Absolute jet energy scale determination at the CMS detector for LHC run 2 at $\sqrt{s} = 13$ TeV using an optimized processing setup

Christoph Alexander Heidecker

MASTERARBEIT

An der Fakultät für Physik des Karlsruher Institut für Technologie (KIT)
Institut für Experimentelle Kernphysik (IEKP)

Abgabedatum: 20.01.2017

Referent: Prof. Dr. Günter Quast
Korreferent: Prof. Dr. Ulrich Husemann
Betreuender Mitarbeiter: Dr. Stefan Wayand
## Contents

1 Introduction .......................................................... 3

2 Physics at the Large Hadron Collider .......... 5
   2.1 Experimental setup ........................................... 5
   2.2 Compact Muon Solenoid at the Large Hadron Collider .......... 6
   2.3 Jet measurement in the CMS experiment ........................ 7
      2.3.1 Origin of jets ........................................ 8
      2.3.2 Reconstruction of jets ................................ 9
      2.3.3 Pileup mitigation ..................................... 10
   2.4 Jet energy calibration ........................................ 11
      2.4.1 General approach ...................................... 11
      2.4.2 Absolute residual corrections .......................... 13
   2.5 Computing environment in high energy physics .............. 14
      2.5.1 Workflow requirements .................................. 15
      2.5.2 Computing resources in use ............................. 16

3 High-throughput computing in high energy physics 19
   3.1 HTCondor as batch system .................................. 19
      3.1.1 Functionality of HTCondor ............................ 20
      3.1.2 Local HTCondor setup using docker containers .......... 21
   3.2 High Throughput Data Analysis utilizing coordinated caching for batch system usage .............................. 23
      3.2.1 Functionality and extension of HTDA .................. 23
      3.2.2 Advantages of HTDA for the calibration approach .......... 23
   3.3 CephFS as extension of the caching ......................... 25
      3.3.1 Structure of the local Ceph file system ................. 26
      3.3.2 Advantages of the Ceph setup for the SSD caching approach 27
   3.4 Future benefits of an expanded caching .................... 30
## Contents

4 Absolute residual corrections using $Z +$ jet events 33
   4.1 Balancing methods .............................................. 33
       4.1.1 Transverse momenta balance .......................... 34
       4.1.2 Missing Transverse Energy Projection Fraction .......... 34
   4.2 Event selection .................................................. 34
       4.2.1 Muon and electron selection for $Z$ boson reconstruction . 35
       4.2.2 Jet selection ............................................... 36
       4.2.3 $Z +$ jets topology selection ............................ 36
       4.2.4 Event re-weighting ....................................... 37
       4.2.5 $E_T^{\text{miss}}$ Event filter ............................... 38
       4.2.6 Muon and electron corrections ............................ 38
       4.2.7 Background estimation .................................... 38
   4.3 Distributions of kinematic variables .......................... 41
       4.3.1 Lepton characteristics .................................... 41
       4.3.2 $Z$ boson characteristics ................................ 43
       4.3.3 Jet distributions .......................................... 44
       4.3.4 Missing transverse energy distributions .................. 45
   4.4 Response measurements .......................................... 46
       4.4.1 MPF and $p_T$-balance .................................... 46
       4.4.2 Response estimation ....................................... 47
       4.4.3 Cross-checks .............................................. 49
       4.4.4 Time-dependency of the jet response ..................... 51
   4.5 Additional studies ............................................... 52
       4.5.1 Replacement of AOD format with miniAOD ................. 52
       4.5.2 Analogous approach using PUPPI ........................ 53

5 Conclusion 59

A Appendix 61

List of Figures 87

List of Tables 93

Bibliography 95
Chapter 1

Introduction

The Large Hadron Collider (LHC) [1] is the most powerful machine to provide scattering events for particle physics studies. Currently, the run 2 of the LHC reaches new dimensions of providing proton-proton collisions. In 2016, an enormous increase of the luminosity at $\sqrt{s} = 13$ TeV provides new challenges of recording data, and reconstructing particles. At the Compact Muon Solenoid (CMS) [2], one of the detectors located at the LHC, the amount of recorded proton-proton collisions is increased dramatically by a factor of about twenty compared to the previous year. This enormous increase of data poses great challenges for the precise reconstruction of particles and the calibration of the reconstructed physics objects. One of the most distorted objects in the reconstructed events is a jet, a bunch of nearby particles. These jets originate from single quarks or gluons, which play an important role in typical particle physics analyses. The jet energy response of the CMS detector is calibrated immediately during data taking to provide preliminary jet energy correction factors which are needed by nearly all CMS analyses. Here, the enormous amount of data presents a challenge in the provision of continuous response corrections in short processing periods. The approach of the jet energy corrections is done by balancing the jet response of the detector against a well-known reference object. For this purpose, the $Z$ boson which can be reconstructed very precisely from two muons or two electrons is used in this thesis. This balancing approach combines the power of different sub-detectors using the well-defined $Z$ boson to provide correction factors for the jet response. Afterwards, the correction factors presented in this thesis are combined with the other results of the CMS jet energy calibration workgroup to compensate for other mis-measurements and detector effects. The results shown in this thesis allow a preliminary insight into new data recorded in 2016. To process these enormous datasets iteratively within short turnaround periods, an optimized setup called High Throughput Data Analysis (HTDA) developed at the Karlsruhe Institute of Technology (KIT) is used. This setup provides a coordinated caching on solid state drive (SSD) storages which enables high input rates and fast reprocessing.
cycles. During this master thesis, the High Throughput Analysis setup is tested, optimized and expanded. This aims at preparing a stable and well operating setup for future calibration approaches and other analyses depending on high data throughput. Future data taking of the CMS detector will produce even larger datasets which will need to be processed as quickly as possible. Hence, processing solutions like the expanded caching technology described in this thesis will be essential for future analyses.
Chapter 2

Physics at the Large Hadron Collider

For the understanding of the studies done during this master thesis, a basic introduction is done first. This includes the experimental setup, the theoretical principles of the measurement and the workflows of the data analysis.

2.1 Experimental setup

The Large Hadron Collider (LHC) [1] as currently the most powerful scattering experiment was realized at the European Organization for Nuclear Research (CERN). This ring accelerator brings two opposite rotating proton bunches up to currently 13 TeV center-of-mass energy and lead ion bunches up to currently 2.76 TeV center-of-mass energy. This needs a large complex of different pre-accelerators [3] to prepare the bunches for the final boost to the full kinetic energy within the LHC accelerator. The LHC complex also allows switching between accelerating proton or lead ion bunches which are chained to a so-called beam. Whereas the proton mode allows a higher collision energy and therefore the production of new particles with higher masses, the lead mode allows the creation of a shortly existing quark-gluon plasma with quasi-free partons. After reaching the full design energy, the opposite rotating beams are crossed at four collision points. At these collision points, every 25 ns two bunches collide and produce a large amount of particles which are measured by one of the four different detectors. Two generalized detectors Compact Muon Solenoid (CMS) [2] and A Toroidal LHC ApparatuS (ATLAS) [4] check for consistency of the standard model of particle physics and the Higgs mechanism. These detectors cover all spatial directions with a cylindrical design, the Large Hadron Collider beauty (LHCb) [5] covers only the heavy forward region to study b quark physics.

A Large Ion Collider Experiment (ALICE) [6] is specialized for the measurement of heavy ion collisions. Since this thesis is based on proton-proton collision events
measured in 2016 at the CMS detector at 13 TeV center-of-mass energy, the further chapters concentrate on the LHC proton mode recorded at the CMS detector.

2.2 Compact Muon Solenoid at the Large Hadron Collider

The jet energy correction described within this master thesis needs a detailed understanding of the detector structure and functionality. Since the CMS detector was designed for multi-purpose measurements, it consists of different sub-detectors which measure different particles and their momenta or energies. Figure 2.1 shows the

![Diagram of CMS detector](image)

**Figure 2.1:** One quadrant of a longitudinal section of the CMS detector [7]. The silicon tracker which records the tracks of charged particles surrounds the collision point in the center of the detector. Whereas the electromagnetic calorimeter (ECAL) measures the energy of photons and electrons, the hadronic calorimeter (HCAL) summarizes the energy of hadron cascades. To measure the momentum of charged particles, a magnetic field is produced by surrounding solenoid magnet. The outer muon chambers (MB/ME) are integrated into the magnetic return yoke (RB/RE). The polar angular distribution is given in degrees or the pseudorapidity $\eta = -\ln(\tan(\frac{\Theta}{2})$
layered design of the CMS sub-detectors in one quadrant of the longitudinal section. The beam line along the z-axis meets the collision point at the origin of the figure. Surrounding this collision point, the silicon tracker detects the tracks of charged particles using their ionisation-effect in the semiconductor. An inner pixel layout supplies an increased precision directly at the beam line and an outer strip layout supplies a larger spatial volume. Using the magnetic field of the detector, the momentum of the recorded particles is determined from the bend of the tracks. To measure the energy of the particles, this tracker is surrounded by two calorimeters. The electromagnetic calorimeter converts the energy of an incoming photon or electron by bremsstrahlung, pair production, and ionization into a cascade of low-energy photons and electrons. Measuring the energy deposits from the scintillation effect of the cascade in crystal lead tungstate as scintillator and absorber, the energy of the primary electron or photon is determined. The hadronic calorimeter produces a cascade from incoming neutral or charged hadrons using hadronic interactions with nonmagnetic steel and brass plates as absorber material. Alternating the absorber material with scintillator material allows to extract the scintillation energy and determine the hadron energy. These three sub-detectors are surrounded by a solenoid which supplies the magnetic field for the track bending. In the outer return yoke, the muon detection system [8] is integrated which consists of drift tubes (DTs), cathode strip chamber (CSCs) and resistive plate chambers (RPCs) as shown in figure 2.1. Since muons interact weakly with the detector material, they pass the detector layers almost exclusively. Therefore, an extra outer sub-detector measures the tracks of muons and helps to identify muon tracks recorded by the silicon tracker.

For calibration purposes like done in this thesis, a combination of all sub-detectors of the CMS increases the precision. Therefore, the spatial coverage of the single sub-detectors is important. The CMS detector covers the full radial azimuth angular φ around the beam pipe. Hence, the limiting factor is the range of the polar angular Θ which is usually described by the pseudorapidity \( \eta = -\ln(\tan(\Theta/2)) \). Since the silicon tracker and the muon chambers cover the region \( \eta < 2.4 \), the calibration is also limited by this value. Furthermore, the transition of the electronic and hadronic calorimeter between the barrel region and the end-cap region around \( \eta = 1.4 \) influences the calibration approach.

2.3 Jet measurement in the CMS experiment

For a better understanding of the approach of the jet energy calibration, a look at the jet measurement at CMS is necessary. This includes the origin of the jets and the reconstruction process at CMS [9]. There, the information of the different sub-detectors is combined to get a precise reconstruction of the primary particles.
using the theoretical concepts of the interaction processes.

2.3.1 Origin of jets

Due to the high energy of the proton-proton collisions inside the CMS detector, a high amount of unbound quarks is produced in the final state. Since the standard model effect of confinement forbids the existence of colored unbound states of quarks, these unbound quarks need to be grouped into color-neutral hadrons. The potential

![Figure 2.2: The effect of confinement (a) leads to generation of additional color-neutral quark pairs due to the increasing potential of the strong force with the distance [10]. By recombining of all participating quarks into color-neutral hadrons, a single unbound quark generates a cascade of hadrons. This so-called hadronization effect (b) forms a jet of particles, which can be clustered after measurement using a jet algorithm [10].](image)

of the strong force between the initial state quarks increases with the distance. After reaching the production threshold for a new pair of quarks, an additional color-neutral quark pair can be produced as shown in figure 2.2. These additional color-neutral pairs of quarks combine with the final state quarks into bound states of color-neutral hadrons. In each event with contributing quarks or gluons, this process is performed repeatedly and a bunch of hadrons is formed. These hadrons decay on the fly into more stable hadrons and leptons, but also radiate gauge bosons. Both, the hadronization and the subsequent decays lead to a cascade of hadrons, leptons and gauge bosons. This so-called jet includes the complete energy of the original quark or gluon. Hence, the original gluon or quark characteristics can be determined from the jet features. Therefore reconstruction methods are needed to identify and characterize jets in events.
2.3.2 Reconstruction of jets

The particle-flow event reconstruction \cite{11} classifies all measured objects of an event using the information of all CMS sub-detector systems. This aims at a full reconstruction of all objects registered by the silicon tracker, both calorimeters, and the muon system. Afterward, these objects are identified as muons, electrons, photons and charged or neutral hadrons using the combined information of all sub-detectors. Furthermore, jets which contain cascades of hadrons, electrons and photons are clustered to determine the original quark or gluon characteristics. Therefore, the CMS particle-flow jet identification criteria\cite{12, 13} are used. This approach uses a jet clustering algorithm to identify and cluster jets. At the CMS experiment, the anti-$k_t$ algorithm \cite{14} is used. As an improved sequential algorithm, the anti-$k_t$ algorithm summarizes all particles with a dynamic radius. Therefore, a distance $d_{ij}$ is calculated for all particle pairs in the event.

$$d_{ij} = \min \left( p_T^{2n}, p_T^{2n} \right) \frac{\Delta^2_{ij}}{R^2} \quad (2.1)$$

With the chosen parameters $n = -1$ and $R = 0.4$. Additionally, $\Delta^2_{ij} = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ is calculated using the rapidity $y$ and azimuth angular $\phi$ of the particles. In the next step, the minimal value of the calculated distances is found.

$$d_{iB} = p_T^{2n} \quad (2.2)$$

If this value is below the threshold value $d_{iB}$ calculated from the $p_T$ of the particle $i$, the particle pair is merged into a new pseudo-particle. But if the calculated distance is above the threshold value, the particle $i$ is defined as a jet and removed from the input list. This approach is repeated until no further jet can be clustered.

