.6 % which is taken as a systematic uncertainty for this effect.
- **Trigger Systematics.** The efficiency of the single muon trigger can be measured on data. The method is described in [118] and the uncertainty on the efficiency was determined to be 0.2 %. Both Monte Carlo based studies and studies using embedding are affected by this systematic error. In embedding this uncertainty also enters in the selection of $Z \rightarrow \mu\mu$ events. However, since both muons can activate the trigger, the additional effect is very small.
- **τ identification efficiency.** The efficiency of hadronic τ identification is subject to an uncertainty of 7 % [117]. As with the τ -jet energy scale this affects both Monte Carlo simulation and embedding since the τ leptons in embedding are simulated.
In fact the τ identification efficiency is currently the largest systematic error for various τ -based analyses. Another use case of the embedding method, apart from predicting the $Z \rightarrow \tau\tau$ background for Higgs searches, is to reduce this uncertainty: given lepton universality, $N(Z \rightarrow \mu\mu) \approx N(Z \rightarrow \tau\tau)$, the ratio of observed number of embedded events to data events (in a mass region where the Higgs boson is already excluded) after correcting for all other systematic effects can be used to determine the hadronic τ identification efficiency by assuming that any remaining difference in the event numbers is due to the τ identification efficiency.
- **$Z \rightarrow \mu\mu$ selection impurities.** The $Z \rightarrow \mu\mu$ event selection is very pure, i.e. background contributions are low. The selection described in [118] has a background fraction of 0.43 ± 0.02 %. For embedding, there are a few differences, described in Section 6.2. Since the cut on the invariant di-muon mass is removed in the selection, the purity is expected to be slightly worse. However, Monte Carlo studies have shown that it is still smaller than 1 %.
- **Statistical uncertainty.** Strictly speaking the statistical uncertainty is not a systematic error source, however there is one important difference between Monte Carlo simulation and embedding: since embedding is a data driven method the statistical precision is limited by the amount of data taken by the detector. In simulation however, the number of events is virtually unlimited as long as there are enough computing resources available.

These systematic uncertainties are also summarized in Table 6.1. It can be seen that with the embedding method some of the large error sources such as luminosity or pile-up are traded for other, smaller ones. This cannot only be used to reduce the overall

Systematic error source	Monte Carlo	Embedding	Impact on event yield
Luminosity	☑	☐	4 %
Pile-up	☑	☐	not quantified
Jet energy scale	☑	☐	probably small
τ -jet energy scale	☑	☑	4.6 %
Trigger systematics	☑	☑	0.2 %
τ identification efficiency	☑	☑	7 %
$Z \rightarrow \mu\mu$ selection impurities	☐	☑	< 1 %
Statistical uncertainty	☐	☑	vanishes with more data

Table 6.1: List of systematic uncertainty sources for the prediction of the event yield in a typical $Z \rightarrow \tau\tau$ analysis. It is also indicated whether an analysis is affected by each source when it uses $Z \rightarrow \tau\tau$ events from Monte Carlo simulation or embedding, respectively.

systematic error of an analysis but it can also serve as a partially independent cross check.

6.2 $Z \rightarrow \mu\mu$ Selection

The first step in producing embedded events is to select a sample of $Z \rightarrow \mu\mu$ events for the particle replacement. The procedure is described in [118]: two isolated global muons are required both of which fulfill quality criteria equivalent to the muon in the $\mu + \tau$ -jet selection described in Section 5.3.2.

The only difference is that for embedding the invariant di-muon mass is not required to be between 60 GeV and 120 GeV. The reason for this is that the primary application of the embedding method is to predict the number of $Z \rightarrow \tau\tau$ background events for a Higgs search, replacing the need for a $Z \rightarrow \tau\tau$ Monte Carlo sample. However, the Higgs signal is expected to be on top of the tail of the Z peak, so it is essential to describe the tail of the mass distribution correctly also for masses well above 120 GeV.

The lower boundary has also been abandoned because this cut leads to distortion of other spectra. For example the distribution in Figure 6.2a shows the p_T of the muon in the $\mu + \tau$ -jet analysis on $Z \rightarrow \tau\tau$ Monte Carlo events (blue) and on embedding applied on $Z \rightarrow \mu\mu$ Monte Carlo events (red) where the invariant mass cut was applied. As can be seen especially in the ratio plot there is a deficiency of events in the low and high p_T regions. Figure 6.2b shows where this effect comes from in the $Z \rightarrow \mu\mu$ selection. The p_T distribution of the two muons is compared on Monte Carlo level for different steps of the selection: obviously the invariant mass cut introduces the distortion which then propagates to the $\mu + \tau$ -jet analysis after the embedding procedure.

6 The Embedding Technique

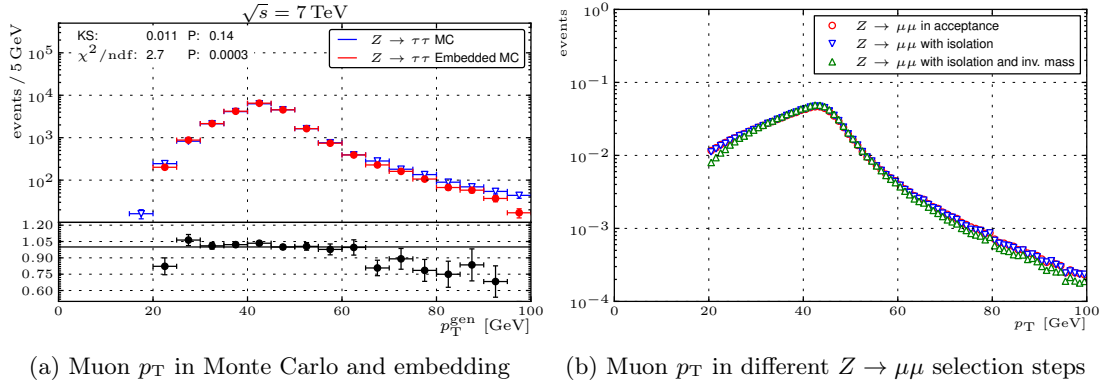


Figure 6.2: Plot (a) compares the p_T of the muon in the $\mu + \tau$ -jet final state for a $Z \rightarrow \tau\tau$ Monte Carlo sample and embedding applied to a $Z \rightarrow \mu\mu$ Monte Carlo sample on generator level. Plot (b) shows the muon p_T in different steps of the selection of $Z \rightarrow \mu\mu$ events. It can be seen that the invariant mass cut in the $Z \rightarrow \mu\mu$ selection introduces a bias into the p_T spectrum which propagates to the $\mu + \tau$ -jet selection after embedding.

6.3 Particle Replacement

The next step in the embedding process is to replace the two muons in a $Z \rightarrow \mu\mu$ event by simulated τ leptons. The lowest possible level where this can be achieved is the level of digitized detector output which consists of individual tracker hits and energy deposits in the calorimeters. At digitized output level the description of data from Monte Carlo simulation and from the detector is the same.

To remove the muons, all associated tracker hits and calorimeter entries have to be removed from the event. Then, a separate $Z \rightarrow \tau\tau$ event is generated with the PYTHIA Monte Carlo generator, the τ leptons are decayed with TAUOLA and the detector response is simulated. Finally, the tracker hits and calorimeter deposits are merged into the original event and the reconstruction algorithms, including E_T^{miss} calculation and Trigger algorithms, are run.

This way of producing the hybrid event is technically very challenging because it is difficult to locate the exact calorimeter hits which belong to the muon: if there are other particles near the muon their contribution must not be removed of course. This method has been implemented and verified on Monte Carlo [78]. However, applying it on data makes the merging of detector hits into the original event much more difficult because the alignment of the individual detector components are different in reality than in simulation: the tracking detector is not centrally aligned around the beampipe but it is slightly shifted [126]. The alignment differences are known and accounted for when reconstructing data events, but they cannot be simulated properly in the Monte Carlo detector simulation.

A different approach performs the embedding on the level of reconstructed particles,

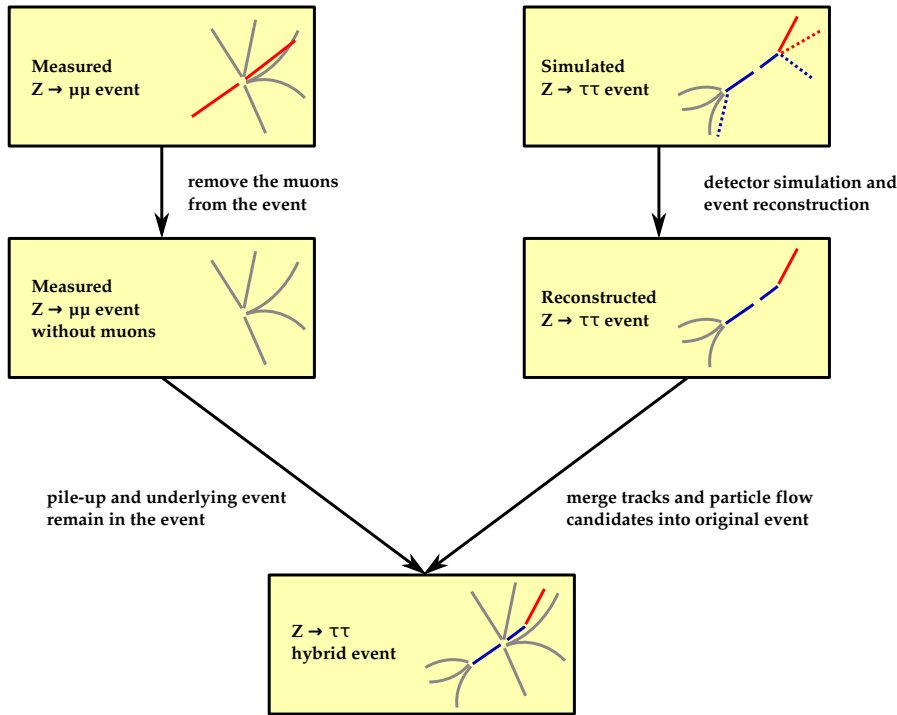


Figure 6.3: Embedding procedure on PARTICLE FLOW level: The full detector simulation and the reconstruction algorithms are run on the separate $Z \rightarrow \tau\tau$ event. The muons are removed from the original data event and then all tracks and PARTICLE FLOW candidates are merged into it.

namely PARTICLE FLOW candidates. This avoids the technical difficulties of the aforementioned procedure. Figure 6.3 visualizes the process: in a fully reconstructed event the PARTICLE FLOW muons are removed. As before a separate $Z \rightarrow \tau\tau$ event is generated and passed through the full detector simulation. This time, the PARTICLE FLOW algorithm is also applied to the separate event and then only the reconstructed tracks and PARTICLE FLOW candidates are merged into the original event. Eventually the PARTICLE FLOW algorithm is re-run on the hybrid event to identify individual PARTICLE FLOW objects from the candidates and to re-compute E_T^{miss} . The advantage of this method is that its application is independent from the actual detector geometry and possible on AOD data (see Section 3.4).

The simplifications of the embedding procedure on PARTICLE FLOW level comes at a small cost, however. Since the trigger decision is not based on PARTICLE FLOW the trigger simulation can only be performed on the separate $Z \rightarrow \tau\tau$ event. For triggers which require only a single muon, a single τ -jet or both a muon and a τ -jet this is not a problem because they are not heavily influenced by other event content. There are also triggers which require an isolated muon or τ -jet, though. Such triggers will have higher efficiencies on embedded events because there is no other event content in the

6 The Embedding Technique

separate $Z \rightarrow \tau\tau$ event which might lead to rejection of the event in data. This is not a showstopper, but care must be taken and possibly correction factors need to be applied.

Another caveat is that only PARTICLE FLOW objects can be used in analyses which use embedded samples. Calorimeter information is not merged into the original event. This is only a small limitation since PARTICLE FLOW already combines information from all subdetectors in an optimal way.

When replacing the muons by τ leptons in both methods the τ momenta have to be adapted for the higher τ rest mass. The goal is to produce an event which would have resulted if the Z decayed into two τ leptons instead of two muons. The four-vector of the two muons are added to reconstruct the four-vector of the Z boson. This can be used to boost the muon four-vectors into the Z boson rest frame. In the rest frame the three-momenta of the τ leptons are scaled so that the relation

$$E_\tau^2 - p_\tau^2 = m_\tau^2 \quad (6.2)$$

is fulfilled for each of them. Solving by p_τ gives

$$p_\tau = \sqrt{E_\tau^2 - m_\tau^2} = \sqrt{\left(\frac{1}{2}E\right)^2 - m_\tau^2} \quad (6.3)$$

where E is the energy of the Z in its rest frame. Since the Z can be off-shell E can differ from the Z rest mass.

6.4 Direct Normalization

Once an embedded $Z \rightarrow \tau\tau$ sample has been produced from a data sample with a certain integrated luminosity \mathcal{L} it can be used to predict the number of $Z \rightarrow \tau\tau$ events in the data sample. This can be done in two different ways.

The shape of the mass distribution in the embedded sample can be fitted to the data in a region where no Higgs signal is expected. The raising edge of the Z peak fulfills this requirement: in the visible mass distribution all bins up to 50 GeV can be used for the fit and in the SVfit mass distributions all bins up to 90 GeV. This method suffers from additional statistical uncertainty in the signal region: In addition to the normal uncertainty from limited statistics there is an extra uncertainty on the normalization constant which originates from the statistical uncertainty in the fitting region.

The second way for normalization involves studying all factors that introduce a difference between an embedded sample and a data sample and correcting for them. This requires more work than the first method because all effects that contribute to such a deviation must be understood. The method does not suffer from an additional statistical error in the signal region, however, the determination of the individual correction factors introduces systematic uncertainties.

It is expected that with sufficient number of data events available, the first method gives lower errors because statistical uncertainties can be reduced with increasing amount of data. However, with low statistical precision, such as with the 2010 CMS data, the

second method is more promising which is why it is studied in the following. In any case both methods can be used to cross check each other. Such a cross-check was also performed within this study. If both methods yield similar errors their results could even be combined to significantly improve the overall prediction.

In the following individual correction factors and efficiencies for prediction of the $Z \rightarrow \tau\tau$ event yield on the percent level are discussed.

