Alma Mater Studiorum · University of Bologna

School of Science
Department of Physics and Astronomy
Master Degree in Physics

THE CMS MUON HIGH-LEVEL TRIGGER AT HL-LHC: PERFORMANCE AND OPTIMISATION

Supervisor:
Dr. Carlo Battilana

Submitted by:
Luca Ferragina

Co-supervisor:
Prof. Francesco Giacomini
Dr. Felice Panaleo

Academic Year 2023/2024

A chi ci ha creduto dall'inizio alla fine.
Alla mia famiglia.

Abstract

The High-Luminosity upgrade for the Large Hadron Collider at CERN will increase its instantaneous luminosity up to a factor of 7.5 over design. This poses significant challenges for the experiments both from the hardware and software points of view. The detectors will be subject to much more radiation, requiring careful consideration of the radiation hardness of the to-be-installed detectors and the damage accumulated by the current ones. From the software point of view, data acquisition and physics reconstruction will have to cope with hundreds of superimposed proton-proton collisions at every bunch crossing. Therefore, all the experiments will undergo a major upgrade to improve their detector, triggering solutions, and reconstruction software. This Master's Thesis presents the foundations for a novel approach to Muon reconstruction at the CMS experiment. Since the muon tracking algorithms have demonstrated excellent performance during previous Runs, the new algorithms aim to maintain or improve the current physics performance, while reducing the computational load. This is achieved by taking full advantage of the upgraded first-level hardware trigger and eliminating some of the redundancy present in the current reconstruction workflow. Both physics and computing performance have been evaluated using Monte Carlo simulated samples targeting expected HL-LHC conditions. The results show close-to-current physics performance with up to about 40% improvement in the timing of some of the modified reconstruction modules. Furthermore, the number of fake tracks produced is also reduced by about 30-40% depending on the reconstruction step, decreasing the complexity of the reconstruction as a whole.

Chapter 1 provides a general overview of the Large Hadron Collider, its experiments and future upgrades.

Chapter 2 describes the CMS experiment at the LHC, with a focus on the Muon system.

Chapter 3 deals with the triggering and reconstruction of Muons at the CMS experiment.

Chapter 4 presents the original results of this thesis: the optimization of the Online Muon reconstruction is evaluated both from the physics and computational performance points of view.

Contents

1 The Large Hadron Collider **3**
- 1.1 Superconducting magnets . . . 5
- 1.2 RF cavities . . . 6
- 1.3 Vacuum system . . . 7
- 1.4 The main LHC experiments . . . 8
  - 1.4.1 ALICE . . . 8
  - 1.4.2 LHCb . . . 9
  - 1.4.3 ATLAS . . . 9
  - 1.4.4 CMS . . . 10
  - 1.4.5 Other experiments at the LHC . . . 10
- 1.5 High-Luminosity LHC . . . 12

2 The CMS experiment **14**
- 2.1 Concept and structure . . . 15
- 2.2 Inner tracking system . . . 17
  - 2.2.1 Pixel detector . . . 17
  - 2.2.2 Silicon strip tracker . . . 18
- 2.3 Electromagnetic calorimeter . . . 19
- 2.4 Hadronic calorimeter . . . 22
- 2.5 Muon system . . . 23
  - 2.5.1 Drift Tubes . . . 25
  - 2.5.2 Cathode Strip Chambers . . . 27
  - 2.5.3 Resistive Plate Chambers . . . 29
  - 2.5.4 Gas Electron Multipliers . . . 30
- 2.6 The Phase-2 upgrade . . . 31

3 Muon Trigger and Reconstruction at CMS HLT in Phase-2 **34**
- 3.1 The L1 Muon Trigger . . . 34
  - 3.1.1 Barrel region . . . 37
  - 3.1.2 Endcap region . . . 38
  - 3.1.3 Overlap region . . . 40

3.1.4 Dedicated Trigger for Displaced Muons . . . 41
3.2 HLT Muon Reconstruction . . . 44
3.3 Offline Displaced Muon Reconstruction . . . 47

4 Optimizing the Online Muon Reconstruction 49
4.1 The Standalone Muon Reconstruction optimization . . . 51
4.2 The Tracker and Global Muon Reconstruction optimization . . . 54
4.2.1 L3 reconstruction Inside-Out first . . . 56
4.2.2 L3 reconstruction Outside-In first . . . 57
4.3 Results . . . 58
4.3.1 Physics performance . . . 59
4.3.2 Timing and computing performance . . . 73
4.4 Future work . . . 79

Conclusions 81

Appendices 83

A Muon Triggering and Reconstruction Glossary 83

B Code and data availability 85

Bibliography 86

Chapter 1

The Large Hadron Collider

The Large Hadron Collider (LHC) is a circular superconducting hadron accelerator and collider installed in the existing 27 km long tunnel built in the late 80s for the CERN LEP machine [1]. In addition to the main circular tunnel, two transfer tunnels link the LHC to the CERN accelerator complex. The latter acts as the main injector for proton beams circulating in the LHC. Figure 1.1 shows a schematic diagram of the CERN accelerator complex in 2022. The original tunnels and civil engineering structures have been fully utilised when moving from LEP to LHC, with additional modifications required. The LHC features 4 interaction points for the 4 main experiments hosted in the accelerator facility: ATLAS [2], CMS [3], ALICE [4] and LHCb [5]. Broadly speaking, the infrastructure for ATLAS and CMS was built anew for the LHC, while underground and surface structures for ALICE and LHCb are largely reused LEP-era buildings and facilities.

The CERN council approved the LHC project in December 1994 as a two-stage machine starting from a center-of-mass energy of 10 TeV to be later increased to 14 TeV. However, in 1996 the final approval was given for the construction of a single-stage machine with a center-of-mass energy of 14 TeV, which was to benefit strongly from the experience accumulated by the physics and CERN communities during the successful operation of LEP.

The LHC is a particle-particle collider, therefore it is composed of two rings where protons circulate in opposite directions, unlike particle-antiparticle colliders, like LEP, where the circulating beams can share the same ring. The tunnel facility built for LEP allowed up to eight interaction points, for geometrical reasons and to compensate for the higher synchrotron radiation losses typical of electrons and positrons. Since the LHC does not have the same synchrotron radiation problems as LEP, only four of the experimental caverns have been instrumented.

The beams are accelerated and focused by a complex system of superconducting electromagnets (section 1.1) and travel in extreme vacuum conditions (section 1.3).

The LHC aims to precisely measure Standard Model physics and reveal Beyond the

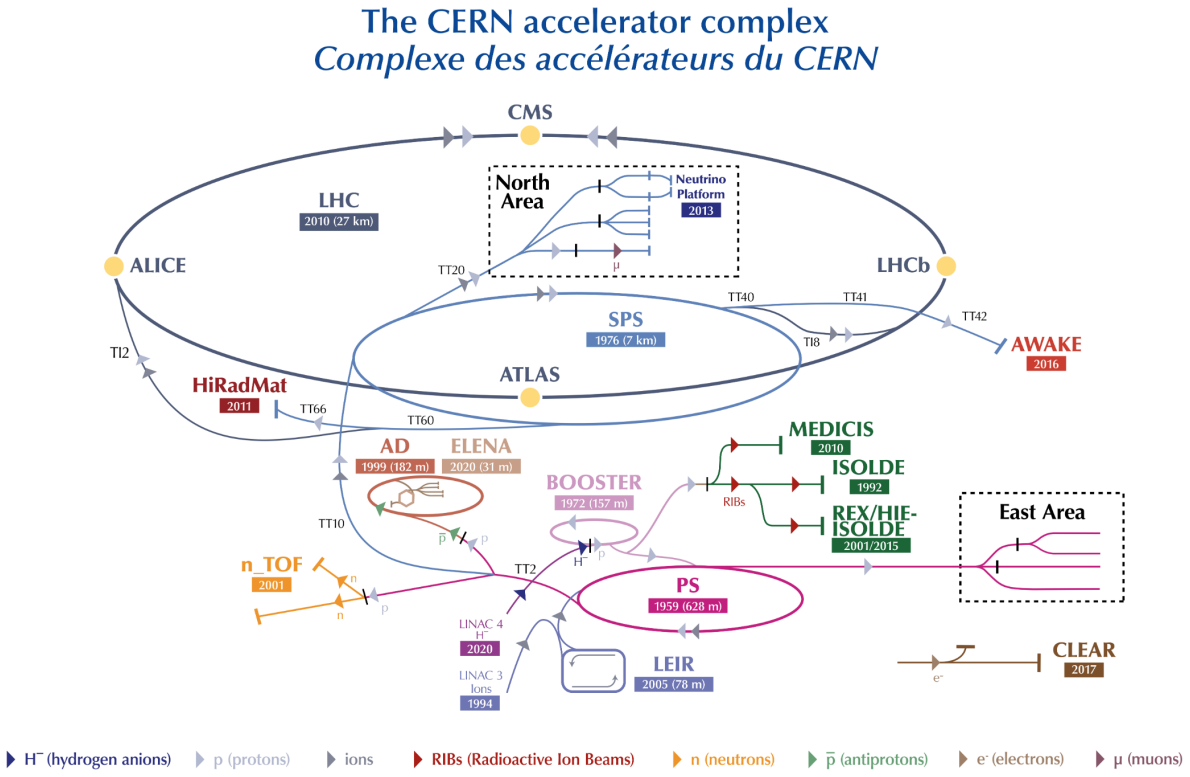


Figure 1.1: *The CERN accelerator complex in 2022 [6].*

Standard Model phenomena. The number of events per second generated by the LHC is given by:

$$N_{event} = L\sigma_{event} \tag{1.1}$$

where L is the machine luminosity and σ_{event} is the cross section for the event under study. The machine luminosity depends on multiple parameters including the geometrical characteristics of the beam and the interaction. Being related to the number of events produced, a high luminosity is required when studying rare events.

The LHC hosts four main experiments:

- two high-luminosity, general-purpose experiments: ATLAS [2] and CMS [3];
- two low-luminosity, specialized experiments: ALICE [4] and LHCb [5]. The former aimed at lead-lead collisions exploration and the latter meant for B-physics measurements.

Both high-intensity experiments aim for an instantaneous luminosity $L = 10^{34}$ cm^{2}s^{-1} for proton-proton operation. On the other hand, LHCb aims at a lower instantaneous luminosity $L = 10^{32}$ cm^{2}s^{-1} for B-sector precision measurements. Finally, the dedicated ion runs for the ALICE experiment aim at a peak luminosity of $L = 10^{27}$ cm^{2}s^{-1} for ion-ion collisions. Integrated luminosity, obtained by integrating the instantaneous luminosity over the operation time of the LHC, gives a measure of how much physics data is collected during a specific time interval. Integrated luminosity has the dimensions of an inverse area and is usually expressed using inverse barns (b^{-1}) with 1 b = 10^{-28} m^2 Each of the main experiments is further described in section 1.4, with a particular focus on the CMS experiment.

The LHC successfully delivered 7 (8) TeV center-of-mass proton-proton collisions from 2009 (April 2012) to the end of Run 1 in 2013, corresponding to a total integrated luminosity of about 30 fb^{-1}. It then started operations back up after the first long shutdown (LS1) in 2015, operating with a 13 TeV center-of-mass energy. LHC operation led to a series of measurements and discoveries, culminating in the confirmation of the Higgs Boson existence in 2012, observed by both the CMS and ATLAS experiments [7, 8].

To further improve the LHC discovery potential, a major upgrade is foreseen between the end of the 2020s and the beginning of the 2030s aimed at increasing the luminosity of the accelerator by a factor of 5 to 7.5 beyond its design value. This is known as the High-Luminosity LHC (HL-LHC) upgrade and is further discussed in section 1.5

1.1 Superconducting magnets

The LHC relies on superconducting magnets to accelerate protons as well as correct their trajectories to make them collide head-on in the four interaction points [9]. The

LHC magnet system makes use of technology proven in previous accelerators with NbTi superconductors cooled by helium at temperatures below 2 K producing fields above 8 T. The electromagnets use a current of more than 11000 amperes to produce the magnetic field, at such currents a superconducting coil is necessary to remove energy losses due to electrical resistance.

The proton acceleration chain develops through the whole CERN accelerator chain where particles are sped up in a series of interconnected accelerators, being shot in the next one when they reach the maximum energy that one part of the chain can achieve. In this context, more than 50 types of magnets are necessary both for the acceleration itself and for bending the trajectory of the particles into the complex path of the acceleration chain.

There are two main categories of magnets:

- Lattice magnets are used to keep the particle beams stable and precisely aligned. Dipole magnets are used to bend the trajectory of the beams in the circular orbit of the LHC, with increasing magnetic fields corresponding to tighter turns. The proton beams circulating the LHC are composed of multiple bunches of protons since they are more likely to collide in greater numbers if they are grouped together when they reach one of the detectors. Because of this, quadrupole magnets are used to tighten the beam by focusing it either vertically or horizontally. These magnets are usually used in pairs to have a net focusing effect in both the vertical and horizontal direction in the end.

- Insertion magnets take over when the beams enter the detectors. These magnets squeeze the particles closer together to prepare them to collide with particles coming from the opposite direction. A system of three quadrupole magnets, called an inner triplet, is used to achieve a significant reduction of the beam width: from 0.2 mm down to 16 μm across. The particle beams are separated again by dipole magnets after colliding, with supporting magnets to reduce the spread of the particles due to the collisions. When it is time to dispose of a beam because bunches have been exhausted or a malfunction requires an instant shutdown, the beams are deflected from the LHC along a straight line leading to a beam dump. A special magnet reduces the beam intensity by a factor of 100000 before the beam collides with a block of concrete and graphite composite. Finally, insertion magnets are also responsible for cleaning the beams removing stray particles to avoid them coming in contact with the sensitive components of the LHC and the experiments it hosts.

1.2 RF cavities

Radiofrequency (RF) cavities are responsible for the acceleration and storage of the proton beam injected in the LHC [10]. Such cavities are metallic chambers containing

an electromagnetic field modelled in a way such that charged particles injected into it receive an electrical impulse that accelerates them.

The LHC hosts a total of 16 RF cavities, working in a superconducting state to increase the energy of the beam injected into the LHC by a factor of more than 14 from about 450 GeV to more than 6.5 TeV in about 20 minutes of operation.

Each cavity is driven by an intensity-modulated electron beam with a frequency of 400 MHz, making the resulting electromagnetic field an oscillating one. Therefore, the timing of the arrival of the particles is important. Such precise cavity modulation allows the RF system to accelerate and separate protons into bunches. Since particles are accelerated by the force due to the varying electromagnetic field, the ideally timed proton, with exactly the right energy, feels a net zero force when the LHC is running at nominal energy, while protons with slight differences in energy are accelerated or decelerated sorting particles into bunches. At regime conditions, each proton beam is divided into 2808 bunches, each containing about 10^{11} protons. At full luminosity, bunches circle the entire LHC circumference at a 40 MHz frequency, giving a time separation between bunches of 25 ns, resulting in about 600 million collisions per second.

1.3 Vacuum system

The LHC has three separate vacuum systems: the insulation vacuum for the cryogenically cooled magnets, the beam pipes vacuum where the proton bunches circulate, and the vacuum used to insulate the helium distribution line. The LHC vacuum system is made up of 104 km of piping divided into 50 km dedicated to thermal insulation with the remaining 54 km being the beam pipes through which the LHC beams travel.

The two insulating vacua are used to thermally isolate the cryomagnets and the helium distribution line which are kept at 1.9 K. These vacua do not have to be better than 10^{-1} mbar.

However, the requirements are significantly more stringent for the beam vacuum, driven by the need to increase the beam lifetime and reduce the background at the interaction points. In this case, the pressure inside the pipes must drop down to around 10^{-10} to 10^{-11} mbar: a vacuum similar to the one found on the surface of the Moon. The beam pipes are made up of 48 km of arc sections, kept at 1.9 K to allow the superconducting magnets to bend the beam trajectory as discussed in section 1.1, and 6 km of straight sections, kept at room temperature. In the arcs, the ultra-high vacuum is achieved by cryogenic pumping of 9000 m^3 of gas. Since the beam pipes are cooled to extremely low temperatures, the gas condenses and sticks to the walls of the beam pipe. About two weeks of pumping are required to bring the pressure down from atmospheric pressure to around 10^{-10} mbar. As far as the straight sections are concerned, they make use of a non-evaporable "getter coating", designed at CERN, that absorbs residual molecules when heated. This coating is effective for removing all gases, excluding methane and

noble gases, which are instead removed by pumps. Moreover, being at room temperature, these sections allow the "bakeout" of all components at 300°C. Bakeout consists of heating the beam pipes from the outside in order to improve the quality of the vacuum.

1.4 The main LHC experiments

The LHC has been designed to have four interaction points along its circumference. Each of the interaction points houses one of the main experiments, featuring its own detector. There is a macroscopic difference between the four main LHC experiments:

- two of them, CMS and ATLAS, are general-purpose experiments, designed to push the limits of our understanding of the Standard Model and directly search for new physics;
- the other two experiments, ALICE and LHCb, are specialised machines, specifically designed for a narrower purpose.

In addition to these experiments, the LHC hosts a series of smaller experiments, usually tied to the operation of one of the aforementioned main experiments. Some of these experiments, which tend to focus on much more specific and narrower physics measurements, are presented in section 1.4.5.

In the following sections, each of the main LHC experiments is briefly introduced, while a major focus is reserved for the CMS experiment, presented in more detail in Chapter 2.

1.4.1 ALICE

A Large Ion Collider Experiment (ALICE) [4, 11] specialises in heavy ion collisions. It is designed to study the strong interaction at extreme energy densities, where quarks and gluon condensate to form an exotic state of matter known as "quark-gluon plasma". In ion-ion collision at the LHC, extreme energies and temperatures are reached, freeing the quarks that are normally confined within the protons and neutrons in nuclei. This mixture of free quarks and gluons, which are normally responsible for keeping the quarks together, is known as quark-gluon plasma and is a fundamental cornerstone of the theory of strong interactions: quantum chromodynamics (QCD). The ALICE collaboration studies the quark-gluon plasma as it expands and cools, observing how it progressively gives rise to the particles that constitute the matter of our universe today. The ALICE collaboration includes almost 2000 scientists from 174 physics institutes in 40 countries (updated April 2022). Some specifications of the ALICE detector are shown in table 1.1.

Dimensions	length: 26 m, height: 16 m, width: 16 m
Weight	10 000 tons
Design	central barrel plus single-arm forward muon spectrometer
Cost of materials	115 MCHF
Location	St. Genis-Pouilly, France (LHC Point 2)

Table 1.1: *ALICE detector specifications.*

1.4.2 LHCb

The Large Hadron Collider beauty (LHCb) [5, 12] experiment focuses on the study of the bottom quark to investigate slight differences between regular matter and antimatter. LHCb is not a 4π detector, meaning that the interaction point is not entirely surrounded by active detectors, the experiment instead focuses on forward particles, those thrown forwards by the collision in one direction. The first subdetector, used to identify the vertex where the proton-proton interaction took place, is mounted extremely close to the collision point, while the others follow one behind the other for a length of about 20 meters. About 1565 scientists, engineers and technicians from 20 countries make up the LHCb collaboration (updated March 2022). Some specifications of the LHCb experiment are reported in table 1.2

Dimensions	length: 21 m, height: 10 m, width: 13 m
Weight	5600 tons
Design	forward spectrometer with planar detectors
Cost of materials	75 MCHF
Location	Ferney-Voltaire, France (LHC Point 8)

Table 1.2: *LHCb detector specifications.*

1.4.3 ATLAS

A Toroidal LHC ApparatuS (ATLAS) [2, 13] is, together with CMS, one of the two general-purpose experiments at the LHC. As such, it is meant to measure and investigate a wide range of physics, from the Higgs boson to Dark Matter candidates and exotic states beyond the Standard Model. It shares the same scientific goals as the CMS experiment but is built taking advantage of different technical solutions and magnet system design. More than 5500 scientists from 245 institutes in 42 countries work on the ATLAS experiment (updated March 2022). Table 1.3 shows some of the specifications of the ATLAS experiment.

Dimensions	length: 46 m, height: 25 m, width: 25 m
Weight	7000 tons
Design	barrel plus endcaps
Cost of materials	540 MCHF
Location	Meyrin, Switzerland (LHC Point 1)

Table 1.3: *ATLAS detector specifications.*

1.4.4 CMS

Compact Muon Solenoid (CMS) [3, 14] is the other general-purpose experiment hosted at the LHC. The experiment is built around a large solenoid magnet. In particular, the magnet is made up of a cylindrical coil of superconducting cable that generates a 4 T magnetic field. The CMS experiment is one of the largest international scientific collaborations in history, involving about 5500 particle physicists, engineers, technicians, students and support staff from 241 institutes in 54 countries (updated May 2022). Some characteristics of the detector are reported in table 1.4 while the experiment is discussed in more detail in Chapter 2.