After the jet clustering, the energy of the jets usually does not correspond exactly to the energy of the original gluon or quark. This is due to a number of reasons: On the one hand side, not all energy of the jet particles is deposited in the electromagnetic or hadronic calorimeter. On the other hand side, possibly not all particles originating from the primary scattered parton are clustered by the jet algorithm. But also particles not originating from the primary interaction are included into the jet clustering. The next chapter takes a more detailed look into these so-called pileup particles and their influence on the jet measurement. Furthermore, detector effects like dead pixels or other material defects can lead to a miss-measurement of the jet. Concluding all these possible distortions of the jet characteristics there is a clear need for calibrating jets for further analysis.
2.3.3 Pileup mitigation

Due to the increased instantaneous luminosity of the LHC a high amount of additional particles originating from scatterings beside the hard process is produced. The so-called pileup events and underlying events [15] influence the measurement of the jets in each event and lead to an additional offset on the measured jet energy. Beside the

![Diagram of jet energy distortion due to pileup events](image)

**Figure 2.3**: An event with two jets and additional particles which distort the jet energy. Additional collisions beside the primary vertex, so-called in-time pileup events, produce particles which are clustered into the jets. The beam remnants and multi-parton interactions lead to so-called underlying events which also leave energy deposits in the clustered jets. Finally, calorimeter entries remaining from a former bunch crossing, so-called out-of-time pileup events, distort the measured jet energy.

hard process of a proton-proton collision a high amount of additional proton-proton collisions in the same bunch crossing exist. These so-called in-time pileup events lead to additional energy in the reconstructed jets. As figure 2.3 demonstrates, the particles of additional vertices can be clustered into the jets of the hard process. Additionally, the beam remnants and multi-parton interactions can also produce energy deposits which are clustered into the jets. These so-called underlying events can lead also to a distortion of the jet energy. Furthermore, the energy deposits of the calorimeters have a cool-down phase, which can be higher than the collision frequency of 25ns. Therefore, the remaining energy deposits of the previous collision
are also clustered into the measured jet. This so-called out-of-time pileup also leads to a miss-measurement of the jet energy.

These three types of jet energy distortion by side effects of the hard process are summarized under the name pileup \([16]\) for the further analysis.

The jet energy distortion by these pileup events needs to be reduced for all analyses. In this calibration, the pileup mitigation is done via two different methods. Before applying the jet clustering, the Charged Hadron Subtraction (CHS) \([17]\) or the PileUp Per Particle Identification (PUPPI) \([18]\) is used to reduce the amount of pile up particles in jets. The CHS method uses the tracker information of the charged particles to reject pile up particles originating from vertices different to the primary vertex. However, PUPPI uses the power of a combined variable defined on a per particle basis do identify pile up objects and subtract them. Since CHS is in use for many years and therefore well checked, it is used as the standard method in CMS analyses. But also PUPPI is used in parallel to examine its advantage for the calibration of current CMS datasets with their increased amount of pileup events. The results of this analysis can be seen in chapter 4.5.2.

### 2.4 Jet energy calibration

As mentioned in the last chapter, all jets clustered by the jet algorithm need to be calibrated. This is done via a multi-layer approach by the CMS jet energy correction and resolution group. Combining the power of different teams from different institutions all over the world with various procedures stable and well-checked jet energy corrections are published for CMS analysis. In the following, the different steps of the calibration are outlined to get an overview of the whole process and to demonstrate the importance of the part done at the KIT especially during this master thesis.

#### 2.4.1 General approach

The jet energy correction is divided into four obligatory stages \([19, 20]\) with additional optional stages as shown in figure 2.4. All the four stages correct for different distortions in Monte Carlo simulations or measured data. The aim of the calibration approach is to get a factor for the correction of the raw jet four-momentum \(\mathbf{P}\), which is needed for later analyses.

\[
\mathbf{P}^{\text{corr}} = C \cdot \mathbf{P}^{\text{raw}} \\
= C_{\text{pileup}}(p_T^{\text{raw}}) \cdot C_{\text{response}}(p_T, \eta) \cdot C_{\text{residual}}(\eta) \cdot C_{\text{residual}}(p_T) \cdot \mathbf{P}^{\text{raw}}
\]  

\(2.3\)\( \quad \)\(2.4\)
This correction factor contains the factors of all the calibration steps: A pileup offset-correction, a Monte Carlo estimated response correction, a relative residual correction and an absolute residual correction.

At the beginning, an offset-correction is applied to correct for pile-up and electron noise of the detector. This offset is determined from a QCD dijet sample processed with and without pileup overlay to determine a correction factor. Additionally, a first residual correction (RC) compares the pileup profiles of Monte Carlo simulation and data. A data-driven sample is obtained by applying a random trigger to get a distribution of pileup events.

Afterward, the nonlinear calorimeter response and detector miscalibration are corrected by a simulation based approach. The Monte Carlo truth and simulated detector response are compared using a QCD dijet dataset. Correction factors are determined by comparing the $p_T$ and $\eta$ dependencies of the Monte Carlo truth with their simulated detector output.

At last, the residual corrections compare the Monte Carlo simulations directly with data. The relative residual corrections use a dijet dataset to link jets in the forward region to jets in the central region. Hence, corrections relative to the $\eta$ value of the jet are calculated. Finally, the absolute residual corrections compare a jet to a well-determined reference object. Therefore, a $\gamma +$ jets, a $Z +$ jets, and a multijet dataset are used to get an absolute correction factor of the jet $p_T$. All these channels use the well-defined central region of the CMS detector to get the needed well-measured reference objects. By combining the absolute residual corrections with the relative residual corrections, jets outside the central region are also well corrected. Hence, we get a proper calibration in the whole detector.

This thesis concentrates on the deployment of the absolute residual corrections using
the Z + jets channel done at the KIT. Hence, a deeper look into this part of the calibration is taken in the following section.

2.4.2 Absolute residual corrections

As described above, the absolute residual corrections compare the leading jet with a well-measured reference object. These two objects are balanced in the transverse plane due to the negligible transverse momentum of the LHC beams crossing inside the CMS detector. Hence, the transverse momentum of both objects is the same, but they are opposite orientated in the radial $\Phi$ plane. There are various possible reference objects with different benefits shown in table 2.1.

<table>
<thead>
<tr>
<th>Channel</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$ + jets</td>
<td>• covers high $p_T$ range</td>
</tr>
<tr>
<td></td>
<td>• deploys high statistics</td>
</tr>
<tr>
<td></td>
<td>• uses electromagnetic calorimeter to reduce jet energy scale uncertainties</td>
</tr>
<tr>
<td>$Z (\rightarrow \mu \mu)$ + jets</td>
<td>• suits perfectly for calibration at lower $p_T$ values</td>
</tr>
<tr>
<td></td>
<td>• is well-defined due to $Z$ boson mass requirement</td>
</tr>
<tr>
<td></td>
<td>• uses silicon tracker and muon system to reduce jet energy scale uncertainties</td>
</tr>
<tr>
<td>$Z (\rightarrow e e)$ + jets</td>
<td>• suits perfectly for calibration at lower $p_T$ values</td>
</tr>
<tr>
<td></td>
<td>• is well-defined due to $Z$ boson mass requirement</td>
</tr>
<tr>
<td></td>
<td>• uses silicon tracker and electromagnetic calorimeter to reduce jet energy scale uncertainties</td>
</tr>
<tr>
<td>Multijets</td>
<td>• suits perfectly for calibration at higher $p_T$ values</td>
</tr>
<tr>
<td></td>
<td>• allows additional comparison between forward and central detector region</td>
</tr>
<tr>
<td></td>
<td>• uses hadronic calorimeter to reduce jet energy scale uncertainties</td>
</tr>
</tbody>
</table>

Table 2.1: Overview of the different channels used for the absolute residual corrections.

The calibration workgroup at the KIT concentrates on the $Z +$ jets channel with the decay modes into two muons or into two electrons. Using the two leptons for the $Z$ boson reconstruction within a fixed mass region leads to a well-defined reference object. This $Z$ boson is reconstructed precisely at the low $p_T$ region. Hence, this channels suit
perfectly for a precise calibration of low $p_T$ jets. The muon decay channel allows to include the muon chambers of the CMS detector as an additional sub-detector into the balancing. This leads to a reduction of the jet response uncertainties. The electron decay channel brings additional statistics into the absolute residual corrections. Both channels use event selections to get the above mentioned ideal topology for the balancing. As shown in figure 2.5, a jet opposite to the reconstructed $Z$ boson is required to balance the $p_T$ values. The correction factors for the jet response are calculated using two different balancing methods described in chapter 4.1. Usually, events contain more than one jet. These sub-leading jets distort the balancing of the recoil object and the leading jet. To reduce their influence, an extrapolation to zero second-jet activity is done. This allows determining the jet response correction for the ideal event topology.

Finally, the jet response corrections of all the mentioned channels are combined using a global fit. This step combines the benefits of each channel to get a precise correction factor, which is independent of the single channel characteristics. This combination of the independent channels reduces the uncertainties of this jet energy correction factor further.

2.5 Computing environment in high energy physics

To perform an analysis like the $Z +$ jets calibration described above, a huge amount of computing power is needed. This is provided by connected computing centers all over the world, by national and regional computing centers and also local computing
resources at the institutes. Whereas the large computing Grids [24] are used by multiple experiments, the local resources as property of the institute are used by their members only. The following chapter will describe briefly the workflow requirements of the \( Z + \) jets calibration as well as what computing resources could be used for what kind of jobs. In general, the workflow approach of the \( Z + \) jets calibration is based on a generalized framework [25] developed by the KIT group. This allows processing the analysis within different steps requiring different computing requirements. Hence, different types of computing resources optimized for different analysis steps could be used to reduce the processing time.

### 2.5.1 Workflow requirements

The typical workflow of the \( Z + \) jets calibration consists of three different processing steps. Typically, these steps have different requirements regarding the computing resources. Important factors are the CPU usage, the memory usage and the network traffic produced by the applied tools. As figure 2.6 shows, the first step transforms the necessary datasets into a less storage-consuming data format for further analysis. The official dimuon and dielectron datasets are provided by various Grid storage systems [26] of LHC computing centers all over the world. Hence, a so-called skimming tool needs to locate the needed files, transforms them into a more user-friendly and less storage-filling format and copies them to a working directory a lot easier accessible.
by end-users. Therefore, worker nodes with a huge bandwidth to the LHC computing centers are needed to handle the selection step. So-called Tier-2 and Tier-3 resources which process user jobs on remote worker nodes meeting all the necessary requirements. To optimize the processing, the skimming is done in parallel using multiple worker nodes at once. Typically, few thousand jobs are started at the same time. Each job processes a small part of the huge number of input files. Afterward, the output files are usually copied to local computing resources for further processing.

In a next step, these files are reprocessed to apply selection cuts, energy corrections, luminosity scaling and necessary reweighting steps. For every change of the selection or correction, this step needs to be repeated. Since this processing step is usually input-output limited, it requires the flexibility to iterate quickly over the same dataset with different settings. Therefore, the processing machines need fast access to the datasets stored on storage servers due to their large size. In general, the datasets need to be reprocessed at least every two weeks to further improve the jet energy corrections by using the most recent datasets.

Finally, a plotting tool is used to determine the correction factors. This includes, for example, the calculation of the balancing methods, the extrapolation to zero second-jet activity or the cross-check plots. Usually, this is done on local server machines with direct user access via secure shell (ssh). This step only requires access to the output files of the second step and a small amount of CPU, memory and storage for the final graphical outputs or ROOT [27] files.

### 2.5.2 Computing resources in use

The Worldwide LHC Computing Grid (WLCG) [26, 28] consists of computing nodes and storage elements [29]. These are all managed in three hierarchical layers the so-called tiers as shown in figure 2.7. The Tier 0 centers located at CERN in Geneva, Switzerland, and Budapest, Hungary collect the raw data of the detectors. Furthermore, it is responsible for a first reconstruction and the distribution of the datasets to the Tier 1 data centers. During LHC shutdowns it is also used for the reprocessing of the datasets. The thirteen Tier 1 centers store one shared copy of the raw datasets and provide reconstructed and raw datasets to all scientific users. It is also used for large-scale reprocessing and distribution of the datasets to the Tier 2 centers. Since this two important layers concentrate on the data provisioning, they are only partially usable for end-users. A required dataset provided by the Tier-1 centers can easily be accessed in read-only mode for processing using the protocols Xrootd [31, 32] and SRM [33]. Only a small part of the Tier 1 computing resources can be used by end-users for analysis processing. The Tier 2 layer consists of more than 160 computing resources of universities or other scientific institutions.
Figure 2.7: The hierarchical structure of the Worldwide LHC Computing Grid [30] including institutional resources allows to process different types of purposes restricted to different layers.

Access to the different computing centers is possible for LHC users all over the world. These computing possibilities are extended by local arrangements of the participating institutions which are summarized in the Tier 3 layer. There, also the access of end-users is limited to the members of the corresponding institution.

For this analysis, the Tier 1 and Tier 3 resources were mainly used for the data processing. The computing nodes of the Tier 1 GridKa at KIT [34] and the NEMO Cluster in Freiburg [35, 36] were used for the skimming and preparation step. At GridKa Tier 1, the number of cores used for processing of a user job depends on the fair share between the different experiments and the users. During the processing of this thesis, a maximum of about 1600 cores has been used at the same time. The Freiburg NEMO Cluster as a Tier 3 resource of the EKP at KIT provided a test cluster with 800 cores in the first place. After the launch of the production cluster [37], the number of physical cores increased to 15000 cores, hyper-threading is currently not enabled. The high bandwidth to the Tier 1 Grid-enabled storage systems [29] allows fast access to the required dataset for processing. This leads to a highly parallel and fast processing of the skimming and preparation step.

The step of selection and correction was processed locally using four computing nodes optimized for fast access to the datasets. To get a high input rate for the processing a caching algorithm was developed by the EKP workgroup [38, 39]. This coordinated caching provides copies of files from file servers on local high-speed solid state drive (SSD) storage and adds the data locality to the local batch system. Using this setup
the iterated processing of the second layer is delivered in a short time.

Finally, the login nodes are used to calculate the jet energy correction and plot histograms for cross-checks. These login nodes are regular servers with Scientific Linux 6 [40, 41] as operating system and provide interactive logins for the users via ssh. This allows the user to plot and process the small sized tuples derived from the skimming and selection workflow.
Chapter 3

High-throughput computing in high energy physics

This chapter concentrates on the hardware and software setups used for the first two steps of the processing approach. Since the absolute residual corrections are calculated event by event without any cross correlation, the processing can be done highly parallel. This allows processing the high amount of data more effective using the combined power of many compute nodes. Therefore, the processing is split into so-called jobs which are completely separated from each other. Each one handles a certain amount of the data using the calibration tool. An HTCondor batch system [42, 43] located at the institute is used to schedule the execution of the programs at so-called worker nodes. HTCondor distributes the jobs to the specified machines, controls the progress of each job and informs the user about the processing progress. The job itself fetches the required data and writes the output back to a specific output storage. Within the submission of the jobs, machines specialized for high-throughput computing are chosen. These support a coordinated SSD caching system named High-Throughput Data Analysis (HTDA) [38] developed by Max Fischer [39] for former calibration processing. To process the large size of the 2016 datasets, this caching system was tested and extended during this master thesis. Furthermore, a distributed object store [44], Ceph [45], was installed on the same machines. This helps to get the large amount of data stored as near as possible to the worker nodes to get a fast access speed.