6.4.1 $\tau\tau \rightarrow \mu + \tau$ -jet Branching Ratio

The two simulated τ leptons can be forced to decay into a particular final state before embedding. Since the analysis exploits the $\mu + \tau$ -jet final state, all τ pairs are forced to decay into this state to increase statistical precision. Therefore, a correction factor corresponding to the branching ratio of $\tau\tau \rightarrow \mu + \tau$ -jet must be applied when comparing embedded events to normal $Z \rightarrow \tau\tau$ events. The branching ratio for $\tau^\pm \rightarrow e^\pm$ and $\tau^\pm \rightarrow \mu^\pm$ are taken from the PDG [30]. For the combined branching ratio this gives

$$k_{\text{corr.}}^{\text{BR}} = 2 \cdot \text{BR}(\tau^\pm \rightarrow \mu^\pm) \cdot (1 - \text{BR}(\tau^\pm \rightarrow e^\pm) - \text{BR}(\tau^\pm \rightarrow \mu^\pm)) = 0.2250. \quad (6.4)$$

6.4.2 $Z \rightarrow \mu\mu$ Selection Efficiency

The $Z \rightarrow \mu\mu$ selection is not fully efficient within the kinematic acceptance of $p_T > 20$ GeV and $|\eta| < 2.1$. For example, the muon quality requirements can fail for a small set of true $Z \rightarrow \mu\mu$ events and some muons might not fulfill the isolation requirements.

However, in order to predict the number of $Z \rightarrow \tau\tau$ events from the measured number of $Z \rightarrow \mu\mu$ decays, the events which fail quality or isolation cuts need to be corrected for. The efficiency of the selection has been studied in [118]. Taking into account that only one of the two muons is required to activate the single muon trigger the fraction of $Z \rightarrow \mu\mu$ events observed in simulation is

$$\epsilon = (1 - (1 - \epsilon_{\text{HLT}})^2) \cdot \epsilon_{\text{iso}}^2 \cdot \epsilon_{\text{trk}}^2 \cdot \epsilon_{\text{sa}}^2 = 0.918 \pm 0.018. \quad (6.5)$$

This number is only valid for events with invariant di-muon mass between 60 GeV and 120 GeV. For an embedded sample this cut was abandoned (see Section 6.2), so the efficiency needs to be re-evaluated with respect to this.

This has been performed on a $Z \rightarrow \mu\mu$ POWHEG Monte Carlo sample, yielding

$$\epsilon_{\text{acc.}}^{\mu\mu} = 0.883 \pm 0.018 \quad (6.6)$$

where the error has been taken from Equation 6.5. On Monte Carlo level, a different systematic error was determined by computing the same number with a PYTHIA sample. The difference between this result, $\epsilon_{\text{acc.}}^{\mu\mu} = 0.885$, and the POWHEG number, is taken as a systematic error on Monte Carlo level.

Generator	All events	Inaccessible	Correction factor
POWHEG	10920	161	0.9855 ± 0.0012
PYTHIA	10341	154	0.9853 ± 0.0012

Table 6.2: The number of events not accessible to embedding because of restriction of the phase space in the $Z \rightarrow \mu\mu$ selection or because of $\tau\tau$ decay mode selection.

6.4.3 Phase Space Restriction

The selection of the $Z \rightarrow \mu\mu$ events imposes a restriction of the τ lepton phase space: since the muons are required to have $p_T > 20$ GeV and $|\eta| < 2.1$, the same will be the case for the embedded τ leptons. In the $\mu + \tau$ -jet analysis described in Section 5.3.2, only $|\eta| < 2.3$ must be fulfilled for the τ -jet so that τ -jets with $2.1 < |\eta| < 2.3$ can be observed in data and Monte Carlo events but not in embedded events. For both the muon and the τ -jet the cut is tightened to 2.0 to circumvent this problem and also to avoid edge effects: otherwise in data or simulation a τ -jet with η slightly greater than 2.1 could be reconstructed with η slightly less than 2.1 and thus pass the selection. This can happen because of finite detector resolution or the pseudorapidity of the visible decay products being different than the one of the τ lepton itself.

In principle the same problem exists with the muon p_T : in the $Z \rightarrow \mu\mu$ selection both muons are required to have $p_T > 20$ GeV, however, the p_T cut on the muon in the $\mu + \tau$ -jet analysis is at 15 GeV. In this case the cut is not tightened, though, because the muon carries away only a fraction of its mother particle's transverse momentum. The fraction of τ leptons with p_T between 15 GeV and 20 GeV but with the muon from the τ decay above 15 GeV is very small. The overall number of muons between 15 GeV and 20 GeV is relatively high so that when the cut were tightened the overall statistical precision would be reduced by about one third.

Furthermore, all generated τ pairs are forced to decay into a muon and a τ -jet to increase statistical precision for the $\mu + \tau$ -jet analysis. This way, contributions from other decay channels which are misidentified as $\mu + \tau$ -jet are not taken into account.

To account for these effects, two classes of events are defined on Monte Carlo level: The first event class contains all events with a muon transverse momentum above 20 GeV on generator level and which are genuine $\mu + \tau$ -jet decays. Such events are accessible to embedding studies. The second class contains events for which one of the two conditions is not fulfilled. These events exist in Monte Carlo or data samples, but not in an embedded sample; they are said to be inaccessible to embedding. After the full $\mu + \tau$ -jet selection, about 1.5 % of all events belong to the second class. Table 6.2 shows the exact numbers for PYTHIA and POWHEG Monte Carlo. The difference between the two is taken as the systematic error on the number and the statistical error is given by the Clopper-Pearson confidence interval [127].

Figure 6.4 shows important kinematic distributions of the events which are not accessible to the embedding procedure. In comparison with all events in the Monte Carlo

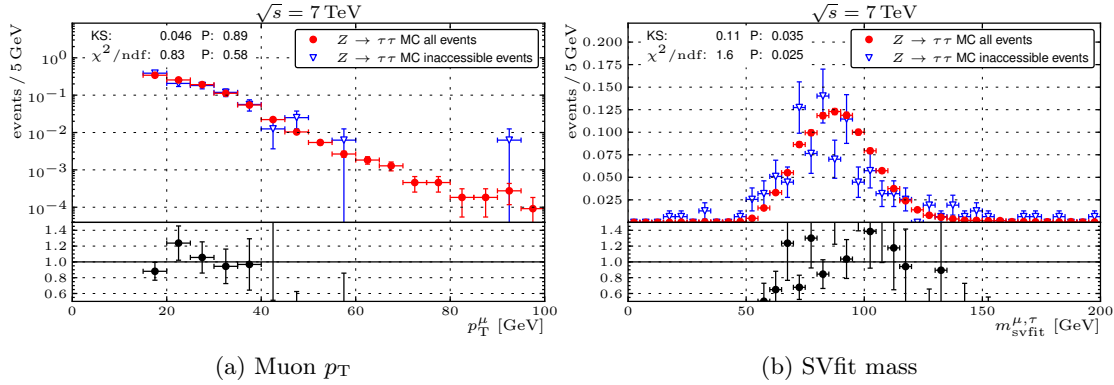


Figure 6.4: Comparison of the muon p_T distribution (a) and the SVfit di- τ mass (b) for all events (red) and for events that could not be generated by embedding (blue). Both distributions are scaled to unity. Within the limited statistics the distribution of inaccessible events is not distorted.

sample it can be seen that the shape of the distributions looks the same so that correcting for this effect with a simple scaling factor does not distort relevant distributions.

The final correction factor used is

$$k_{\text{corr.}}^{\text{phase space}} = 0.9855 \pm 0.0012 (\text{stat.}) \pm 0.0002 (\text{syst.}). \quad (6.7)$$

6.4.4 Muon Isolation Efficiency

There are two effects which can cause a difference in the muon isolation variable, $I_{\text{rel}}^{\text{PF}}$:

- The $Z \rightarrow \mu\mu$ selection requires both muons to be isolated. Even though the isolation criterion is weaker than the one used for the muon in the $\mu + \tau$ -jet analysis the fraction of isolated muons is higher in an embedded sample than in a regular data or Monte Carlo sample. Since the reduced number of $Z \rightarrow \mu\mu$ events due to the isolation requirement is already corrected for as described in Section 6.4.2 this effect leads to a higher number of $Z \rightarrow \tau\tau$ events in an embedded sample.
- In embedded events, the muons in the original $Z \rightarrow \mu\mu$ event can radiate photons due to bremsstrahlung or synchrotron radiation. This can lead to a distortion of the transverse momentum or invariant mass spectra. However, when the transverse momentum of the radiated photon is high enough (“hard photon”) then it will spoil the muon isolation quantity and therefore prevent the event from being used for embedding at all (see Section 6.2). Radiation of soft photons (low transverse momentum) does not alter the muon transverse momentum distribution significantly but can affect the $I_{\text{rel}}^{\text{PF}}$ distribution in the subsequent $\mu + \tau$ -jet analysis in such a way that muons will be less isolated, thus events are more likely to be dropped.

6 The Embedding Technique

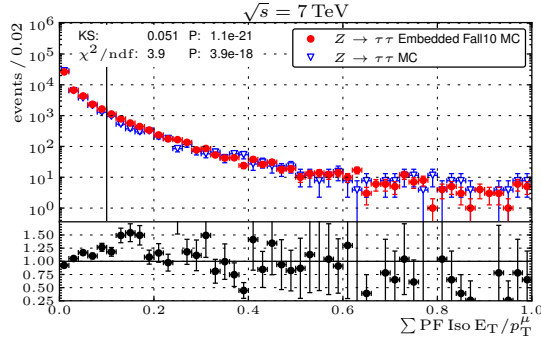


Figure 6.5: Comparison of the muon isolation variable for $Z \rightarrow \tau\tau$ (blue) Monte Carlo events and for $Z \rightarrow \mu\mu$ Monte Carlo events with embedding applied (red). Especially from the first few bins which contain many events it can be concluded that muons in embedded events tend to be less isolated. This can be explained by additional photons radiated by the original muons that have been replaced by τ leptons. The black line indicates the selection cut.

To quantify the overall impact of the two effects, the efficiency of the muon isolation criterion is studied as last step of the selection process in an embedded sample and in a regular $Z \rightarrow \tau\tau$ Monte Carlo sample. The muon isolation requirement contributes to the rejection of other final states faking a $\mu + \tau$ -jet event. Therefore, the phase space correction factor determined in the previous section is different for the stage of the analysis where the muon isolation cut has not been applied. In order to obtain the efficiency of the cut and to be able to compare it properly to simulation the raw event numbers are scaled with the respective phase space correction factor. For the selection where no muon isolation is applied it is determined the exactly same way as described in the previous section.

Figure 6.5 shows the distribution of $I_{\text{rel}}^{\text{PF}}$ on POWHEG $Z \rightarrow \tau\tau$ Monte Carlo events and on POWHEG $Z \rightarrow \mu\mu$ Monte Carlo events with embedding applied. The two effects described above explain why the two distributions do not agree well. Especially in the bins with low but nonzero isolation an excess in embedded events can be observed which leads to the conclusion that the efficiency of the cut is lower on embedded events than on regular Monte Carlo events. Table 6.3 shows the event yield before and after the isolation cut for POWHEG and PYTHIA Monte Carlo and embedded events.

To correct for this effect, another correction factor is introduced which is given by the ratio of the efficiencies between normal Monte Carlo events and embedded events. The systematic error on the correction factor again is given by the difference between POWHEG and PYTHIA. The final number is

$$k_{\text{corr.}}^{\mu \text{ isolation}} = 0.978 \pm 0.015 \text{ (syst.)} \quad (6.8)$$

where the systematic error is much larger than the statistical error so the latter is neglected. The reason for the significant difference between the POWHEG and PYTHIA

Data sample	Before isolation	With $I_{\text{rel}}^{\text{PF}} < 0.1$	Efficiency
$Z \rightarrow \tau\tau$ (POWHEG)	11903	10920	0.917
embedded $Z \rightarrow \tau\tau$ (POWHEG)	45923	41283	-
corrected for phase space	46723	41900	0.897
$Z \rightarrow \tau\tau$ (PYTHIA)	11439	10341	0.904
embedded $Z \rightarrow \tau\tau$ (PYTHIA)	12085	10883	-
corrected for phase space	12311	11047	0.897

Table 6.3: Number of events after all $\mu+\tau$ -jet selection steps before and after the isolation cut. The numbers are given for regular Monte Carlo events and embedded Monte Carlo events for both POWHEG and PYTHIA. The statistical precision of the PYTHIA embedded sample is reduced to about 1/4 of the original size however.

efficiencies still has to be investigated.

6.4.5 Trigger Efficiency on Data

The efficiency of the single muon High Level Trigger (HLT) is different on samples of simulated events and on data. Therefore additional correction factors need to be applied when embedding is used with data. The different efficiency enters both in the $Z \rightarrow \mu\mu$ selection and in the $Z \rightarrow \tau\tau \rightarrow \mu + \tau$ -jet selection.

The numbers for the trigger efficiency on both data and Monte Carlo level are known from [118].

For $Z \rightarrow \mu\mu$ only one of the two muons needs to fire the trigger so the correction factor will turn out rather small. The probability for a trigger with efficiency ϵ to *fail* identifying either muon is given by $(1 - \epsilon)^2$. The correction factor is defined by the ratio of the efficiencies on data and Monte Carlo simulation, so

$$k_{\text{corr.}}^{\mu\mu \text{ trig.}} = \frac{1 - (1 - \epsilon_{\text{MC}})^2}{1 - (1 - \epsilon_{\text{Data}})^2} = 1.004 \pm 0.0003 \quad (6.9)$$

where the error is obtained by Gaussian error propagation. The correction factor is greater than unity as in data fewer $Z \rightarrow \mu\mu$ events are detected than on Monte Carlo level and the goal is to estimate the full number of such events in order to apply the $N(Z \rightarrow \mu\mu) = N(Z \rightarrow \tau\tau)$ assumption.