Dimensions	length: 21 m, height: 15 m, width: 15 m
Weight	12 500 tons
Design	barrel plus endcaps
Cost of materials	500 MCHF
Location	Cessy, France (LHC Point 5)

Table 1.4: *CMS detector specifications.*

1.4.5 Other experiments at the LHC

The LHC hosts a series of smaller experiments other than the four main ones mentioned above. These smaller experiments are generally focused on a narrower physics measurement programme and exploit the same interaction point and collisions as one of the main experiments.

- The Large Hadron Collider forward (LHCf) [15, 16] uses particles thrown forward by collisions in the LHC to simulate cosmic rays. Cosmic rays are charged particles coming from outer space that naturally and constantly interact with our planet's atmosphere triggering a cascade of particles, some of which can reach ground level. Having access to cosmic-ray-like conditions in a controlled environment can help to calibrate large-scale cosmic-ray experiments. LHCf is made up of two detectors along the LHC beamline, at 140 meters on either side of the ATLAS interaction

point. Each of the two detectors weighs only 40 kilograms and measures 30 cm long, 80 cm high, and 10 cm wide.

- MAssive Timing Hodoscope for Ultra-Stable neutraL pArticles (MATHUSLA) [17, 18] is a proposed experiment at LHC. It is a dedicated large-volume detector that would be installed on the surface above CMS or ATLAS. Such a detector would act as a displaced vertex detector to search for beyond the Standard Model long-lived particles (LLPs) that could decay meters away from the primary interaction vertex, thus outside the sensitive volume of the main experiments. MATHUSLA would also act as a cosmic ray telescope at CERN.

- MilliQan [19] is a sub-detector experiment tied to the CMS experiment. This experiment aims at detecting and measuring millicharged particles: particles with charges much smaller than that of the electron. The detector is installed in a tunnel 33 m away from the CMS interaction point, with 17 m of rock shielding to reduce beam backgrounds.

- Monopole and Exotic particle Detector At the LHC (MoEDAL) [20, 21] searches directly for the magnetic monopole, a hypothetical particle with either a "north" or a "south" magnetic charge, but not both. The MoEDAL detector exploits the same interaction point as LHCb. It is composed of two main sub-modules. First, its tracking capabilities are granted by 125 Nuclear Track Detectors in the form of 47 m^2 of plastic layers. Secondly, it has about a tonne of trapping detectors meant to register and capture exotic particles. Exotic particles could form a tiny trail as they traverse the Nuclear Track Detectors, breaking long-chain molecules in the plastic. In preparation for data taking during run 3 of the LHC, the MoEDAL detector has been upgraded to MoEDAL-MAPP. The additional detector, MAPP (MoEDAL Apparatus for Penetrating Particles), aims to extend MoEDAL's physics reach by providing sensitivity to millicharged particles and long-lived exotic particles.

- The Total, elastic and diffractive cross-section measurement (TOTEM) experiment [22, 23] measures the total proton-proton cross-section and studies elastic and diffractive scattering at the LHC. TOTEM detectors are spread across almost half a kilometre around the CMS interaction point. The TOTEM experiment is made up of almost 3000 kg of equipment, including four particle "telescopes" and 26 "roman pots" detectors. The telescopes use Cathode Strip Chambers and Gas Electron Multipliers to track the particles produced in collision at the CMS interaction points. The Roman pots are special detectors equipped with silicon sensors used to perform measurements of scattered protons and with the unique ability to move sensors both vertically and horizontally.

- The ForwArd Search ExpeRiment (FASER) [24, 25] is designed to search for light and extremely weakly interacting particles. The existence of such exotic particles,

albeit not yet confirmed, is predicted by a multitude of models of physics beyond the Standard Model. FASER is located along the beam trajectory, 480 m away from the ATLAS detector, as such it is ideally positioned to detect particles emitted in the very forward region, invisible to the ATLAS experiment due to the need to have a hole in the detector to allow the beam pipe to cross it. FASER also has a sub-detector, FASERν, specifically designed to detect neutrinos. This detector could provide valuable new data, since no neutrino produced at the LHC has ever been detected, despite the large number that are produced in each beam crossing and the high energies that they carry.

1.5 High-Luminosity LHC

The High-Luminosity Large Hadron Collider project (HL-LHC) [26] aims to enhance the discovery potential of the LHC. The Large Hadron Collider has successfully produced proton-proton collisions with center-of-mass energy 7-8 TeV during Run 1 in the early 2010s, then up to almost 14 TeV during Run 2 and 3 from spring 2015. However, to enhance the discovery potential of the four main experiments hosted at the LHC, the accelerator itself needs a major upgrade to extend its operability and substantially increase the luminosity it can achieve. A more powerful LHC allows rarer events to be detected and studied, pushing our understanding of the Standard Model forward. It is expected that the engineering work to upgrade the LHC will allow the machine to increase the luminosity from the current value of about 10^{34} cm^{-2}s^{-1} to 5×10^{34} cm^{-2}s^{-1}. The margin applied when developing the new components should even allow to operate the machine up to a peak instantaneous luminosity of 7.5×10^{34} cm^{-2}s^{-1}. This represents an increase in peak instantaneous luminosity by a factor of 5 to 7.5 with respect to the current LHC design. The upgrade relies on a series of innovations that push accelerator technology forward. Among these, there are 11-12 T superconducting magnets, compact superconducting cavities for beam control, and new technology for beam collimation.

The foreseen increase in luminosity is tied to a corresponding increase in pile-up, i.e. the number of proton-proton collision events per bunch crossing, with averages expected to hit from 140 to 200: a substantial increase over the current average of about 70 pile-up in ATLAS and CMS. Pile-up increasing poses significant challenges for both the detectors and the reconstruction of the current experiments, which will all undergo significant upgrades before the beginning of Run 4. Some of the changes planned for the CMS experiments are described in Chapter 2, with more details on the muon system discussed in Chapter 3.

Figure 1.2 shows the timeline for HL-LHC, currently scheduled to start data taking after the third LHC Long Shutdown (LS3) after the end of Run 3, whose starting has been pushed back by the COVID-19 pandemic.

Figure 1.2: *Timeline of LHC operations from Run 1 to HL-LHC. After the third long shutdown period (LS3), the HL-LHC is supposed to start taking data.*

Chapter 2

The CMS experiment

The Compact Muon Solenoid (CMS) [3, 14] detector is a multi-purpose experiment currently operating at the Large Hadron Collider (LHC) at CERN. The detector was designed to study various physics phenomena in both the electroweak and strong sectors of the Standard Model, as well as explore exotic possibilities beyond our current understanding of fundamental physics. Among the physics objectives of the experiment are:

- the exploration of the electroweak sector;
- search for the Higgs boson, which was discovered in 2012 [7, 8], including measuring its properties;
- precision measurements of the Standard Model particles and interactions;
- flavour physics;
- heavy-ion physics;
- searches for new physics beyond the Standard Model.

During the LHC Run 1, between 2009 and 2012, CMS recorded a total integrated luminosity of about 30 fb^{-1}. Run 2 followed between 2015 and 2018, after the first long shutdown (LS1). In this case, the LHC provided a centre-of-mass energy of 13 TeV, total integrated luminosity of about 165 fb^{-1}, and peak instantaneous luminosities up to 2×10^{-34} cm^{-2} s^{-1}. During LS1 the first set of detector upgrades, referred to as Phase-1, was implemented. The LHC Run 3 started in 2022, it is scheduled to end by 2026 with a total integrated luminosity collected of about 250 fb^{-1} at centre-of-mass energy 13.6 TeV. During LS3, scheduled to start in 2026 at the end of Run 3, CMS will undergo a major upgrade, referred to as Phase-2, to prepare for data taking at the HL-LHC, designed to deliver instantaneous luminosities up to 7.5×10^{-34} cm^{-2} s^{-1} at a centre-of-mass energy

of 14 TeV. At the end of HL-LHC, a total integrated luminosity of 3000 fb^{-1} will be collected, more than a factor 10 improvement over Run 3.

CMS adopts a right-handed coordinate system centered at the nominal collision point with the y-axis pointing upward, and the x-axis pointing towards the centre of the LHC. Therefore, the z-axis sits along the beam direction. The azimuthal angle ϕ is measured in the $x - y$ plane within a $0 \leq \phi \leq 2\pi$ range. In this plane, the radial coordinate is denoted by r. The polar angle θ is measured from the z-axis within a $0 \leq \theta \leq \pi$ range. A schematic representation of this coordinate system is shown in figure 2.1.

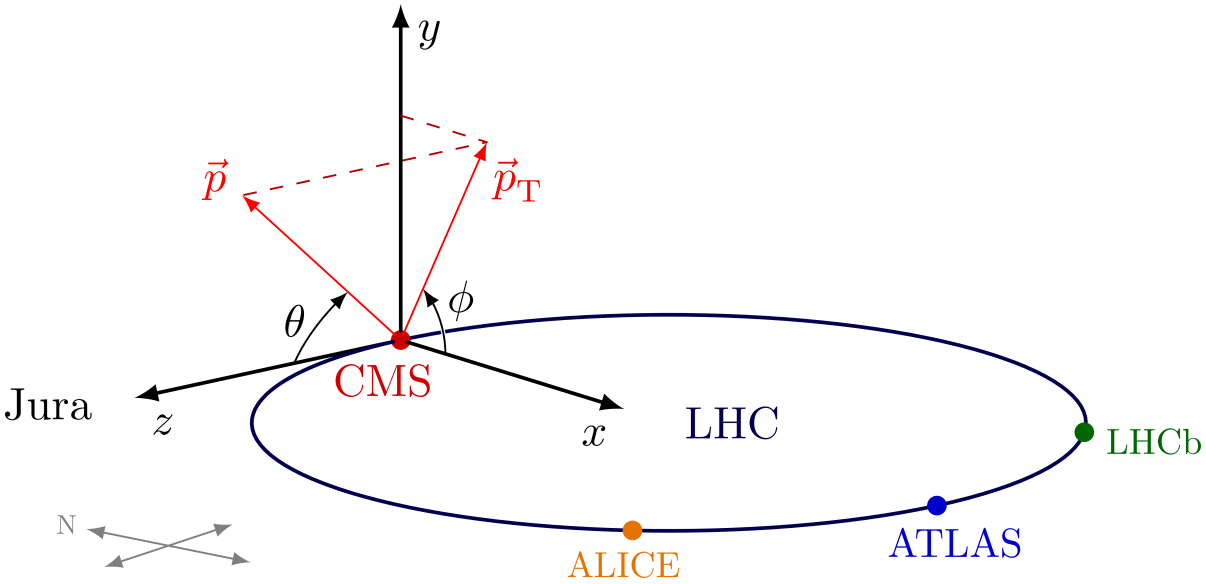


Figure 2.1: *CMS coordinate system.*

The pseudorapidity η is often used in place of the polar angle and it is defined as:

$$\eta = -ln\left(tan\frac{\theta}{2}\right) \tag{2.1}$$

In the following, the CMS detector and all its subsystems are described, with a particular focus on the muon system. The planned upgrades for Phase-2 are also mentioned.

2.1 Concept and structure

The CMS detector is designed to be nearly hermetic with a cylindrical shape and a multiple-layer design. It has an overall length of 22 m, a diameter of 15 m and weights 14000 tons. Figure 2.2 shows a schematic view of the CMS detector.

One of the most important factors that influenced the CMS design is the requirement of measuring muon momentum precisely. Since a large bending power is required to

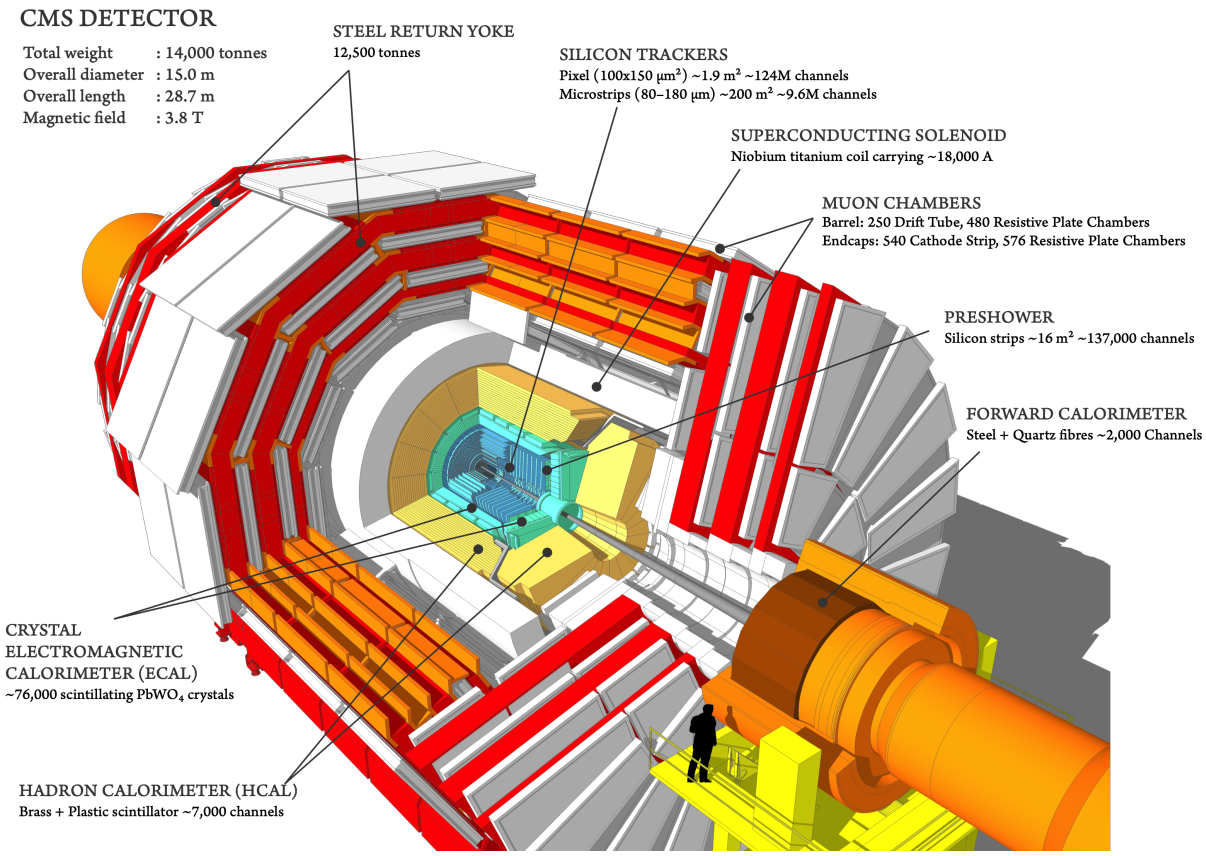


Figure 2.2: *Schematic drawing of the CMS detector.*

measure the momentum of high-energy charged particles, the key feature of CMS is a superconducting solenoid 12.5 m long and with an internal diameter of 6 m, which provides a 3.8 T magnetic field. Most of the detectors are inside the magnetic volume, with only the muon chambers being outside the solenoid. Starting from the layer closest to the beam pipe and moving radially towards the outside, the CMS detector is made up of:

- a silicon pixel and a strip tracker to reconstruct charged particles and secondary vertices from the decays of very short-lived particles;
- a lead tungstate crystal ($PbWO_4$) electromagnetic calorimeter (ECAL) that measures the energy deposited by electrons and photons;
- a brass and scintillator hadronic calorimeter (HCAL) that measures the energy deposited by hadrons;
- gas-ionization detectors embedded in the steel flux-return yoke of the solenoid to track muons that punch through the calorimeters.

The LHC provides a bunch crossing every 25 ns with a mean pile-up of about 75 in Run 3. This translates to more than a billion proton-proton interactions per second. Therefore, a fast event-selection chain is needed to be able to quickly decide whether to store a particular collision for further analysis or discard it. This is what the CMS triggering system does. It is a two-tiered system with a first-level (L1) trigger implemented in custom hardware processors and a second level, known as High-Level Trigger (HLT), implemented in software on a cluster of commercial processors running a version of the offline CMS event reconstruction optimized for timing [27]. The L1 trigger can select events at a rate of about 100 kHz with a latency of 4 μs [28], using information from the calorimeters and muon chambers. The HLT was originally designed to reduce the event rate to about 1 kHz before data storage but it is currently operating at a rate of about 5 kHz in Run 3.

2.2 Inner tracking system

The inner tracking system measures the trajectories of charged particles and locates the primary and secondary vertices of interaction. Momentum is also measured from the trajectory since the tracks of charged particles are bent by the magnetic field: the greater the momentum, the larger their curvature radius. The inner tracking system is made up of two modules: the pixel detector and the silicon strip tracker. Both modules are entirely silicon-based and are meant to have a low degree of interference with particles and a high radiation tolerance.

2.2.1 Pixel detector

The pixel is responsible for vertex location and, as such, it is the closest detector to the interaction point. Currently, the pixel detector is made up of four barrel layers and three discs in each endcap, totalling 124 million readout channels. It provides four-point tracking for charged particles, ensuring good performance even above the design luminosity of the LHC. Each layer is divided into small units, the pixels: n-in-n type silicon sensors with dimensions 150 $\times$ 100 μm. A charged particle crossing one of the pixels deposits enough energy to produce an electron-hole pair in the silicon sensor. The resulting signal is received by an amplification and readout chips.

The two most important parameters for pixel performance are hit efficiency and position resolution since both strongly affect the ability to perform pattern recognition and b tagging.

The hit efficiencies measured at an instantaneous luminosity of 2×10^{34} cm^{-2} s^{-1} are 97, 98, 99, and 99.5% for the barrel layers 1-4, respectively. For the endcap disks the average efficiency is 99% [29].

The position resolution is measured using the "triplet" method [29], where the expected hit position in a layer is extrapolated from two other layers. The position resolution for the third layer is 11 μm in the $r-\phi$ direction and 24.3 μm in the z direction. The inner layers have slightly lower position resolution mainly due to radiation damage. For the forward disks, the resolution is 11.9 μm in the r direction and 21.0 μm in the z direction. Being the closest detector to the beam pipe, pixel sensors suffer most from radiation damage induced by the thousands of charged particles produced in each interaction. To maintain high efficiency and resolution, the voltage bias applied to the silicon pixels is periodically increased from a nominal value of 150 V right after installation up to 800 V in the first layer and 600 V for all the others.

2.2.2 Silicon strip tracker

The Silicon Strip Tracker (SST) is made up of ten cylindrical layers of silicon strip sensors in the barrel and nine disks in either endcap. Together with the pixel detector, it measures the trajectories of charged particles up to a pseudorapidity of $\eta = 2.5$. The tracker sensors are segmented into long, thin, strips used to measure the trajectories of charged particles and provide a hit resolution of 20 μm for particles that cross them perpendicularly. In total, the SST has 9.3 million silicon micro-strips corresponding to 198 m^2 of active silicon area distributed over 15148 modules. Most layers use single-sided p-on-n micro-strip sensors used to measure the r and ϕ coordinates in the barrel, while ϕ and z are measured by the modules in the endcaps. In four layers in the barrel and three rings in the endcaps, double-sided modules are used to add a course measurement of an additional coordinate: z in the barrel and r in the endcaps. Figure 2.3 shows a schematic view of one quadrant of the CMS inner tracking system.