3.1 HTCondor as batch system

As described above, an HTCondor batch system is used to manage the execution of jobs at the Institut für Experimentelle Kernphysik of the KIT. These jobs contain typical analysis workflows like the calibration workflow described above. To satisfy
the requirements of jobs which use certain tools and access specific data pools, all jobs are executed in a defined software environment. This environment is an individually configured and extended Scientific Linux 6 (SL6) [40, 41], which need to be deployed to all login nodes and worker nodes.

To understand the advantages of the whole setup for high-throughput analyses, a detailed insight into the HTCondor batch system is needed.

### 3.1.1 Functionality of HTCondor

The HTCondor setup [46] is divided into three management parts: A login node is used to submit jobs, worker nodes execute the jobs and a central manager conveys between both. A user is able to use a login node to develop and test tools and programs directly or submit these programs to the HTCondor scheduler for batch system execution. This alerts the central manager about the new waiting jobs. The central manager also collects information about the worker nodes including information about the load of the machines, and available free slots to execute jobs. On the side of the worker nodes, the HTCondor setup provides so-called job slots which allocate parts of the machine resources like a given number of CPU cores or a certain amount of memory. Using all information, the central manager tries to match the waiting jobs at the login node to the free slots at the worker nodes providing the proper resources. The allocation is reported to the HTCondor instance at the login node.

**Figure 3.1:** HTCondor handles the process of job submission at the login node, transfer via a Central manager and execution at worker node. Furthermore, a HTDA manager collects information about needed files and coordinates the SSD caching.
3.1 HTCondor as batch system

node which claims the needed resources and transmits the job to the worker node. This corresponding worker node starts the execution process within the predefined software environment and sends monitoring information back to the login node.

On the one hand, the approach of the HTCondor batch system allows a flexible setup of worker nodes and the integration of remote sites. For example, this was utilized for a dynamic integration of remote worker nodes at the NEMO cluster in Freiburg [47, 48]. On the other hand, it also allows adapting various extensions to the HTCondor system. Interactions with the condor scheduler provide the possibility to extract information about the job like the datasets needed for the job execution. This allows transmitting information of the needed files for the job execution to the HTDA manager. The HTDA manager coordinates the caching of the files on the SSDs of the worker nodes. Furthermore, it matches the jobs to the worker node, which has already most of the files needed for the processing cached. This is done via an automatic manipulation of the HTCondor job matching to route specific computing tasks to optimized machine types like one of the HTDA nodes.

3.1.2 Local HTCondor setup using docker containers

The local HTCondor setup at the institute uses different setups to provide various batch system slots suitable for different job requirements. Since not all worker nodes provide the needed fully configured Scientific Linux 6 setup directly, abstraction layers are needed to support full compatibility. Machine types with the same configuration are summarized within a so-called cloud site. Whereas the nodes at the Nemo cluster in Freiburg [47, 48] boot virtual machines with a fully configured Scientific Linux 6 via an OpenStack management [49], other machines like the Desktop cluster [50] and the HTDA nodes [38, 39] use docker container [51] to satisfy the requirements of the software environment.

During this master thesis, the High Throughput Data Analysis system was migrated from a direct Scientific Linux 6 setup to a CentOS 7 [52] setup. This approach avoids some compatibility and stability issues concerning AUFS [53], an overlay file system used for the SSD caching, and the conservative Scientific Linux 6 kernel. Newer kernel versions like the ones deployed by the operating system CentOS 7 support the full features of the HTDA in a stable environment. The new operating system was chosen because Scientific Linux and CentOS are part of the Redhat Linux community, which guarantees a similar provisioning of software. To support the full computing environment of the analysis workflows, the HTCondor jobs have to be executed within an SL6 environment. Since a direct support of docker container via HTCondor was announced at the HTCondor week 2015 for versions 8.3.6 and above [54], this technology was implemented into the local setup to provide the SL6 environment.
3 High-throughput computing in high energy physics

<table>
<thead>
<tr>
<th>Cloud site</th>
<th>Configuration</th>
</tr>
</thead>
</table>
| Nemo cluster\ Freiburg [47, 48]     | • use an OpenStack manager [49] to boot dynamically virtual machines as needed via a batch system for the purpose of accounting the resource utilization  
• virtual machines contain fully configured Scientific Linux 6 setup |
| Desktop cloud [50]                  | • use free resources on desktop computers with the operating system Ubuntu  
• uses docker containers [51] to provide a Scientific Linux 6 setup |
| High Throughput nodes [38, 39]      | • use a CentOS 7 setup to provide SSD caching via HTDA for faster job execution  
• uses also docker containers [51] to provide a Scientific Linux 6 setup |

Table 3.1: Overview of HTCondor cloud sites at the institute.

for the jobs. Tests on machines of the desktop cloud showed that virtual machines produce an enormous amount of CPU overhead and reserve an enormous amount of memory from the beginning without direct usage. The new technology of docker container provides the same execution environment without the CPU overhead and the memory reservation. The performance of Docker containers was tested extensively and compared with the performance of virtual machines in the context of a master thesis [50]. Docker containers are a lightweight approach of offering the environment of a specific operating system. They execute the specified software within an isolated and encapsulated process. This process uses the Linux kernel of the host system and adds the additional libraries of the required operating system to provide the software environment. The encapsulation of the process consumes negligible overhead of CPU due to the usage of the same Linux kernel, which is already loaded. Since the process loads only the libraries which are needed, the memory reservation is reduced to a minimum. These advantages make docker containers optimal for the provisioning of a high energy physics environment on worker nodes.

The HTCondor setup was tested excessively with all kind of jobs including CPU intensive or high throughput jobs. Within this testing, all results of this master thesis used the docker technology to process the step of selection and correction of the analysis. There were no failures or other issues, which are produced by the usage of
3.2 High Throughput Data Analysis utilizing coordinated caching for batch system usage

docker containers. The HTCondor setup combined with docker worked as expected and offered the processing possibilities to other particle physics communities located at the institute like the AMS [55] workgroup. Each group can process their tools within different software environments provided by different docker containers. This makes the HTCondor approach with the usage of docker containers a powerful setup for even larger scales for different science communities.

3.2 High Throughput Data Analysis utilizing coordinated caching for batch system usage

This section concentrates on the advantages of the HTDA setup to the calculation of the absolute residual corrections. As mentioned above, the setup was improved and extended during this master thesis. Furthermore, it was tested excessively and iteratively using the large amount of storage needed for the datasets of LHC run 2. To understand the advantages of the HTDA caching for the dataset processing, the functionality of the SSD caching system needs to be understood first.

3.2.1 Functionality and extension of HTDA

The HTDA system collects information about processed files to be cached on SSDs for further processing. This is done via an extension of the HTCondor scheduler which interacts with a database located at the HTDA manager. During this step, the description of a job which will be executed at the HTDA worker nodes is scanned for dataset files needed for the processing. Information about the needed files is gathered by the HTDA manager and stored in a SQL database. The HTDA manager itself locates already cached files and sends information about the most suitable machine slot to the HTCondor scheduler. Afterwards, HTCondor assigns the waiting job to a corresponding machine slot. Furthermore, the HTDA manager also schedules the caching of new files suitable for repeated job execution on the same data.

On the side of the worker nodes, an overlay file system called AUFS transparently provides the files stored at a remote storage server as well as the cached files on an SSD in a single file system. The HTDA client gets instructions from the HTDA manager to store new files on SSD or removes old ones.

3.2.2 Advantages of HTDA for the calibration approach

As mentioned above, the whole setup was migrated to a more stable CentOS 7 operating system in the context of this master thesis. During this step, the experimental
setup was integrated into the service management of CentOS. This includes the integration of the configurations, scripts, and logging into a common Linux system service setup. The next step is the migration of the HTDA management into a virtual service machine or docker container. This allows an integration of the experimental HTDA setup into the common administrative workflow at the institute. All the migration steps aim at providing an SSD caching setup which is easy to install, manage and control.

As mentioned above, the migration step required the usage of docker containers to provide the correct computing environment of high energy physics. As figure 3.2 shows, the new technology was also tested in combination with the HTDA setup. For HTCondor jobs running on the bare operating system of the machine, the storage servers are mounted via the NFS protocol and combined via an overlay file system with the SSD caches. This configuration works also for docker containers, which can access the overlay file system AUFS directly. Hence, the migration to docker containers is also supported by the HTDA caching approach. Tests show that there are no restrictions accessing the HTDA caches from inside docker containers compared to direct access.

Beside this migration, some additional extensions are applied to the local setup. Since the size of the datasets has increased dramatically during LHC run 2 and will increase further for future data taking, the caching has been expanded. The former 512GB SSDs replaced by new 1TB SSDs could serve a doubled storage capacity for caching. Therefore, a total size of 4TB can be cached which corresponds to nearly half of the

![Figure 3.2: HTDA](image_url)
3.3 CephFS as extension of the caching

datasets used for the processing of the absolute residual corrections. Tests of the total read rate of the current HTDA setup described in the following section 3.3.2 confirm the high input rates of the SSD caching. Additionally, a setup with two SSD caches per worker nodes was tested and can be installed to reach doubled performance for future processing of analyses. Furthermore, the integration of special SSDs directly connected via the so-called PCI-Express standard to the main board can reach a read speed up to 2.5 GB/s per SSD which would increase the current read speed by a factor of 5. For that case, possible limits of the analysis software, the CPU speed and the memory need to be checked.

Since the current SSD setup is adapted to the number of CPU cores, the CPU speed, and the installed memory, it provides enormous improvements for high throughput computing. Therefore, the processing time of the selections and corrections step is further reduced from about 5 hours to one hour using the SSD caching. Possible extensions may be necessary in the future to make the local HTDA instance ready for coming data processing which need to handle constantly increasing dataset sizes.

3.3 CephFS as extension of the caching

The high amount of data presents also the storage capacity with huge challenges. During processing, the analysis software needs to have fast access to the datasets which means high write rates and especially high read rates. As the introductions to the analysis software described above, the complete dataset used for the results of this thesis for the electron and muon decay channel consists of around 7.5 TB need to be stored after the skimming level. This amount of data increases further at the end of 2016 and will also increase by factors during the next years. Hence, new possibilities to store data on larger scales for local processing at the institute need to be found. The hardware and software configuration of the HTDA worker nodes offers the possibility of testing a distributed file system.

At the beginning, GlusterFS [56] was chosen for testing. This distributed file system provides high read and write rates [57] and an easy to handle setup and configuration. During the installation, the major part of the settings is chosen automatically. This gives the administrator assistance on the one hand, but also limits the possibilities of the setting. There, the administrator can decide between some pre-configured settings and precise adjustments of the GlusterFS to the hardware setup. This includes, for example, the distribution of replicas on the hard drives and the nodes. Some extended tests showed some compatibility issues with the local HTDA setup. These issues include some incompatibilities [58] between GlusterFS and the overlay file system AUFS used by the co-ordinated SSD caches. By now, there are no patches or other updates available to solve these issues. Hence, this distributed file system
In addition to GlusterFS, another distributed file system named Ceph was tested. This distributed file system is based on so-called object storage which manages the stored data in unstructured objects. Each stored file is divided into small parts, which are distributed to all worker nodes providing storage. These objects containing metadata and data information are identified by global unique identifiers. In contrast to a usual file system, an object storage has no hierarchical structure what allows providing massive amounts of inexpensive, highly scalable, and partially self-healing storage. Ceph as an object storage also provides high scalability up to very large setups and the possibility of a precise configuration. The installation of Ceph is not as easy as GlusterFS but deploys a large amount of settings to configure. This includes the data management, monitoring or data access. To understand the advantages of a distributed file system, a detailed look into the local configuration of the Ceph cluster is done in the following section.

3.3.1 Structure of the local Ceph file system

The local Ceph installation consists of the four HTDA worker nodes with 10 times 4 TB hard disks. As figure 3.3 shows, each of the 10 hard drives provides a so-called object storage device (OSD) which organizes the storage of data into containers.
called objects. Within the cluster of OSDs provided by 10 hard disc drives (HDDs) installed at four worker nodes, an arbitrary number of Ceph pools can be created. In this test case, a metadata and a data pool were created and merged to a Ceph file system. Whereas the metadata pool stores the metadata of the Ceph file system with three replicas, the data pool contains the main storage for storing the files with two replicas. The number of replicas is configured manually and can be changed under full operation. All metadata is managed by a metadata server (MDS) which is also responsible for storage interface compatibility. An inactive backup server at a second worker node was installed at the local setup. This takes over if the first metadata server crashes or goes out of order. Furthermore, three monitoring servers (MON) manage the mapping of the whole file system and control the cluster state. They also deploy access points to the Ceph file system for external machines.

Another important feature of the Ceph file system is the replacement of a HDD under operation. Using the Ceph tool to mark a Ceph OSD for removal triggers the Ceph management to copy all data located at this Ceph OSD to the residual Ceph file system. This step is possible if the HDD is still working correctly and the remaining storage capacity is big enough to store all data. Afterwards, the HDD can be removed completely from the Ceph setup or replaced by a new one. The case of a corrupted HDD can be covered by using replicas within the Ceph file system. Then, the lost data is recovered from the replica parts distributed over all remaining Ceph OSDs. Testing this Ceph features showed an easy handling for administrators without any troubles. The main part of the replacement or recovering process is done automatically by the Ceph management. This feature contributes to the stable behavior of the local setup.

### 3.3.2 Advantages of the Ceph setup for the SSD caching approach

For the local setup, the integration of the Ceph file system into the HTDA system was tested. Hence, the Ceph file system was mounted on each node into the overlay file system of the SSD caching. Tests show that this distributed file system shows no compatibility issues with the overlay file system. This allows testing the advantages of the distributed object store compared to a common network storage. During these tests, the advantages of the SSD caching are tested additionally.