In the $\mu+\tau$ -jet analysis, a correction is required since the trigger decision for the hybrid event is simulated during the embedding process. Since there is only a single muon to activate the trigger, the ratio of the trigger efficiencies directly yields the corresponding correction factor:

$$k_{\text{corr.}}^{\mu+\tau\text{-jet trig.}} = \frac{\epsilon_{\text{Data}}}{\epsilon_{\text{MC}}} = 0.9672 \pm 0.0020. \quad (6.10)$$

	Correction	Events	Stat. Uncert.
selected $Z \rightarrow \mu\mu$ events		431 387	656
selected $\mu + \tau$ -jet events after embedding		41 173	202
correct for $Z \rightarrow \mu\mu$ efficiency	$1/\epsilon_{\text{acc.}}^{\mu\mu}$	46 612	229
correct for $\tau\tau \rightarrow \mu + \tau$ -jet	$k_{\text{corr.}}^{\text{BR}}$	10 557	52
correct for non-reachable phase space	$1/k_{\text{corr.}}^{\text{phase space}}$	10 712	53
correct for different isolation efficiency	$1/k_{\text{corr.}}^{\mu \text{ isolation}}$	10 960	54
predicted number of $\mu + \tau$ -jet events		10 960	54
number of observed $\mu + \tau$ -jet events		10 920	104

Table 6.4: Application of the correction factors step by step to an embedded Monte Carlo sample. Within statistical errors the number of observed events can be predicted very accurately on the percent level.

Since the trigger efficiency is higher on Monte Carlo level than on data, the number of embedded events is corrected to lower values.

6.5 Closure Test

Before applying embedding to data, it is advisable to verify the method with samples of simulated events: given a $Z \rightarrow \mu\mu$ Monte Carlo sample with embedding and taking all the correction factors from the previous section into account it is expected that, within statistical and systematic errors, the number of $\mu + \tau$ -jet events in a $Z \rightarrow \tau\tau$ Monte Carlo sample can be predicted. This verification is done with POWHEG Monte Carlo samples where both samples are scaled so that they both correspond to the same equivalent integrated luminosity.

Table 6.4 shows the number of embedded events passing the $\mu + \tau$ -jet selection and the event number after applying each of the correction factors step by step. The trigger efficiency correction is not performed at this point since it is only relevant when using embedding with data. After the final correction the number agrees with the one obtained from $Z \rightarrow \tau\tau$ Monte Carlo simulation which is given in the last row for reference.

Figure 6.6 shows the visible mass and the SVfit mass distributions for both samples with direct normalization after applying all correction factors. It can be inferred that not only the total event count matches between Monte Carlo samples and embedded Monte Carlo samples, but also the shape of these two important distributions. Distributions of other quantities also agree nicely; they are shown in Appendix C.3.

The systematic error on the total correction factor was obtained by adding the systematic errors of all individual correction factors in quadrature. It is 1.7 %, mostly dominated by the systematic uncertainty on the muon isolation efficiency correction.

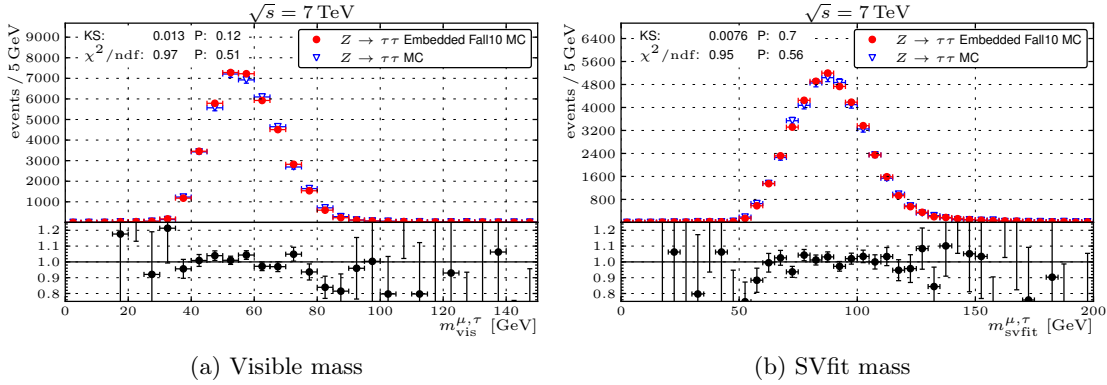


Figure 6.6: Comparison of the visible di- τ mass distribution (a) and the SVfit di- τ mass (b) for Monte Carlo events (blue) and embedded Monte Carlo events with all correction factors applied (red). Both distributions agree nicely in shape and in overall normalization, verifying the correctness of the method.

6.6 Application on Data

The same procedure is now applied on measured $Z \rightarrow \mu\mu$ events from data and compared to measured $Z \rightarrow \tau\tau \rightarrow \mu + \tau$ -jet candidate events. The data are taken from the 2010 CMS data taking period and correspond to 36 pb^{-1} .

In data, there are background contributions from non- $Z \rightarrow \tau\tau$ processes. These are estimated by scaling the corresponding simulated samples according to the integrated luminosity, the same way it was performed in Section 5.3.2. As a consequence, the systematic uncertainty on the luminosity applies to those background contributions. The dominating signal contribution from embedded $Z \rightarrow \tau\tau$ events does not suffer from this uncertainty as outlined before.

Table 6.5 presents the number of embedded data events after all correction factors have been applied and after the Monte Carlo expectation for background contributions has been added. The systematic uncertainties taken into account are summarized in Table 6.6 and added in quadrature. For the background Monte Carlo contribution only the luminosity error is used (4%). Summing up all the numbers gives for the expected number of $Z \rightarrow \tau\tau \rightarrow \mu + \tau$ -jet events in data

$$N_{\tau\tau \rightarrow \mu + \tau\text{-jet}}^{\text{exp.}} = 391 \pm 14 \text{ (stat.)} \pm 24 \text{ (syst.)} \pm 4 \text{ (lumi.)}. \quad (6.11)$$

This number agrees very well with the number actually observed in data,

$$N_{\tau\tau \rightarrow \mu + \tau\text{-jet}}^{\text{obs.}} = 359 \pm 19 \text{ (stat.)}. \quad (6.12)$$

As with the Monte Carlo closure test, Figure 6.7 shows the visible mass and SVfit mass distributions for data events and embedded data events. The latter includes the back-

	Correction	Events	Stat. Uncert.
selected $Z \rightarrow \mu\mu$ events		14 863	122
selected $\mu + \tau$ -jet events after embedding		1 081	33
correct for $Z \rightarrow \mu\mu$ efficiency	$1/\epsilon_{\text{acc.}}^{\mu\mu}$	1 224	37
correct for $\tau\tau \rightarrow \mu + \tau$ -jet	$k_{\text{corr.}}^{\text{BR}}$	277	8
correct for non-reachable phase space	$1/k_{\text{corr.}}^{\text{phase space}}$	281	9
correct for different isolation efficiency	$1/k_{\text{corr.}}^{\mu \text{ isolation}}$	288	9
correct for HLT inefficiency in $Z \rightarrow \mu\mu$	$k_{\text{corr.}}^{\mu\mu \text{ trig.}}$	289	9
correct for HLT inefficiency in $Z \rightarrow \tau\tau \rightarrow \mu + \tau$ -jet	$k_{\text{corr.}}^{\mu+\tau\text{-jet trig.}}$	280	9
predicted number of $\mu + \tau$ -jet events		280	9
Background expectation from Monte Carlo		111	11
number of observed $\mu + \tau$ -jet events		359	19

Table 6.5: Application of the correction factors step by step to an embedded data sample. The difference in the number of predicted and observed events can be attributed to statistical and systematic uncertainties (see text).

Effect	Systematic uncertainty
τ identification efficiency	7 %
τ -jet energy scale	4.6 %
$Z \rightarrow \mu\mu$ selection efficiency	1.8 %
Correction factors (mostly driven by $k_{\text{corr.}}^{\mu \text{ isolation}}$)	1.5 %

Table 6.6: Summary of the systematic error sources for the expected number of data events in the embedded sample of $Z \rightarrow \tau\tau$ events.

ground expectation from Monte Carlo simulation. In both cases, the shapes of the two distribution agree nicely. Again, additional distributions are available in Appendix C.4.

6.7 Normalization With a Fit to the Mass Distribution

In Section 6.4 it was mentioned that another method for normalizing the embedded sample is a fit to the non-signal region of the mass distribution. Such a fit was made both for the closure test and for application on data with the visible mass distribution. In a given bin i of the distribution two quantities are chosen as follows:

$$N_1^i = N_{\text{data}}^i \quad (6.13)$$

$$N_2^i = \alpha \cdot N_{\text{emb}}^i + N_{\text{bkg}}^i \quad (6.14)$$

6.7 Normalization With a Fit to the Mass Distribution

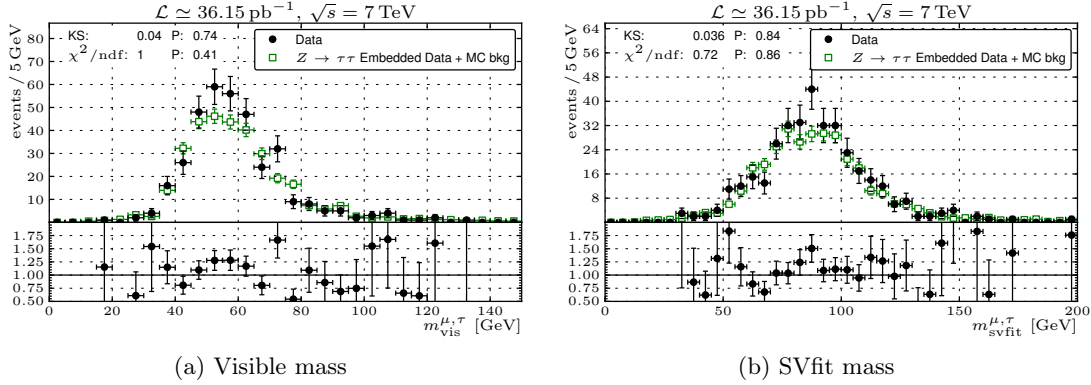


Figure 6.7: Comparison of the visible di- τ mass distribution (a) and the SVfit di- τ mass (b) for data events (black) and embedded data events (green) where the background expectation from Monte Carlo simulation has been included in the embedded events.

Source	Direct Normalization	Rising Edge Fit
Monte Carlo	0.2655 ± 0.005	0.2435 ± 0.015
Data	0.2585 ± 0.007	0.1681 ± 0.084

Table 6.7: Total correction factors with their uncertainties for both the direct normalization and the rising edge fit normalization. The low statistical precision leads to high uncertainties for the numbers obtained with the fit.

where N_1^i is simply the number of data events in that bin. N_2^i is the number of embedded events multiplied with a scaling factor α , plus the Monte Carlo expectation for background contributions. In simulation, N_1^i equals the normal $Z \rightarrow \tau\tau$ Monte Carlo sample and the background term of N_2^i is zero. The scaling factor α is now chosen so that χ^2 , which is defined as follows, becomes minimal:

$$\chi^2 = \sum_{i \in m_{\mu+\tau\text{-jet}} < 50 \text{ GeV}} \frac{(N_1^i - N_2^i)^2}{(N_{1,\text{err}}^i)^2 + (N_{2,\text{err}}^i)^2} \quad (6.15)$$

where $N_{1,\text{err}}^i$ and $N_{2,\text{err}}^i$ are the statistical errors of the two numbers.

Table 6.7 shows the fit results and compares them to the correction factors obtained with the direct normalization procedure. Gaussian error propagation of the uncertainties of the individual correction factors lead to uncertainty on the direct normalization numbers. The uncertainty for the fit results comes directly from the fit output.

Even for the Monte Carlo case where there are about 10,000 events passing the $\mu+\tau$ -jet selection the direct normalization still yields a more accurate result. The relatively high

error on the correction factor obtained with the fitting procedure on data is due to the low statistical precision of the 36 pb^{-1} sample. With increasing integrated luminosity it should get nearer to and eventually pass the number obtained with direct normalization, provided that the shape of the visible mass distribution agrees between data events and embedded events.

6.8 Conclusions and Outlook

In this chapter, the embedding method for estimating the $Z \rightarrow \tau\tau$ contribution has been introduced and demonstrated to work on 2010 CMS data in the $\mu + \tau$ -jet channel. However, the method as such can be used for all other channels the same way. For example it has been exploited in the $\tau\tau \rightarrow \mu\mu$ channel to estimate a systematic error on the selection efficiency [108].

The embedding method is not constrained to transform $Z \rightarrow \mu\mu$ events into $Z \rightarrow \tau\tau$ ones. Choosing $Z \rightarrow \mu\mu$ as source events proves useful because the signature is very clean and muons can be measured very well by the CMS detector. Besides a transformation into $Z \rightarrow \tau\tau$, it is also possible to create $W \rightarrow \tau\nu$ events. In this case the different vector boson mass, production rate and kinematic properties need to be corrected. One could also use $W \rightarrow \mu\nu$ events as source events, however in this case the neutrino energy is not known and the background contamination will be higher.

The method will benefit from verification on more data. In 2011 the LHC machine is generally better understood and much higher luminosity has been achieved. An integrated luminosity of more than 1 fb^{-1} was already delivered by mid-2011. Due to the higher event rates, more sophisticated triggers requiring both a muon and a τ -jet (so-called “cross-triggers”). Also, an isolation criterion can already be demanded on the trigger level. Such triggers cannot be used for embedding because there is no trigger simulation for cross triggers available at the time of this writing and because the trigger decision is made on the separate $Z \rightarrow \tau\tau$ events where both objects are intrinsically highly isolated, leading to a different trigger efficiency than on data. Instead, the trigger efficiencies must be studied and corrected for. An initial look into the early 2011 CMS data shows promising results, however many of the correction factors studied in Section 6.4 need to be re-evaluated: the $Z \rightarrow \mu\mu$ efficiency is likely to be different because of higher pile-up in an average event and because of a different trigger efficiency of the double muon trigger. With increasing statistical precision, the normalization method where a mass distribution of the embedded sample fitted to data in a non-signal region becomes superior to studying all correction factors individually. With the 2010 CMS data, however, direct normalization turned out to produce far more accurate results.

Summary and Conclusions

The Large Hadron Collider at CERN has opened up a new era in High Energy Physics. It allows to observe physics processes at unprecedented energies under laboratory conditions. Its main goal is the discovery of the Higgs boson which would complete the Standard Model of Particle Physics. However, also the exclusion of the Standard Model Higgs boson over the full mass range as well as the discovery of new particles or interactions are in reach of the LHC and might even be more interesting results.