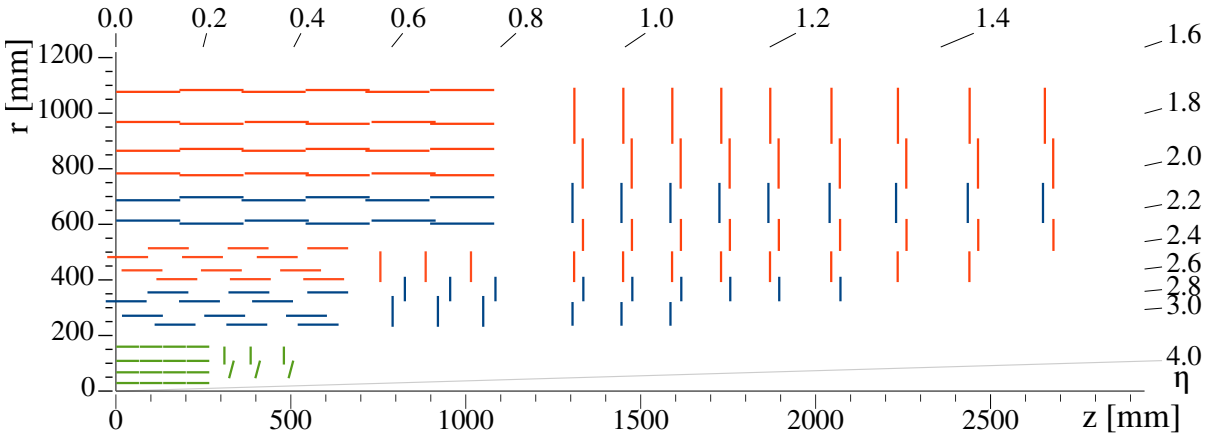


Figure 2.3: *Schematic view of one quadrant of the CMS inner tracking system in the $r-z$ view. The pixel detector is shown in green, while single-sided and double-sided tracking modules are represented as red and blue segments, respectively.*

Some key performance metrics are momentum resolution and tracking efficiency. The CMS tracker can detect and track particles with transverse momentum p_T as low as 50 MeV within $\eta < 2.5$. Tracks with momentum around 100 GeV have an impact parameter resolution of about 10 μm and a p_T resolution close to 1%. Furthermore, an important aspect of the Tracker is the hit efficiency: the detection efficiency for a particle traversing a strip. The measurement for hit efficiency is performed on high-purity tracks with trajectories close to sensor edges being ignored to avoid inactive regions. The efficiency is determined from the fraction of traversing tracks with a hit in a module anywhere within a range of 15 strips from the expected position [30]. The average SST hit efficiency measured during Run 2 is about 99.5%, with some variation between layers.

2.3 Electromagnetic calorimeter

The Electromagnetic CALorimeter (ECAL) detects, absorbs, and measures electrons and photons. It is made of 75848 lead tungstate ($PbWO_4$) crystals: 61200 in the barrel and 7324 in each endcap, covering a pseudorapidity range $\eta < 3$. The 23 cm deep crystals, having high density and short radiation length, correspond to about 25 radiation lengths X_0 with fine granularity. Moreover, they constitute an optimal choice for a radiation-hard, fast calorimeter to cope with bunch crossings every 25 ns with thousands of charged particles produced each second. The crystals emit blue-green scintillation light in fast, short, photon-bursts in amounts proportional to the energy deposited by electrons or photons. About 80% of the scintillation light is emitted in 25 ns, reducing the light contamination from different bunch crossings. The emitted light is collected and amplified by photodiodes attached directly to the back of each crystal. Figure 2.4 shows one of the crystals with the photodiode attached.

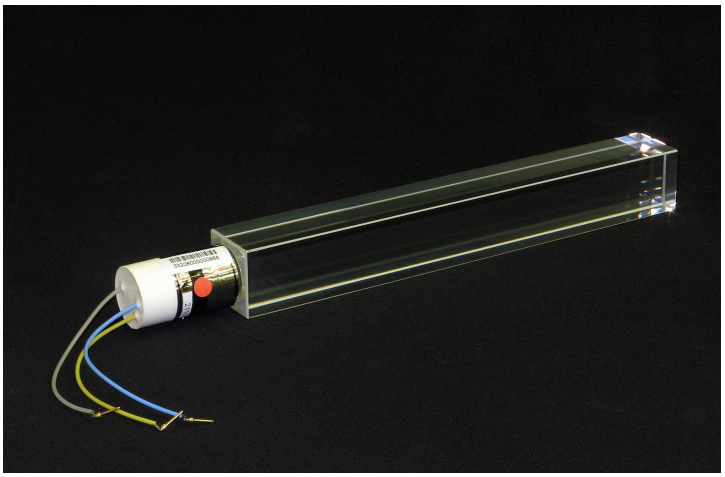


Figure 2.4: *A $PbWO_4$ crystal with photodetector attached.*

Electrons and photons are identified and measured from energy deposits using algorithms to cluster the deposits in each crystal and constrain the size and shape of each cluster to what is expected for electrons or photons. ECAL also contributes to the measurements of jets (via their electromagnetic component) and missing transverse momentum, mostly carried by neutrinos that are not detected at all in CMS. The electron momentum is estimated by combining information from the tracker and energy deposits in the calorimeter. This method results in a momentum resolution for electrons with $p_T \approx 45$ GeV in a range between 1.6 and 5%. The momentum resolution is generally better in the barrel and can vary depending on the energy loss experienced by the electron in the layers in front of the ECAL (namely, the pixel and the tracker).

One peculiar characteristic of the crystals used in the ECAL is that their transparency changes when they are irradiated [31]. This leads to the creation of colour centres that absorb some of the scintillation light, modifying the expected response of the calorimeter. This process is dynamic and depends on the dose absorbed by the crystals. Partial recovery occurs spontaneously at room temperature when the crystals are not irradiated for prolonged periods. However, since the scintillation process is not altered by the transparency changes, it is possible to correct the response using a reference light signal. To this aim, the ECAL has been equipped with a dedicated laser monitoring system meant to apply correction to the response on a crystal-to-crystal basis. Figure 2.5 shows the long-term evolution of the ECAL response to laser light during Run 1 and Run 2.

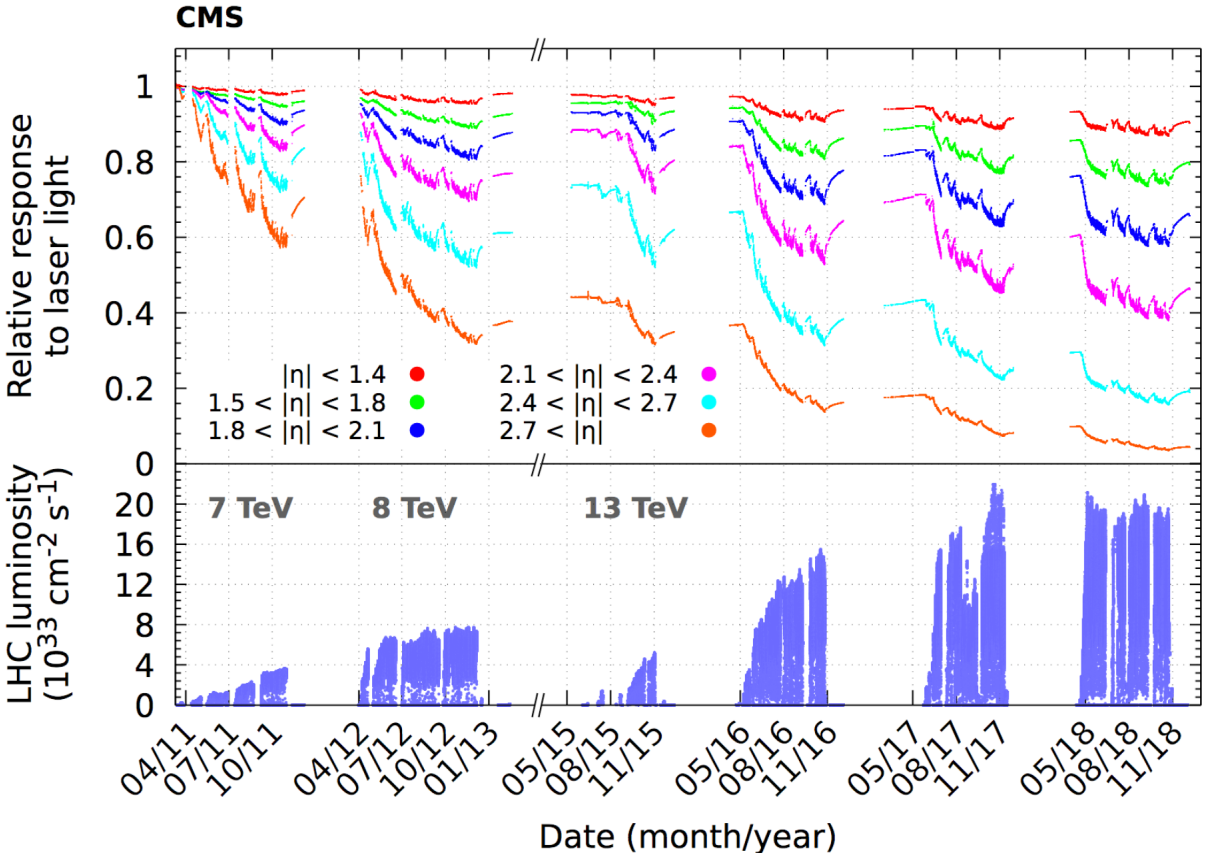


Figure 2.5: *Relative response to laser light injected into the ECAL crystals, measured by the laser monitoring system, averaged over all crystals in bins of $|\eta|$. The response change observed in the ECAL channels is up to 13% in the barrel and reaches up to 62% at $|\eta| \approx 2.5$, the limit of the CMS inner tracker acceptance. The response change is up to 96% in the region closest to the beam pipe. The recovery of the crystal response during the periods without collisions is visible. These measurements, performed every 40 minutes, are used to correct the physics data. The lower panel shows the LHC instantaneous luminosity as a function of time.*

2.4 Hadronic calorimeter

The Hadron CALorimeter (HCAL) is designed to measure the energy of both charged and neutral hadrons. It also contributes to the identification of hadrons, electrons, and photons as well as the reconstruction of jets and missing transverse momentum together with the ECAL. The HCAL is made up of four subdetectors for different coverage regions: the hadron barrel (HB), hadron endcap (HE), hadron outer (HO), and hadron forward (HF). The HB and HE are located inside the solenoid magnet, between the ECAL and the muon system. This constraints the amount of absorber material that can be used to stop the hadronic showers. Because of this, the HO is placed outside the solenoid to complement the measurements of HB. Finally, the HF modules are placed at 11.5 m from the interaction point on either side to extend the pseudorapidity coverage from $|\eta| < 3$ to $|\eta| < 5.2$. The HCAL is designed to have good hermeticity, covering almost the entire 4π solid angle. When combining information from the entire CMS detector, the jet energy resolution typically amounts to 15–20% at 30 GeV, 10% at 100 GeV, and 5% at 1 TeV. Figure 2.6 shows a schematic view of the HCAL with the four major subdetectors highlighted in different colours. The HB and HE cover the pseudorapidity regions $|\eta| < 1.392$ and $1.305 < |\eta| < 3.0$, respectively. The HO provides a measurement of the shower tails in the region $|\eta| < 1.26$, and the HF covers $3.0 < |\eta| < 5.2$.

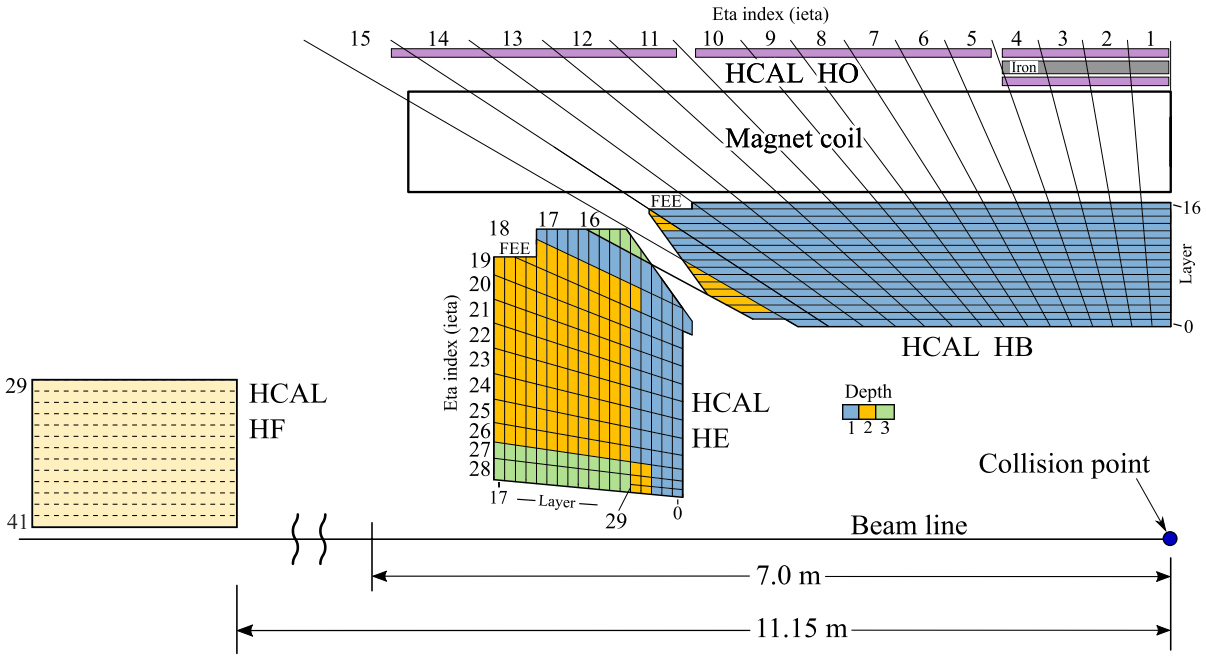
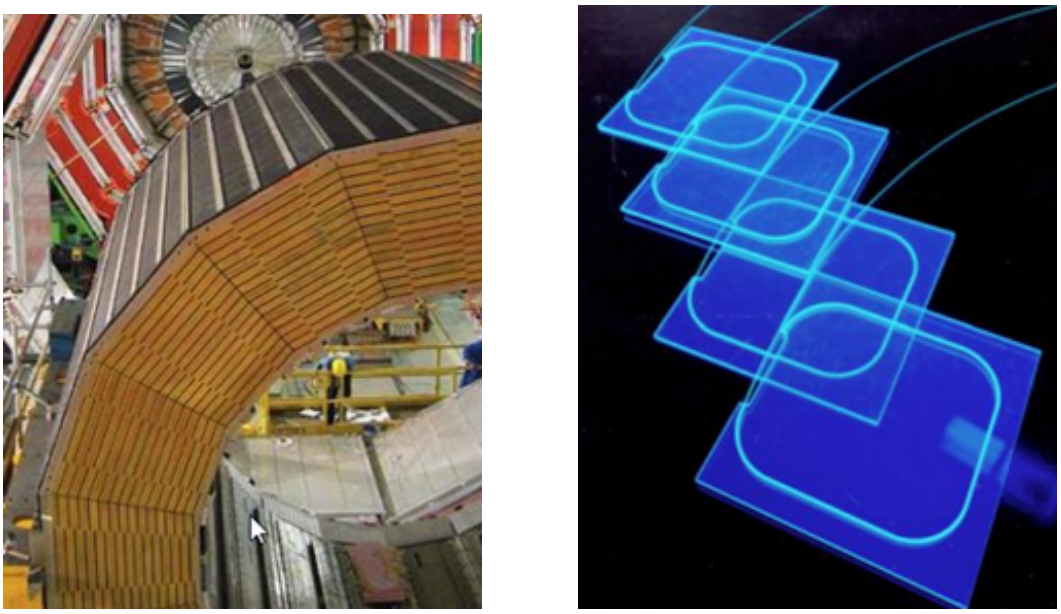


Figure 2.6: *Schematic view of the hadron calorimeter.*

The HB and HE are sampling plastic scintillating calorimeters which use brass as the absorber. The HB absorber and scintillating tiles are shown in figure 2.7(a) and (b),

respectively. The plastic scintillating tiles produce blue light that is shifted to green by wavelength shifting fibres embedded in the calorimeter so that it can be collected by silicon photomultipliers.



(a) *(b)*

Figure 2.7: *(a) brass absorber used in the hadron barrel calorimeter. (b) scintillating tiles with wavelength shifting fibres used as the active material in the barrel, endcap and outer hadron calorimeters.*

The HF is a sampling calorimeter with a steel absorber. Instead of plastic scintillators, it uses plastic-clad quartz fibres that produce Cherenkov light as the active elements.

2.5 Muon system

The muon (μ) is an elementary particle, belonging to the family of leptons. It has an electric charge of -1 (+1 for antimuons), a spin of 1/2, and a mass about 200 times higher than the electron mass at about 105 MeV. High-energy muons are produced as a result of proton-proton collision either directly or, more often, as decay products of intermediate products. Unlike most other detectable particles, including electrons, photons and most hadrons, muons cross the whole CMS experiment almost without interaction, being detected only in the tracker and in the most external group of detectors that form the muon system.

Muon detection enables to recognise signatures of interesting processes over the high background currently present at the LHC and expected at HL-LHC. Other than being

easier to identify than other particles thanks to their ability to traverse most of the detector almost unaffected, muons also provide the best mass resolution for events that result in final states with leptons at the Electroweak symmetry breaking scale.

The muon system has three core functions: muon identification, momentum measurement, and triggering. Good momentum resolution and triggering capabilities are granted by the high-field solenoidal magnet. While the triggering capabilities are discussed in more detail in Chapter 3, the momentum resolution achieved using information from all relevant CMS detectors is shown in figure 2.8. For muons with $p_T < 200$ GeV the tracker gives the most precise momentum measurement because they are more subject to multiple scattering in the calorimeters and the iron return yoke in the muon system, thus spoiling the momentum reconstruction based on the trajectory measured in the muon chambers. Muons with higher p_T benefit from the combination of the inner tracker and the muon system.

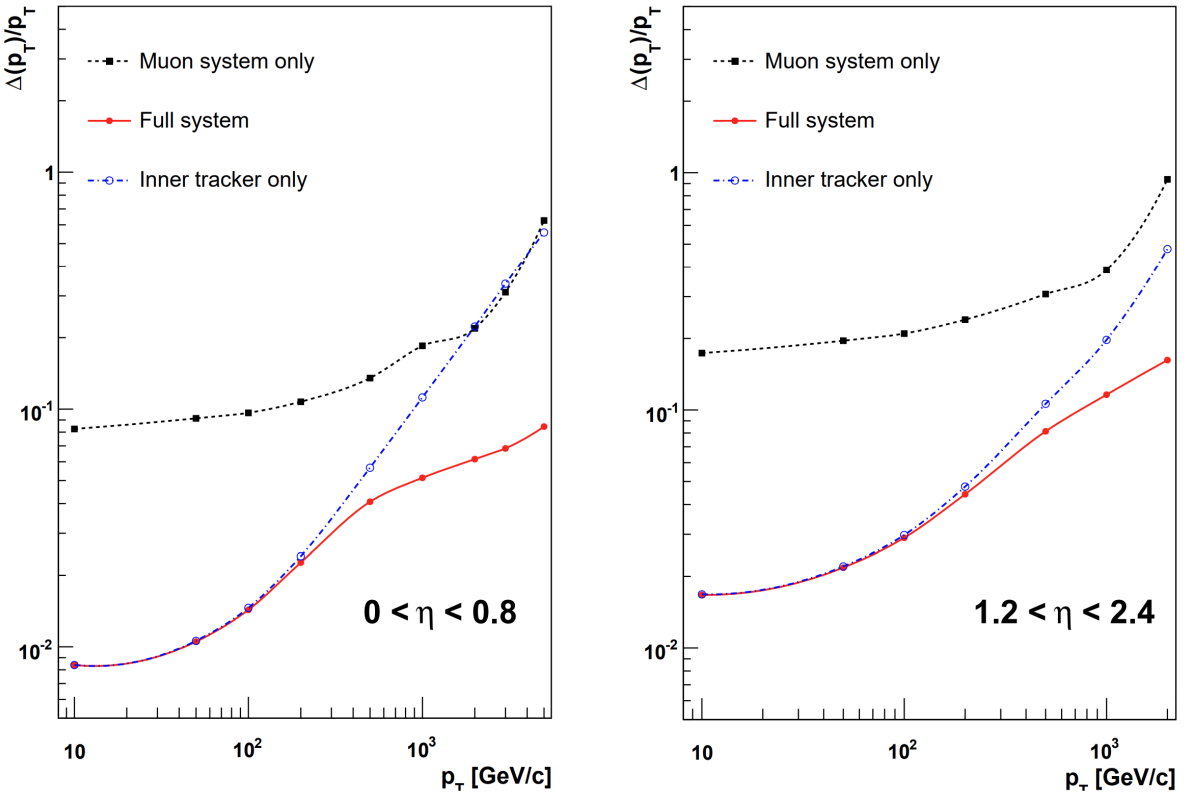


Figure 2.8: *Muon transverse momentum resolution as a function of the transverse momentum (p_T) using only the muon system (black), only the inner tracking system (blue), and both (red). The left plot shows the performance in the barrel ($|\eta| < 0.8$); the right plot shows the performance in the endcap $1.2 < |\eta| < 2.4$.*

The CMS muon system is designed to be able to reconstruct the momentum and the

charge of muons over the entire kinematic range of the LHC. Therefore, it comprises four types of gaseous detectors: the drift tubes (DTs), the cathode strip chambers (CSCs), the resistive-plate chambers (RPCs) and the recently added gas electron multiplier (GEM). Each of these detectors is detailed in a section in the following. Being placed outside of the solenoidal magnet, the muon system was driven to have a cylindrical barrel section and 2 planar endcap regions. Figure 2.9 shows a schematic representation of the CMS muon system at the start of Run 3.