Looking at the Ceph setup without activated HTDA, the pure Ceph file system is tested for stability and performance. For that purpose, the complete read rate of the Ceph file system derived from all four worker nodes in parallel is measured. This read rate is compared to the complete read rate of all four HTDA caches and the read speed of all four worker nodes accessing a network storage. The resulting read rates are measured with deactivated caches and buffers. For a simulation of dataset
processing, ten files, each one with 10GB of data, are accessed by different streams in parallel. To check the stability of the three different scenarios for parallel processing in the HTCondor batch system, the read rates are measured for different numbers of parallel streams. This allows simulating a parallel processing of batch jobs with currently maximal 84 job slots which are all accessing the needed datasets. The results of this read tests are illustrated in figure 3.4.

Looking at the values of the Ceph file system, an interesting trend is observed. This file system shows a stable read rate even for highly parallel accesses. While this behavior is expected for SSDs, the read rate of an HDD usually suffers from highly parallel file accesses. Measuring the network traffic of a worker node during read accesses shows that the maximum transfer rate of the network is reached. The network bandwidth of the machines providing Ceph needs to be shared between the metadata server, monitoring servers and the Ceph file system itself. Since all stored objects are distributed over all worker nodes, accessing a file needs accesses to all corresponding OSDs on all worker nodes. This effects a huge increase of the network traffic. Since enough HDDs are integrated into the Ceph storage system, the restricted read rate of the HDDs is not the limiting factor anymore. Therefore, the local Ceph setup is directly limited by the network infrastructure. Further improvements can only be reached by replacing the network infrastructure and installing enough HDDs to provide the necessary combined transfer rate. During read and write accesses, only a small CPU overhead produced by the Ceph management is observed. Since

**Figure 3.4:** Tested absolute read rates of the setup. For comparison reasons and applicability for batch system usage, the complete input rates of all four worker nodes are measured for different numbers of parallel streams.
only 75% all hyper-threaded cores are usually allocated for batch job processing, there are enough free resources for system and management tasks. Hence, the Ceph management does not influence the processing of the analysis software.

Looking at the HTDA setup, each of the four worker nodes is currently equipped with one SSD for caching. The SSD caching shows a stable read rate independent of the number of streams which enables stable high input rates for batch job processing. The current setup with four SSDs in total provides the highest input rates compared to the Ceph file system and the network storage. This setup provides enough transfer rate for fast reprocessing of the calibration approach. Maybe, in the future, the increasing sizes of the needed datasets will make an extension of the HTDA setup with second SSDs per worker node necessary to get acceptable processing times. Testing also the read speed of setup using two SSDs per worker node shows that in this case the summarized input rate of the worker nodes is doubled. Hence, the HTDA setup provides the best chance for scaling up.

The network storage itself provides only low read rates which clearly indicates the need of a caching approach like the HTDA setup. Here, also a Ceph file system would improve the storage access for users.

The maximum performance can be achieved by using the Ceph file system as storage system for the datasets needed for processing the calibration workflow. Activating the coordinated caching allows caching half of the datasets. During processing, the cached files are accessed using the full read speed of the SSDs. For all uncached files, the Ceph file system provides higher read rates compared to common network storage servers. This allows combining the optimized read rates of the HTDA setup and the Ceph file system which leads to very high input rates for processing. Tests with the calibration datasets stored on the Ceph file system and on the SSD caches via HTDA showed a further reduction of the processing time by a factor of two. This allows processing the 7.5TB of data in about 30 minutes. Hence, new correction results are provided within a short time after data taking.

Usually, the datasets are copied sequentially from remote sites to local network storage servers. Therefore, the write rate of the local Ceph file system is tested. Here, a file with 10GB of data is written in one stream to the storage possibilities. The results of the test are shown in table 3.2 The measured read speed of the storage providers was significantly above their write speed. Whereas the read rate of the Ceph file system is limited by the network infrastructure, the measured write rate of \((340.0 \pm 7.1)\,\text{MB/s}\) is clearly below this limit. Since a single HDD was measured to have a transfer rate of \((186.5 \pm 0.5)\,\text{MB/s}\) in single steam write and read tests, the write rate should not differ from the read rate. Further tests showed that the limiting factor is the management overhead of the Ceph file system which leads to a reduction of the possible transfer rate.
Comparing the Ceph file system transfer rates with the access speed of two common storage servers located at the institute, a significant increase of the read and write speed is achieved. Hence, a Ceph file system which uses the combined power of all included machines outperforms the common network storage servers. To profit from these benefits, a Ceph configuration needs to be taken into account for further purchases of storage server hardware. Furthermore, a second layer of caching could be implemented in the future to also provide fast transfer rates for files which are not cached due to the limited storage capacity of the SSDs.

Overall, the Ceph file system runs stable with extensions like the HTDA setup. In our case, the limiting factor of the Ceph file system is the network connection in combination with the Ceph management overhead. However, comparing the limitations with common network storage servers, an enormous increase of read and write speed can be observed. A combination of the very fast accessible HTDA caching with a fast accessible Ceph layer provides an optimized environment for high throughput computing.

### 3.4 Future benefits of an expanded caching

Combining the power of the HTDA caching with docker containers and a Ceph file system enables a performance boosted setup for high-throughput computing. This enables a fast and efficient way of processing large datasets. At the LHC, the sizes of the recorded datasets increase constantly. Other particle physics experiments see the same trends to larger datasets. Hence, systems like the coordinated caching on SSDs via the HTDA setup are absolutely necessary for future analysis purposes.

The currently used HTDA setup is flexible enough to serve also users of other experiments. Integrating of docker containers into the HTDA setup was an important step to provide this new possibility. Each scientific group can create its own docker container, which contains their software environment. For example, the AMS workgroup located at the local institute is now able to run their jobs on the same machine setup.
as the CMS users. In this case, they can also cache their datasets on the SSDs using the HTDA approach.

Extensive tests with the coordinated caching in a stable environment showed very high and stable throughput rates combined with large reductions of the processing time. This enables fast processing of analysis workflows and prompt provisioning of results like done for the absolute residual jet energy corrections. Furthermore, the integration of the HTDA setup into common administrative workflows makes the transfer to other machine setups easier. This includes the expansion of the SSD caching by adding new worker nodes or providing a completely new setup somewhere else. Future studies will show whether this setup can also be easily transferred to the virtual machines at the Nemo cluster in Freiburg easily.

The expansion of the caching setup using a distributed file system provides new possibilities. The testing of Ceph as an example of a highly configurable distributed file system done during this thesis show a stable and performant behavior. These characteristics can also be used for future approaches of storing a massive amount of data in particle physics on local machines. A second caching layer could be realized in the future by using this distributed file systems to expand the SSD caching with fast accessible HDDs further. Since SSD storage is still cost-intensive and restricted in storage capacity, this approach provides a cheap possibility of expanding the cache. Although the read rate of the Ceph file system is lower than the SSDs of the HTDA setup, a combination of both tools combines the power of both and provides further optimized input rates for processing.

Summarizing the advantages of all the computing approaches tested in the context of this thesis, we have a highly performant setup. This setup provides the possibility to process the large datasets needed for the $Z + \text{jets}$ calibration iteratively and with short turnaround times.
Chapter 4

Absolute residual corrections using $Z + \text{jet}$ events

As described in chapter 2.4.2, the absolute residual corrections use the combined corrections provided by different channels. The Karlsruhe CMS group concentrates on the $Z + \text{jets}$ channel with the electron and muon $Z$ boson decay modes. As part of the calibration process done at the KIT [59], this thesis uses the 2016 datasets for calibration purposes. This chapter includes the jet response extrapolation additionally to studies for further improvements and extensions of the calibration.

4.1 Balancing methods

In the calibration process, two different jet response estimators are used, which are sensitive to different effects. These balancing methods compare a jet to a $Z$ boson as a reference object. Both methods respect different event characteristics to balance. The $p_T$ balance method compares the $p_T$ of the $Z$ boson to the $p_T$ of the jet, whereas the Missing $E_T$ Projection Fraction method takes into account the whole event environment. Comparing both response estimators allows detecting points of failure more easily. The balancing approach is done for Monte Carlo simulation and data separately to determine their different behaviors. After the balancing, an extrapolation to the optimal back-to-back topology is done.
4 Absolute residual corrections using \(Z + \text{jet}\) events

4.1.1 Transverse momenta balance

This balance method compares the transverse momentum of the leading jet to the transverse momentum of the reference object to get the detector response \(R\).

\[
R_{\text{obs}}^{\text{leading jet}} \left( p_{T}^{Z} \right) = \frac{p_{T}^{\text{leading jet}}}{p_{T}^{Z}} = R_{p_{T}}
\]  

(4.1)

In this case, the reference object corresponds to the \(Z\) boson determined from the reconstructed two muons or two electrons.

4.1.2 Missing Transverse Energy Projection Fraction

The MPF not only takes into account the \(p_{T}\) values of the reference object and the leading jet, but also the whole event environment. Ideally, the transverse momenta of the two objects are perfectly balanced and the transverse momenta neutralize each other.

\[
\vec{p}_{T}^{Z} + \vec{p}_{T}^{\text{leading jet}} = 0
\]  

(4.2)

For reconstructed objects on detector level, these transverse momenta are scaled by the detector responses \(R_{Z}\) and \(R_{\text{leading jet}}\). In this approach, the missing transverse energy including deviations of the detector responses is taken into account.

\[
R_{\text{obs}}^{Z} \cdot p_{T}^{Z} + R_{\text{obs}}^{\text{leading jet}} \cdot p_{T}^{\text{leading jet}} = -E_{T}^{\text{miss}}
\]  

(4.3)

The \(Z\) boson is well-described by the precise measurement of the leptons. Hence, the \(Z\) boson does not contribute to the missing energy with \(R_{Z} = 1\) and the detector response can be determined by:

\[
R_{\text{obs}}^{\text{leading jet}} = 1 + \frac{E_{T}^{\text{miss}} \cdot p_{T}^{Z}}{\left( p_{T}^{Z} \right)^{2}} = R_{MPF}
\]  

(4.4)

Both response methods are susceptible to pile up contributions. But in contrast to the \(p_{T}\) balance method, the MPF method is not biased by final state radiation and various systematic biases. Hence, this MPF method is the usual CMS response definition, which is compared to the \(p_{T}\) balance response to recognize mis-calibration as soon as possible.

4.2 Event selection

This master thesis is based on parts of the 2016 datasets of the CMS detector with \(\sqrt{s} = 13\) TeV. In detail, dimuon and dielectron samples with 27.22 fb\(^{-1}\) integrated
luminosity were available for the jet energy response calibration done in this thesis. Furthermore, a dilepton Monte Carlo sample is processed for the simulation comparison using the same applied event selection. Kinematic cuts are applied to get a clean dataset with less amount of background events. Due to the fact that two very similar channels, $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$, are used, the used selections differ only at some points. The applied selection cuts are aligned to trigger-thresholds and the $Z$ boson characteristics.

4.2.1 Muon and electron selection for $Z$ boson reconstruction

Both event selections are based on a high-level trigger (HLT) with additional soft cuts. In this case, the high-level triggers select two leptons with different $p_T$ thresholds. The $Z \rightarrow \mu\mu$ channel uses a trigger stream with minimum $p_T$ cuts at 17 GeV and 8 GeV for the leading and second leading muon (HLT_Mu17_Mu8 [60]). Analogously, the $Z \rightarrow ee$ selection is based on a trigger stream with minimum $p_T$ cuts at 17 GeV and 12 GeV the leading and second leading electron (HLT_Ele17_Ele12 [60]).

Furthermore, the following cuts are applied for the muon or electron selection.

**Muon selection**

- tight muon ID [61, 62]
- tight muon isolation [63]
- $2 \leq N_\mu \leq 3$
- $p_T^\mu > 20$ GeV
- $\eta^\mu < 2.3$

**Electron selection**

- tight electron ID [64]
- $2 \leq N_e \leq 3$
- $p_T^e > 25$ GeV
- $\eta^e < 2.4$

To apply additional lepton quality requirements, this analysis requires tight muon IDs [61, 62] or electron IDs [64] which select a subset of the particle-flow muons or electrons using a predefined set of characteristics. CMS studies showed an efficiency of more than 96% for muons with a $p_T$ above 10 GeV applying the tight muon id [65]. Analogously, an efficiency of more than 80% was observed for electrons with a $p_T$ above 40 GeV applying the tight electron ID [66]. Furthermore, the muon selection also requires an isolation criterion [63] to reject muons which are embedded in jets. Isolation criteria are not applied to the electron channel since cross-checks show they reduce the statistics dramatically which leads to large uncertainties. To reduce the amount of mis-measured leptons and the influence of trigger thresholds further, minimum $p_T$ cuts are applied. Additionally, muons or electrons are selected from the $\eta$ range of the silicon tracker, and the electromagnetic calorimeter. The reconstruction
uses these lepton pairs to determine the characteristics of the required Z boson. To get well-reconstructed Z boson events, some additional cuts are required.

- $p_T^Z > 30 \text{ GeV}$
- $|m_{Z}^{\text{reconstructed}} - 91 \text{ GeV}| < 20 \text{ GeV}$

The signal of two leptons can also be produced via a photon or a Z boson or other background processes described in chapter 4.2.7. Requiring a reconstructed Z boson mass close to the theoretical value mainly accepts signal events. Hence, the background processes which would be dominant for lower reconstructed $m_Z$ are suppressed.

Additional to the lepton quality cuts, some jet and Z topology cuts are applied as described below.

### 4.2.2 Jet selection

After the pileup subtraction done via the Charged Hadron Subtraction or PileUp Per Particle ID described in chapter 2.3.3 and the jet clustering described in chapter 2.3.2, the jet energy corrections [67] are applied usually. Due to the calculation of the absolute residual corrections within this analysis, these correction factors are excluded.

In Addition, some quality cuts for the jets are applied as listed below.

- Loose JetID [12, 13]
- $N^{\text{jets}} > 0$
- $p_T^{\text{leading jet}} > 12 \text{ GeV}$
- $|\eta^{\text{leading jet}}| < 1.3$

First, at least one jet with a loose jet ID [12, 13] as cut-based jet quality criteria is required. Here, a selection efficiency of 99.82% with a fake rate of 7.97% was denoted [68]. The jets are categorized by their $p_T$ values which means that the jet with the highest $p_T$ is chosen as the leading jet. A minimum $p_T$ cut is applied to reduce the amount of faked jets mainly located in the lower $p_T$ region. The valid region of the jets is chosen due to the layout of the hadronic calorimeter. Here only the barrel region is used to get well-measured jets.