Since the start of the physics program the operation of the LHC so far has exceeded all expectations. More than 1 fb^{-1} was delivered to the LHC experiments by mid-2011. The processing of this enormous amount of data as well as Monte Carlo sample production and physics analyses put a high load on the Worldwide LHC Computing Grid. In order for individual Grid sites to keep operating smoothly, sophisticated monitoring tools are installed to supervise the Grid site's components. The HAPPYFACE Project, a meta monitoring solution, allows to evaluate a center's overall status by looking at a single website, dramatically reducing the manpower needed for computing shifts and allowing non-experts to perform such shifts. Within the scope of this thesis, the HAPPYFACE core architecture was technically improved to further ease module development and many additional modules have been developed. As a result, several Grid sites consider deploying HAPPYFACE for their monitoring.

The $H \rightarrow \tau\tau$ channel is a dominant decay mode of the Higgs boson if its mass is low. Such a light Higgs boson is favored by the Standard Model. The $\tau\tau \rightarrow \mu + \tau\text{-jet} + \nu\nu\nu$ decay channel has a clean event signature and a high branching ratio. In 2010 and 2011 CMS data many $Z \rightarrow \tau\tau$ decays were observed in this channel and used for commissioning the Higgs search. However, $Z \rightarrow \tau\tau$ is also the largest background contribution to a possible Higgs signal with the mass of the resonance being the only feasible discriminant. In order to obtain a high statistical significance of a potential signal, a good di- τ mass reconstruction and low uncertainties on the background are essential. While the "SVfit" method provides an excellent mass reconstruction with a much improved resolution compared to previous methods, the estimation of $Z \rightarrow \tau\tau$ background contributions was a major topic of this thesis.

Background contributions can be estimated from Monte Carlo studies, however this introduces large systematic uncertainties since the exact pile-up conditions are not known and since certain processes such as hadronization and parton showers depend on heuristic models. To reduce such systematic effects, methods were developed to estimate background contributions from data. The embedding technique is such a method that can be applied for $Z \rightarrow \tau\tau$ events. The basic idea is that, due to lepton universality, the $Z \rightarrow \mu\mu$ decay behaves exactly the same as the $Z \rightarrow \tau\tau$ decay, both in kinematical properties and

Summary and Conclusions

in absolute frequency. However, $Z \rightarrow \mu\mu$ events can be selected with very high purity and efficiency from data. In the embedding method, the muons from such events are removed and replaced by simulated τ leptons. All other event content, especially the soft contributions that are hard to simulate, remain in the original event. Such hybrid events not only allow the prediction of the shape of various distributions such as the visible mass or the SVfit mass, but also the total number of $Z \rightarrow \tau\tau$ events to expect within a certain data sample can be estimated. This requires studying many systematic effects that introduce differences between the actual data sample and the embedded sample, such as the efficiency of the muon selection or photon radiation by the removed muons.

The feasibility of this absolute normalization of an embedded sample was presented in this thesis for the first time. It was shown that an accuracy on the percent level can be achieved in simulation, and perfect agreement within statistical errors was observed on 2010 data. Similar studies for the much larger 2011 data sample are ongoing.

The embedding technique is a significant contribution to the Higgs search in CMS. It enhances the statistical significance of a potential discovery, improves exclusion limits and allows to reduce correlations between different decay modes in a CMS-wide combination.

A Additional HappyFace Modules

A.1 List of HappyFace Modules

The following is a complete list of all modules available in the HAPPYFACE core distribution. Some of them are discussed in detail in Section 4.3.6 while others are described briefly in this appendix.

Name	Short Description	Page
<code>CMSFileConsistencyCheck</code>	Compares file size on the SE vs. DBS or PHEDEX	-
<code>CMSPhedexAgents</code>	Monitors whether PHEDEX agents are up and running	104
<code>CMSPhedexBlockReplicas</code>	Shows incomplete PHEDEX dataset blocks	-
<code>CMSPhedexErrorLog</code>	Report of failed PHEDEX transfers	105
<code>CMSPhedexLinks</code>	Status of PHEDEX transfer links between sites	-
<code>CMSPhedexPhysicsGroups</code>	Displays space usage of physics groups at a T2	-
<code>CMSsiteReadiness</code>	Reads and displays the CMS site readiness status	-
<code>DashboardDatasetUsage</code>	Shows the usage of the datasets at a site	106
<code>dCacheDataManagement</code>	Matches DBS information with DCACHE	59
<code>dCacheDatasetRestoreLazy</code>	Monitoring of DCACHE staging requests	107
<code>dCacheInfoPool</code>	Monitoring of DCACHE pools	108
<code>dCacheTransfers</code>	Monitoring of transfers between DCACHE pools	109
<code>JobsDist</code>	Shows distribution of jobs at the site	110
<code>JobsEfficiencyPlot</code>	Shows efficiency of jobs for each user at the site	111
<code>JobsStatistics</code>	Shows statistics of jobs at the site	58
<code>PhedexStats</code>	Shows statistics of all PHEDEX transfers to/from a site	-
<code>PhpPlotCMSPhedex</code>	Shows plots of PHEDEX transfer rates	-
<code>PhpPlotDashboardJobSummary</code>	Shows summary plot of all jobs at the site	-
<code>PhpPlotDashboard</code>	Shows plots from DASHBOARD for the site	-
<code>RSSFeed</code>	Shows an RSS feed on the HAPPYFACE website	60
<code>SAM</code>	Shows status of SAM tests.	112
<code>Summary</code>	Shows a summary of multiple HAPPYFACE instances	61

Table A.1: Available HAPPYFACE modules with a short description. Some modules are described in more detail on the given pages.

A.2 Module Descriptions

A.2.1 CMSPhedexAgents



Phedex Agents (Prod_KIT)

Fri, 29. Jul 2011, 17:16 - [Show module information](#)

agent	label	last report	critical
FileDownload	mss-migrate	00d:00h:06m	yes
BlockDownloadVerifyInjector	mgmt-blockverifyinjector	00d:00h:49m	no
BlockDownloadVerify	blockverify	00d:00h:02m	yes
FileIssue	mgmt-issue	00d:00h:00m	no
FileExport	exp-pfn	00d:00h:34m	yes
FileRouter	mgmt-router	00d:00h:06m	no
FileRemove	download-remove	00d:00h:02m	yes
FilePump	mgmt-pump	00d:00h:02m	no

details

agent	label	host	directory	version
FileDownload	mss-migrate	cms-kit.gridka.de	/home/cmssgm/phedex /instance/Prod_KIT /state	PHEDEX_4_0_0
BlockDownloadVerifyInjector	mgmt-blockverifyinjector	vocms02.cern.ch	/data/ProdNodes /Prod_Mgmt/state	PHEDEX_4_0_0
BlockDownloadVerify	blockverify	cms-kit.gridka.de	/home/cmssgm/phedex /instance/Prod_KIT /state	PHEDEX_4_0_0
FileIssue	mgmt-issue	vocms02.cern.ch	/data/ProdNodes /Prod_Mgmt/state	PHEDEX_4_0_0
FileExport	exp-pfn	cms-kit.gridka.de	/home/cmssgm/phedex /instance/Prod_KIT /state	PHEDEX_4_0_0
FileRouter	mgmt-router	vocms02.cern.ch	/data/ProdNodes /Prod_Mgmt/state	PHEDEX_4_0_0
FileRemove	download-remove	cms-kit.gridka.de	/home/cmssgm/phedex /instance/Prod_KIT /state	PHEDEX_4_0_0
FilePump	mgmt-pump	vocms02.cern.ch	/data/ProdNodes /Prod_Mgmt/state	PHEDEX_4_0_0

Figure A.1: The “CMSPhedexAgents” module shows the execution time and result of the latest PHEDEX agent runs. The PHEDEX agents make sure that various operations required for PHEDEX transfers to and from the Grid site work correctly.

A.2.2 CMSPhedexErrorLog

**Transfer Errors from T1_DE_KIT_Buffer (debug)**Fri, 29. Jul 2011, 15:46 - [Show module information](#)

failed transfers	3
failed transfers details	
failed transfers due to destination	2
failed transfers due to source	0
failed transfers due to transfer	1
failed transfers due to unknown reasons	0
fraction of destination errors	67%
fraction of source errors	0%
fraction of transfer errors	33%

details		
node	failed transfers	error content
T2_UK_London_IC	1	TRANSFER error during TRANSFER phase: [NO_PROGRESS] No markers indicating progress received for more than 300 seconds
T2_PT_LIP_Lisbon	2	DESTINATION error during TRANSFER_PREPARATION phase: [INVALID_PATH] Invalid SURL specified!

Figure A.2: The “CMSPhedexErrorLog” module shows failed PHEDEX transfers to a from the Grid site. In this case, outgoing transfers from T1_DE_KIT are shown. There are three failed transfers, however neither of the failures are due to problems at the transfer source. From the KIT point of view the transfers work correctly. Therefore, the module does not report a problem. The details table contains more specific information about the failed transfers.

A.2.3 DashboardDatasetUsage

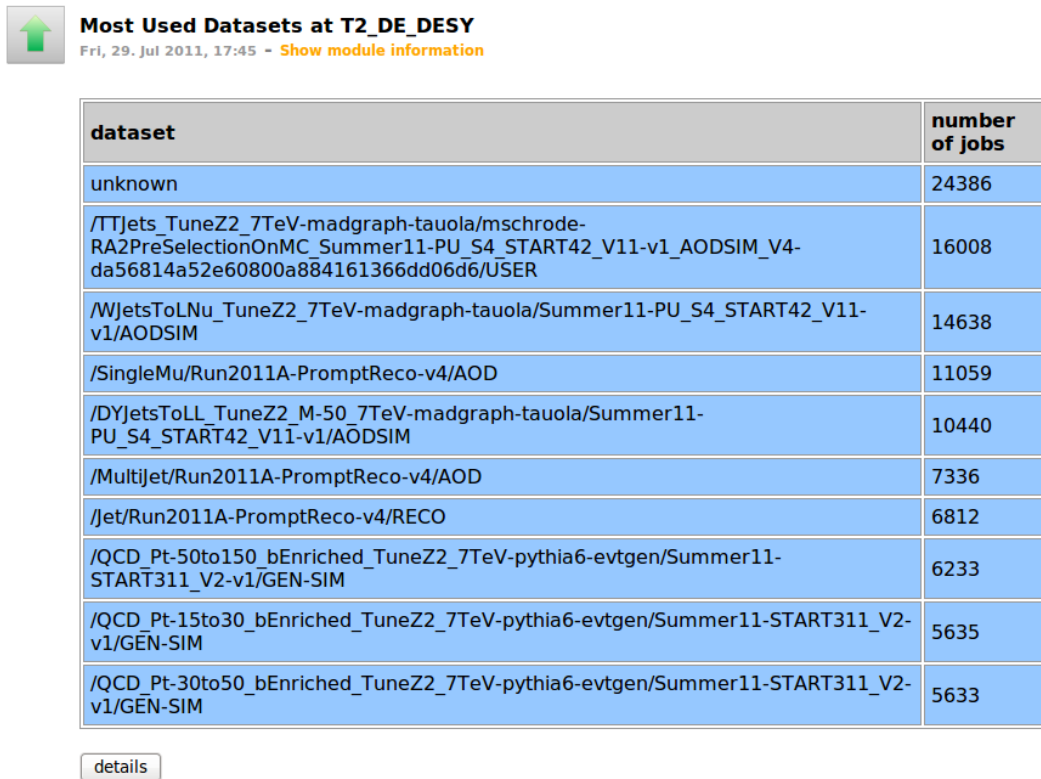


Figure A.3: The “DashboardDatasetUsage” module shows the most often requested datasets at a center. The module does not do any rating yet but its output can be useful to verify that the access to these datasets is smooth (for example by creating replicas in DCACHE).

A.2.4 dCacheDatasetRestoreLazy

**dCache Dataset Restore Monitor (Lazy)**Fri, 29. Jul 2011, 15:46 - [Show module information](#)

Total number of stage requests	29
... with status Pool2Pool:	2
... with status Staging:	27
Stage request with problems	0
... with status Waiting:	0
... with status Suspended:	0
... with status Unknown:	0
... time limit hit (24:00:00)	0
... retry limit hit (2)	0

Figure A.4: The “dCacheDatasetRestoreLazy” module shows staging requests of a DCACHE instance. Staging means transferring a file from tape storage to disk storage. The module reports if there are staging requests that take very long (indicating that they are stuck) or requests that did not succeed after several retries.

A.2.5 dCacheInfoPool

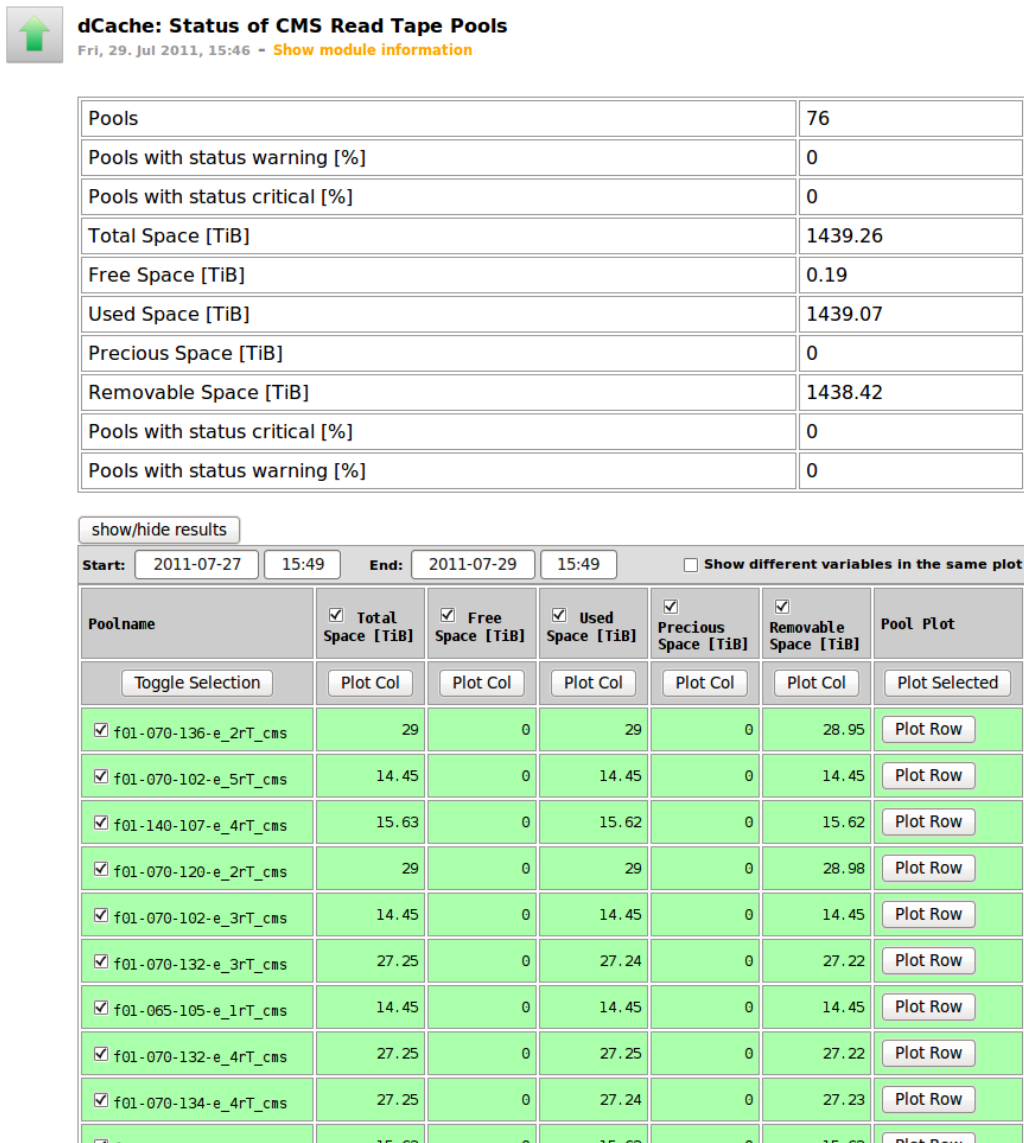


Figure A.5: The “dCacheInfoPool” module shows the status of a group of DCACHE pools. It shows occupied, available and total space of the whole group and for the individual pools. This allows to verify that the workload management between the pools is working correctly. The module reports a critical status when the free space of the pool group falls below a specified threshold.