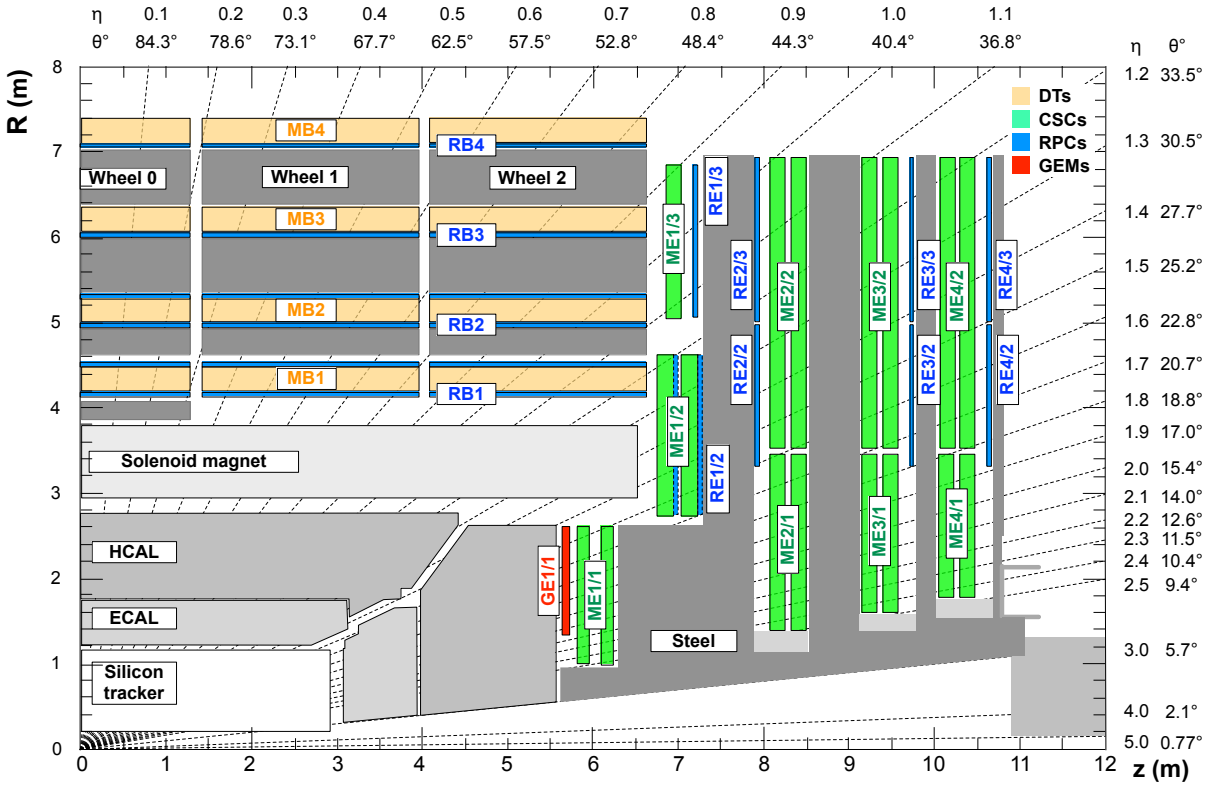


Figure 2.9: *Schematic view of a CMS detector quadrant in the $r - z$ plane at the start of Run 3. The various muon stations are shown in different colours: drift tubes (DTs), with labels MB, are light yellow; cathode strip chambers (CSCs), with labels ME, are green; resistive plate chambers (RPCs), with labels RB and RE, are light blue; gas electron multipliers (GEMs), with labels GE, are in red. The M stands for muon, B denotes barrel, and E endcap. The dark grey areas represent the magnet yokes.*

2.5.1 Drift Tubes

The low expected rate in the barrel and the low strength of the magnetic field in the region were the main factors motivating the decision to employ drift tube chambers [32].

The barrel detector is organised in 4 stations forming concentric cylinders around the beamline. The 3 inner cylinders contain 60 drift chambers each, with the other cylinder having 70. The basic DT detector unit is a rectangular drift cell with a transverse size of 4.2×1.3 cm^2. Each chamber is filled with an Ar and CO_2 gas mixture (85% and 15% respectively) and has an anode wire in the centre, while the borders act as the cathode and contain field-shaping strips. These strips create an electric field such that the drift of ionization electrons produced by the crossing of a chamber by a muon is almost uniform. The trajectory of the muon candidate is then determined from the arrival time of the currents generated on the anode wires. The transversal dimension of the chambers is 21 mm, corresponding to a drift time of 380 ns: a value small enough to produce low occupancy without needing specialised electronics for multiple-hit cases. Figure 2.10 shows the schematic representation of a DT cell.

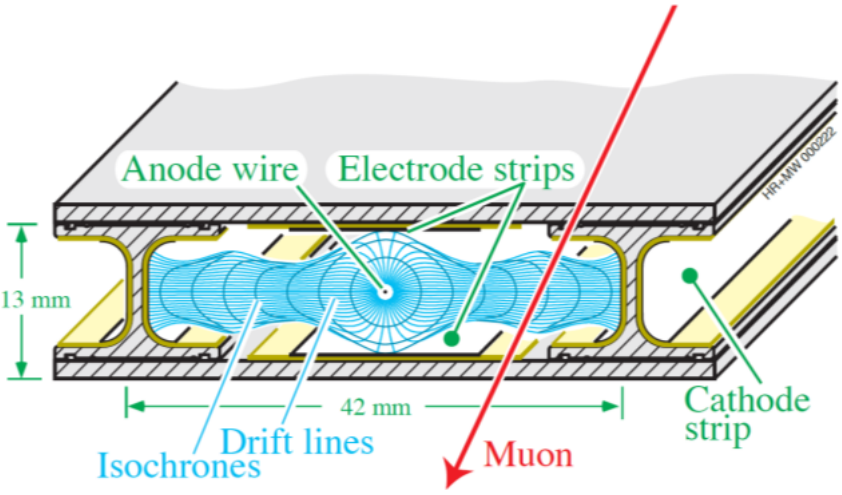


Figure 2.10: *Layout of a CMS DT cell.*

Within a chamber, cells are placed parallel to each other to form layers (L), with groups of four layers forming superlayers (SL), as shown in figure 2.11. Each DT chamber contains two SLs that measure the muon trajectory in the bending plane ($r-\phi$). Chambers from the three innermost stations (MB1-3) are equipped with two additional SLs to measure the position along the longitudinal ($r-z$) plane as well.

The full muon barrel system contains 250 DT chambers covering a pseudorapidity range $|\eta| < 1.2$. They are arranged in five wheels placed, parallel to each other, along the z axis. Within each wheel, there are four concentric station rings (MB1-MB4), segmented into 12 sectors along ϕ, with each sector covering about a 30° window.

The performance measured during Run 1 and Run 2 shows an offline efficiency higher than 99% for the reconstruction of track segments [33]. These segments are characterised by spatial and time resolutions around 100 μm and 2 ns, respectively. The efficiency of reconstructing a standalone DT segment in the trigger (also called a trigger primitive)

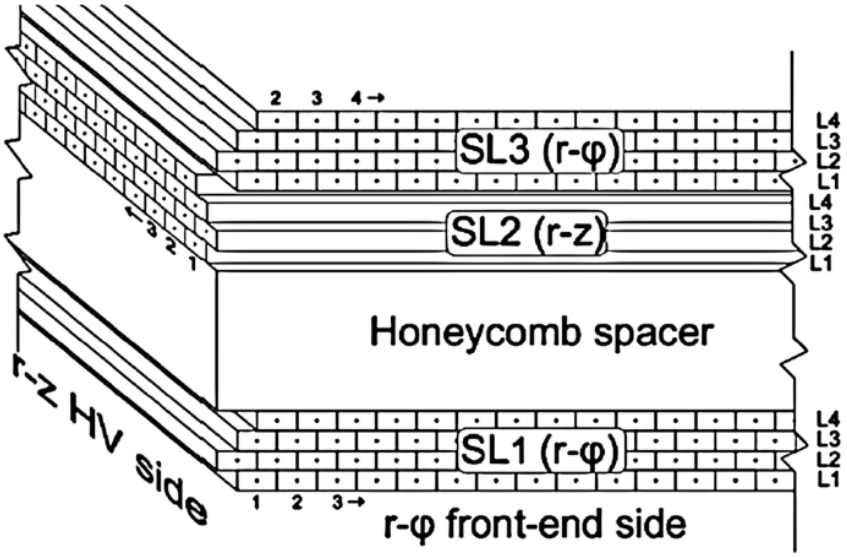


Figure 2.11: *Schematic view of a DT chamber*

and correctly identifying its bunch crossing (BX) of origin is above 95%. The position (direction) resolution of the DT trigger segments is approximately 1 mm (5 mrad).

2.5.2 Cathode Strip Chambers

The cathode strip chambers (CSCs) are multiwire proportional chambers comprised of 6 anode wire plates positioned among 7 cathode panels. This arrangement produced 6 gas gaps each having a plane of radial cathode strips and a plane of anode wires running perpendicular to the strips, as shown in figure 2.12. Having 6 gaps per chamber results in a short drift length, therefore a fast signal collection, suited for the high occupancy expected in the endcaps. The strips are arranged to run radially in the CMS coordinate system to be able to measure the muon position in the plane perpendicular to the colliding beam axis (r-ϕ). The anode wires run azimuthally (along ϕ) and measure the radial coordinate, while a precise measurement of ϕ is obtained from charges induced on the cathode strips as shown in figure 2.13. The gas mixture used to fill the gaps is 40% Ar, 50% CO_2, and 10 % CF_4. Argon is the working gas that gets ionized by charged particles, while CO_2 acts as the quencher needed to achieve large gas gains, and CF_4 is used to prevent anode ageing.

The muon endcap system consists of 540 trapezoidal CSC modules, arranged into four disks placed perpendicularly to the beam direction at increasing distance from the interaction point (ME1-4). The first station is divided into three rings, while the other stations are divided into just two rings. Each camber covers 10 or 20° sectors in ϕ and all chambers, except for the third ring of the first muon station, overlap by five

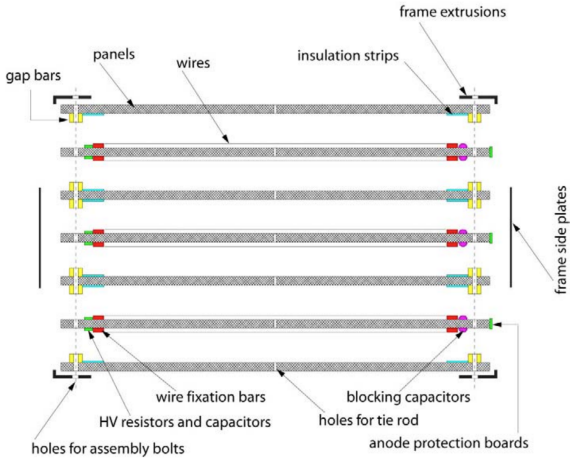


Figure 2.12: *Exploded schematic view of an entire cathode strip chamber with 6 gas gaps and 7 cathode plates.*

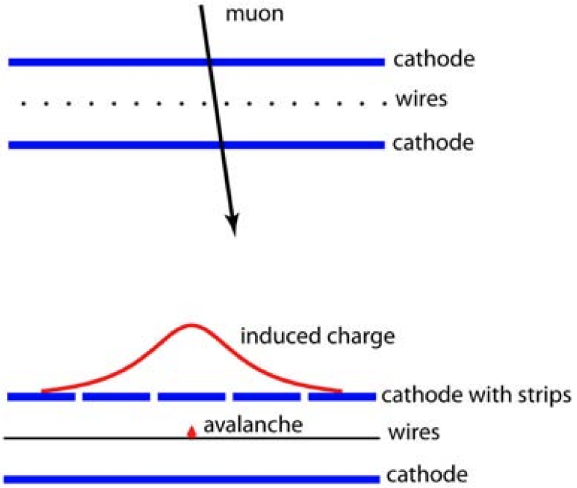


Figure 2.13: *Schematic view of a single gap in a CSC chamber. By interpolating charges induced on cathode strips by the avalanche positive ions produced near a wire, it is possible to obtain a precise localisation of an avalanche along the wire direction.*

strips at each edge, and thus cover the whole ϕ range without gaps. A muon in the pseudorapidity range $1.2 < |\eta| < 2.4$ crosses 3 or 4 CSCs. In the barrel-endcap overlap region, $0.9 < |\eta| < 1.2$, muons are detected both by CSCs and DTs. CSC detectors can measure the position and arrival time of a candidate with high precision and are therefore useful for muon identification and triggering. The typical position and time resolutions achieved by the CSC detectors in CMS are 50-140 μm, depending on chamber type, and 3 ns per chamber, respectively.

2.5.3 Resistive Plate Chambers

Resistive Plate Chambers (RPC) are made up of two parallel detecting layers separated by a thin, gas-filled, gap. Their main characteristic is the ability to provide a coarse position measurement associated with a precise time measurement, the latter with a resolution comparable to that of scintillators. RPCs are capable of tagging the timing of a ionising event in a time much shorter than the 25 ns that pass between consecutive LHC bunch crossings. Therefore, the muon trigger can exploit information from the RPCs to identify the relevant bunch crossing for any muon track candidate, even in conditions with high rates and backgrounds. CMS RPCs consist of double-gap modules, each with two gaps operated in avalanche mode, with common readout strips in the middle, as shown in figure 2.14. This design allows each gap to operate at a lower gain (lower HV) while maintaining a higher total efficiency with respect to a single-gap configuration.

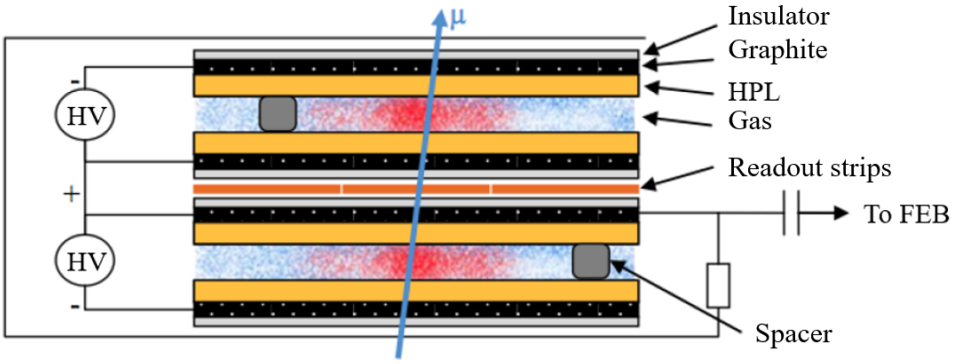


Figure 2.14: *Schematic of the double-gap RPCs used in CMS.*

In the barrel, there are 6 layers of RPCs: one on either side of the first two muon stations and one in front of the third and fourth stations. The redundancy in the first two stations allows the trigger algorithm to always perform the reconstruction with information from 4 layers, even for low p_T particles, which may stop inside the iron yoke. The endcap region is instrumented with 4 layers of RPCs to cover the region up to $|\eta| = 1.9$.

2.5.4 Gas Electron Multipliers

HL-LHC will substantially increase the maximum hit rate in the forward region of CMS. This represents a challenge for the muon system, which must remain radiation-hard, have a high rate capability, and maintain optimal muon reconstruction efficiency while minimizing the number of misidentified tracks and keeping the L1 trigger rate under control. These are the main reasons that led to the decision to adopt gas electron multiplier detectors in the forward region in addition to CSCs and RPCs. One GEM station (GE1) has already been installed before the start of Run 3 and it covers the region $1.55 < |\eta| < 2.18$. This is the first of three GEM stations that will be installed for HL-LHC.

The key feature of a GEM is a thin foil made of a perforated, insulating polymer and surrounded by conductors. The CMS triple GEM detector comprises four gas gaps separated by three GEM foils, as shown in figure 2.15. A voltage difference is applied between the foils producing strong electric fields in the holes. When the gas volume is ionized by the passage of a charged particle, electrons are accelerated by the electric fields in the holes and read out on strips.

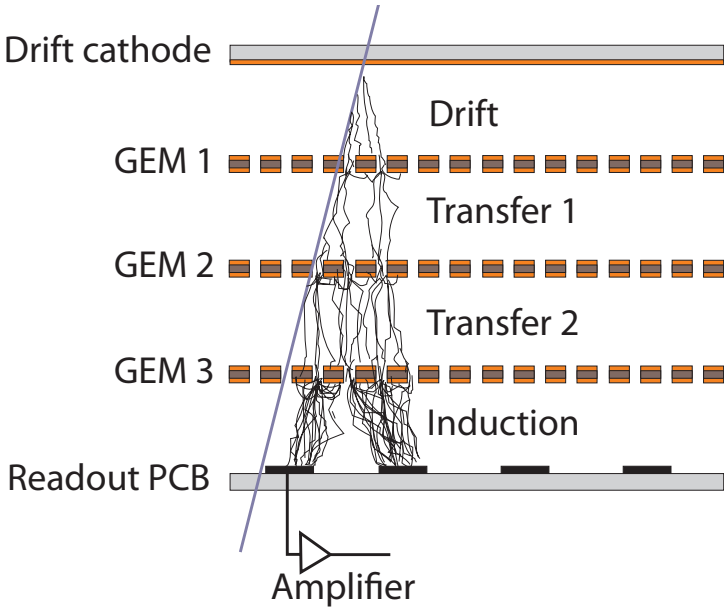


Figure 2.15: *Blowout of the CMS triple GEM detector with the perforated GEM foils and readout planes.*

The triple GEM layout supports a charge amplification factor of up to a factor of several 10^5, while limiting the probability of electrical breakdown or discharge, providing good timing and spatial resolution. The amplified charge induces a measurable signal on the readout electrode, which is segmented to provide positional information. The gas

mixture was chosen to be 70% Ar and 30% CO_2.

Before installing the new GEM detector for Run 3 several key performance parameters have been measured at test beams. Among them, the most relevant ones are the single-hit efficiency and the time resolution [34]. The measured efficiency was greater than 98% with a time resolution of less than 10 ns.

2.6 The Phase-2 upgrade

The Phase-2 upgrade for CMS is meant to revisit all areas of the experiment to improve its capabilities to take advantage of the new physics opportunities offered by the HL-LHC. As discussed in section 1.5, the accelerator upgrade will provide an unprecedented physics reach, thanks to a substantial increase in instantaneous luminosity, giving access to rarer physics processes to be measured. However, this increased reach comes with significant challenges for both the detector and the reconstruction. Increasing the luminosity also boosts the probability of having multiple overlapping proton-proton collisions in a single bunch crossing (pile-up). During Run 3, the average pile-up is between 60 and 80, a number destined to more than double during HL-LHC operation, going up to an estimated mean of 200 in Run 5. Therefore, the detectors need to improve their capabilities to isolate and measure the products of interesting proton-proton collision, while coping with even higher radiation doses [35].

One of the core features of this upgrade is the substitution of radiation-damaged detectors with new, more performing ones. In this context the Si pixel and tracker currently installed at the heart of the CMS experiment, will be replaced with a new silicon-based system 10 times more radiation hard and with higher granularity. This upgrade will grant better tracking and identification performance, while also giving new input information to the L1 Trigger in the form of the momentum measured by the tracker. This precise measurement will be fed at 40 MHz to the L1 Trigger, allowing it to make better decisions. Some advantages of this approach are described in more detail for specific muon triggering algorithms in section 3.1.

Another core feature of the upgrade is the mitigation of the increase in pile-up on the reconstruction performance. In general, this translates to a requirement of improved time resolution for almost all subdetectors, mostly achieved thanks to new front-end electronics. However, a fundamental capability of pile-up mitigation will be provided by a new Timing Layer installed between the tracker and the electromagnetic calorimeter. This layer allows 4D (spatial + time) reconstruction of tracks and vertices with a time resolution of 30-50 ps. The timing layer is made up of two different modules: the Barrel Timing Layer (BTL) and the Endcap Timing Layer (ETL). The former, meant to cover the pseudorapidity region $|\eta| < 1.5$, is constructed from 40 mm thick, Cerium-doped LYSO crystals read out by Silicon Photomultipliers [36]. The latter extends the pseudorapidity coverage up to $|\eta| < 3$ and uses ultrafast, large-pitch, Si Low-Gain

Avalanche Detectors (LGADs) operated at -30 °C.