### 4.2.3 Z + jets topology selection

Finally, some cuts are necessary to guarantee the expedience of the events for the balancing of the jet and the Z boson. Therefore, the Z boson should be spatially opposite to the leading jet. Furthermore, the $p_T$ of the sub-leading jets should be less than 30% of the Z boson $p_T$. 

36
4.2 Event selection

- $|\Delta \eta_{\text{leading jet, } Z - \pi}| < 0.34$
- $\frac{p_T^{\text{second jet}}}{p_T^{Z}} < 0.3$

These cuts lead to a dataset suitable for the balancing. Furthermore, this approach reduces the amount of background events as described later in chapter 4.2.7.

4.2.4 Event re-weighting

The massive increase of the instantaneous luminosity in 2016 leads to a large amount of pile up events as described above. Since the exact distribution of pile up events can only be determined during data taking, the Monte Carlo prediction has to be reweighted.

A direct measurement of the pileup contribution is done via the number of additional pile-up interactions $n_{PU}$. This distribution can be derived from the instantaneous luminosity measurement during the LHC run. For Monte Carlo simulation, the distribution was defined at the generation. This calibration uses the official CMS recommendation for pileup estimation [69].

An indirect measurement of the pile up contribution can be done via the total number of primary vertices $n_{PV}$ for each event. These values are determined after the full reconstruction of the events and before applying any quality cuts. Figure 4.1 shows the $n_{PV}$ distribution after the pile up reweighting. The jet energy corrections and resolutions group discussed the influence of the observed large deviations between data and Monte Carlo simulation. Since the cross-check plots in the figures A.21 A.22 show no significant $n_{PV}$ dependency of the jet response, the deviations can be neglected.

![Figure 4.1: $n_{PV}$ distribution after pile up reweighting and $E_T^{\text{miss}}$ distribution for $Z \rightarrow \mu\mu$.](image)
4 Absolute residual corrections using $Z + \text{jet}$ events

4.2.5 $E_T^{\text{miss}}$ Event filter

To reduce the amount of mis-measured events and to get a cleaner dataset for the calibration, a selection depending on $E_T^{\text{miss}}$ can be applied. Events with a high $E_T^{\text{miss}}$ can be produced when several parts of the event reconstruction fails which can be checked by looking at the functionality status of the corresponding detector parts. Hence, a rejection of these events should lead to a clear and well-measured $Z + \text{jet}$ dataset. Looking at the $E_T^{\text{miss}}$ distributions shown in figure 4.1, the amount of events with high $E_T^{\text{miss}}$ values is negligible. Therefore, we currently do not apply an additional $E_T^{\text{miss}}$ filter, but we will switch to the official requirements for an $E_T^{\text{miss}}$ cut after they are published for the 2016 runs.

4.2.6 Muon and electron corrections

Muons are precisely reconstructed by combining the information of the inner tracker and the outer muon chambers. Therefore, it is not explicitly necessary to use muon corrections for this calibration step. The reconstruction of electrons considers the inner tracker and the electromagnetic calorimeter. Hence, electrons are not measured as precisely as muons due to a more difficult extraction of the signal from the residual event. For the current calibration, no electron corrections are applied either. Plots in chapter 4.3.2 show that both channels suit perfectly for the balancing approach. So the muon and electron corrections are not absolutely necessary. Further studies will show whether an improvement can be achieved by applying these corrections.

4.2.7 Background estimation

Looking at the possible backgrounds of the used $Z + \text{jets}$ channel, there are mainly the W boson production, the $t\bar{t}$ production, the diboson production WW and WZ, background via QCD processes and the Drell-Yan-background via photons. The corresponding cross sections shown in table A.1 of the appendix for a center-of-mass energy of $\sqrt{s} = 13$ TeV are illustrated in figure 4.2. All these processes can lead to two leptons and additional jets in the final state. These so-called fake particles do not originate from the signal process, but imitate its signatures. The W boson production as the process with the highest cross section is only taken into account if an additional fake lepton is reconstructed. This reduces the amount of W events dramatically. Since we require a range of 15 GeV around the Z boson mass the main part of the W boson, $t\bar{t}$ and WW production is rejected. To pass the selection, both leptons need to have suitable four-vectors for the Z boson reconstruction in the required mass range. Furthermore, the WW production either needs to be in a full-leptonic decay with additionally radiated jets or a semi-leptonic decay with an additional
4.2 Event selection

Figure 4.2: Cross sections for the main background processes for the $Z + \text{jets}$ events.

fake lepton. Here, the branching ratios reduce the amount of $WW$ production events passing the selection. Looking at the second leading cross section, the $t\bar{t}$ process produces two $b$ jets in the final state. These $b$ jets have a reduced influence into the calibration due to the extrapolation to zero second-jet activity. Since these two $b$ jets originate from the $t$ quark decays, they do not have a negligible $p_T$. Hence, cutting on $\alpha = \frac{p_T^{\text{second jet}}}{p_T} < 0.3$ leads to a rejection of the $t\bar{t}$ background process. Finally, the diboson productions $WZ$ and $ZZ$ can pass the event selection. Both processes are suppressed by large factors in the cross section. Furthermore, the $ZZ$ process needs to decay into two muons or electrons with additional jets to pass the selection criteria. Looking at the $WZ$ production, the $Z$ boson needs to decay leptonically whereas the $W$ boson decays hadronically. These restrictions additionally reduce the amount of $WZ$ and $ZZ$ production faking the signal process via the corresponding branching ratios.

In order to confirm that the event selection reduces the amount of background events dramatically, a comparison between two cuts is done. The figures 4.3 and 4.4 show the reconstructed mass of $Z$ boson with different range cuts. The left plots require a $Z$ boson mass between 84 and 98 GeV, whereas the right plots show the range required by the above-mentioned selection. Here, the missing regions are dominated by the background processes. Hence, the cut plots contain a higher amount of signal events compared to the uncut plots. To compare the $Z$ boson mass peaks a Voigt-function was fitted. This function is a convolution of a Gaussian and Lorentz function. The smearing of the $Z$ boson mass peak is caused by the decay width of the $Z$ boson and the limited detector resolution. Therefore, the shape of the peak can be described.
4 Absolute residual corrections using $Z +$ jet events

Figure 4.3: $M^Z$ distribution of the $Z$ boson for the $Z \rightarrow \mu\mu$ channel with a fitted Voigt function combined with a linear background estimation for different $Z$ mass ranges.

by a convolution of a Gaussian and a Lorentz function. Here, the Gaussian function implicates the detector resolution and the Lorentz function the decay time of the $Z$ boson. Since the Drell-Yan production at the LHC can not separate between lepton pairs produced via a photon ($\gamma^*$) or via a $Z$ boson, the complete dilepton datasets are used for the calibration. To check the influence of the production via a photon a rough estimation of the fraction including photon production and the background processes described above needs to be done. In the selected region around the mass of the $Z$ boson, the fraction of background events can be approximated by a linear function added to the Voigt-function. The pure $Z$ boson production dominates only in the region around the mass peak. Since away of the mass peak the number of events becomes negligible, also the fraction of background events is negligible. To get a more stable fit the Lorentz width was fixed to a value obtained from a fit of a Lorentz function to the generated mass distribution of the $Z$ boson. Looking at the top right of the plots, the fitted peak value and the Gaussian peak width are shown. Whereas the muon channel indicates a good agreement between Monte Carlo simulation and the data, a clear difference between simulation and data is observed for the electron channel. This is a first indication for a miss-calibration of the electrons for the 2016 LHC run.

Looking at the different cut regions of the $Z$ boson mass, the smaller region contains a higher amount of signal events compared to taking a larger region into account. Comparing both cut versions, one determines a slight difference between the Gaussian widths and a negligible difference between the $Z$ boson mass values. This comparison shows that the amount of all background processes has a weak influence on the $Z$
4.3 Distributions of kinematic variables

The following chapters consider the kinematic distributions of all reconstructed particles. This allows a consistency check of the characteristics in Monte Carlo simulation and data. For the above-mentioned balancing of the Z boson and the jets, all participating particles need to be well-measured and consistent with the simulation. First, the characteristics of the leptons and the reconstructed Z boson are studied. Also, the characteristics of the leading jets and the subleading jet are checked for consistency. The $p_T$ and the $\eta$ distributions are good indicators of miss-measurement or wrong Monte Carlo simulation. Hence, the following chapters have a look at exemplary distributions to check for any differences between simulation and data. The full set of plots can be found in the appendix.

4.3.1 Lepton characteristics

It is most important that the Z boson as a reference object for the balancing is well-described by data and Monte Carlo simulation. Hence, it is necessary to observe the measured particles used at the Z boson reconstruction. For the channels $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$ used for this calibration a look at the muon or electron kinematics

**Figure 4.4:** $M_Z$ distribution of the Z boson for the $Z \rightarrow ee$ channel with a fitted Voigt function combined with a linear background estimation for different Z mass ranges.

boson characteristics. Therefore, the background processes can be ignored for the calibration and further studies.
4 Absolute residual corrections using Z + jet events

is important. Furthermore, the reconstructed Z boson kinematics is checked for consistency.

Muon characteristics

For the $Z \rightarrow \mu\mu$ channel, a look at the kinematic distributions of the muons is necessary. This is exemplarily done for the negatively charged muon. Hence, their $p_T$ and $\eta$ distributions are shown in figure 4.5. The positively charged muons nearly look identically as shown by the plots in the appendix. A look at the shown distributions of the muon indicates a good agreement between data and Monte Carlo within their uncertainties. Hence, this channel should also show a good description of the reconstructed Z boson shown in the next chapter.

Electron characteristics

The $Z \rightarrow ee$ channel is analogously studied via the kinematic variables of the negatively charged electrons. Figure 4.6 shows the $p_T$ and $\eta$ distributions of these electrons. Here, no significant disagreement between data and Monte Carlo simulation can be determined within the uncertainties. The analogous plots for the positively charged electrons are included in the appendix. These look very similar to the distributions shown at this point and show no significant differences either. Hence, the electrons are well-defined and suitable for the Z boson reconstruction.
4.3 Distributions of kinematic variables

**Figure 4.6:** $p_T$ and $\eta$ distributions of the negative charged electrons for the $Z \rightarrow ee$ channel.

### 4.3.2 Z boson characteristics

After studying the lepton distributions, the kinematics of the reconstructed Z boson are checked for deviations between data and Monte Carlo. Here, the same kinematic variables are studied as done for the individual leptons for both channels. The corresponding plots of the $p_T^Z$ and $\eta^Z$ are shown in figure 4.7 exemplarily for the $Z \rightarrow \mu\mu$ channel. Both channels show dominating Monte Carlo uncertainties for high $p_T$ bins. An increase of the Monte Carlo statistics and a reduction of the uncertainties can be achieved by using a higher amount of simulated events. For future calibrations

**Figure 4.7:** $p_T$ and $\eta$ distributions of the Z boson for the $Z \rightarrow \mu\mu$ channel.
with a higher number of measured events, the number of generated events should be increased to reduce the influence of the Monte Carlo uncertainties on the correction uncertainties. Since the detector simulation in Monte Carlo was not updated to the 2016 layout of the CMS detector by now, this small effect is observed which has less influence on the balancing approach. Furthermore, the detector response of the Monte Carlo simulation needs to be calibrated for the current setup of the detector during runtime. Future studies will use a re-reconstructed version of the datasets which include a more precise reconstruction of the particles as can be done during data taking. There the effect should be reduced. However, this effect does not influence the calibration much, because only the absolute value of the variable $\eta$ is taken into account. For the absolute value of $\eta$, this deviation is negligible. The reconstructed mass of the $Z$ boson displayed in figure 4.3 and figure 4.4 of chapter 4.2.7 show a clear mass peak. As restricted by the above-mentioned $Z$ boson cuts, a range of 40 GeV around the theoretical $Z$ boson mass value is chosen for the calibration. In addition to the above-mentioned background studies, a small shifting of the $Z$ boson mass can be determined for the $Z \rightarrow \mu\mu$ channel. This effect has no influence on the calibration approach and will be corrected during the global fit combining all channels.

4.3.3 Jet distributions

Looking at the jets selected for calibration, the same kinematic variables can be studied to detect mis-measurement or wrong Monte Carlo simulation. Therefore, figure 4.8 shows the $p_T$ and $\eta$ distributions of the leading jet for the muon channel. The ratio of Monte Carlo simulation and data for $p_T$ leading jet illustrate deviations,
which are corrected by the jet energy corrections determined in this thesis. Especially for the low $p_T$ region, this deviation increases dramatically. Hence, we see a clear need for the absolute jet energy corrections. Analogous to the $p_T$ distribution, the $p_T^{\text{leading jet}}$ also indicates low statistics at high values. In summary, it can be stated that the measured leading jets are suitable for balancing against the $Z$ boson to get jet energy correction factors.

After having a look at the reference object and the jet to calibrate, the residual jets of the chosen events have to be studied. These sub-leading jets highly influence the $p_T$ of the leading jets and therefore the balancing. The second jet distribution is exemplarily shown in figure 4.9 for an estimation of all sub-leading jets with lower $p_T$ values. Both the $p_T^{\text{second jet}}$ and the $\eta^{\text{second jet}}$ show large differences between Monte Carlo simulation and data. There, absolute jet energy corrections are much more needed to get precise descriptions of the jet characteristics.

### 4.3.4 Missing transverse energy distributions

The missing transverse energy projection fraction (MPF) includes the whole $E_T^{\text{miss}}$ of the event. Hence, it is also necessary to check for any deviations in the $E_T^{\text{miss}}$ distributions. As shown in figure 4.1, the missing transverse energy distribution shows a small shift between data and simulation. A miss-measurement of the jet energy directly leads to a shifting of the missing transverse energy. Since the absolute residual corrections correct the jet energy, also the $E_T^{\text{miss}}$ is indirectly corrected by this calibration step. This small shift in the missing transverse energy also affects the MPF balancing as described in the next chapter.
4.4 Response measurements

As described above, balancing methods are used to estimate the jet response. The largest influence on the jet response comes from the sub leading jets and the missing transverse energy of the event. For the determination of mis-configurations or mis-measurements both balancing methods are compared. Furthermore, the values of both balancing methods are calculated for different sub-leading jet activity. These values are used to do an extrapolation to the jet response for the ideal event topology of only one jet and a Z boson.