A.2.6 dCacheTransfers

**dCache Transfers**Sat, 23. Jul 2011, 16:01 - [Show module information](#)

Total number of transfers	54
Speed average [KiB/s]	218
Standard deviation of speed distribution [KiB/s]	61
Number of transfers with warnings	46
... due to time limit	0
... due to speed limit	46
Number of critical transfers	1
... due to time limit	0
... due to speed limit	1

show/hide results

pnfsID	Pool	Host	Status	Time	Trans. [GiB]	Speed [KiB/s]
0000934F16E3869543A58274858E83125ADA	f01-070-118-e_4rT_cms	c01-028-117.gridka.de	WaitingForDoorTransferOk	0 d 02:06:13	0.9	121
00009049EB456BB049388EB196509A00B05E	f01-070-120-e_4rT_cms	c01-028-117.gridka.de	WaitingForDoorTransferOk	0 d 00:26:30	0.2	152
00000B19EFBC2C0745C0A25173EF871CAEC5	f01-070-120-e_2rT_cms	c01-009-106.gridka.de	WaitingForDoorTransferOk	0 d 02:22:37	1.3	162
0000BE91E07D963A4B809CA06DD4144FD5C1	f01-070-118-e_4rT_cms	c01-001-123.gridka.de	WaitingForDoorTransferOk	0 d 00:45:09	0.4	166
00009552E0D47A0B496482E3F594D88E72C9	f01-070-132-e_4rT_cms	c01-009-104.gridka.de	WaitingForDoorTransferOk	0 d 00:20:29	0.2	169
00007A4369F34CF1459F90987F2BE210CF68	f01-070-136-e_4rT_cms	c01-013-126.gridka.de	WaitingForDoorTransferOk	0 d 01:35:02	0.9	169
0000065B4E159BB147F3B6BF4AE66A6D42C8	f01-070-119-e_2rT_cms	c01-001-124.gridka.de	WaitingForDoorTransferOk	0 d 00:45:09	0.4	171
000053B8D2A4F4D7438F876F09FEB1353BCC	f01-070-132-e_3rT_cms	c01-013-120.gridka.de	WaitingForDoorTransferOk	0 d 00:51:26	0.5	173
0000E8E6FA0D2E784835B1A75910928A6A39	f01-070-132-e_4rT_cms	c01-009-106.gridka.de	WaitingForDoorTransferOk	0 d 00:36:18	0.4	173
000016159C9188C5453791F6067376B21800	f01-070-120-	c01-016-179.gridka.de	WaitingForDoorTransferOk	0 d	0.5	174

Figure A.6: The “dCacheTransfers” module shows the status of transfers between different DCACHE pools. This includes transfers of incoming files from tape write pools to tape read pools to make them available to the outside but also transfers between tape read pools (so-called replicas). The module issues a warning if transfers take very long or are unusually slow (both are indications for hanging transfers).

A.2.7 JobsDist



Figure A.7: The “JobsDist” module shows the distribution of a certain variable (in this case the walltime) for all jobs running at the center. Other variables that can be plotted include job efficiency and CPU time. The input file of this module is the same as the one for the “JobsStatistics” module.

A.2.8 JobsEfficiencyPlot

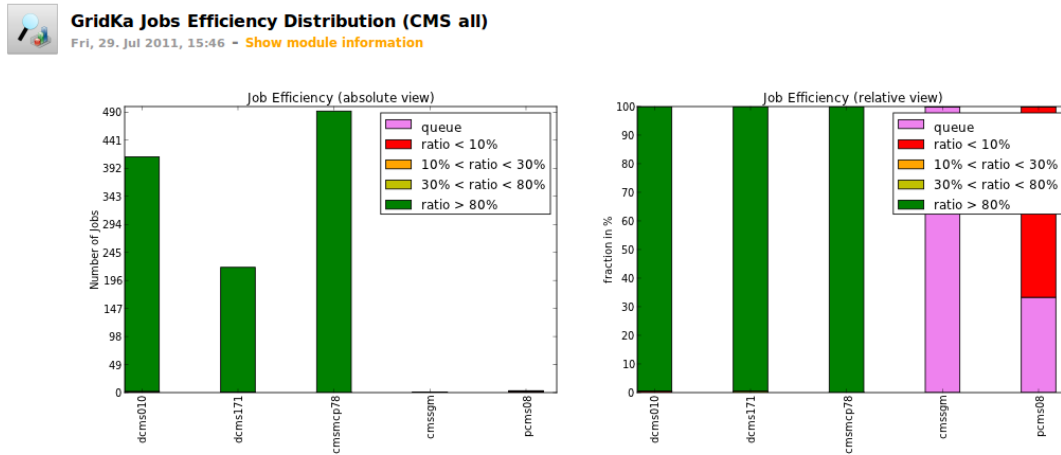


Figure A.8: The “JobsEfficiencyPlot” module is a plot module (no rating) showing a stacked bar graph of the number of jobs for different users. The fraction of different colors in the bars correspond to different job efficiencies. This plot allows to easily attribute inefficient jobs to a particular user. The input file of this module is the same as the one for the “JobsStatistics” module.

A.2.9 SAM



GridKa SAM OPS Table

Fri, 29. Jul 2011, 15:46 - [Show module information](#)

Group status:

ComputingElements	ce-2-fzk.gridka.de, ce-3-fzk.gridka.de, ce-4-fzk.gridka.de
CREAM ComputingElements	cream-1-fzk.gridka.de, cream-2-fzk.gridka.de, cream-3-fzk.gridka.de, cream-4-kit.gridka.de, cream-5-kit.gridka.de
StorageElements	cmssrm-fzk.gridka.de

Individual service status:

CE	ce-2-fzk.gridka.de
CE	ce-3-fzk.gridka.de
CE	ce-4-fzk.gridka.de
SRMv2	cmssrm-fzk.gridka.de
CREAMCE	cream-1-fzk.gridka.de
CREAMCE	cream-2-fzk.gridka.de
CREAMCE	cream-3-fzk.gridka.de
CREAMCE	cream-4-kit.gridka.de
CREAMCE	cream-5-kit.gridka.de
SRMv2	dgridsrm-fzk.gridka.de

error/warning results

Element Type	Element Name	Status	Test Name	Test Time
SRMv2	dgridsrm-fzk.gridka.de	warn (0.5)	SRMv2-org.cms.SRM-V0Del	2011-07-29 13:34:52
SRMv2	dgridsrm-fzk.gridka.de	na (0.5)	SRMv2-org.cms.SRM-GetPFNFromTFC	2011-07-29 13:34:52
SRMv2	dgridsrm-fzk.gridka.de	warn (0.5)	SRMv2-org.cms.SRM-V0Put	2011-07-29 13:34:52
SRMv2	dgridsrm-fzk.gridka.de	warn (0.5)	SRMv2-org.cms.SRM-V0Get	2011-07-29 13:34:52

successful results

Figure A.9: The “SAM” (Site Availability Monitoring) module visualizes the output of SAM tests. SAM tests are special jobs that are sent to all CEs of a Grid site. These jobs verify that basic functionality such as copying files to the site’s SE work properly and that certificates have not expired.

B Datasets Used

This section lists the datasets used for the physics analyses in this thesis.

B.1 Simulation

Events	Cross section [pb]	Equivalent integrated luminosity [pb ⁻¹]
FALL10 Production .../Fall10-E7TeV_ProbDist_2010Data_BX156_START38_V12-v1/GEN-SIM-RECO		
/DYToTauTau_M-20_CT10_TuneZ2_7TeV-powheg-pythia-tauola/...		
1,994,719	1,666	1,197
/DYToMuMu_M-20_CT10_TuneZ2_7TeV-powheg-pythia/...		
1,998,931	1,666	1,200
/WPlusToMuNu_CT10_TuneZ2_7TeV-powheg-pythia/...		
1,997,318	6,152	325
/WMinusToMuNu_CT10_TuneZ2_7TeV-powheg-pythia/...		
1,996,548	4,286	466
/WPlusToTauNu_CT10_TuneZ2_7TeV-powheg-pythia-tauola/...		
1,995,871	6,152	324
/WMinusToTauNu_CT10_TuneZ2_7TeV-powheg-pythia-tauola/...		
1,994,870	4,286	465
/TTTo2L2Nu2B_7TeV-powheg-pythia6/...		
996,022	65.83	15,130
/QCD_Pt-20_MuEnrichedPt-15_TuneZ2_7TeV-pythia6/...		
28,315,088	84,679	334
SUMMER11 Production .../Summer11-PU_S3_START42_V11-v2/AODSIM		
/DYToTauTau_M-20_TuneZ2_7TeV-pythia6-tauola/...		
2,032,536	1,666	1,220
/DYToMuMu_M-20_TuneZ2_7TeV-pythia6/...		
2,148,325	1,666	1,290
/WToMuNu_TuneZ2_7TeV-pythia6/...		
5,413,258	10,438	519
/WToTauNu_TuneZ2_7TeV-pythia6-tauola/...		
5,500,000	10,438	527
/TT_TuneZ2_7TeV-pythia6-tauola/...		
1,089,625	157.5	6,918
/QCD_Pt-20_MuEnrichedPt-10_TuneZ2_7TeV-pythia6/...		
8,797,418	75,300	117

Table B.1: Datasets used for simulated events.

B.2 Data

Certified events	Run range	Integrated luminosity [pb^{-1}]
2010 Dataset		
/Mu/Run2010A-Nov4ReReco_v1/RECO 20,868,540	132440 - 146239	3.1
/Mu/Run2010B-Nov4ReReco_v1/RECO 28,195,692	146240 - 149442	32.9
2011 Dataset		
/SingleMu/Run2011A-May10ReReco-v1/AOD 16,990,276	160403 - 163261	44.4
/TauPlusX/Run2011A-May10ReReco-v1/AOD 11,395,477	163269 - 164236	156
/TauPlusX/Run2011A-PromptReco-v4/AOD 15,383,324	165088 - 167913	887

Table B.2: Datasets used for measured events.

C Additional Plots

C Additional Plots

C.1 $\tau^+\tau^-$ Final States in 2010 CMS Data

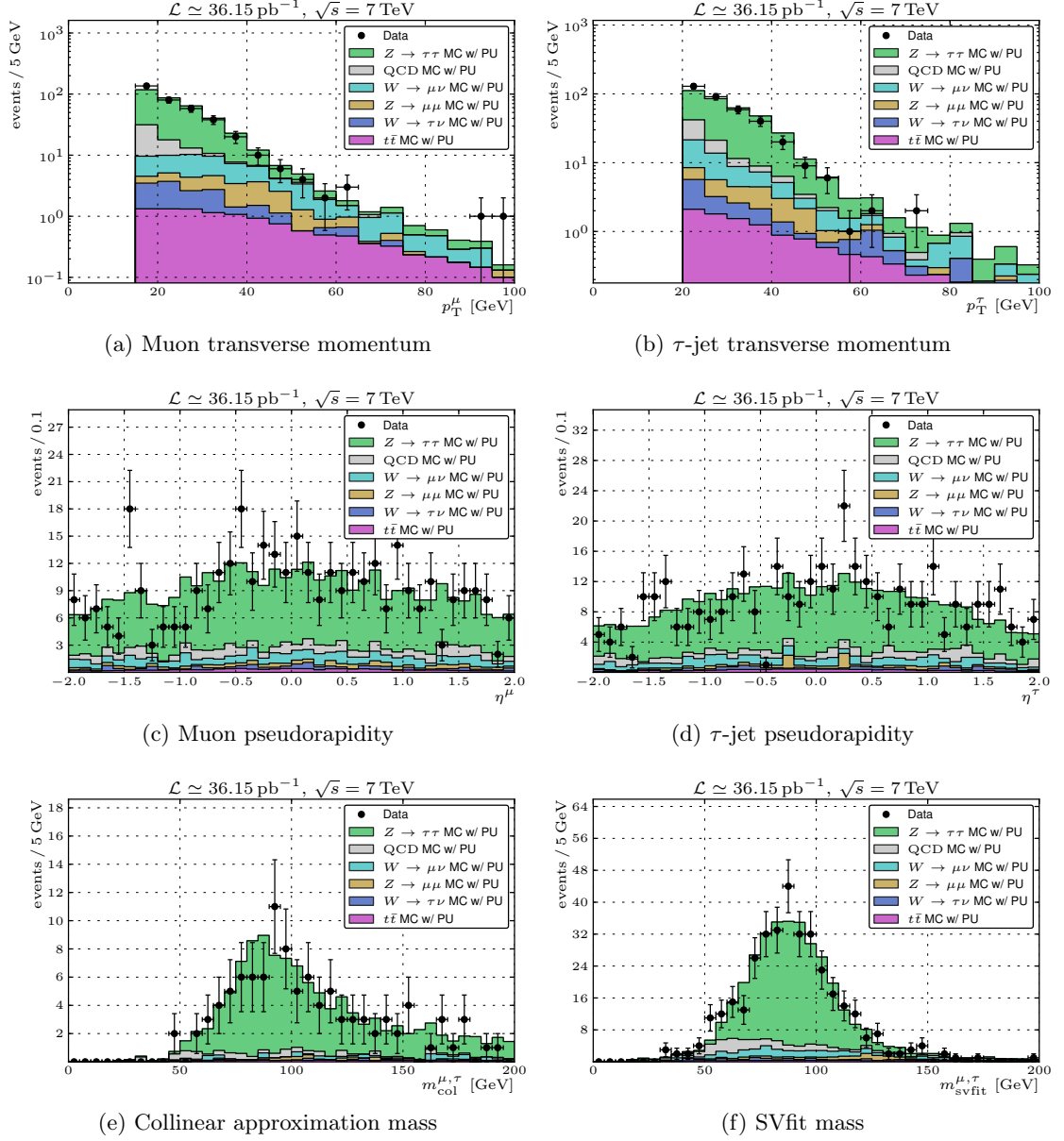


Figure C.1: Various kinematic variables of the muon and the τ -jet in 2010 CMS data after the full $\mu + \tau$ -jet selection as described in Section 5.3.2.