As far as calorimeters are concerned, the electromagnetic calorimeter in the barrel will maintain the same crystals and readout electronics, with the front-end electronics being upgraded to achieve a better time resolution (up to 30 ps for 30 GeV electrons). The operational temperature will also be lowered from 18 °C (Run 3) to 9 °C to mitigate radiation damage. The endcap calorimeters, both electromagnetic and hadronic, will be entirely substituted by a new High-Granularity Calorimeter (HGCAL). HGCAL is designed to provide precise spatial and timing measurements to produce better-quality clusters and increase the reconstruction quality for jets while aiding in particle identification and isolation [37]. The new calorimeter uses mixed technology for its electromagnetic and hadronic sections. The former is made up of 6 million Si cells, covering the area closer to the interaction point. Behind the silicon sector, there are 250000 scintillator tiles with Si photomultipliers attached on the back for the readout, as shown in figure 2.16.

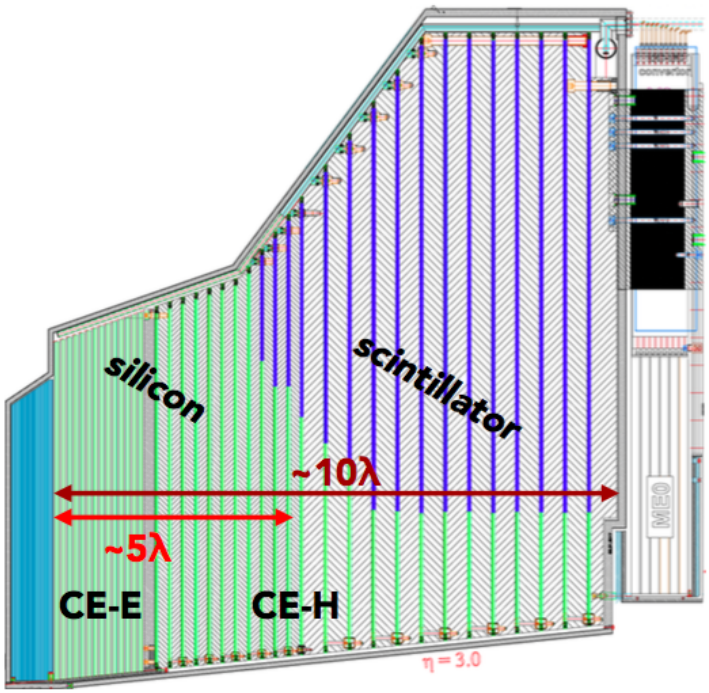


Figure 2.16: *Schematic view of the HGCAL design.*

The electromagnetic section uses a mixture of Pb, Cu, and Cu-W absorbers interlayed among 26 active layers which correspond to 27.7 radiation lengths (X_0) and about 1.5 hadronic interaction lengths (λ). The hadronic section uses steel absorbers and is made up of 7 silicon layers and 14 mixed silicon-scintillator layers, amounting to a total depth corresponding to more than 10λ.

Finally, the muon detectors will remain largely unchanged, apart from upgraded

electronics to improve their performance at different stages and cope with the higher rates [38]. Thanks to these upgrades, the time granularity of the RPCs' readout will decrease from 25 to 1.5 ns. Moreover, new GEMs will be installed in the very forward region increasing the redundancy of CSC and RPC information while also extending the coverage to $|\eta| < 2.8$. Two more rings of GEMs will be installed in each endcap, in addition to the first ring already installed for Run 3. A more detailed discussion of the muon system upgrades and how they impact the trigger and reconstruction is presented in Chapter 3.

Chapter 3

Muon Trigger and Reconstruction at CMS HLT in Phase-2

3.1 The L1 Muon Trigger

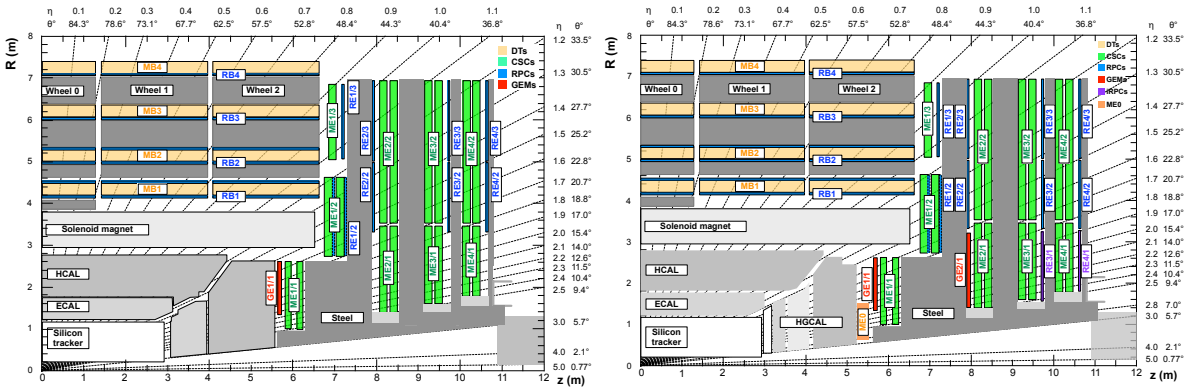
Currently, the CMS hardware trigger (L1) identifies Standalone Muon candidates using only muon system trigger information. This information is referred to as Muon trigger primitives (TP) and it consists of electronic encoding of the parameters of a segment (in DT or CSC) or a hit (in RPC), from each muon station. The L1 trigger track finders use these primitives to define Standalone L1 Muon tracks with 20-30% momentum resolution, estimated from the curvature of the reconstructed tracks and largely limited by multiple scattering in the calorimeters and the iron return yoke, especially at low momentum.

The existing CMS hardware muon trigger is designed to find muon tracks using DT primitives in the barrel region (roughly $|\eta| < 0.8$), CSC trigger primitives in the endcap region(roughly $|\eta| > 1.2$, and a combination of both kinds of primitives in the so-called overlap region where a single muon could realistically cross both DT and CSC stations. Additionally, RPC hit clusters are integrated with DT and CSC trigger primitives, so eventual RPC hits contribute to the track-finding algorithms in both barrel and endcap regions. RPC information is used to increase the efficiency of trigger primitives and, in the barrel, improve their timing. In the overlap region, all three subsystems contribute to the L1 track finding. A schematic representation of a $r - z$ quadrant of the current muon system is shown in figure 3.1(a) where all the subsystems are highlighted and the distinction between barrel, overlap, and endcap regions is clearly visible.

For Phase-2, the CMS trigger two-layered approach will be largely unchanged, with a hardware L1 Trigger [39] and a software High-Level Trigger (HLT) [40].

The L1 Muon trigger will take advantage of the new detectors installed for the Phase-2 upgrade that are shown in Figure 3.1(b). In particular, the endcap region in Phase-2 will see the addition of Gas Electron Multiplier (GEM) chambers in the forward region to

extend the coverage up to $|\eta| < 2.8$ from the current limit $|\eta| < 2.4$. Moreover, together with the improved Resistive Plate Chambers (iRPCs), these new detectors will improve background rejection and generally increase the performance of the muon trigger system as a whole in high-pile-up conditions. Finally, the addition of L1 Tracker information to the Muon trigger system will also contribute to both the aforementioned improvements.



(a) *(b)*

Figure 3.1: *(a) $r-z$ quadrant of the CMS muon system with z parallel to the beamline and increasing from left to right, r increasing upwards. The interaction point is in the lower left corner. The Drift Tubes (DTs) are labelled MB (Muon Barrel) and shown in light yellow, the Cathode Strips Chambers (CSCs) are labelled ME (Muon Endcap) and shown in light green, and the iron return yoke is shown in dark grey. Resistive Plate Chambers (RPCs) are mounted both in the barrel and in the endcap where they are labelled RB and RE, respectively. The first Gas Electron Multiplier (GEM) detector installed for Run-3 as a testbed for the Phase-2 upgrade is also shown in red. (b) Same $r-z$ section of the CMS muon system with the added detectors for Phase-2. Improvements mainly focus on the high-η region, where GEM stations have been added and are shown in orange and red. The endcap region also benefits from new improved Resistive Plate Chambers (iRPC) shown in purple.*

The target maximum trigger rate for L1 will be increased from the current 100 kHz to 750 kHz with a corresponding increase in trigger latency from the current 3.6 μs to 12.5 μs. The requirements for the Phase-2 L1 Standalone Muon trigger are:

- Keep thresholds of prompt muon triggers close to current levels, allowing single muon L1 p_T thresholds as low as, or lower, than what is in use in Run-3, around 20-25 GeV, even at the maximum design luminosity of HL-LHC;
- Provide good enough spatial resolution for efficient and pure matching with the Track Trigger and to provide better efficiency for ultra-high momentum muons;

- Provide Standalone Muon triggering capabilities for triggering on event topologies not covered by the Track Trigger system, such as long-lived neutral particles decaying into displaced muons.

Figure 3.2 shows the architecture of the Phase-2 L1 Muon Trigger.

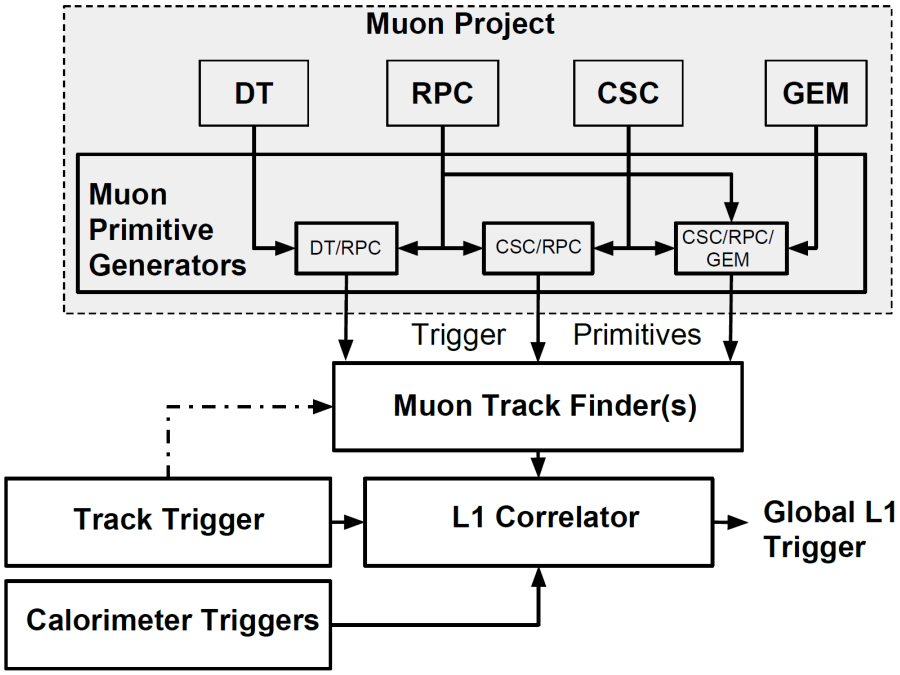


Figure 3.2: *Phase-2 L1 muon trigger data flow diagram.*

The L1 trigger is made up of multiple sub-modules, each of the detectors produces trigger primitives that are then combined and fed to the Muon Track Finders to produce Standalone Muon candidates. Moreover, for the first time, in Phase-2 the Muon Trigger information will be combined with information from the Track Trigger via a L1 Correlator. The quality of the muon TPs themselves will be enhanced by the addition of GEM and RPC forward chambers as well as improved timing resolution for DT and RPC. In particular, hits from the new GEM chambers will be combined from the beginning with trigger primitives from the nearby CSCs to increase efficiency and improve the quality of the measurement of the local muon direction and position. The addition of Track Trigger information to the L1 muon trigger will greatly increase the momentum resolution for muons coming from the interaction region. Such capability is integral to the design of the new silicon Outer Tracker. Since, in High-Luminosity conditions, many track candidates are found by the Tracker Trigger in each bunch crossing, matching these tracks with the cleaner Standalone Muon tracks coming from the L1 Muon Trigger provides well-identified muons with precisely measured momentum. The Track Trigger can also

be directly combined with trigger primitives before the track finding in the muon chambers takes place. This produces L1 objects referred to as L1 Tracker Muons, mirroring the offline reconstruction and improving the efficiency for very low p_T muons, especially in the barrel [41].

In the following, the barrel, endcap, and overlap regions are discussed separately focusing on hardware upgrades as well as the new possibilities offered by combining all available information into a single muon trigger candidate. All performance results shown for these regions were reported in the Technical Design Report for the Phase-2 Upgrade of the CMS L1 trigger [39], where detailed comparisons are reported and an exact definition of all measured quantities is provided. Displaced leptons are a promising physics channel for the HL-LHC, as such the upgrade of the muon system for Phase-2 has been designed to provide the information necessary to be able to reconstruct these final states. Specific triggers dedicated to the reconstruction of displaced leptons are also discussed.

3.1.1 Barrel region

The Phase-2 upgrade for CMS does not include any new muon detectors in the barrel region. However, the electronics of existing DT and RPC detectors will be upgraded. As far as the RPCs are concerned, the upgrade aims at improving the time resolution. In the DT system, the electronics will be replaced, including a new trigger primitive generator. The entire electronic system will be moved in the off-detector backend, allowing the DT trigger to use the full detector resolution, improving both timing and position measurement resolution by about a factor of 5. Moreover, the same backend will also receive data from RPC. The combination of DT and RPC information at the trigger level will provide benefits such as improving the bunch crossing identification and providing a general boost in quality to the original primitive. These benefits have already been demonstrated in Phase-1 conditions [42].

Moreover, having access to hits from both detectors in the same backend allows to retain efficiency in case single hit DT efficiency were to degrade due to ageing. The best performance will still be obtained when both subsystems provide hits, but the capability to generate trigger primitives in each chamber from DT or RPC alone is fundamental to guarantee the operation of the muon trigger in case of failures in either of the subsystems.

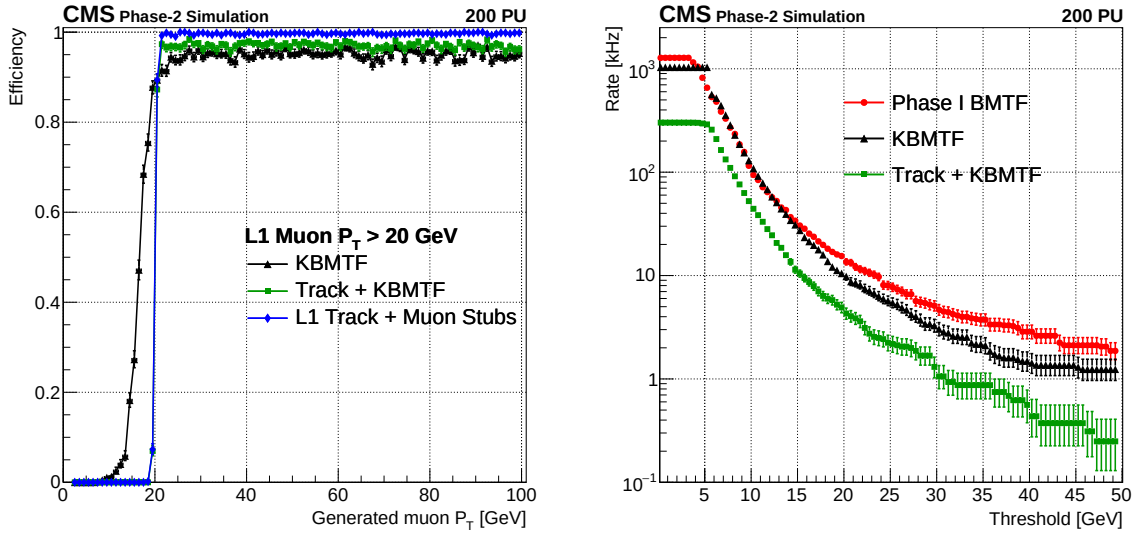
Finally, the Kalman Barrel Muon Track Finder (KBMTF) uses the well-known Kalman Filter algorithm to produce L1 Muon tracks starting from the outermost muon station where a trigger primitive was found, propagating inwards.

The performance of the triggering algorithms is assessed by measuring the L1 efficiency and trigger rate.

The efficiency of the Phase-1 trigger is compared to the new approach merging information from the tracker and muon chambers in figure 3.3(a). The improved p_T resolution of the algorithm merging tracker and muon trigger information implemented for Phase-2

is demonstrated by the sharp efficiency turn-ons. The efficiency of the new algorithm at the plateau is above 99%, demonstrating a substantial improvement in reducing the inefficiency by about a factor of 5 compared to the Standalone trigger, where the plateau approaches 95%. The improvement comes from regaining muon tracks in the gaps between the barrel muon detector wheels which are lost in the standalone muon trigger.

Figure 3.3(b) shows the comparison between Phase-1 and Phase-2 trigger rate. For a typical single muon threshold of 20 GeV, the expected rate at 200 average pileup events is about 4.5 kHz, which is less than 1% of the available Phase-2 L1 budget.



(a) *L1 trigger efficiency in the barrel.*

(b) *L1 trigger rate in the barrel.*

Figure 3.3: *(a) L1 trigger efficiency as a function of the generated muon p_T with a L1 threshold of 20 GeV. The Phase-1 Barrel Muon Track Finder Algorithm (KBMTF) is shown in black, the Phase-2 equivalent is shown in green, while the algorithm taking advantage of information both from the tracker and from the muon chambers is shown in blue. (b) Single muon rate as a function of the L1 p_T threshold for the Phase-1 Barrel Muon Track Finder (red), Phase-2 equivalent (black), and and L1 Tracker Muon approach (green).*

3.1.2 Endcap region

The current implementation of trigger logic in the endcap adds RPC hits directly in the same backend as CSC ones, allowing to execute a more sophisticated track finding algorithm: the Endcap Muon Track Finder (EMTF). RPC clusters/strips and the raw input data from the CSC, called local charged tracks (LCTs) are used as inputs for the

EMTF. LCTs contain a coarse description of where the charged candidate was detected through the endcap, station, sector and chamber numbers of the CSC where it originated. These numbers encode information about the position in ϕ and θ relative to the chamber itself. As shown in figure 3.1(b), in each endcap there are four stations along the z-direction (ME1-4) and six sectors which cover the full 2π ϕ range. Each sector is further divided into nine chambers which can detect a total of two LCTs each.

The EMTF algorithm uses the LCTs in a given event to determine a set of outputs:

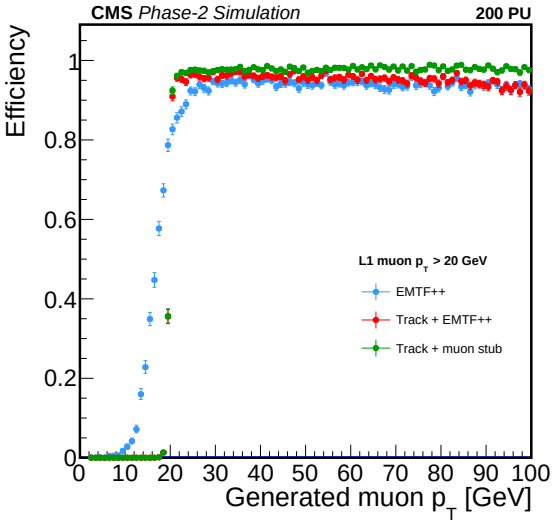
- precise relative position in ϕ and θ relative to a sector;
- the deflection angles between stations $\delta\phi$ and $\delta\theta$;
- a quality index;
- a list of the LCTs used to build the track;
- estimates of the charge and p_T of the candidate.

The current track finding method uses LCTs to form "extrapolation pairs", with RPC hits used to substitute missing LCTs. By building such a pair, their associated position information results in three-dimensional spacial information that can be tested for compatibility with a muon produced in the primary vertex. When a match is found, the extrapolation pairs are grouped together to form a track. The track finding algorithm developed for the Phase-2 upgrade is able to analyze all 18 possible LCTs/RPC hits from all four stations in each endcap to form a track via pattern recognition directly at the L1 trigger level.