4.4.1 MPF and $p_T$-balance

A first look at the distributions of the balancing methods $p_T$ balance and MPF helps to check their applicability. As figure 4.10 exemplarily shows, both methods show a good agreement between data and Monte Carlo simulation within their uncertainties. Only a small shift between the Monte Carlo prediction and data can be detected for the MPF response which is negligible. This shift comes from the small shift in the missing transverse energy described in chapter 4.3.4. Whereas the MPF method is symmetric and has a Gaussian shape, the $p_T$ balance method is not as symmetric as expected. This could be caused by the huge amount of sub-leading jets at the increased center-of-mass energy of $\sqrt{s} = 13$ TeV.

Figure 4.10: $p_T$ balance and MPF distributions for the $Z \rightarrow \mu\mu$ channel.
4.4.2 Response estimation

The ideal event topology of this calibration is a jet opposite to the Z boson. Due to final state radiation (FSR) or underlying event, there are additional sub-leading jets, which lead to out of cone energy loss of the leading jet barely described in simulation. This effect leads to a distortion of the jet response calculated by the balancing methods. To eliminate this effect, an extrapolation to no second jet activity is done. As shown in figure 4.11 this extrapolation is done for data and simulation and both balancing methods separately.

\[
\alpha = \frac{p_T^{\text{second jet}}}{p_T^Z} \tag{4.5}
\]

For \(\alpha \to 0\), which means to go left on the x-axis, there are no second or other sub-leading jets. Hence, there the Z boson and the leading jet can be balanced against each other perfectly. This ideal event topology has to be achieved using a linear fit to the jet response versus \(\alpha\) and extrapolating to zero. The ratio between data and Monte Carlo simulation at this point corresponds to the correction value for the jet response.

![Figure 4.11: \(p_T\) balance and MPF extrapolation without applied absolute residual correction factors for the \(Z \to \mu\mu\) and \(Z \to ee\) channel.](image)

This extrapolation approach uses the fraction \(\alpha\) as an estimation of the second jet activity.
energy calibration. Here, also the statistical uncertainties of the jet energy corrections illustrated by the colored bands are determined.

As the plots show, the $p_T$ balance method has a high dependency on the second jet activity $\alpha$. The MPF method includes the whole event environment by taking into account $E_T^{\text{miss}}$ in addition to the $p_T$ values of the particles. Therefore, this method is almost independent of the second jet activity $\alpha$ since it shows almost a constant behavior. The extrapolated correction factors of the $Z \rightarrow \mu\mu$ and the $Z \rightarrow ee$ channels are determined from the shown extrapolation. For the determination of the absolute residual jet energy correction values, the extrapolated values for the ratio between data and Monte Carlo simulation are inverted.

Muon channel

- $f_{\text{PT balance}} = 1.0103 \pm 0.0049$
- $f_{\text{MPF}} = 1.0187 \pm 0.0052$

Electron channel

- $f_{\text{PT balance}} = 1.0316 \pm 0.0066$
- $f_{\text{MPF}} = 1.0186 \pm 0.0063$

Looking at the extrapolation plots of the muon channel, a good agreement between Monte Carlo simulation and data is observed. Furthermore, the muon channel provides consistent correction factors within their statistical uncertainties comparing both balancing methods. These good agreements indicate an accurate reconstruction of the event environment and a precise determination of the correction factors. This also includes the low statistical uncertainties of the calculated correction factors with around 5%. For the balancing approach, no systematic uncertainties are considered so far. Since the electron channel shows a large discrepancy between Monte Carlo simulation and data, a mis-calibration of the electrons is determined. For the provisioning of correction values suitable for the combination, the electron correction need to be adapted to the 2016 setup of the CMS detector. Looking at the results of both balancing methods for this channel, a large discrepancy is observed which confirms the above-mentioned mis-calibration of the electrons. In general, the combination of the different channels can correct for this distortions by applying a systematic correction factor to each channel. In this case the large uncertainties of the electron channel reduce the influence of the mis-calibration on the final correction factors.

After providing the $Z +$ jet correction factors to the CMS jet calibration workgroup, these values are taken into a global fit. Here, all channels used for this calibration step are combined to global correction factors for the different $p_T$ bins using the MPF method. For this purpose, the two $Z+$ jets channels and the multijet and the $\gamma+$ jets channel [70] are used. For 0.6 fb$^{-1}$ of the 2016 data, the results shown in figure 4.12 of the global fit were published. This figure shows the jet energy response ratio between data and Monte Carlo simulation for all four channels. Fitting a global function to the entries depending on the transverse momentum of the jet, a smooth
Figure 4.12: Jet energy response after the global fit [71] which combines the balancing approaches of the two $Z^+$ jets channels and the multijet and the $\gamma^+$ jets channel. The jet energy scale uncertainties derived from the global fit are shown as an orange band around the jet energy correction function. For comparison reasons the results of LHC run 1 are also shown.

The function is derived which describes the absolute jet energy correction factor. In this step, also the systematic uncertainties are taken into account which are ignored at the former balancing approaches. Here, also the distortions of the electron channel are corrected, which leads to a higher systematic uncertainties and less influence on the resulting correction factors. The muon channel however provides the consistency to contribute at the global fit highly. Furthermore, a global jet energy scale uncertainty is determined by combining the uncertainties of the channels and the systematic uncertainties. The curve derived from 2016 datasets shows a bend around $p_T^{\text{jet}} \approx 300$ GeV, which was not observed in run 1 datasets shown by a reference line. Thresholds chosen for the noise suppression in the calorimeter were changed for the run 2 which lead to this bend. Future data taking and a re-processing of the old datasets may lead to jet energy corrections without this effect and lower jet energy scale uncertainties.

4.4.3 Cross-checks

To guarantee the stability and quality of the calibration, the balancing methods are checked with applied correction factors. The extrapolation plot is used where the values of Monte Carlo simulation and data should coincide. Furthermore, the balancing methods are plotted against the variables $|\eta^{\text{leading\ jet}}|$, $n_{\text{PV}}$ and $p_T^Z$. This
Figure 4.13: $p_T$ balance and MPF extrapolation with applied correction factors for the $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$ channel.

helps to find unexpected dependencies or deviations of the balancing methods. All the cross-check plots are shown in the figures A.19, A.20, A.21, A.22, A.23 and A.24 of the appendix. Since the dependency of the leading jet pseudo-rapidity $|\eta|$ indicates an interesting feature, figure 4.14 is analyzed exemplarily. For low $|\eta|$ values, the absolute residual corrections correct the differences between data and Monte Carlo simulation very well. But going into the forward region including the end-caps and the heavy forward region, the corrections do not change much. This
effect can be explained by the fact that the absolute residual corrections only use the central region for the calibration. The former relative residual corrections should weight the forward jets using jets from the central region. Hence, the shown feature has to be checked in all calibration steps to get more precise jet energy calibrations. Concluding all these cross-checks, the absolute residual corrections lead to a better description of data by Monte Carlo simulations. This also leads to a more precise calibration of the measured jet energy. Whereas almost all cross-check plots show dependencies between the $p_T$ balance method and the used variables, the MPF method does not show any large dependency. This shows clearly that the MPF method is the favored balancing method, which is more resistant against many influences.

### 4.4.4 Time-dependency of the jet response

During the CMS data taking periods B to G, some corrections are updated to the detector description of 2016. Hence, we expect a small time-dependency of the two balancing methods for all run periods. To estimate the influence of this effect, a cross-check for the balancing output versus the variables $|\eta^{\text{leading jet}}|$, $n_{\text{PV}}$ and $p_T^Z$ is done for the different run periods. As an example for this cross-check, the $n_{\text{PV}}$ dependency of the two balancing methods is studied in figure 4.15 for any time-dependency. Run B, C and D are summarized due to their nearly identical setup, which leads also to no significant differences in their jet response. The same is done for run E and F. All other cross-check plots are shown in the appendix at figure A.25, A.26, A.27, A.28, A.29, and A.30. For all these cross-check plots no absolute residual corrections are applied to determine the differences of the datasets which are used for the calibration. Here, a clear time-dependency of the $p_T$ balance and MPF methods is recognized. Whereas the early runs indicate a larger discrepancy of about 5% to the Monte Carlo prediction, run G shows a better coincidence with the simulation. This effect is caused by blocked readouts of the silicon tracker strips during data taking. Therefore, residual charge in the silicon tracker strips coming from former events led to distorted track measurements and influenced the reconstruction of the events. For run G the acceleration voltage which empties the silicon tracker strips was increased to reduce this effect. To reduce the influence of this effect for run B till F, a re-reconstruction of the events was started and will be available for future calibrations. For the current absolute jet energy corrections, the Jet Energy Resolution and Correction team of the CMS collaboration decided to deploy jet energy corrections separately for each of the run periods. Hence, the configuration effects are taken into account individually by calculating the absolute residual corrections separately. Providing these sets of corrections should lead to a higher precision of the jet energy for further analyses. This will be checked for future calibrations within further studies.
4 Absolute residual corrections using $Z+jet$ events

4.5 Additional studies

Except for the calibration, also some systematic studies are done by the Karlsruhe $Z+Jet$ calibration group within this master thesis. These studies include the comparison between using miniAOD instead of AOD as basic format and the advantage of using PUPPI jets instead of CHS jets.

4.5.1 Replacement of AOD format with miniAOD

The AOD format is used for the $Z+jet$ calibration because it offers the possibility to get the uncorrected jet, muon, and electron characteristics directly and suitable for the calibration approach. Since the integrated luminosity increased largely at LHC run 2, the amount of data which need to be processed for this calibration increased proportionally as well. To reduce the runtime dominated by the input and output of the datasets, it was necessary to switch to a less storage consuming data format. The miniAOD as a reduced AOD format offers less information, but also about 90% less disc space to process. To guarantee the quality of the calibration using the miniAOD instead of the AOD format, some cross-checks need to be done.

Therefore, the same cross-check plots described above are used to look for deviations between both formats. Exemplarily, the $p_T^Z$ dependency of the balancing methods is chosen in figure 4.16. This variable provides high statistics at lower $p_T^Z$ values and lower statistics at higher values. Especially for the data points the differences between the miniAOD and AOD is too small to see because the points superpose in
4.5 Additional studies

Figure 4.16: Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $n_{PV}$ for the $Z \rightarrow \mu\mu$ channel.

nearly all cases. Since both Monte Carlo simulations show significantly lower statistics than the data samples, the differences are dominated by the statistical uncertainties. The minimal differences between the simulated datasets agree well within the given uncertainties. All other cross-check plots which can be found in the appendix at the figures A.31, A.32, A.33, A.34, A.35, and A.36, show the same behavior. They all show no significant differences between the miniAOD and AOD datasets exceeding the uncertainties.

In general, the shown comparisons between miniAOD and AOD format result in negligible differences. Due to this very good agreement of the datasets shown in this study, the miniAOD format will be used for further calibrations and studies. This allows us to reduce the needed amount of computing power or network traffic and process also the heavily increasing datasets.

4.5.2 Analogous approach using PUPPI

As described in the introduction, the standard pileup subtraction is done via charged hadron subtraction. PUPPI can bring an advantage by reducing only objects, which are strongly classified as pileup. Hence, using PUPPI can lead to a higher amount of events passing the pileup selection and lead to higher statistics at the calibration. This can reduce the statistical uncertainties of the correction factors.

Before deploying PUPPI as the standard method, the calibration approach using the algorithm has to be checked for consistency. Therefore, all studies done for the calibration using CHS are redone using the PUPPI algorithm. This includes a closer
4 Absolute residual corrections using $Z +$ jet events

look at the $p_T$ and $\eta$ distributions of the leading and second-leading jets, the $Z$ boson characteristics and the missing transverse energy $E_{\text{miss}}^T$. The following plots concentrate on the $Z$ boson decay mode into muons. Plots corresponding to the electron channel can be found in the appendix.

**Figure 4.17:** $p_T$ and $\eta$ distributions of the leading jet for the $Z \to \mu\mu$ channel comparing CHS and PUPPI algorithm.

Beginning with the leading jet characteristics shown in figure 4.17, the $p_T$ and $\eta$ distributions using the PUPPI or the CHS datasets leads to nearly the same distributions. At the first sight, the second jet characteristics in figure 4.18 for PUPPI

**Figure 4.18:** $p_T$ and $\eta$ distributions of the second jet for the $Z \to \mu\mu$ channel comparing CHS and PUPPI algorithm.
jets deviate from the CHS method. But most of the deviations are affected by the shape comparison done in these plots. This shape comparison equates the integrals of all distributions to the Monte Carlo simulation. Nevertheless, the plots indicate that the PUPPI algorithm is more effective for low $p_T$ jets. Hence, the PUPPI graph shows no depression in the low $p_T$ region of the leading jet unlike the graph of the CHS dataset. This excess of low $p_T$ events influences the normalization of the shape comparison and leads to a lower graph at higher $p_T$ bins for the PUPPI dataset. Looking at the $\eta$ region, the deviation of both methods can be explained by the acceptance region of the PUPPI algorithm. The currently used PUPPI version only supports pileup rejection in the tracker region. As shown in the figures, the zero entries at $\eta > 2.4$ in combination with the shape comparison lead to the excess of the PUPPI dataset in the central region.

![Figure 4.19: Z mass and $\alpha$ distributions for the $Z \rightarrow \mu\mu$ channel comparing CHS and PUPPI algorithm.](image)

The mass of the $Z$ boson shown in figure 4.19 as an example of the $Z$ boson characteristics looks very similar. Hence, applying the PUPPI algorithm also leads to a $Z$ boson measurement which is suitable for the balancing approach. Looking at the $\alpha$ distribution shown in figure 4.19, the full advantage of using PUPPI can be seen. Comparing CHS and PUPPI within the shape comparison, we see that PUPPI leads to a large number of events in the low $\alpha$ region passing the selection. This can be used to reduce the uncertainties of the jet response extrapolation to $\alpha \rightarrow 0$ which directly leads to lower jet energy scale uncertainties.