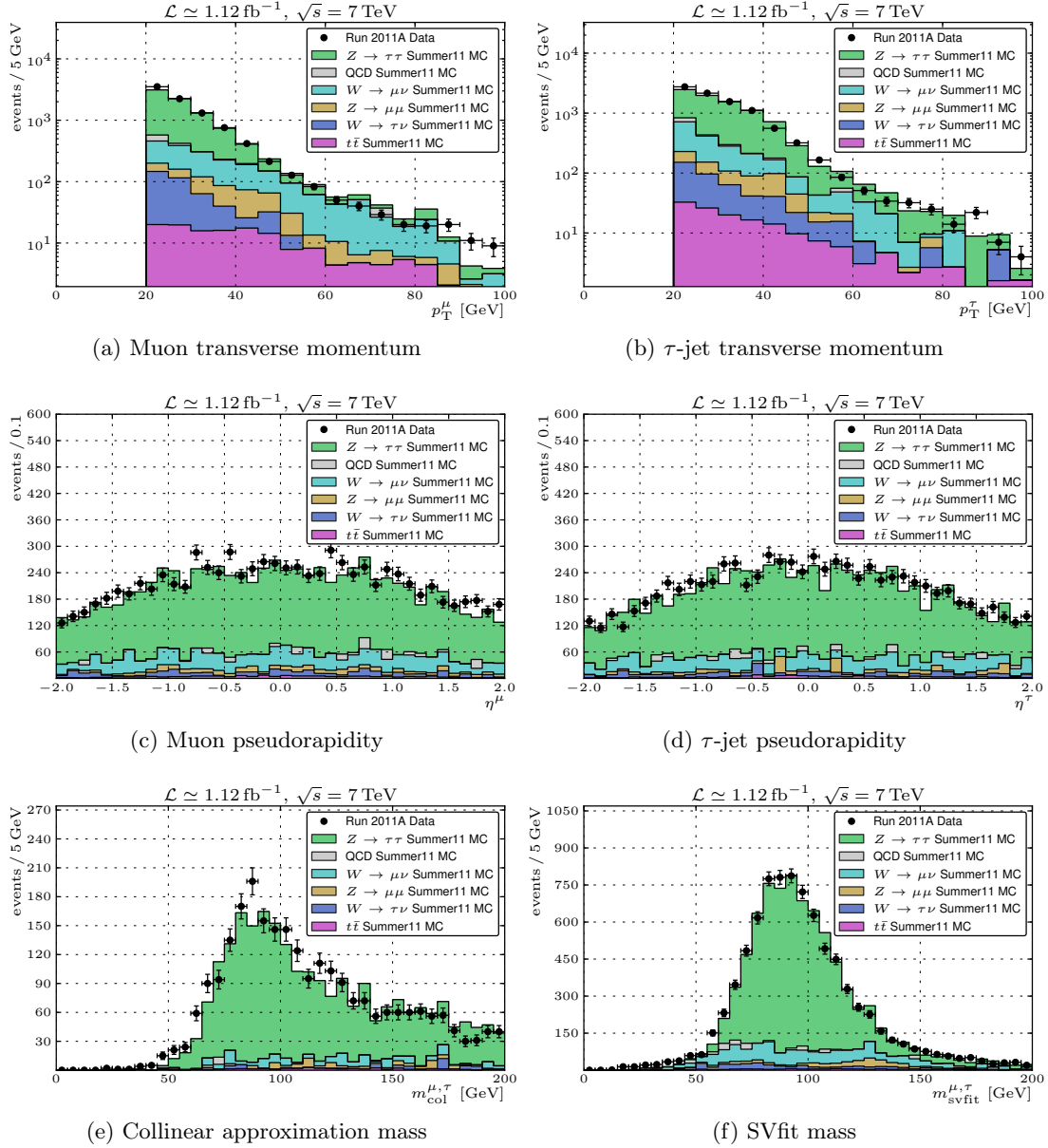
C.2 $\tau^+\tau^-$ Final States in 2011 CMS Data

Figure C.2: Various kinematic variables of the muon and the τ -jet in 2011 CMS data after the full $\mu + \tau$ -jet selection as described in Section 5.4.

C.3 Embedding Monte Carlo Closure Test

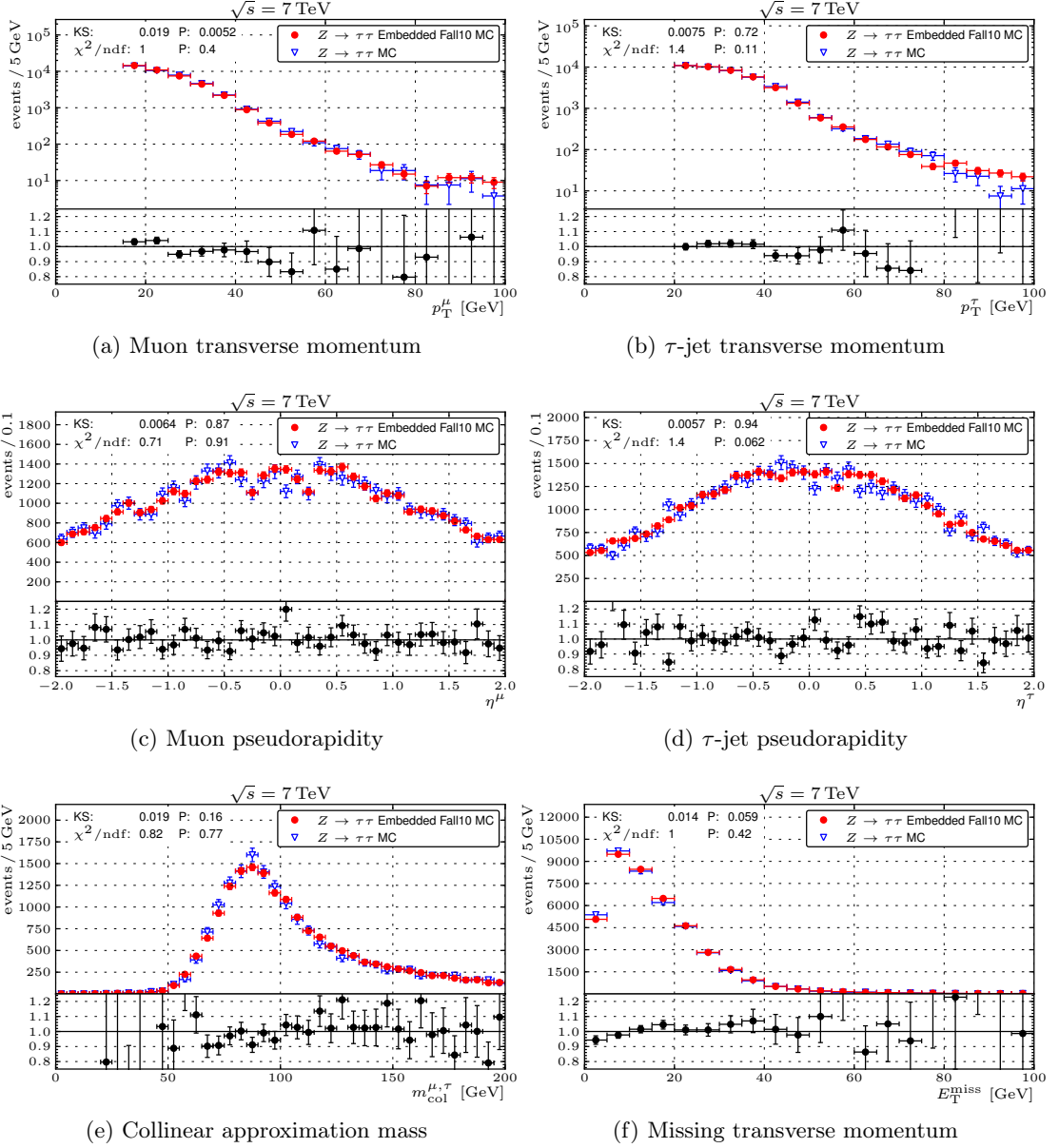
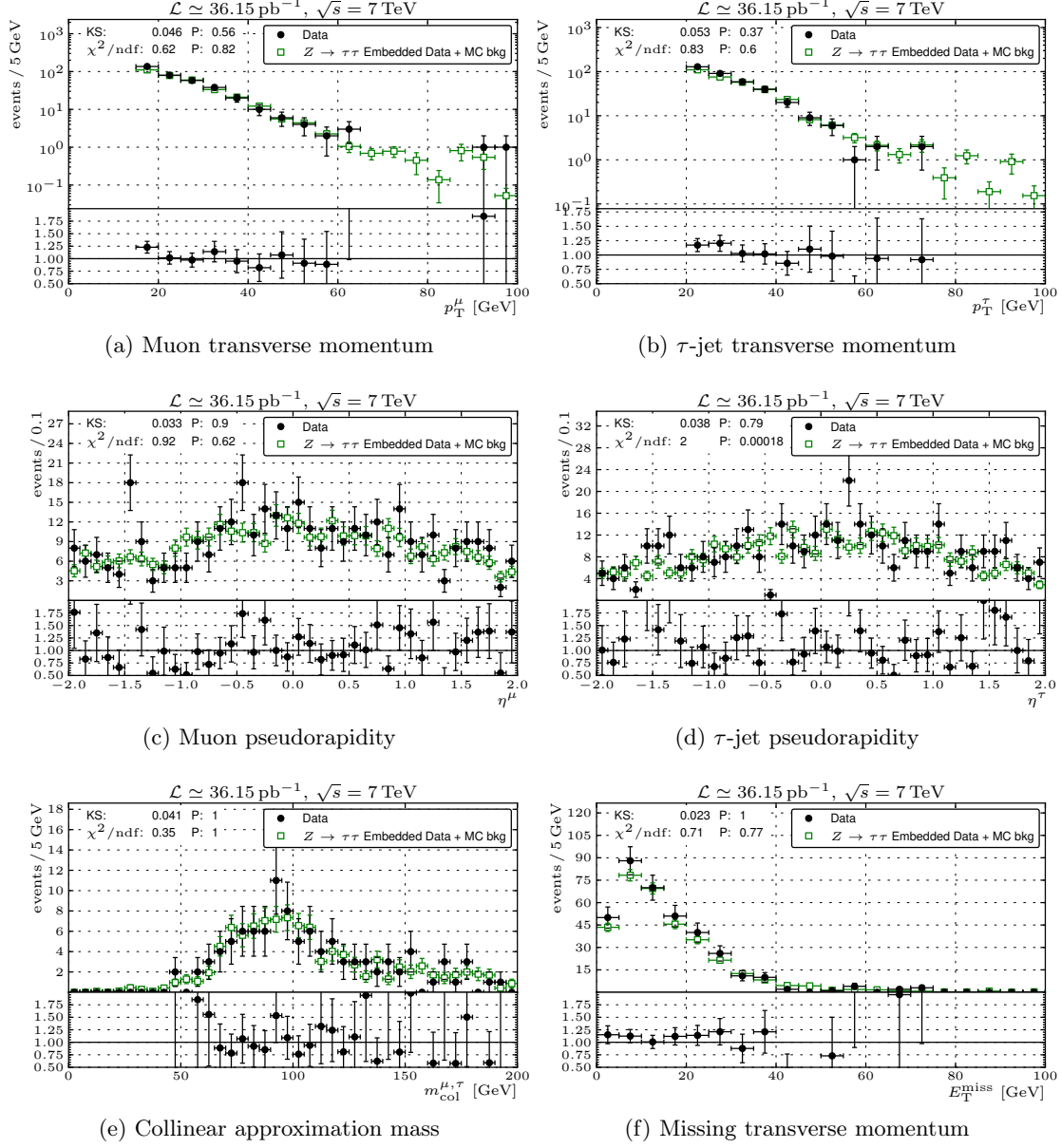


Figure C.3: Various kinematic variables of the muon and the τ -jet for normal and embedded $\mu + \tau$ -jet events on Monte Carlo level.

C.4 Embedding on 2010 CMS Data

Figure C.4: Various kinematic variables of the muon and the τ -jet for normal and embedded $\mu + \tau$ -jet events on 2010 CMS data.