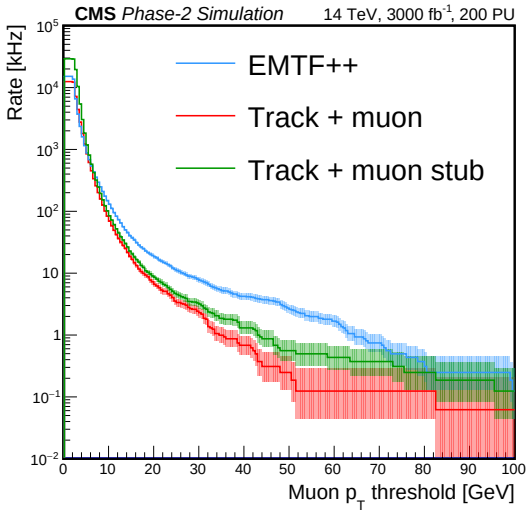
With the Phase-2 upgrade, the EMTF will benefit from the addition of new detectors in the very forward region: the improved RPCs and GEMs. The additional hits recorded in these detectors will allow the algorithm to recover the efficiency losses due to acceptance gaps at high η.

The efficiency as a function of the generated p_T of the improved EMTF algorithm (EMTF++) and the new approach exploiting both EMTF and tracker tracks is shown in figure 3.4(a) for a L1 muon p_T threshold of 20 GeV. As expected the p_T turn-on is significantly sharper, with a corresponding slightly higher efficiency for muons above the selected p_T threshold.

The rate shown in figure 3.4 demonstrates a significant reduction with respect to the EMTF++ algorithm. The expected rate at a $p_T = 20$ GeV threshold is about 10 kHz.



(a) L1 trigger efficiency in the endcap.



(b) L1 trigger rate in the endcap.

Figure 3.4: *(a) L1 trigger efficiency as a function of the generated muon p_T with a L1 threshold of 20 GeV. The improved Endcap Muon Track Finder (EMTF++) algorithm is shown in light blue, the algorithm matching a tracker track to the EMTF results is shown in red, while the algorithm taking advantage of information both from the tracker and from the muon chambers is shown in blue. (b) Single muon rate as a function of the L1 p_T threshold for the EMTF++ algorithm (light blue), L1 tracker tracks + EMTF (red), and L1 tracker tracks + muon trigger primitives (green).*

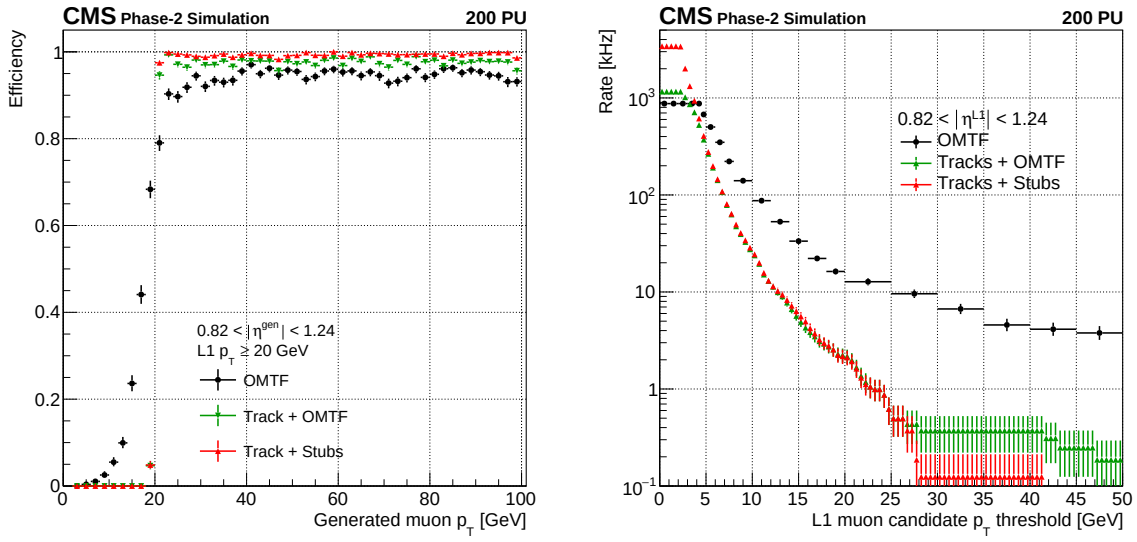
3.1.3 Overlap region

The overlap region of the muon system is defined as the η range where DT, RPC and CSCs overlap: $0.83 < |\eta| < 1.24$. Because of this condition, the orientation of the muon chambers and the magnetic field are not uniform, resulting in challenges for the muon trigger and reconstruction.

Currently, a dedicated version of the Muon Track Finder algorithm is used to identify candidates in the overlap region. This algorithm takes in input information from 18 detectors: 6 DT detectors (3 for ϕ and 3 for θ measurements), 8 RPC detectors (5 in the barrel and 3 in the endcap), and 4 CSC detectors.

For Phase-2 no specific hardware improvement is foreseen in the overlap region. However, the improvements for barrel and endcap regions will benefit the overlap as well. In particular, having access to a finer DT spacial resolution and directly adding hits from the RPCs to the ones recorded in DT will significantly improve the track finding capabilities when a few chambers are hit in the barrel. Similarly, the improved efficiency in the endcap due to the CSC-RPC coincidence will also be implemented in the overlap.

Figure 3.5 shows the comparison of the performance of the current standalone Overlap Muon Track Finder (OMTF) with the ones of the algorithm matching L1 tracker tracks to OMTF muons, and the new direct matching of L1 tracker tracks with trigger primitives in the muon chambers. As expected, the efficiency, shown as a function of generated muon p_T, increases with respect to both the OMTF Standalone reconstruction and the L1 tracker track matching to OMTF Muons. The largest increase occurs at the edges of the overlap region. Moreover, the use of L1 tracker tracks reduces the rate considerably, more than a factor of 5 at a L1 threshold of 20 GeV.



(a) *L1 trigger efficiency in the overlap.*

(b) *L1 trigger rate in the overlap.*

Figure 3.5: *(a) L1 trigger efficiency as a function of the generated muon p_T with a L1 threshold of 20 GeV. The current Standalone Overlap Muon Track Finder (OMTF) is shown in black, the algorithm matching L1 tracker tracks with OMTF Muons is shown in green, while the approach matching L1 tracker tracks directly with muon trigger primitives is shown in red. (b) Single muon rate as a function of the L1 p_T threshold for the OMTF (black), L1 tracker tracks + OMTF (green), and L1 tracker tracks + muon trigger primitives (red).*

3.1.4 Dedicated Trigger for Displaced Muons

High-momentum muons that are produced with a significant displacement from the primary vertex are prime candidates for beyond the Standard Model physics and require specific Standalone Muon triggering capabilities. As previously discussed, the Phase-2 muon trigger for prompt muons will match Standalone Muon objects in the muon cham-

bers with L1 Track Trigger objects, containing information from the pixel and tracker. This provides a better momentum measurement, with the higher quality information from the inner tracker replacing the momentum estimate calculated in the muon chambers. The combined performance of these two systems has been shown to be excellent for prompt muons. However, the L1 Track Trigger is inefficient for displaced tracks [35].

Long-lived particles (LLPs) are yet unobserved particles beyond the Standard Model that can travel a substantial distance between the interaction point where they are produced and their decay point, thus presenting specific and recognizable signatures. LLPs are featured in many BSM models and their observation and measurement could provide insight into many central questions that the SM cannot answer. Considering, as an example, long-lived bosons that couple directly to Standard Model particles, they could decay into electrons or muons thus producing displaced lepton signatures in the CMS detector. In general, the decay to muons might not be the only decay channel, or even the most common one, but it can be the most triggerable, thanks to the relatively low occupancy in the muon system and the Standalone Muon reconstruction efficiency. The current muon system can reconstruct muons with a transverse displacement with respect to the beam spot up to about 350 cm and a transverse impact parameter up to about 100 cm. In contrast, the L1 Track Trigger can only reconstruct prompt muons. Not having access to L1 Track Trigger information however limits the precision of the momentum measurement: the L1 muon p_T reconstruction assumes that all tracks originate in the interaction points. Dropping this constraint reduces resolution and increases the trigger rate. Therefore, there is a need to optimize the measurement of muon direction with information from a single station.

All p_T assignment algorithms in use in the L1 trigger rely on the measurement of the track bending angle in at least two stations. A proof of concept has already been implemented in the barrel [38]. It shows good performance for Displaced Muons with high efficiency from $p_T \approx 15$ GeV as shown in figure 3.6.

Displaced triggering in the endcap is substantially more challenging. A combination of the thinness of the chambers and the weakness of the magnetic field renders the p_T measurement less accurate, especially in the very forward region. With current endcap detectors, only one direction is well measured, while at least two are required for a suitable trigger. However, a muon trigger for Displaced candidates has been implemented in the endcap, relying on:

- inclusion of the new GEM detectors for Phase-2 to provide the second good measurement of muon direction;
- an improved CSC trigger primitive position and direction measurement;
- a hybrid algorithm that combines direction- and position-based p_T measurements;
- a L1 Track Trigger veto. Every L1 track in a cone of $\Delta R = 0.12$ evaluated at the second muon station around a muon candidate is rejected.

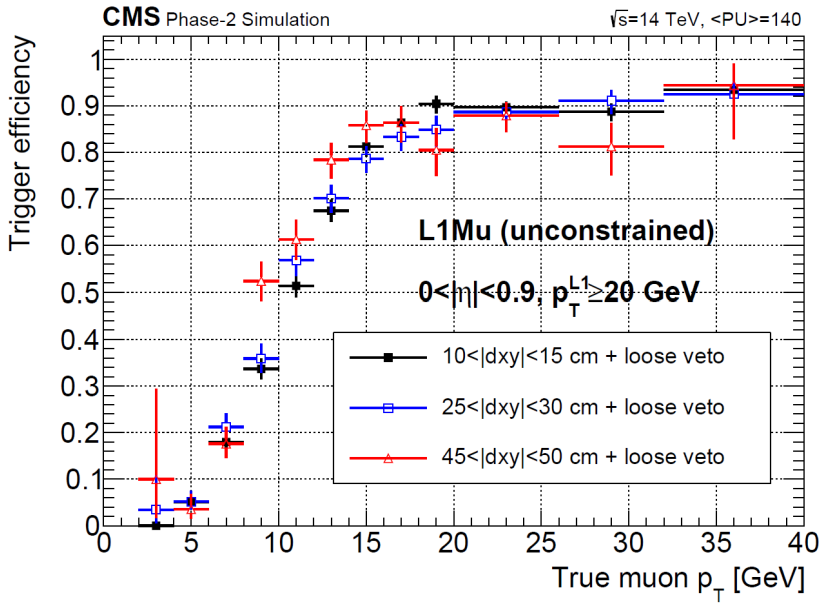


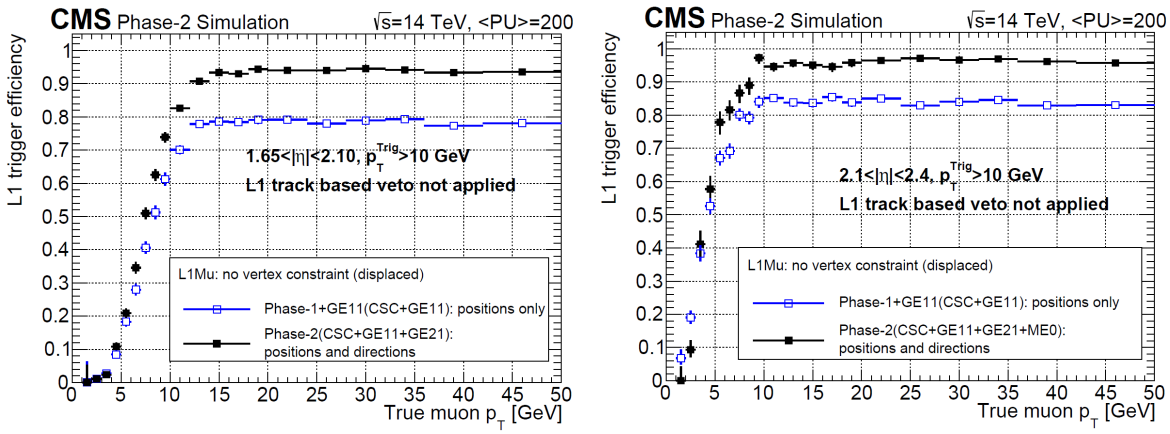
Figure 3.6: *L1 Muon trigger efficiency versus true muon p_T measured with the displaced algorithm in the barrel.*

Figure 3.7 shows the performance of the algorithm implemented for Displaced Muon triggering in the endcap in the regions $1.65 < |\eta < 2.10$ (a) and $2.1 < |\eta < 2.4$ (b). The results show that these algorithms have high efficiency both in the barrel and endcap regions, for a wide range of Displaced Muon p_T, with an acceptable trigger rate.

Finally, taking into account the triggering logic previously discussed, the design of the Phase-2 L1 muon trigger will be as follows:

- L1 Standalone Muon reconstruction generates two p_T measurements for each muon identified in the muon chambers: one prompt (with a constraint on the beam spot) and one displaced.
- L1 muons are matched with information from the L1 Track Trigger:
  - If the match is successful, the L1 Track Trigger p_T is used and the candidate produced is a prompt L1 Tracker Muon.
  - If the match fails and the muon passes the L1 tracks veto, the non-prompt L1 Muon p_T is used to produce a Displaced Muon candidate.

This logic is compatible with the trigger logic shown in figure 3.2 and achieves good results for both prompt and Displaced Muons.



(a) *(b)*

Figure 3.7: *L1 trigger efficiency versus true muon p_T for the endcap Displaced Muon algorithm in the region $1.65 < |\eta| < 2.10$ (a) and $2.1 < |\eta| < 2.4$ (b). No L1 Track Trigger veto applied.*

3.2 HLT Muon Reconstruction

In the current CMS reconstruction, tracks from muons are reconstructed both in the inner tracker (*Tracker track*) and in the muon system (*Standalone Muon track*). Figure 3.8 shows a transversal slice of the CMS detector.

Focusing on the Muon track (light blue line), multiple reconstructed muon objects can be distinguished at the trigger level:

- **L1 Standalone Muon**: hardware trigger tracks built using only the trigger primitives in the muon chambers;
- **L1 Tracker Muon**: hardware trigger track built by matching a Tracker trigger track with one or more trigger primitives in the muon chambers;
- **L2 Standalone Muon**: track built using only information from the muon chambers.
- **L3 Tracker Muon (L3 IO track)**: muon object identified combining tracker tracks with one or more DT or CSC segments. A track reconstructed inside-out, using only information from the inner tracker is referred to as L3 IO track;
- **L3 Global (or combined) Muon (L3 OI tracks)**: muon object built by matching a L2 Standalone Muon with a tracker track reconstructed propagating from the outer tracker towards the pixel detector. This object contains a tracker track referred to as L3 OI track;

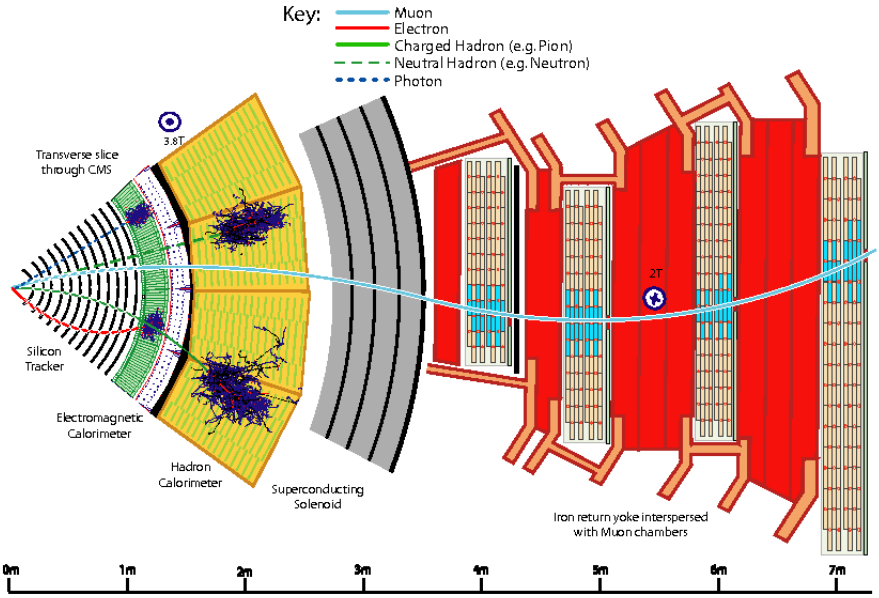


Figure 3.8: *Schematic view of a transversal slice of the CMS detector in the barrel region. The different sub-detectors of CMS can be identified, from left to right: inner tracker (shown with curved black lines), electromagnetic calorimeter (the green region), hadronic calorimeter (yellow region), and muon chambers (red regions).*

- **HLT Muon**: the final type of reconstructed object produced by the HLT after reconstruction and identification combining the two kinds of L3 tracks and adding the Muon ID variables.

The current reconstruction workflow starts from the information provided by the L1 hardware trigger as well as hits in the muon chambers. Track segments are built in individual muon chambers using a linear fit to the position of the reconstructed hits in each of the layers of the chambers (12 or 8 in the case of DT, 6 in the case of CSC). All the permutations of pairs of segments produced in the muon chambers are used to generate L2 Standalone Muon seeds consisting of position and direction vectors and an initial estimate of the muon transverse momentum (p_T). This collection of seeds is then matched with the information from the L1 hardware trigger, namely the L1 Tracker Muons. The resulting matching collection is used as a starting point for the track fits in the muon system which uses the Kalman-filter [43] technique with information from DTs, CSCs, RPCs, and GEMs. Momentum resolution is improved with a beam-spot constraint in the fit for collision data when dealing with prompt muons, while Displaced Muons require dedicated triggering and reconstruction algorithms discussed in the following.

Two L3 approaches are used in CMS, using as starting points either the L1 Tracker Muons or the L2 Standalone Muons:

- *Inside-Out (IO) Tracker Muon reconstruction.* This approach starts from the L1 Tracker Muons found in the inner tracker, considering every track with $p_T > 0.5$ GeV and total momentum $p > 2.5$ GeV as possible muon candidates. All the candidates are extrapolated to the muon chamber taking into account the magnetic field, the expected energy losses in the inner detectors, and multiple Coulomb scattering in the detector material. During the HLT Muon ID, if at least one muon segment geometrically matches the extrapolated track, that track qualifies as a Tracker Muon.

- *Outside-In (OI) Global Muon reconstruction.* Starting from the L2 seeds, a matching tracker track is found by geometrically matching the two reconstructed objects propagated to a common surface. Merging the inner track and the associated track in the muon chambers yields a Global Muon, an approach that can improve the momentum resolution compared to the tracker-only approach in the high p_T region (starting from ≈ 200 GeV).

After the execution of both the IO and OI L3 reconstructions, two collections of L3 inner tracks are produced. Those are then merged together and, out of their combination with L2 Standalone Muons, a new Global Muon collection is built. Finally, the Muon ID takes in input the latter two collections and the muon system reconstruction products to provide, for each muon, a single interface object that exposes a vast set of variables useful for selection purposes (e.g. number of hits per detector, track quality, ...). The HLT Muon ID selection relies on a subset of criteria typically used by Offline Muon IDs [38](pp. 282-285):

- at least 1 valid hit is required in the pixel detector and at least 5 are required in the tracker;

- the extrapolated tracker track must match a segment in the muon chambers in ϕ and η within a 3 cm window;

- if the p_T associated with the track is > 8 GeV, and the extrapolation confirms at least chambers from two stations are crossed, the track is required to match a segment in a further muon station.

This approach allows to maintain high efficiency in the entire pseudorapidity range $|\eta| < 2.4$, with the HLT Muon ID selection retaining more than 99% of the prompt muons identified by the L1 trigger.

Figure 3.9 shows a schematic representation of the entire HLT Muon reconstruction and identification workflow as described.