For the PUPPI algorithm a specific PUPPI $E_T^{\text{miss}}$ is used which is also shown in figure 4.20. The $E_T^{\text{miss}}$ needs to be recalculated, because the changed pileup rejection directly changes the amount of non-classified particles which contribute to the $E_T^{\text{miss}}$. Due to the improved pileup rejection, this $E_T^{\text{miss}}$ should show a better resolution. The
muon channel confirms this expectation with fewer events at higher $E_{\text{miss}}$ values and a sharper peak compared to the CHS selection. But the missing transverse energy $E_{\text{miss}}$ shows a huge deviation between both pile up subtraction methods for the electron channel. Here, a significant amount of events at higher $E_{\text{miss}}$ values is observed, which is caused by a double-counting of the electrons in the PUPPI algorithm. This leads to a miscalculation of $E_{\text{miss}}$ and influences the event environment dramatically. This also affects our final extrapolation shown in figure 4.21, because the calculation of the MPF balancing method includes $E_{\text{miss}}$. For the muon channel, the extrapolations

for both balancing methods coincide at $\alpha \to 0$. The resulting jet energy correction
values and uncertainties show quite a good agreement compared to the extrapolation based on the CHS method. On the other side, the electron channel shows a big disagreement and an incorrect extrapolation of the MPF method. This deviation can be explained looking at the above-mentioned differences in the PUPPI $E_{\text{miss}}^T$. A huge amount of missmeasured events at high $E_{\text{miss}}^T$ values affect the MPF calculation negatively.

Hopefully, this bug in the PUPPI $E_{\text{miss}}^T$ calculation for electrons can be fixed using an updated PUPPI version in further studies.
Chapter 5

Conclusion

With this master thesis I provided preliminary absolute jet energy corrections for 2016 datasets of the CMS detector up to 27.22 fb$^{-1}$ of integrated luminosity. This was achieved even though the size of the recorded datasets was increased heavily which challenged the fast provisioning of correction factors immediately during data taking. For this purpose, a high throughput computing solution was extensively used and extended for future increases of data sizes.

For fast processing of large datasets, like the ones used by the jet calibration done at the CMS experiment, a high read rate is of paramount importance. Therefore, a coordinated caching mechanism using solid state drives (SSDs) was installed at four worker nodes in the local batch system for processing analysis workflows. Here, the so-called High Throughput Data Analysis (HTDA) setup caches needed files identified from former analysis processing for future usage. Since the datasets increased dramatically, even the installation of SSDs with a doubled size does not allow the complete caching of the necessary files. To also provide fast input rates during processing for uncached files, a distributed file system was tested. Such a file system distributes the stored data between all integrated worker nodes providing storage capacity on hard disc drives (HDDs). For batch system usage, a high number of parallel accesses need to be served with high transfer rates. Tests showed the same stability of the high read rate for different number of access streams as expected for SSDs. Hence, the Ceph file system is suitable for batch system usage and can be combined with the HTDA setup. Although the total read rate of the tested Ceph file system is lower than the value of the SSD caches, it outperforms classical network storage servers dramatically. Combining the power of the HTDA caching with the Ceph file system reduces the time needed for processing of the absolute residual corrections enormously by a factor of 10.

This setup enabled the repeated processing of the calibration workflow during data taking including the latest recorded datasets. The calculation of the absolute residual
corrections was done using a balancing approach where the jet was compared to a well-measured reference object. For that purpose a Z boson which can be reconstructed from two muons or electrons was used. The balancing was done via two different method, the $p_T$ balancing and the Missing Transverse Energy Projection Fraction (MPF). These were compared to search for possible mis-measurements. Since additional jets distort the balancing, an extrapolation to a clean back-to-back topology is done as shown in figure 4.11. The latest results with $27.22\text{fb}^{-1}$ of CMS data were presented in this thesis. The Z boson reconstruction using the muon decay channel provided precisely determined characteristics of the reference object which is suiting perfectly for the balancing approach. Hence, this channel provided precisely determined correction factors with small statistical uncertainties. Compared to the muon channel, the Z boson reconstruction using the electron decay mode did not provide the same precision. Here, large deviations between Monte Carlo prediction and data were observed. These are caused by miss-calibrated electrons whose corrections needed to be updated to the CMS setup of 2016 for further improvements. Preliminary absolute residual corrections of the CMS collaboration were determined by combining the two channels provided in the context this master thesis with a $\gamma +$ jets and a multijet channel. A global fit takes into account the systematic uncertainties and corrects for systematic shifts like observed in the electron channel. This approach allows the inclusion of the electron channel, but leads to higher uncertainties for that channel and therefore less importance for the fitting approach. The muon channel however provides the needed requirements to contribute at the precise determination of the absolute residual jet energy correction factors.

For the future correction factors a re-reconstruction of the 2016 datasets and new Monte Carlo simulations with an updated detector description of 2016 were processed. This will contain an improvement of the agreement between data and Monte Carlo simulation. Hence, future results of the global fit will contain the full possible precision and will be provided as final absolute jet energy correction factors with their corresponding jet energy uncertainties.
### Appendix A

#### Table A.1: Overview of inclusive cross sections for $\sqrt{s} = 13$ TeV.

<table>
<thead>
<tr>
<th>Cross section in [pb]</th>
<th>value $\pm \sigma_{\text{stat}} \pm \sigma_{\text{syst}} \pm \sigma_{\text{theo}} \pm \sigma_{\text{stat}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma(pp \to W^+ X) \times B(W^+ \to \ell^+ \nu \ell)$ [72]</td>
<td>11370 $\pm 50$ $\pm 230$ - $\pm 550$</td>
</tr>
<tr>
<td>$\sigma(pp \to W^- X) \times B(W^- \to \ell^- \bar{\nu} \ell)$ [72]</td>
<td>8580 $\pm 50$ $\pm 160$ - $\pm 410$</td>
</tr>
<tr>
<td>$\sigma(pp \to ZX) \times B(Z \to \ell^+ \ell^-)$ [73]</td>
<td>1870 $\pm 2$ $\pm 35$ - $\pm 51$</td>
</tr>
<tr>
<td>$\sigma(pp \to t \bar{t})$ [74]</td>
<td>746 $\pm 58$ $\pm 53$ - $\pm 36$</td>
</tr>
<tr>
<td>$\sigma(pp \to WW)$ [75]</td>
<td>115.3 $\pm 5.8$ $\pm 5.7$ $\pm 6.4$ $\pm 3.6$</td>
</tr>
<tr>
<td>$\sigma(pp \to WZ)$ [76]</td>
<td>40.9 $\pm 3.4$ $^{+3.1}_{-3.3}$ $\pm 0.4$ $\pm 1.3$</td>
</tr>
<tr>
<td>$\sigma(pp \to ZZ)$ [77]</td>
<td>16.7 $^{+2.9}<em>{-2.6}$ $^{+0.7}</em>{-0.5}$ $\pm 0.3$ $\pm 0.8$</td>
</tr>
</tbody>
</table>
Figure A.1: $n_{PV}$ distribution for $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$ after pile up reweighting.

Figure A.2: $E_T^{\text{miss}}$ distribution for $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$. 
Figure A.3: $p_T$ distributions of the positive and negative charged muons for the $Z \rightarrow \mu\mu$ channel.

Figure A.4: $\eta$ distributions of the positive and negative charged muons for the $Z \rightarrow \mu\mu$ channel.
Figure A.5: $p_T$ distributions of the positive and negative charged electrons for the $Z \rightarrow ee$ channel.

Figure A.6: $\eta$ distributions of the positive and negative charged electrons for the $Z \rightarrow ee$ channel.
Figure A.7: $p_T$ and $\eta$ distributions of the $Z$ boson for the $Z \rightarrow \mu\mu$ channel.

Figure A.8: $p_T$ and $\eta$ distributions of the $Z$ boson for the $Z \rightarrow ee$ channel.
Figure A.9: $M^Z$ distribution of the $Z$ boson for the $Z \rightarrow \mu\mu$ channel with and without a fitted Voigt function.

Figure A.10: $M^Z$ distribution of the $Z$ boson for the $Z \rightarrow ee$ channel with and without a fitted Voigt function.
Figure A.11: $p_T$ and $\eta$ distributions of the leading jet for the $Z \rightarrow \mu\mu$ channel.

Figure A.12: $p_T$ and $\eta$ distributions of the leading jet for the $Z \rightarrow ee$ channel.
Figure A.13: $p_T$ and $\eta$ distributions of the second jet for the $Z \rightarrow \mu\mu$ channel.

Figure A.14: $p_T$ and $\eta$ distributions of the second jet for the $Z \rightarrow ee$ channel.
Figure A.15: $p_T$ balance and MPF distributions for the $Z \rightarrow \mu\mu$ channel.

Figure A.16: $p_T$ balance and MPF distributions for the $Z \rightarrow ee$ channel.
Figure A.17: $p_T$ balance and MPF extrapolation without applied absolute residual correction factors for the $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$ channel.

Figure A.18: $p_T$ balance and MPF extrapolation with applied absolute residual correction factors for the $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$ channel.
Figure A.19: MPF (up) and $p_T$ balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus $|\eta|$ for the $Z \rightarrow \mu\mu$ channel.
Figure A.20: MPF (up) and $p_T$ balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus $|\eta^{\text{leading jet}}|$ for the $Z \rightarrow ee$ channel.
Figure A.21: MPF (up) and $p_T$ balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus $n_{PV}$ for the $Z \rightarrow \mu\mu$ channel.
Figure A.22: MPF (up) and $p_T$ balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus $n_{PV}$ for the $Z \rightarrow ee$ channel.
Figure A.23: MPF (up) and $p_T$ balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus $p_T^Z$ for the $Z \rightarrow \mu\mu$ channel.
Figure A.24: MPF (up) and $p_T$ balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus $p_T^Z$ for the $Z \to ee$ channel.
Figure A.25: Time-dependence for AOD CHS datasets using $p_T$ balance and MPF extrapolation versus $|\eta^{\text{leading jet}}|$ for the $Z \rightarrow \mu\mu$ channel.

Figure A.26: Time-dependence for AOD CHS datasets using $p_T$ balance and MPF extrapolation versus $|\eta^{\text{leading jet}}|$ for the $Z \rightarrow e e$ channel.
Figure A.27: Time-dependence for AOD CHS datasets using $p_T$ balance and MPF extrapolation versus $n_{PV}$ for the $Z \to \mu\mu$ channel.

Figure A.28: Time-dependence for AOD CHS datasets using $p_T$ balance and MPF extrapolation versus $n_{PV}$ for the $Z \to ee$ channel.
Figure A.29: Time-dependence for AOD CHS datasets using $p_T$ balance and MPF extrapolation versus $p_T^Z$ for the $Z \rightarrow \mu\mu$ channel.

Figure A.30: Time-dependence for AOD CHS datasets using $p_T$ balance and MPF extrapolation versus $p_T^Z$ for the $Z \rightarrow ee$ channel.
Figure A.31: Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $|\eta^{\text{leading jet}}|$ for the $Z \to \mu\mu$ channel.

Figure A.32: Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $|\eta^{\text{leading jet}}|$ for the $Z \to ee$ channel.
Figure A.33: Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $n_{PV}$ for the $Z \rightarrow ee$ channel.

Figure A.34: Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $n_{PV}$ for the $Z \rightarrow ee$ channel.
Figure A.35: Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $p_T^Z$ for the $Z \to \mu\mu$ channel.

Figure A.36: Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $p_T^Z$ for the $Z \to ee$ channel.
Figure A.37: $p_T$ and $\eta$ distributions of the leading jet for the $Z \rightarrow \mu\mu$ channel with comparing CHS and PUPPI algorithm.

Figure A.38: $p_T$ and $\eta$ distributions of the leading jet for the $Z \rightarrow ee$ channel with comparing CHS and PUPPI algorithm.
Figure A.39: $p_T$ and $\eta$ distributions of the second jet for the $Z \rightarrow \mu\mu$ channel with comparing CHS and PUPPI algorithm.

Figure A.40: $p_T$ and $\eta$ distributions of the second jet for the $Z \rightarrow ee$ channel comparing CHS and PUPPI algorithm.
Figure A.41: Z mass and $\alpha$ distributions for the $Z \rightarrow \mu\mu$ channel comparing CHS and PUPPI algorithm.

Figure A.42: Z mass and $\alpha$ distributions for the $Z \rightarrow ee$ channel comparing CHS and PUPPI algorithm.
Figure A.43: $E_{T}^{\text{miss}}$ and $p_{T}^{Z}$ distributions for the $Z \rightarrow \mu\mu$ channel comparing CHS and PUPPI algorithm.

Figure A.44: $E_{T}^{\text{miss}}$ and $p_{T}^{Z}$ distributions for the $Z \rightarrow ee$ channel comparing CHS and PUPPI algorithm.
2.1 One quadrant of a longitudinal section of the CMS detector [7]. The silicon tracker which records the tracks of charged particles surrounds the collision point in the center of the detector. Whereas the electromagnetic calorimeter (ECAL) measures the energy of photons and electrons, the hadronic calorimeter (HCAL) summarizes the energy of hadron cascades. To measure the momentum of charged particles, a magnetic field is produced by surrounding solenoid magnet. The outer muon chambers (MB/ME) are integrated into the magnetic return yoke (RB/RE). The polar angular distribution is given in degrees or the pseudorapidity $\eta = -\ln(tan(\theta/2))$. 

2.2 The effect of confinement (a) leads to generation of additional color-neutral quark pairs due to the increasing potential of the strong force with the distance [10]. By recombining of all participating quarks into color-neutral hadrons, a single unbound quark generates a cascade of hadrons. This so-called hadronization effect (b) forms a jet of particles, which can be clustered after measurement using a jet algorithm [10].