C Additional Plots

List of Figures

1	Picture of the CMS detector in 2006	ii
2	Screenshot of the HAPPYFACE website	iii
3	Visible mass of $\mu + \tau$ -jet candidates	v
4	Visible mass distributions of embedded events	vi
1.1	Example Feynman diagrams	6
1.2	Potential of the Higgs field	11
1.3	Leading order Feynman diagrams for Higgs production	14
1.4	Branching ratio and decay width of the Standard Model Higgs boson	15
1.5	Fit of the Higgs mass to electroweak precision data	16
1.6	Higgs exclusion limits of the LHC experiments ATLAS and CMS	17
2.1	Aerial view of the LHC area	20
2.2	The LHC injection chain	22
2.3	Schematic overview of the CMS detector	25
2.4	Schematic view of the inner tracking system of CMS	26
2.5	Schematic view of the electromagnetic calorimeter of CMS	28
2.6	Schematic view of the hadronic calorimeter of CMS	29
2.7	The muon system of CMS	30
2.8	Particle identification in CMS	31
2.9	Data reduction system of CMS	32
3.1	PARTICLE FLOW performance on jets and missing transverse momentum	42
3.2	Typical physics analysis workflow	43
4.1	The multi-layered structure of the LHC Computing Grid	46
4.2	Typical work flow of a Grid job	47
4.3	Screenshot of the HAPPYFACE website	51
4.4	Example history plot generated by HAPPYFACE	52
4.5	Work flow in the HAPPYFACE framework	53
4.6	The “JobsStatistics” module in HAPPYFACE	58
4.7	The “dCacheDataManagement” module in HAPPYFACE	59
4.8	The “Summary” module in HAPPYFACE	61
5.1	Example Feynman diagrams for τ lepton decays	65
5.2	Illustration of the signal and isolation cones	66
5.3	Fake rate of tau identification algorithms	68
5.4	Principle of the collinear approximation	69

List of Figures

5.5	Comparison of the three di- τ mass reconstruction algorithms	70
5.6	Visible mass distribution in individual selection steps	76
5.7	Distribution of muon isolation and transverse mass	77
5.8	Number of reconstructed vertices and visible mass in 2011 CMS data . . .	80
6.1	Principle of the embedding technique	83
6.2	Effect of an invariant di-muon mass cut for the embedding selection	88
6.3	Embedding procedure on PARTICLE FLOW level	89
6.4	Important distributions for events not accessible to embedding studies . .	93
6.5	The muon isolation variable for embedded Monte Carlo events	94
6.6	Mass distributions for normal and embedded events on Monte Carlo level	97
6.7	Mass distributions for normal and embedded events on 2010 CMS data . .	99
A.1	The “CMSPhedexAgents” module	104
A.2	The “CMSPhedexErrorLog” module	105
A.3	The “DashboardDatasetUsage” module	106
A.4	The “dCacheDatasetRestoreLazy” module	107
A.5	The “dCacheInfoPool” module	108
A.6	The “dCacheTransfers” module	109
A.7	The “JobsDist” module	110
A.8	The “JobsEfficiencyPlot” module	111
A.9	The “SAM” module	112
C.1	Kinematic properties of 2010 $\mu + \tau$ -jet selection	116
C.2	Kinematic properties of 2011 $\mu + \tau$ -jet selection	117
C.3	Kinematic properties for normal and embedded events on Monte Carlo level	118
C.4	Kinematic properties of normal and embedded events on 2010 CMS data .	119

List of Tables

1.1	The four fundamental interactions	4
1.2	The fermions and their couplings to the fundamental forces	5
2.1	Collider parameters of the LHC and the Tevatron	21
5.1	Properties of the τ^- lepton	64
5.2	Prominent hadronic τ decay modes	66
5.3	$\tau^+\tau^-$ decay channels studied in CMS and their branching ratio	72
5.4	Monte Carlo samples used in this analysis	74
5.5	High Level Triggers used for the $\mu + \tau$ -jet selection in 2010	74
5.6	Event yield after the various selection steps of the $\mu + \tau$ -jet analysis	78
5.7	High Level Triggers used for the $\mu + \tau$ -jet selection in 2011	79
6.1	List of systematic error sources in $Z \rightarrow \tau\tau \rightarrow \mu + \tau$ -jet analyses	87
6.2	Number of events not accessible to embedding due to phase space restrictions	92
6.3	Number of events after the full analysis before and after muon isolation	95
6.4	Application of the correction factors on an embedded Monte Carlo sample	96
6.5	Application of correction factors to an embedded data sample	98
6.6	Summary of systematic error sources in the embedding study	98
6.7	Summary of the total correction factors for an embedded sample	99
A.1	Available HAPPYFACE modules	103
B.1	Datasets used for simulated events	113
B.2	Datasets used for measured events	114

List of Tables

Bibliography

- [1] Ernest Rutherford. The Scattering of the Alpha and Beta Rays and the Structure of the Atom. *Proceedings of the Manchester Literary and Philosophical Society IV*, pages 18–20, 1911.
- [2] Michael E. Peskin and Dan V. Schroeder. *An Introduction To Quantum Field Theory (Frontiers in Physics)*. Westview Press, 1995.
- [3] Peter Dunne. Looking for consistency in the construction and use of Feynman diagrams. *Physics Education*, 36(5):366, 2001.
- [4] Chien-Shiung Wu et al. Experimental Test of Parity Conservation in Beta Decay. *Physical Review*, 106:1413, 1957.
- [5] Abdus Salam. Weak and Electromagnetic Interactions. Originally printed in *Svartholm: Elementary Particle Theory, Proceedings Of The Nobel Symposium Held 1968 At Lerum, Sweden*, Stockholm 1968, 367-377.
- [6] J. Horstkotte, A. Entenberg, R. S. Galik, A. K. Mann, H. H. Williams, W. Koza-necki, C. Rubbia, J. Strait, L. Sulak, and P. Wanderer. Measurement of neutrino-proton and antineutrino-proton elastic scattering. *Phys. Rev. D*, 25(11):2743–2761, Jun 1982.
- [7] T. Saeki. W mass measurement at LEP. 1999.
- [8] Francesco Spano. Standard model electroweak measurements at LEP. 2006.
- [9] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13(9):321–323, Aug 1964.
- [10] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13(16):508–509, Oct 1964.
- [11] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. Global conservation laws and massless particles. *Phys. Rev. Lett.*, 13(20):585–587, Nov 1964.
- [12] Manuel Zeise. Study of Z Boson Decays into Pairs of Muon and Tau Leptons with the CMS Detector at the LHC. IEKP-KA/2011-20, 2011.
- [13] CMS Collaboration. Figures from the CMS physics Technical Design Report, Volume II: Physics Performance. CMS Collection., Jun 2006.

Bibliography

- [14] F. Abe et al. Observation of Top Quark Production in $p\bar{p}$ Collisions with the Collider Detector at Fermilab. *Phys. Rev. Lett.*, 74(14):2626–2631, Apr 1995.
- [15] K. Kodama et al. Observation of tau-neutrino interactions. *Phys. Lett.*, B504:218–224, 2001.
- [16] LHC Higgs Cross Section Working Group, S. Dittmaier, C. Mariotti, G. Passarino, and R. Tanaka (Eds.). Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables. *CERN-2011-002*, CERN, Geneva, 2011.
- [17] Christoph Hackstein. Searches for the Higgs Boson at the LHC Based on its Couplings to Vector Bosons. 2011.
- [18] The LEP Collaborations (ALEPH, DELPHI, L3 and OPAL), the LEP Electroweak Working Group and the SLD Heavy Flavour Group. A Combination of preliminary electroweak measurements and constraints on the standard model. 2004.
- [19] T. Aaltonen et al. Combined CDF and D0 Upper Limits on Standard Model Higgs Boson Production with up to 8.2 fb^{-1} of Data. 2011.
- [20] Martin Goebel for the Gfitter Group. Status of the global fit to electroweak precision data. *PoS*, ICHEP2010:570, 2010.
- [21] The ALEPH Collaboration, The DELPHI Collaboration, The L3 Collaboration, The OPAL Collaboration, The SLD Collaboration, The LEP Electroweak Working Group, and The SLD Electroweak Heavy Flavour Groups. Precision electroweak measurements on the Z resonance. *Physics Reports*, 427:257–454, May 2006.
- [22] ATLAS Collaboration. Combined Standard Model Higgs Boson Searches in pp Collisions at $\sqrt{s} = 7 \text{ TeV}$ with the ATLAS Experiment at the LHC. *ATLAS Note*, ATLAS-CONF-2011-112, 2011.
- [23] CMS Collaboration. Search for Standard Model Higgs boson in pp collisions at $\sqrt{s} = 7 \text{ TeV}$. *CMS Physics Analysis Summary*, CMS-PAS-HIG-11-011, 2011.
- [24] CMS Collaboration. Performance of the b-jet identification in CMS. *CMS Physics Analysis Summary*, CMS-PAS-BTV-11-001, 2011.
- [25] Y. Fukuda et al. Evidence for oscillation of atmospheric neutrinos. *Phys. Rev. Lett.*, 81:1562–1567, 1998.
- [26] Q. R. Ahmad et al. Measurement of the charged current interactions produced by B-8 solar neutrinos at the Sudbury Neutrino Observatory. *Phys. Rev. Lett.*, 87:071301, 2001.
- [27] Oliver Sim Brüning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock. *LHC Design Report*. CERN, Geneva, 2004.

- [28] D Brandt, H Burkhardt, M Lamont, S Myers, and J Wenninger. Accelerator physics at LEP. *Reports on Progress in Physics*, 63(6):939, 2000.
- [29] TeVI Group. Design Report Tevatron 1 project. 1984.
- [30] Nakamura, K. et al. The Review of Particle Physics. *J. Phys.*, G37:075021, 2010.
- [31] K. Aamodt et al. The ALICE experiment at the CERN LHC. *JINST*, 3:S08002, 2008.
- [32] G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008.
- [33] CMS Collaboration. Figures from the CMS Physics Technical Design Report, Volume I: Detector Performance and Software. CMS Collection., Feb 2006.
- [34] A. Augusto Alves et al. The LHCb Detector at the LHC. *JINST*, 3:S08005, 2008.
- [35] A. Valishev. Tevatron accelerator physics and operation highlights. Presented at 2011 Particle Accelerator Conference (PAC'11), New York, NY, 28 Mar - 1 Apr 2011.
- [36] Tevatron Run II Parameters. <http://www-ad.fnal.gov/runII/parameters.pdf>.
- [37] The TOTEM Collaboration. The TOTEM Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3(08):S08007, 2008.
- [38] R D'Alessandro, O Adriani, L Bonechi, M Bonghi, G Castellini, D A Faus, K Fukui, M Grandi, M Haguenaue, Y Itow, K Kasahara, D Macina, T Mase, K Masuda, Y Matsubara, H Menjo, M Mizuishi, Y Muraki, P Papini, A L Perrot, S Ricciarini, T Sako, Y Shimizu, K Taki, T Tamura, S Torii, A Tricomi, W C Turner, J Velasco, A Viciani, and K Yoshida. The LHCf experiment at CERN: motivations and current status. *Nucl. Phys. B, Proc. Suppl.*, 190:52–58, 2009.
- [39] James Pinfold, R Soluk, Y Yao, S Cecchini, G Giacomelli, M Giorgini, L Patrizii, G Sirri, D H Lacarrère, K Kinoshita, J Jakubek, M Platkevic, S Pospíšil, Z Vykydal, T Hott, A Houdayer, Claude Leroy, J Swain, D Felea, D Hasegan, G E Pavalas, and V Popa. Technical Design Report of the MoEDAL Experiment. Technical Report CERN-LHCC-2009-006. MoEDAL-TDR-001, CERN, Geneva, Jun 2009.
- [40] M Bajko, F Bertinelli, N Catalan-Lasheras, S Claudet, P Cruikshank, K Dahlerup-Petersen, R Denz, P Fessia, C Garion, JM Jimenez, G Kirby, Ph Lebrun, S Le Naour, K-H Mess, M Modena, V Montabonnet, R Nunes, V Parma, A Perin, G de Rijk, A Rijllart, L Rossi, R Schmidt, A Siemko, P Strubin, L Tavian, H Thiesen, J Tock, E Todesco, R Veness, A Verweij, L Walckiers, R Van Weelderren, R Wolf, S Fehér, R Flora, M Koratzinos, P Limon, and J Strait. Report of the Task Force on the Incident of 19th September 2008 at the LHC. Technical Report CERN-LHC-PROJECT-Report-1168, CERN, Geneva, Mar 2009.

Bibliography

- [41] R Denz, K Dahlerup-Petersen, F Formenti, K H Meß, A Siemko, J Steckert, L Walkiers, and J Strait. Upgrade of the protection system for superconducting circuits in the LHC. Technical Report CERN-ATS-2009-008, CERN, Geneva, Jul 2009.
- [42] Simon van der Meer. Calibration of the effective beam height in the ISR. Technical Report CERN-ISR-PO-68-31. ISR-PO-68-31, CERN, Geneva, 1968.
- [43] CMS Collaboration. Measurement of CMS Luminosity. *CMS Physics Analysis Summary*, CMS-PAS-EWK-10-004, 2010.
- [44] The CMS Collaboration. The CMS experiment at the CERN LHC. *JINST*, 3 S08004:361, 2008.
- [45] Martin Frey. Development of Highly Radiation Hard Silicon Strip Sesors for Application at the Super Large Hadron Collider. IEKP-KA/2009-18, 2009.
- [46] L. Feld, R. Jussen, W. Karpinski, K. Klein, and J. Sammet. DC-DC buck converters for the CMS Tracker upgrade at SLHC. *Journal of Instrumentation*, 6(01):C01020, 2011.
- [47] CMS Collaboration. Performance of the CMS hadron calorimeter with cosmic ray muons and LHC beam data. *Journal of Instrumentation*, 5:3012–+, March 2010.
- [48] CMS outreach, <http://cmsinfo.cern.ch/outreach/>.
- [49] CMS Collaboration. The TriDAS Project Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger. *CERN/LHCC*, 2002-26, 2002.
- [50] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 05:026, 2006.
- [51] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.
- [52] Bo Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand. Parton Fragmentation and String Dynamics. *Phys. Rept.*, 97:31–145, 1983.
- [53] Michael Heinrich. A Jet Based Approach to Measuring Soft Contributions to Proton-Proton Collisions with the CMS Experiment. IEKP-KA/2010-24, 2011.
- [54] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. *JHEP*, 06:043, 2010.
- [55] Stefano Frixione and Bryan R. Webber. Matching NLO QCD computations and parton shower simulations. *JHEP*, 06:029, 2002.
- [56] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. NLO vector-boson production matched with shower in POWHEG. *JHEP*, 07:060, 2008.