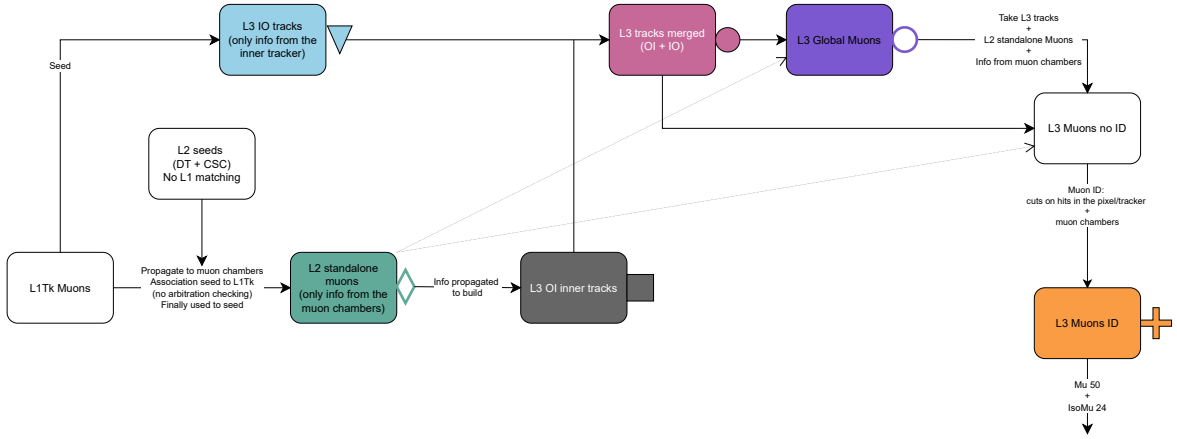
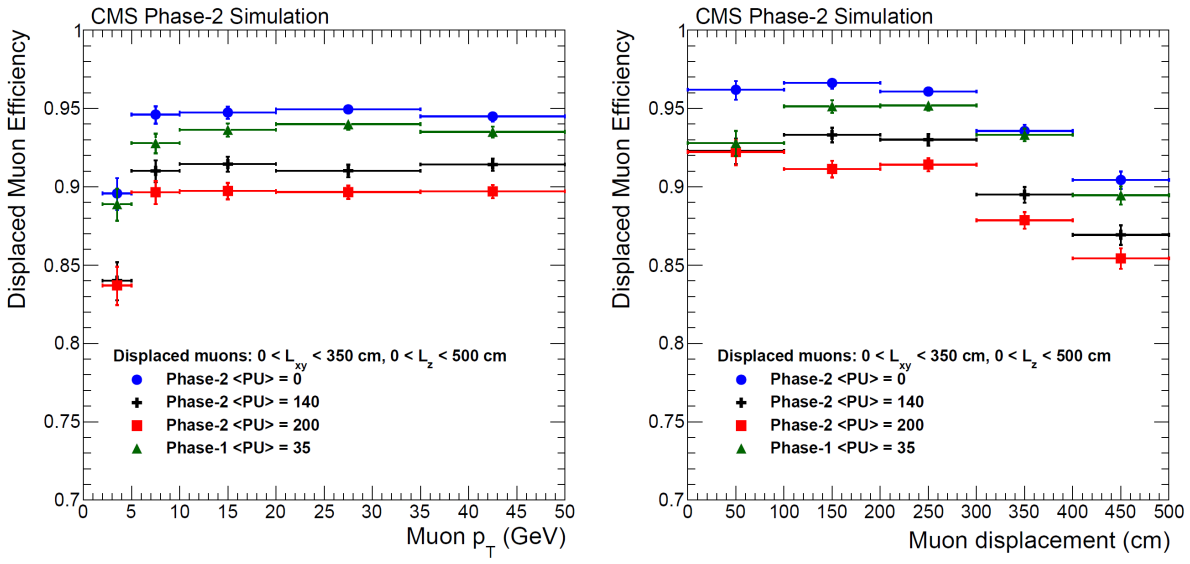


Figure 3.9: *Schematic representation of the current Muon reconstruction and identification workflow for Phase-2. From left to right, the L1 Tracker Muons built by the hardware trigger and the L2 seeds built in the muon chambers are matched and used to seed L2 Standalone Muons. The latter are then used to seed L3 inner tracks in the Outside-In approach, while the L1 Tracker Muons are directly used to seed L3 tracker tracks in the Inside-Out approach. The two L3 inner track collections (Inside-Out and Outside-In) are then merged and used to produce Global Muons by geometrically matching them to Standalone Muon tracks in the muon chambers. Finally, all previously used information is used to build the final L3 Muon collection that undergoes ID and produces the final Muon candidates used for triggering at the HLT.*

3.3 Offline Displaced Muon Reconstruction

Displaced muons represent one of the most promising signatures of new physics that can be probed at HL-LHC. Therefore, a special Standalone Muon reconstruction algorithm was developed for the reconstruction of highly displaced muons [44]. In this reconstruction workflow, the Displaced Standalone (DSA) tracks are reconstructed using only hits in the muon chambers and have no constraints to the beam spot. In addition, the reconstructed segments are not required to point to the event main vertex. The additional hits provided by the new detectors in the forward region naturally benefit the Standalone Displaced Muon reconstruction, as it relies on the standard Standalone Muon reconstruction. The performance of the Displaced reconstruction has been studied in Phase-2 conditions using simulated muons with flat distributions in the transverse impact parameter (0-50 cm) and in transverse momentum (2-50 GeV). Moreover, the study only considers muons that originated before the Muon System, therefore with transverse displacement $L_{xy} < 350$ cm and longitudinal displacement $L_z < 500$ cm. The usual purity requirements are applied, considering only reconstructed muons with more than 75% reconstructed hits matched with the correct simulated muon. To ensure a good enough

momentum measurement, calculated from the sagitta of the reconstructed track, at least two hits in two separate muon stations are also required. The results of this study are shown in figure 3.10. The pile-up conditions expected in Phase-2 reduce the efficiency of the DSA muon reconstruction by about 5% with respect to Run 2 performance. However, the reconstruction remains efficient even for large displacements.



(a) *(b)*

Figure 3.10: *Displaced Standalone Muon reconstruction efficiency as a function of the simulated muon transverse momentum (a) and displacement (b), for the Phase-2 detector in three pile-up scenarios, compared to the performance of the Phase-1 detector.*

The dedicated reconstruction algorithm for displaced muons has not yet been ported to the Phase-2 HLT, while displaced muon triggers are currently being used in Run-3. However, the new Offline displaced reconstruction must be kept in mind throughout the redesign of the Muon trigger for Phase-2.

Chapter 4

Optimizing the Online Muon Reconstruction

As previously discussed in Chapters 1 and 3, HL-LHC will increase the instantaneous luminosity by a factor of 5 (7.5) in Run 4 (5) with respect to LHC design. As a consequence of this increase in luminosity, the average collision pile-up will also increase to approximately 140 in Run 4 and around 200 in Run 5. While this drastic increase in luminosity provides an unprecedented opportunity to push the limits of the Standard Model and probe new phenomena, but it also poses significant challenges in computational and reconstruction aspects.

To face these challenges, the CMS online reconstruction software is undergoing a fundamental rethinking of its algorithms beyond what is imposed by detector upgrades, with the aim of maintaining or improving the physics performance achieved in Run 3 in a much more complex environment and with stricter computational performance requirements. Phase-2 conditions will surpass the expected performance improvements of traditional CPUs used at the HLT, particularly in reconstruction workflows dealing with extremely high occupancies and are most impacted by the increased luminosity and pile-up (i.e. the pixel, tracker, and calorimeters).

In this context, the full pixel and inner tracker reconstruction has already been reimagined to take advantage of heterogeneous computing platforms to achieve higher throughput, better energy efficiency, and improved physics performance [45]. The heterogeneous computing paradigm is based on the concept of a single software that can be executed on multiple devices, therefore matching the chosen device to the task at hand. In this case, since reconstruction workflows deal with independent events wherein tracks can also be reconstructed mostly independently from one another, a heterogenous solution can take advantage of Graphics Processing Units (GPUs) which are innately capable of executing multiple tasks in parallel.

The current timing for a CPU-only execution of the full HLT online reconstruction of a typical Run 3 simulated $t\bar{t}$ event with center-of-mass energy $\sqrt{s} = 13.6$ TeV is shown

in figure 4.1. The plot shows the real execution time, thus also considering the time necessary to read the input data and all memory accesses. Each reconstruction module is shown in a different colour and takes up a section proportional to its execution time. In current conditions, muon reconstruction represents a considerable fraction of the total execution time of the HLT reconstruction.

Element	Time	Fraction
AlCa	3.2 ms	0.5 %
B tagging	1.9 ms	0.3 %
CTPPS	0.0 ms	0.0 %
DQM	1.4 ms	0.2 %
E/Gamma	59.5 ms	10.1 %
ECAL	43.8 ms	7.4 %
Framework	0.0 ms	0.0 %
HCAL	61.1 ms	10.3 %
HLT	6.1 ms	1.0 %
I/O	4.7 ms	0.8 %
Jets/MET	15.8 ms	2.7 %
L1T	7.4 ms	1.2 %
Muons	93.7 ms	15.8 %
Particle Flow	44.3 ms	7.5 %
Pixels	117.4 ms	19.8 %
Taus	8.2 ms	1.4 %
Tracking	71.3 ms	12.0 %
Vertices	5.1 ms	0.9 %
event setup	0.4 ms	0.1 %
idle	1.2 ms	0.2 %
other	45.5 ms	7.7 %
total	*592.0 ms*	*100.0 %*

Figure 4.1: *Timing of the full HLT Online reconstruction on a sample of a few tens of thousands of Run 3 simulated $t\bar{t}$ events [46]. The performance shown refers to the reconstruction being executed only on CPUs, a significant speed-up is obtained when executing the same workflow on GPUs available at the HLT farm. The time shown in the centre of the circle refers to the mean time required to process a typical event in Run 3 conditions. The innermost ring shows the general reconstruction category, with information becoming more and more specific moving towards the outside, finishing in the outermost ring which shows the name of each reconstruction module. Sectors with the GPU label refer to modules that present macroscopic differences when the reconstruction is executed on GPUs that are currently available in the HLT farm.*

While the current muon reconstruction offers overall excellent physics performance, it has been shown that its approach cannot scale up to Phase-2 conditions due to a multitude of factors, including timing and the increase in the fake rate. This work focuses on two main optimisation aspects, tackling the Online Standalone Muon reconstruction as well as the Global and Tracker Muon reconstructions.

Firstly, the current implementation of the Standalone Muon reconstruction does not fully utilize the new information available from the upgraded hardware trigger. This mostly refers to a better estimate of kinematics quantities of the muon, which could allow to shrink the extrapolation windows used in the track finding algorithms, thus producing higher-quality candidates while also reducing the computing requirements.

Furthermore, the current Tracker Muon (IO) track finding detailed in section 3.2 will suffer from the increased occupancy in the inner tracker. Such complexity directly translates into an increased probability of fake matching, that is matching a tracker track to an uncorrelated segment in the muon chambers, resulting in a fake muon. On the other hand, even the Global Muon (OI) reconstruction will face a challenging environment with an increasing probability of failure when propagating towards the pixel detector. This could lead to broken tracks with no information close to the vertex, therefore reducing the overall reconstruction efficiency, for example in events where two muons are produced in the same vertex (dimuons) or close-by muons that might not be correctly separated.

Therefore, there is a real need to revise and eventually redesign the muon reconstruction to face both the computational challenges and the physics reconstruction challenges posed by Phase-2 conditions.

The work presented in the following focuses on optimizing the current muon reconstruction workflow at the HLT taking advantage of new approaches and paving the way for the possibility of integrating new technologies. In particular, there are two main focuses of this first optimization pass of the Phase-2 muon reconstruction: Standalone Muon reconstruction (L2), detailed in section 4.1 and Tracker / Global Muon reconstruction (L3) explored in section 4.2.

4.1 The Standalone Muon Reconstruction optimization

When dealing with Standalone Muon reconstruction, the main aim is to take advantage of the new capabilities offered by the L1 Trigger upgrade. As mentioned in section 3.1, the Phase-2 hardware trigger will combine information from the tracker and muon systems, producing L1 Tracker Muons with much better momentum resolution with respect to the L1 Standalone Muons in use until present. Moreover, the current approach to the Standalone reconstruction, detailed in section 3.2, starts from two collections of seeds:

1. the L1 Tracker Muons produced by the hardware trigger;
2. the Standalone Offline seeds produced with only and all the track segments reconstructed in the muon chambers.

These two collections then need to be geometrically matched to produce a unique

collection of L2 seeds. However, this approach presents a couple of problems exacerbated by Phase-2 conditions:

- The matching is performed by extrapolating each L1 Tracker Muon and L2 Offline seed to a common surface to then check their respective global θ and ϕ coordinates. This process becomes more complex as the occupancy increases, enhancing the probability of wrong matching, meaning that a L1 Tracker Muon is associated with the wrong L2 Offline seed, spoiling the next steps of the reconstruction.
- The current Offline Standalone seeding approach estimates the momentum of each candidate through the trajectory measured in the muon chambers, not taking advantage of the more precise information available thanks to the inclusion of information from the inner tracker in the new L1 Tracker Muons.

Therefore, the creation of seeds represents the starting point for the optimization of the Standalone Muon reconstruction. Wanting to tackle both the highlighted difficulties in Phase-2 conditions, the guiding principles for the redesign of the Standalone seeding approach focused on simplifying the creation of seeds removing the need to create two separate collections to be merged and exploiting the new information present in the L1 Tracker Muons.

The new seeding approach builds a single collection of L2 seeds starting from L1 Tracker Muons in the inner tracker as well as trigger primitives and reconstructed local segments in the muon chambers. In particular, for every L1 Tracker Muon in the event, the trigger primitives used to build it are matched with DT segments in the barrel, CSC segments in the endcap and a combination of both in the overlap region. The primitive-segment match is based on multiple factors:

- The primitive and the segment are required to be in the same muon chamber.
- Global ϕ separation: a matching segment is required to be in a matching window with $\Delta\phi < 0.05$.
- If multiple segments are found inside the ϕ window, the number of hits in each of the found segments is compared, favouring segments with the highest number of hits.
- Finally, only in the barrel, to further discriminate eventual duplicate segments with the same number of hits, a selection in θ is applied, requiring that the angle measured by the L1 Tracker Muon is within a window $\Delta\theta < 0.1$ with respect to the segment.

This process allows to associate at most one trigger primitive to one segment per muon chamber. Furthermore, in the barrel region, if any of the 4 muon stations has

no trigger primitive or the matching has failed, a rough extrapolation from the closest station with a match is performed. This is specifically applied in the barrel region since, in this case, it has been measured that the Offline efficiency (i.e. the production of segments in the muon chambers) is higher than the efficiency of Trigger Primitives. The extrapolation assumes a mostly straight trajectory and opens a 20 cm wide window in the chamber where no matches were found to look for segments. When the extrapolation produces multiple results, the same matching logic as the primitive-segment case is applied. The extrapolation logic is specifically applied to the barrel region due to its reduced occupancy, even in Phase-2 conditions with high collision pile-up. As such, the number of DT segments per event remains in the few tens range, reducing the probability of wrong matching when extrapolating from a nearby station.

Once the primitive-segment matching and the extrapolation have been performed for all the stations, the information is propagated to the second station (MB2 in the barrel, ME2 in the endcap) where the seed is created. The propagation, although not strictly necessary, was implemented in this way to conform to the current convention, present since Run 1. This allows the new seeding module to take advantage of the same track finding algorithms already present while propagating all the matched segments, as well as the precise momentum measurement from the L1 Tracker Muon to proceed with the L3 Outside-In reconstruction as shown in the diagram in figure 3.9.

In the end, at most one L2 seed is created per L1 Tracker Muon, removing the need to match two distinct seeds collections and improving the momentum information propagated to build L3 tracks. The reduction in the number of seeds created is shown in figure 4.2. The result, obtained from a sample of about 15000 simulated $Z \to \mu\mu$ events with average pile-up 200 in Phase-2 conditions, shows the number of: Offline Standalone Muon seeds (light blue), L2 seeds matched with L1 Tracker Muons in the current seeding workflow (pink), and L2 seeds produced in a single pass by the new seeding module (orange). The new seeding module produces much fewer seeds than the current implementation, with the mode being 2 seeds per $Z \to \mu\mu$ event as it is expected when a pair of muons is produced, and the mean being slightly lower than that. This is to be compared with the Offline seeds having a mode of 10 produced per event and a mean slightly higher than that. The number of Offline seeds is then reduced by matching with the L1 Tracker Muons information, producing a distribution more in line with the new seeding module, but still with more seeds than L1 Tracker Muons in some instances. This is instead not possible in the new module where at most one seed is produced per L1 Tracker Muon. The reduction in the number of seeds comes with a close-to-none impact on the physics performance as demonstrated by the L2 Standalone Muons performance discussed in section 4.3.1.

Removing the need to create and then match two separate collections of seeds also allows to resolve ambiguities in the matching resulting from multiple matches as well as reconstruct prompt and displaced muons in a single pass, taking advantage of both L1 Tracker Muons (with a constraint on the vertex and good momentum resolution)

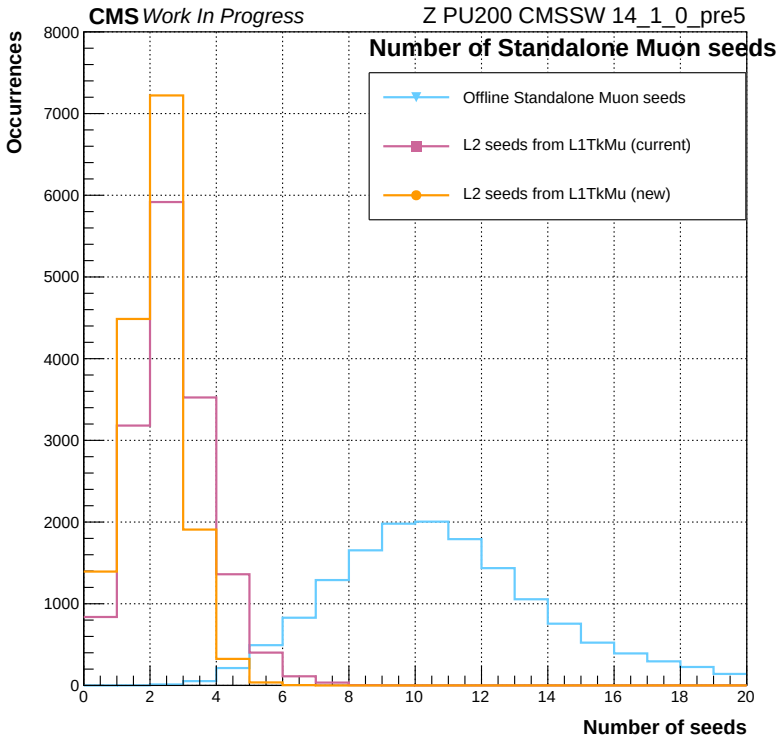


Figure 4.2: *Number of different kind of Standalone Muon seeds per $Z \to \mu\mu$ event with 200 pile-up. The Offline seeds are shown in light blue, the results obtained from matching them with L1 Tracker Muons is shown in pink, and the seeds produced by the new seeding module are shown in orange.*

and L1 Standalone Muons without any constraint on the vertex. In the latter case, the momentum is estimated using the measurement of the curvature of the track in the muon chambers and the track finding and fitting are appropriately modified to deal with segments not pointing at the primary interaction vertex.

4.2 The Tracker and Global Muon Reconstruction optimization

As discussed in section 3.2, when combining information from the inner tracker and the muon chambers, the current reconstruction workflow executes two separate algorithms for each candidate identified either in the inner tracker (as a L1 Tracker Muon) or in the muon chambers (as a L2 Standalone Muon). When starting from the inner tracker, the

reconstruction proceeds Inside-Out aiming to match an inner track with a Standalone Muon or individual segments from the muon chambers if no muon was created. On the other hand, when the reconstruction starts from the Standalone Muons, it proceeds Outside-In, intending to produce a Global Muon, that is a Standalone Muon matched with its corresponding tracker track. The tracker tracks produced by these reconstruction workflows are collectively referred to as L3 Muon tracks, with the former being identified as IO (Inside-Out) and the latter OI (Outside-In). Currently, both the Inside-Out and Outside-In reconstructions are performed for each candidate in every event.

The redundancy in the L3 reconstruction allows it to maintain high efficiency in the whole acceptance region $|\eta| < 2.4$, but it comes at a large computational cost. The inner tracker reconstruction is among the most challenging and time-consuming tasks in the muon reconstruction, as seen by the large sector occupied by the track candidate maker in the timing plot shown in figure 4.1. Moreover, in order to maintain high efficiency at every value of η and p_T, L3 track reconstruction also retains a large number of fakes, produced especially at low p_T. The combination of these two factors implies that the current L3 reconstruction could be refined for Phase-2 conditions, where timing needs to be reduced to face the increased luminosity and the number of fake tracks risks to explode due to the harsher pile-up conditions.