2.3 An event with two jets and additional particles which distort the jet energy. Additional collisions beside the primary vertex, so-called in-time pileup events, produce particles which are clustered into the jets. The beam remnants and multi-parton interactions lead to so-called underlying events which also leave energy deposits in the clustered jets. Finally, calorimeter entries remaining from a former bunch crossing, so-called out-of-time pileup events, distort the measured jet energy.
2.4 Levels of jet energy corrections [21] from a reconstructed jet to a calibrated jet: A Pileup correction to Monte Carlo simulation (MC) with first data comparison within a first residual correction (RC), a calorimeter response correction, an $\eta$ depending relative residual correction via a dijet dataset and an $p_T$ dependent absolute residual correction using $Z + \text{jets}$, $\gamma + \text{jets}$ and multijets datasets. Optional a flavour correction step can be done. .......................... 12

2.5 The absolute residual corrections balance a jet with a $Z$ boson on the opposite site. Therefore the $Z \rightarrow \mu\mu$ (a) and $Z \rightarrow ee$ (b) channels are used. Event displays are produced by the web-interface of the CMS event virtualization tool [22, 23] using run 148031 event 105485643 (a) and run 141960 event 135353826 (b) .......................... 14

2.6 Workflow of the KIT $Z + \text{jets}$ workgroup to determine absolute residual jet energy corrections in three different processing steps. .......................... 15

2.7 The hierarchical structure of the Worldwide LHC Computing Grid [30] including institutional resources allows to process different types of purposes restricted to different layers. .......................... 17

3.1 HTCondor handles the process of job submission at the login node, transfer via a Central manager and execution at worker node. Furthermore, a HTDA manager collects information about needed files and coordinates the SSD caching. .......................... 20

3.2 HTDA. .......................... 24

3.3 Functionality of the Ceph file system setup. A cluster of object storage devices (OSDs) provides the storage for a metadata and a data Ceph pool which are merged to a Ceph file system. This Ceph file system is managed by one metadata server (MDS) with a backup server and three monitoring servers (MON) which provide access to the storage. 26

3.4 Tested absolute read rates of the setup. For comparison reasons and applicability for batch system usage, the complete input rates of all four worker nodes are measured for different numbers of parallel streams. 28

4.1 $n_{PV}$ distribution after pile up reweighting and $E_T^{\text{miss}}$ distribution for $Z \rightarrow \mu\mu$. .......................... 37

4.2 Cross sections for the main background processes for the $Z + \text{jets}$ events. 39

4.3 $M_Z$ distribution of the $Z$ boson for the $Z \rightarrow \mu\mu$ channel with a fitted Voigt function combined with a linear background estimation for different $Z$ mass ranges. .......................... 40

4.4 $M_Z$ distribution of the $Z$ boson for the $Z \rightarrow ee$ channel with a fitted Voigt function combined with a linear background estimation for different $Z$ mass ranges. .......................... 41
4.5 $p_T$ and $\eta$ distributions of the negatively charged muons for the $Z \rightarrow \mu\mu$ channel. ........................................... 42
4.6 $p_T$ and $\eta$ distributions of the negative charged electrons for the $Z \rightarrow ee$ channel. ........................................... 43
4.7 $p_T$ and $\eta$ distributions of the Z boson for the $Z \rightarrow \mu\mu$ channel. . .. 43
4.8 $p_T$ and $\eta$ distributions of the leading jet for the $Z \rightarrow \mu\mu$ channel. . .. 44
4.9 $p_T$ and $\eta$ distributions of the second jet for the $Z \rightarrow \mu\mu$ channel. . .. 45
4.10 $p_T$ balance and MPF distributions for the $Z \rightarrow \mu\mu$ channel. ........ 46
4.11 $p_T$ balance and MPF extrapolation without applied absolute residual correction factors for the $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$ channel. .............. 47
4.12 Jet energy response after the global fit [71] which combines the balancing approaches of the two $Z+$ jets channels and the multijet and the $\gamma+$ jets channel. The jet energy scale uncertainties derived from the global fit are shown as an orange band around the jet energy correction function. For comparison reasons the results of LHC run 1 are also shown. ........................................... 49
4.13 $p_T$ balance and MPF extrapolation with applied correction factors for the $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$ channel. .................. 50
4.14 MPF balancing balancing with (left) and without (right) applied residual corrections versus $|\eta^{\text{leading jet}}|$ for the $Z \rightarrow \mu\mu$ channel. ........................................... 50
4.15 Time-dependence for AOD CHS datasets using $p_T$ and MPF balancing versus $n_{PV}$ without applied correction factors for the $Z \rightarrow \mu\mu$ channel. 52
4.16 Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $n_{PV}$ for the $Z \rightarrow \mu\mu$ channel. .................. 53
4.17 $p_T$ and $\eta$ distributions of the leading jet for the $Z \rightarrow \mu\mu$ channel comparing CHS and PUPPI algorithm. .................. 54
4.18 $p_T$ and $\eta$ distributions of the second jet for the $Z \rightarrow \mu\mu$ channel comparing CHS and PUPPI algorithm. .................. 54
4.19 Z mass and $\alpha$ distributions for the $Z \rightarrow \mu\mu$ channel comparing CHS and PUPPI algorithm. .................. 55
4.20 $E_T^{\text{miss}}$ distributions for the $Z \rightarrow \mu\mu$ channel (left) and $Z \rightarrow ee$ channel (right) comparing CHS and PUPPI algorithm. .............. 56
4.21 $p_T$ and $\eta$ distributions of the leading jet for the $Z \rightarrow \mu\mu$ channel. .... 56
A.1 $n_{PV}$ distribution for $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$ after pile up reweighting. . 62
A.2 $E_T^{\text{miss}}$ distribution for $Z \rightarrow \mu\mu$ and $Z \rightarrow ee$. .................. 62
A.3 short description .................. 63
A.4 short description .................. 63
A.5 $p_T$ distributions of the positive and negative charged electrons for the $Z \rightarrow ee$ channel. .................. 64
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.6</td>
<td><em>η</em> distributions of the positive and negative charged electrons for the <em>Z → ee</em> channel</td>
<td>64</td>
</tr>
<tr>
<td>A.7</td>
<td><em>p_T</em> and <em>η</em> distributions of the <em>Z</em> boson for the <em>Z → μμ</em> channel</td>
<td>65</td>
</tr>
<tr>
<td>A.8</td>
<td><em>p_T</em> and <em>η</em> distributions of the <em>Z</em> boson for the <em>Z → ee</em> channel</td>
<td>65</td>
</tr>
<tr>
<td>A.9</td>
<td><em>M_Z</em> distribution of the <em>Z</em> boson for the <em>Z → μμ</em> channel with and without a fitted Voigt function</td>
<td>66</td>
</tr>
<tr>
<td>A.10</td>
<td><em>M_Z</em> distribution of the <em>Z</em> boson for the <em>Z → ee</em> channel with and without a fitted Voigt function</td>
<td>66</td>
</tr>
<tr>
<td>A.11</td>
<td><em>p_T</em> and <em>η</em> distributions of the leading jet for the <em>Z → μμ</em> channel</td>
<td>67</td>
</tr>
<tr>
<td>A.12</td>
<td><em>p_T</em> and <em>η</em> distributions of the leading jet for the <em>Z → ee</em> channel</td>
<td>67</td>
</tr>
<tr>
<td>A.13</td>
<td><em>p_T</em> and <em>η</em> distributions of the second jet for the <em>Z → μμ</em> channel</td>
<td>68</td>
</tr>
<tr>
<td>A.14</td>
<td><em>p_T</em> and <em>η</em> distributions of the second jet for the <em>Z → ee</em> channel</td>
<td>68</td>
</tr>
<tr>
<td>A.15</td>
<td><em>p_T</em> balance and MPF distributions for the <em>Z → μμ</em> channel</td>
<td>69</td>
</tr>
<tr>
<td>A.16</td>
<td><em>p_T</em> balance and MPF distributions for the <em>Z → ee</em> channel</td>
<td>69</td>
</tr>
<tr>
<td>A.17</td>
<td><em>p_T</em> balance and MPF extrapolation without applied absolute residual correction factors for the <em>Z → μμ</em> and <em>Z → ee</em> channel</td>
<td>70</td>
</tr>
<tr>
<td>A.18</td>
<td><em>p_T</em> balance and MPF extrapolation with applied absolute residual correction factors for the <em>Z → μμ</em> and <em>Z → ee</em> channel</td>
<td>70</td>
</tr>
<tr>
<td>A.19</td>
<td>MPF (up) and <em>p_T</em> balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus *</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>η_{leading jet}</em> for the <em>Z → μμ</em> channel</td>
<td></td>
</tr>
<tr>
<td>A.20</td>
<td>MPF (up) and <em>p_T</em> balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus *</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>η_{leading jet}</em> for the <em>Z → ee</em> channel</td>
<td></td>
</tr>
<tr>
<td>A.21</td>
<td>MPF (up) and <em>p_T</em> balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus *</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>n_{PV}</em> for the <em>Z → μμ</em> channel</td>
<td></td>
</tr>
<tr>
<td>A.22</td>
<td>MPF (up) and <em>p_T</em> balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus *</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>n_{PV}</em> for the <em>Z → ee</em> channel</td>
<td></td>
</tr>
<tr>
<td>A.23</td>
<td>MPF (up) and <em>p_T</em> balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus *</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>p_{ZT}</em> for the <em>Z → μμ</em> channel</td>
<td></td>
</tr>
<tr>
<td>A.24</td>
<td>MPF (up) and <em>p_T</em> balancing (down) balancing with (left) and without (right) applied absolute residual corrections versus *</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>p_{ZT}</em> for the <em>Z → ee</em> channel</td>
<td></td>
</tr>
<tr>
<td>A.25</td>
<td>Time-dependence for AOD CHS datasets using <em>p_T</em> balance and MPF extrapolation versus *</td>
<td>η_{leading jet}</td>
</tr>
<tr>
<td>A.26</td>
<td>Time-dependence for AOD CHS datasets using <em>p_T</em> balance and MPF extrapolation versus *</td>
<td>η_{leading jet}</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>A.27</td>
<td>Time-dependence for AOD CHS datasets using $p_T$ balance and MPF extrapolation versus $n_{PV}$ for the $Z \rightarrow \mu\mu$ channel.</td>
<td>78</td>
</tr>
<tr>
<td>A.28</td>
<td>Time-dependence for AOD CHS datasets using $p_T$ balance and MPF extrapolation versus $n_{PV}$ for the $Z \rightarrow ee$ channel.</td>
<td>78</td>
</tr>
<tr>
<td>A.29</td>
<td>Time-dependence for AOD CHS datasets using $p_T$ balance and MPF extrapolation versus $p_Z^T$ for the $Z \rightarrow \mu\mu$ channel.</td>
<td>79</td>
</tr>
<tr>
<td>A.30</td>
<td>Time-dependence for AOD CHS datasets using $p_T$ balance and MPF extrapolation versus $p_Z^T$ for the $Z \rightarrow ee$ channel.</td>
<td>79</td>
</tr>
<tr>
<td>A.31</td>
<td>Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $</td>
<td>\eta^{leading\ jet}</td>
</tr>
<tr>
<td>A.32</td>
<td>Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $</td>
<td>\eta^{leading\ jet}</td>
</tr>
<tr>
<td>A.33</td>
<td>Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $n_{PV}$ for the $Z \rightarrow ee$ channel.</td>
<td>81</td>
</tr>
<tr>
<td>A.34</td>
<td>Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $n_{PV}$ for the $Z \rightarrow ee$ channel.</td>
<td>81</td>
</tr>
<tr>
<td>A.35</td>
<td>Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $p_Z^T$ for the $Z \rightarrow \mu\mu$ channel.</td>
<td>82</td>
</tr>
<tr>
<td>A.36</td>
<td>Comparing miniAOD versus AOD datasets using $p_T$ balance and MPF extrapolation versus $p_Z^T$ for the $Z \rightarrow ee$ channel.</td>
<td>82</td>
</tr>
<tr>
<td>A.37</td>
<td>$p_T$ and $\eta$ distributions of the leading jet for the $Z \rightarrow \mu\mu$ channel with comparing CHS and PUPPI algorithm.</td>
<td>83</td>
</tr>
<tr>
<td>A.38</td>
<td>$p_T$ and $\eta$ distributions of the leading jet for the $Z \rightarrow ee$ channel with comparing CHS and PUPPI algorithm.</td>
<td>83</td>
</tr>
<tr>
<td>A.39</td>
<td>$p_T$ and $\eta$ distributions of the second jet for the $Z \rightarrow \mu\mu$ channel with comparing CHS and PUPPI algorithm.</td>
<td>84</td>
</tr>
<tr>
<td>A.40</td>
<td>$p_T$ and $\eta$ distributions of the second jet for the $Z \rightarrow ee$ channel with comparing CHS and PUPPI algorithm.</td>
<td>84</td>
</tr>
<tr>
<td>A.41</td>
<td>Z mass and $\alpha$ distributions for the $Z \rightarrow \mu\mu$ channel comparing CHS and PUPPI algorithm.</td>
<td>85</td>
</tr>
<tr>
<td>A.42</td>
<td>Z mass and $\alpha$ distributions for the $Z \rightarrow ee$ channel comparing CHS and PUPPI algorithm.</td>
<td>85</td>
</tr>
<tr>
<td>A.43</td>
<td>$E_T^{miss}$ and $p_T^Z$ distributions for the $Z \rightarrow \mu\mu$ channel comparing CHS and PUPPI algorithm.</td>
<td>86</td>
</tr>
<tr>
<td>A.44</td>
<td>$E_T^{miss}$ and $p_T^Z$ distributions for the $Z \rightarrow ee$ channel comparing CHS and PUPPI algorithm.</td>
<td>86</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Overview of the different channels used for the absolute residual corrections. ........................................ 13

3.1 Overview of HTCondor cloud sites at the institute. ................................................................. 22

3.2 Sequential read and write speed of the Ceph file system compared to a different network storage server with a hardware-RAID 5 configuration using a single stream. ........................................ 30

A.1 Overview of inclusive cross sections for $\sqrt{s} = 13$ TeV. ........................................ 61
Bibliography


[57] Giacinto Donvito, Testing of several distributed file-system (HadoopFS, CEPH and GlusterFS) for supporting the HEP experiments analysis, 2013, url: https://www.gluster.org/ (visited on 03/01/2017).


Acknowledgments

I would like to thank Professor Dr. Günter Quast for the opportunity to work on my master thesis. This includes the provisioning of the interesting topic of the thesis and the helpful support during work.

Further thanks to Professor Dr. Ulrich Husemann for being second reviewer.

Furthermore, I appreciate the provision of the computing resources and the scientific environment by the CMS collaboration and the WLCG collaboration. I acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 39/963-1 FUGG.

I would like to appreciate all the help and support concerning the physics content given by my advisor Dr. Stefan Wayand. Additionally, I would like to thank Dr. Manuel Giffels and Dr. Max Fischer for being contact persons for questions regarding the computing part of this thesis.

The whole workgroup at the IEKP was great in providing a pleasant working atmosphere where questions and ideas could be discussed and solved together. This applies in particular to the stimulating discussions and the active support done by my room mates Günther Erli, Frank Fischer and Matthias Schnepf. I would like to thank the latter for being also a great colleague since the beginning of our studies.
Erklärung der selbständigen Anfertigung der Masterarbeit

Hiermit erkläre ich, dass ich die Masterarbeit mit dem Titel

»Absolute jet energy scale determination at the CMS detector for LHC run 2 at $\sqrt{s} = 13$ TeV using an optimized processing setup«

selbständig und unter ausschließlicher Verwendung der angegebenen Hilfsmittel angefertigt habe.

____________________
Christoph Alexander Heidecker
Karlsruhe, den 20.01.2017