- [57] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. Vector boson plus one jet production in POWHEG. *JHEP*, 01:095, 2011.
- [58] Emanuele Re. Single-top production with the POWHEG method. *PoS*, DIS2010:172, 2010.
- [59] Adam Kardos, Costas Papadopoulos, and Zoltan Trocsanyi. Top quark pair production in association with a jet with NLO parton showering. 2011.
- [60] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. NLO Higgs boson production via gluon fusion matched with shower in POWHEG. *JHEP*, 04:002, 2009.
- [61] Paolo Nason and Carlo Oleari. NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG. *JHEP*, 02:037, 2010.
- [62] Stanislaw Jadach, Zbigniew Was, Roger Decker, and Johann Kuehn. The tau decay library TAUOLA: Version 2.4. *Comput. Phys. Commun.*, 76:361–380, 1993.
- [63] Rene Brun and Fons Rademakers. ROOT – An object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(1-2):81 – 86, 1997. New Computing Techniques in Physics Research V.
- [64] I. Antcheva, M. Ballintijn, B. Bellenot, M. Biskup, R. Brun, N. Buncic, Ph. Canal, D. Casadei, O. Couet, V. Fine, L. Franco, G. Ganis, A. Gheata, D. Gonzalez Maline, M. Goto, J. Iwaszkiewicz, A. Kreshuk, D. Marcos Segura, R. Maunder, L. Moneta, A. Naumann, E. Offermann, V. Onuchin, S. Panacek, F. Rademakers, P. Russo, and M. Tadel. ROOT – A C++ framework for petabyte data storage, statistical analysis and visualization. *Computer Physics Communications*, 180(12):2499 – 2512, 2009.
- [65] Physics Analysis Workstation. <http://paw.web.cern.ch/paw/>.
- [66] CINT, a command line C/C++ interpreter. <http://root.cern.ch/root/Cint.html>.
- [67] A Afaq, A Dolgert, Y Guo, C Jones, S Kosyakov, V Kuznetsov, L Lueking, D Riley, and V Sekhri. The CMS dataset bookkeeping service. *Journal of Physics: Conference Series*, 119(7):072001, 2008.
- [68] JavaScript Object Notation. <http://www.json.org/>.
- [69] Agostinelli, S. et al. G4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250 – 303, 2003.
- [70] CMS Collaboration. Tracking and Vertexing Results from First Collisions. *CMS Physics Analysis Summary*, CMS-PAS-TRK-10-001, 2010.

Bibliography

- [71] Wolfgang Adam, R Frühwirth, Are Strandlie, and T Todor. Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC. Technical Report CMS-NOTE-2005-001. CERN-CMS-NOTE-2005-001, CERN, Geneva, Jan 2005.
- [72] J.-R. Vlimant. Track Reconstruction with the CMS Tracking Detector. In *Proceedings of HCP07, Elba Island, Italy*, 2007.
- [73] R Frühwirth, Wolfgang Waltenberger, and Pascal Vanlaer. Adaptive vertex fitting. Technical Report CMS-NOTE-2007-008. CERN-CMS-NOTE-2007-008, CERN, Geneva, Mar 2007.
- [74] S. Catani, Yu. L. Dokshitzer, M. H. Seymour, and B. R. Webber. Longitudinally-invariant k_{\perp} -clustering algorithms for hadron-hadron collisions. *Nuclear Physics B*, 406(1-2):187 – 224, 1993.
- [75] Stan Bentvelsen and Irmtraud Meyer. The cambridge jet algorithm: features and applications. *The European Physical Journal C - Particles and Fields*, 4:623–629, 1998. 10.1007/s100520050232.
- [76] Gavin P. Salam and Gregory Soyez. A practical Seedless Infrared-Safe Cone jet algorithm. *JHEP*, 05:086, 2007.
- [77] CMS Collaboration. Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET. *CMS Physics Analysis Summary*, CMS-PAS-PFT-09-001, 2009.
- [78] Manuel Zeise. Modelling of background from Z boson decays for the Higgs boson search at the LHC in the channel $H \rightarrow \tau\tau$. IEKP-KA/2008-10, 2008.
- [79] Christoph Eck, J Knobloch, Leslie Robertson, I Bird, K Bos, N Brook, D Düllmann, I Fisk, D Foster, B Gibbard, C Grandi, F Grey, J Harvey, A Heiss, F Hemmer, S Jarp, R Jones, D Kelsey, M Lamanna, H Marten, P Mato-Vila, F Ould-Saada, B Panzer-Steindel, L Perini, Y Schutz, U Schwickerath, J Shiers, and T Wenaus. *LHC computing Grid: Technical Design Report. Version 1.06 (20 Jun 2005)*. Technical Design Report LCG. CERN, Geneva, 2005.
- [80] G L Bayatyan, Michel Della Negra, Lorenzo Foà, A Hervé, and Achille Petrilli. *CMS computing: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 2005. Submitted on 31 May 2005.
- [81] A Scheurer and the German CMS Community. German contributions to the CMS computing infrastructure. *Journal of Physics: Conference Series*, 219(6):062064, 2010.
- [82] WLCG Tier Sites. <http://lcg.web.cern.ch/LCG/public/tiers.htm>.

- [83] Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid. <http://lcg.web.cern.ch/LCG/mou.htm>, 2009.
- [84] Armin Scheurer. Algorithms for the Identification of b-Quark Jets with First Data at CMS. IEKP-KA/2008-19, 2008.
- [85] Fabrizio Pacini. Job Description Language (JDL) Attributes Specification. <https://edms.cern.ch/document/590869/1>, 2005.
- [86] Mathias de Riese, Patrick Fuhrmann, Tigran Mkrtchyan, Michael Ernst, Alex Kulyavtsev, Vladimir Podstavkov, Martin Radicke, Neha Sharma, Dmitry Litvinintsev, Timur Perelmutov, and Ted Hesselroth. *The dCache Book*.
- [87] J. Rehn. Phedex high-throughput data transfer management system. In *Proceedings of CHEP06, Mumbai, India*, 2006.
- [88] J Andreeva, S Belov, A Berejnoj, C Cirstoiu, Y Chen, T Chen, S Chiu, M D F D Miguel, A Ivanchenko, B Gaidioz, J Herrala, M Janulis, O Kodolova, G Maier, E J Maguire, C Munro, R P Rivera, R Rocha, P Saiz, I Sidorova, F Tsai, E Tikhonenko, and E Urbah. Dashboard for the LHC experiments. *Journal of Physics: Conference Series*, 119(6):062008, 2008.
- [89] Volker Büge, Viktor Mauch, Günter Quast, Armin Scheurer, and Artem Trunov. Site specific monitoring of multiple information systems – the HappyFace Project. *Journal of Physics: Conference Series*, 219(6):062057, 2010.
- [90] Viktor Mauch. Development of a Meta Monitoring System for Grid Sites and Effects of the Simulation of Hadronic Jets on the Reconstruction Efficiency in the Vector Boson Fusion Channel $H \rightarrow \tau\tau$. IEKP-KA/2008-27, 2008.
- [91] The Python Programming Language. <http://www.python.org>.
- [92] SQLite database engine. <http://www.sqlite.org>.
- [93] PHP: Hypertext Preprocessor. <http://www.php.net>.
- [94] SQLAlchemy Object Relational Manager. <http://www.sqlalchemy.org>.
- [95] PHP Data Objects (PDO). <http://www.php.net/manual/en/book.pdo.php>.
- [96] MySQL database engine. <http://www.mysql.com>.
- [97] PostgreSQL database engine. <http://www.postgresql.org>.
- [98] Apache HTTP Server. <http://httpd.apache.org>.
- [99] SubVersion, a version control system. <http://subversion.apache.org>.

Bibliography

- [100] HappyFace Documentation. https://ekptrac.physik.uni-karlsruhe.de/trac/HappyFace/wiki/Version_2.
- [101] Plans for Tier-1 monitoring. <https://twiki.cern.ch/twiki/bin/view/CMS/CmsTier1MonitoringProject>, 2010. [CMS members only].
- [102] The Chimera name server. <http://trac.dcache.org/projects/dcache/wiki/Chimera>.
- [103] Universal Feed Parser. <http://feedparser.org>.
- [104] The CMS Collaboration. CMS Physics Technical Design Report, Volume II: Physics Performance. *Journal of Physics G: Nuclear and Particle Physics*, 34(6):995, 2007.
- [105] Zong-ru Wan. A Search for New Physics with High Mass Tau Pairs in proton anti-proton collisions at $s^{**}(1/2) = 1.96$ -TeV at CDF. 2005. Ph.D. Thesis (Advisor: John Conway).
- [106] Jan-e Alam, Bedangadas Mohanty, Sanjay K. Ghosh, Sarbani Majumder, and Rajarshi Ray. Heavy lepton pair production in nucleus-nucleus collisions at LHC energy - a case study. 2011.
- [107] M. Bona et al. An Improved Standard Model Prediction of $BR(B \rightarrow \tau\nu)$ and its Implications for New Physics. *Phys. Lett.*, B687:61–69, 2010.
- [108] A. Bethani, A. Burgmeier, A.B. Meyer, A. Raspereza, M. Rosin, G. Schott, G. Quast, R. Walsh, and M. Zeise. Measurement of $\sigma(pp \rightarrow Z) \cdot Br(Z \rightarrow \tau\tau)$ in the dimuon channel with CMS in pp collisions at 7 TeV. CMS AN-2010/446.
- [109] CMS Collaboration. Study of tau reconstruction algorithms using pp collisions data collected at $\sqrt{s} = 7$ TeV. *CMS Physics Analysis Summary*, CMS-PAS-PFT-10-004, 2010.
- [110] M. Bachtis, L. Bianchini, M. Edelhoff, E. Friis, T. Fruboes, S. Gennai, S. Maruyama, A. Mohammadi, A. Nayak, A. Savin, J. Swanson, and C. Veelken. Performance of tau reconstruction algorithms with 2010 data in CMS. CMS AN-2011/045.
- [111] CMS Collaboration. Tau identification in CMS. *CMS Physics Analysis Summary*, CMS-PAS-TAU-11-001, 2011.
- [112] C.M. Bishop. Neural networks and their applications. *Review of Scientific Instruments*, 65(6):1803–1831, 1994.
- [113] M. Bachtis, J. Swanson, A. Savin, S. Dasu, and W. Smith. Study of di-tau spectrum using $\mu\tau$ and $e\tau$ final states with CMS detector at $\sqrt{s} = 7$ TeV. CMS AN-2010/387.

- [114] S. Baffionia, F. Beaudetteb, D. Benedettib, J. Brasonc, G. Daskalakisd, E. Di Marcoe, C. Campagnarif, C. Charlota, S. Harperg, P.D. Kalavasef, J.D. Kellerh, D. Lelasi, P. Meridianib, M. Pieric, I. Puljaki, N. Ranieric, R. Rompotisl, C. Rovellie, R. Salernoa, M. Sanic, C. Seezl, Y. Siroisa, Y. Tuc, and Yagilc. A. Electron Identification in CMS. CMS AN-2009/178.
- [115] J. Conway, E. K. Friis, M. Squires, C. Veelken, G. Cerati, S. Malvezzi, R. Manzoni, E. Re, and J. Kolb. Search for MSSM neutral Higgs $\rightarrow \tau^+\tau^-$ Production using the TaNC Tau id. algorithm. CMS AN-2010/460.
- [116] Volker Blobel and Erich Lohrmann. *Statistische und numerische Methoden der Datenanalyse*. Teubner Verlag, 1 edition, 1998.
- [117] CMS Collaboration. Measurement of the inclusive $Z \rightarrow \tau^+\tau^-$ cross section in pp collisions at $\sqrt{s} = 7$ TeV. *CMS Physics Analysis Summary*, CMS-PAS-EWK-10-013, 2010.
- [118] J. Alcaraz et al. Measurement of the Inclusive W and Z Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV: Update with full 2010 statistics. CMS AN-2010/395.
- [119] CMS Collaboration. Measurement of the W and Z inclusive production cross sections at $\sqrt{s} = 7$ TeV with the CMS experiment at the LHC. *CMS Physics Analysis Summary*, CMS-PAS-EWK-10-002, 2010.
- [120] Matteo Cacciari. FastJet: A Code for fast $k(t)$ clustering, and more. 2006.
- [121] Harris Drucker and Corinna Cortes. Boosting decision trees. In *Neural Information Processing Systems*, pages 479–485.
- [122] Matthias Wolf. Statistical Combination of Decay Channels in Searches for the Higgs Boson at the LHC. IEKP-KA/2010-10, 2010.
- [123] CMS Collaboration. Absolute luminosity normalization. *CMS Detector Performance Summary*, CMS-DP-2011-002, Mar 2011.
- [124] CMS Collaboration. Jet Energy Calibration and Transverse Momentum Resolution in CMS. *CMS Physics Analysis Summary*, CMS-PAS-JME-10-011, 2010.
- [125] M. Bluj, A. Burgmeier, T. Früboes, G. Quast, and M. Zeise. Modelling of $\tau\tau$ final states by embedding τ pairs in $Z \rightarrow \mu\mu$ events. CMS AN-2011/020.
- [126] CMS Collaboration. Studies of Tracker Material. *CMS Physics Analysis Summary*, CMS-PAS-TRK-10-003, 2010.
- [127] C. J. CLOPPER and E. S. PEARSON. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934.

Bibliography

Acknowledgments

First and foremost, my thank goes to my advisor Prof. Dr. Günter Quast for steering me into the right directions, for letting me the freedom to try out own ideas and for sharing with me his never ending wisdom. For accepting to be co-referee for this thesis, my further thank goes to Prof. Dr. Wim de Boer.

Special thank goes to Dr. Manuel Zeise for all the great teamwork in the past year and who never grew tired in answering all questions I had.

I am grateful to all my colleagues for the great atmosphere in the working group, for proofreading this thesis and for many inspiring discussions. In particular I want to thank Dr. Volker Büge for introducing me into the group, Dr. Armin Scheurer, Dr. Andreas Oehler, Dr. Michael Heinrich, Dr. Oliver Oberst, Dr. Christoph Hackstein, Fred Stober, Joram Berger, Thomas Hauth, Thomas Müller, David Kernert, Stefan Riedel and Timo Doll.

I thank Dr. Alexei Raspereza and Agni Bethani for the fruitful collaboration with the DESY group.

Finally, my thank goes to my family who always encouraged me to do what I felt like, and especially my father who passed away too soon. I am grateful to my mother, my brother and my friends for helping me to cope with these hard times.

Acknowledgements

Hiermit versichere ich, die vorliegende Arbeit selbstständig verfasst
und nur die angegebenen Hilfsmittel verwendet zu haben.

Armin Burgmeier

Karlsruhe, den 1. August 2011