In most cases, efficiency and track quality for L3 IO and OI tracks are comparable. Therefore, the first pass of optimization for the L3 reconstruction is the implementation of a module that allows the execution of either of the two L3 reconstructions first, allowing the second one to be executed only for candidates that do not produce a good-enough L3 track after the first pass. The module should be flexible enough to choose which reconstruction to execute first, check the quality of the produced tracks and produce a collection of seeds for the second L3 reconstruction to be executed containing only the objects reconstructed from inputs that were not used already. This approach also allows to retain only good-quality tracks after the first reconstruction pass, further reducing the fake rate in the merged L3 collection and easing the work done by the Muon ID by anticipating the cuts applied to all L3 inner tracks on the required number of hits in the pixel (at least one) and in the tracker (at least 6).

The selection module was designed with the flexibility to choose which L3 reconstruction to execute first as a feature. This has two main long-term benefits:

- allows to choose the L3 reconstruction to execute first based on multiple performance metrics including efficiency, timing, fake rate, and p_T resolution;
- the choice can be re-evaluated based on detector ageing or relevant changes throughout data taking that may affect the reconstruction performance.

In the following, the two possibilities of executing either Inside-Out or Outside-In reconstruction first are discussed in detail. Their performance is evaluated and compared with the current, fully redundant, implementation.

4.2.1 L3 reconstruction Inside-Out first

Figure 4.3 shows the muon reconstruction workflow when L3 Inside-Out reconstruction is executed first.

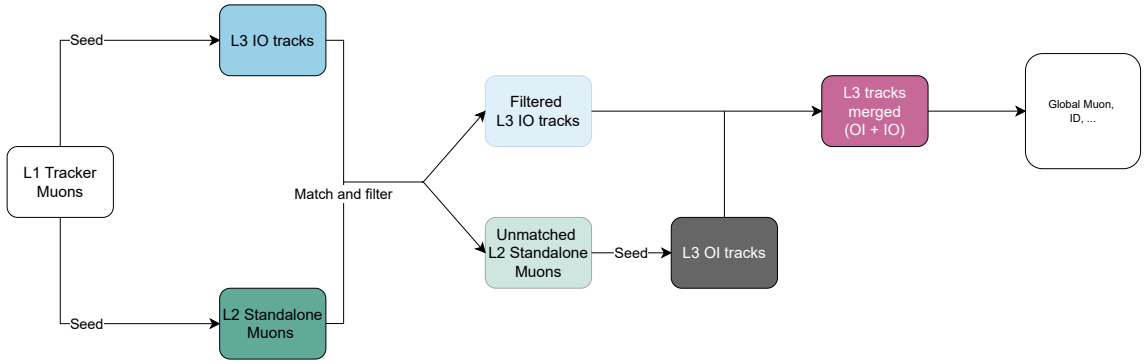


Figure 4.3: *Schematic diagram of the muon reconstruction workflow in the case where L3 Inside-Out reconstruction is executed first.*

L1 Tracker Muons are the common starting point, used to seed both L2 Standalone Muons and L3 Inside-Out tracks. The latter two collections produced are then matched and filtered. In particular:

- The quality of each L3 track is assessed, requiring at least 1 hit in the pixel and 6 hits in the tracker (same criteria as HLT Muon ID), as well as a maximum $\chi^2/d.o.f = 5$ to consider the track of good-enough quality. All tracks that pass the quality selection are added to the filtered collection.
- The L1 Tracker Muon used to seed each L2 Standalone Muon is geometrically matched with the L3 IO tracks that passed the quality selection, requiring $\Delta R < 0.02$.
  - If the match is successful, the L2 Standalone Muon will not be used again.
  - If the match is unsuccessful, the L2 Standalone Muon will be used to seed L3 OI tracks.

This approach allows the reconstruction to take advantage of the superior inner track quality, and information about the vertex coming from the pixel of the L3 Inside-Out tracks, resorting to Outside-In reconstruction only for candidates that were not correctly reconstructed the first time. Moreover, using the filtered L3 Inside-Out tracks, together with the Outside-In collection to create the L3 merged tracks reduces the fake rate and

makes the later Muon ID pass simpler, having already filtered the large majority of the tracks using some of the same criteria (e.g. the number of hits in the pixel and tracker).

4.2.2 L3 reconstruction Outside-In first

Figure 4.4 shows the muon reconstruction workflow when L3 Outside-In reconstruction is executed first.

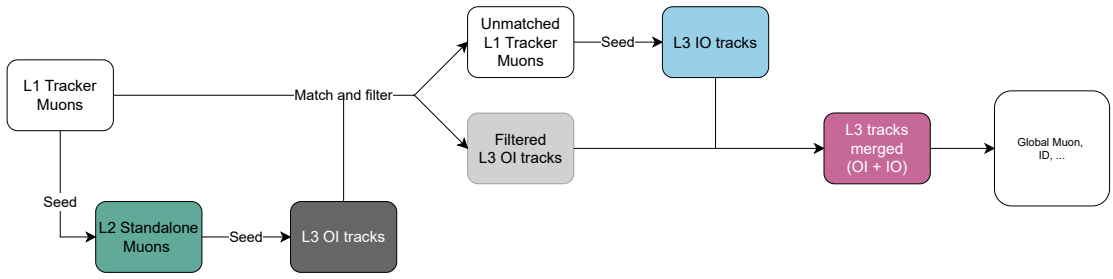


Figure 4.4: *Schematic diagram of the muon reconstruction workflow in the case where L3 Outside-In reconstruction is executed first.*

In this case, L1 Tracker Muons are only used to directly seed L2 Standalone Muons as discussed in section 4.1. The L2 muon collection is then used as the starting point for L3 Outside-In reconstruction. The L3 tracks are filtered, requiring some minimum quality criteria and each L1 Tracker Muon in the event is matched with the filtered L3 tracks. In particular:

- The quality of each L3 track is assessed, requiring at least 1 hit in the pixel and 6 hits in the tracker (same criteria as HLT Muon ID), as well as a maximum $\chi^2/d.o.f = 5$ to consider the track of good-enough quality. All tracks that pass the quality selection are added to the filtered collection.
- A match is considered successful if a filtered L3 Outside-In track is found in a $\Delta R < 0.02$ window with respect to a L1 Tracker Muon, calculated using the global coordinates of both objects.
  - If a successful match is found, the corresponding L1 Tracker Muon will not be used again.

– If the matching is unsuccessful, the associated L1 Tracker Muon will be reused to seed the L3 Inside-Out tracks in the next reconstruction step.

This reconstruction approach reduces the complexity of inner muon track reconstruction, resorting to the most complex and intensive Inside-Out reconstruction only for candidates that were not correctly reconstructed using information from the muon chambers. As such, the fake rate is reduced in all steps of the reconstruction and the Muon ID has to apply its cuts on significantly fewer candidates.

4.3 Results

The development of the new reconstruction modules has been based on two core pillars: achieve a physics performance as close as possible to Run 3, and improve computing resources usage.

The first item has been the main guiding principle for any change done, especially when developing the new Standalone seeding logic. To assess physics performance, a validation workflow targeting all intermediate reconstruction objects has been created and used throughout development, taking advantage of the validation module offered by the reconstruction software. Therefore, relevant validation plots comparing the reconstruction to the simulated data were produced for most reconstructed objects. Furthermore, the performance of specific objects that are generally not plugged into the validation (e.g. L1 Tracker Muons), was assessed by saving and analyzing detailed information through custom ROOT trees. Some of the key performance metrics measured are:

- the $\chi^2/d.o.f$ of reconstructed tracks;
- the efficiency as a function of η, defined using the muon track validator. The efficiency is calculated as the fraction of reconstructed muons that are associated with a simulated muon passing a given simulation-to-reconstruction matching criteria (i.e. at least 75% of the hits used in the reconstructed object were produced by the correct simulated muon);
- the efficiency as a function of the muon transverse momentum p_T;
- the fake rate as a function of η for each type of reconstructed object. The fake rate is defined as the ratio between the number of fakes (i.e. reconstructed objects not matching a simulated one) over the total number of reconstructed objects. Since in the simulated data no information about the pile-up particles is retained, this definition of fake rate makes other muons not coming from the main generated event of interest fakes;
- the resolution of the transverse (d_{xy}) and longitudinal (d_z) impact parameters;

- the transverse momentum resolution;
- the segment or hit multiplicity, meaning the number of hits/segments used to produce a track or seed;
- the multiplicity of each kind of reconstructed object (also referred to as the complexity of a given object).

For some of these metrics, relevant plots are shown for both the current implementation and the modified versions implemented for this thesis work in section 4.3.1. Comparison plots for specific objects are also reported.

The timing has been tackled mainly by reducing the redundancy present in the current schema as described in section 3.2. Moreover, the CMS reconstruction software integrates a utility to measure the timing of the various reconstruction modules. This utility has been used extensively to test the changes. However, the timing measured in this way does not paint the full picture, which is discussed in more detail in section 4.3.2

4.3.1 Physics performance

Physics performance is mostly evaluated via validation plots to compare the performance of different reconstructed objects or the same objects before and after the changes applied as part of this work. In the following, the impact of changes on Standalone Muons and Muons seeds as well as Global Muons and Tracker Muons tracks are discussed in their respective sections. All the physics results shown refer to a sample of about 15000 simulated $Z \to \mu\mu$ events at $\sqrt{s} = 14$ TeV with pile-up 200 (HL-LHC conditions). The relatively low number of events is due to the fact that each event undergoes the full HLT reconstruction chain, meaning that processing a factor of 10 more events for statistical stability would require a significant amount of execution time (in the order of a few days). Therefore, this figure was chosen to allow for faster iteration while still providing usable validation results.

Figure 4.5 shows the efficiencies of multiple reconstructed objects (as highlighted in the reconstruction workflow reported in figure 3.9) in the current implementation (top), Inside-Out reconstruction first (bottom left), and Outside-In reconstruction first (bottom right). In the latter two cases, the L3 tracks reconstructed during the second pass (L3 OI, and L3 IO tracks, respectively) show what seems to be much worse efficiency. This is not the case, since the second pass is only trying to reconstruct objects missed by the first one. Therefore, those efficiencies should be seen as compensating for eventual inefficiencies of the first pass. L3 OI tracks show a drop in efficiency in the range $0.7 \leq |\eta| \leq 1.3$ which is not surprising since this region corresponds to the overlap for the muon system where matches between Standalone Muon tracks have overall lower quality.

Figure 4.6 shows a similar comparison focusing on the reconstruction efficiency as a function of the transverse momentum. Once again the current implementation is shown

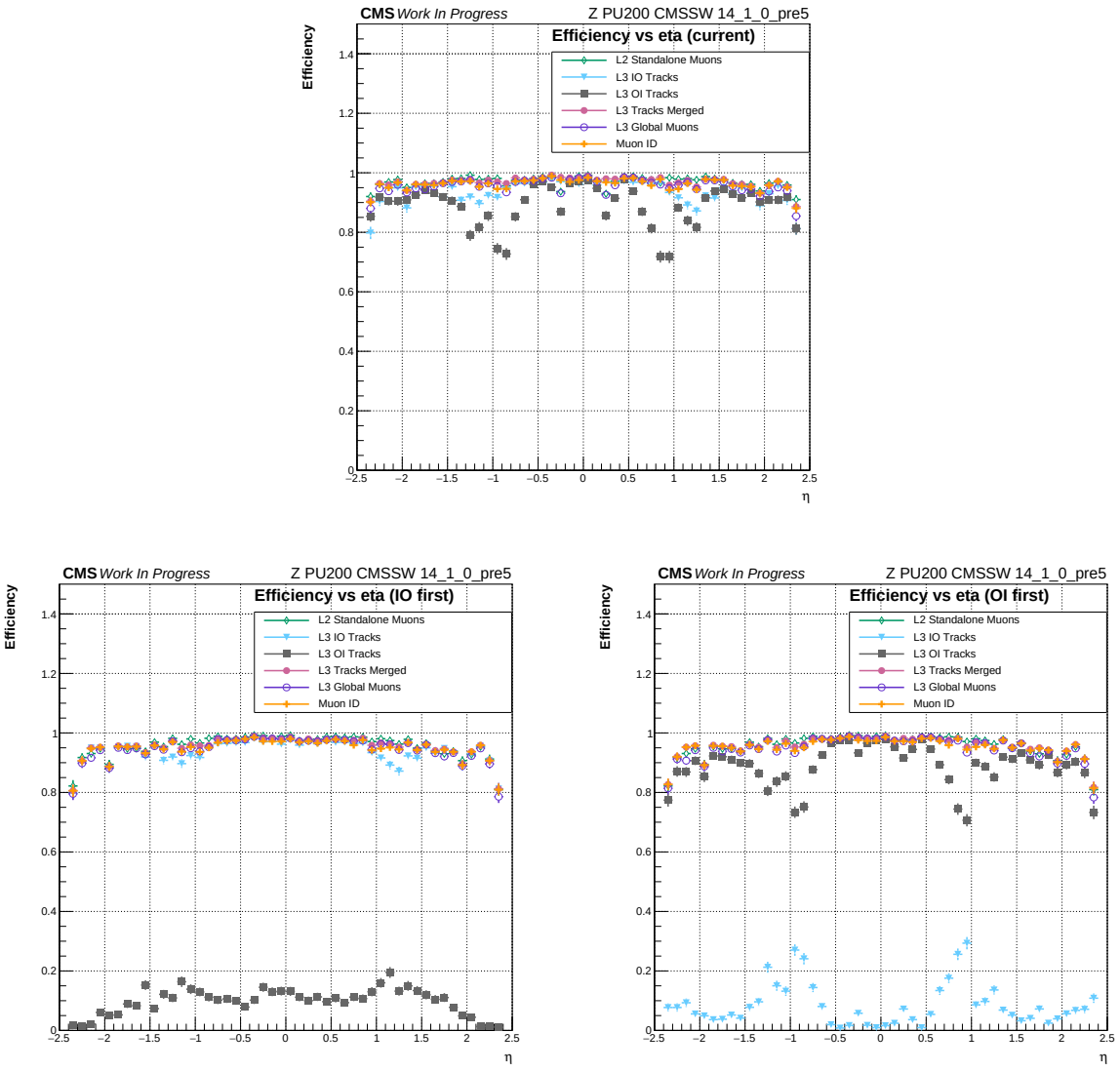


Figure 4.5: *Efficiency as a function of η for multiple reconstructed objects: Standalone muons (green), L3 Tracker tracks (IO) (light blue), L3 Muon tracks (OI) (grey), L3 tracks merged (pink), Global Muons (purple), Muon ID (orange). The top plot shows the current implementation, the bottom left plot shows the L3 Inside-Out reconstruction first, and the bottom right displays the L3 Outside-In reconstruction first.*

as the top figures, with the two modified reconstructions Inside-Out first and Outside-In first shown on the bottom left and bottom right, respectively. As already stated, in the modified reconstructions, the L3 tracks corresponding to the second pass are expected to show a sharp drop in efficiency, since they only attempted if the first pass failed or did not produce a reconstructed track of good-enough quality. As a general trend, most reconstructed objects reach high efficiency at a transverse momentum close to 5 GeV, with L3 Tracker Muons (L3 OI tracks) lagging slightly behind due to the strict requirements of having hits in at least two muon chambers. Since this requirement is not shared by other candidates, that can be reconstructed even if hits are recorded in a single chamber, the L3 OI tracks are less efficient at low p_T where multiple scattering is dominant thus a significant number of muons only crosses a single muon chamber and the propagation of the tracks tends to fail more often. As a side note, requiring hits in at least two muon chambers is necessary for the L3 OI track seeds, since the curvature measured between the chambers is used to estimate the momentum of the muon.

Figure 4.7 shows the comparison between the same reconstructed objects and workflows, measuring the fake rate as a function of η. As a general feature, the fake rate of the modified reconstruction workflows tends to be lower for mostly all reconstructed objects, except L3 Tracker tracks in the Outside-In first case. In this specific instance, it looks like the first pass can reconstruct most of the actual muons, leading to the second pass being almost entirely made up of fakes.

Table 4.1 shows the multiplicities of reused objects in more detail. The Inside-Out first approach manages to reconstruct all tracks in a single pass in more than 80% of the events in the sample. In the remaining fraction of events, the mean number of Standalone Muons to be reused is slightly above 1 per event.

The Outside-In first workflow requires a second pass more often, in about 75% of the sample events. When a second pass is necessary, the mean number of L1 Tracker Muons to be reused is slightly lower than 2.5 per event.

Reconstruction	Total events	Events with no reused object	Percentage	Mean number of reused objects (only for events where necessary)
Inside-Out first	15376	12545	81.6%	1.083 ± 0.018
Outside-In first	15376	3683	24.0%	2.464 ± 0.005

Table 4.1: *Comparison of the number of events that do not require a second reconstruction pass for the Inside-Out first and Outside-In first reconstruction workflows. The mean number of objects to be reused is also reported and computed considering only events where a second pass is necessary.*

More detailed comparisons are discussed in the following.

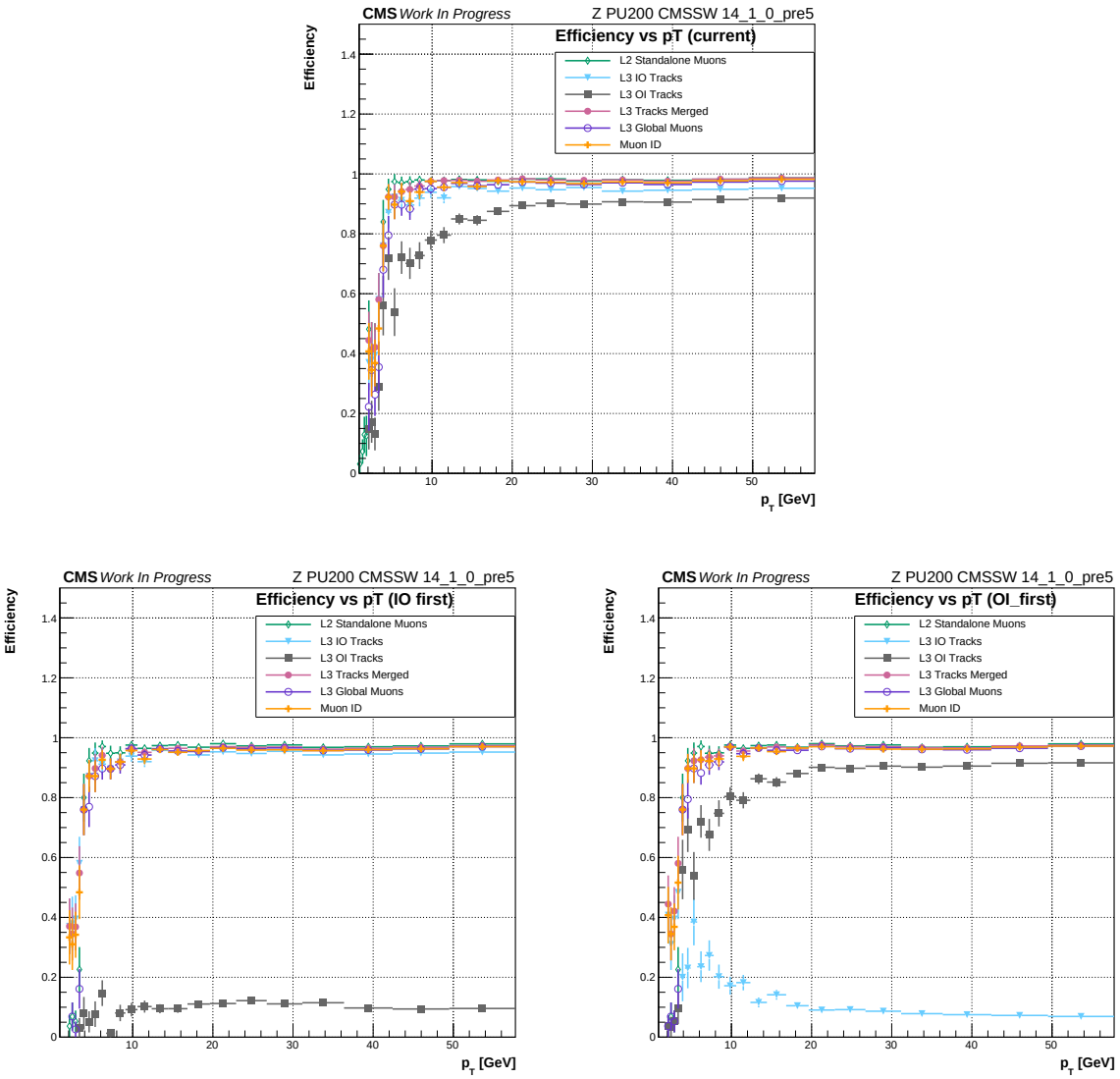


Figure 4.6: *Efficiency as a function of p_T for multiple reconstructed objects: Standalone muons (green), L3 Tracker tracks (IO) (light blue), L3 Muon tracks (OI) (grey), L3 tracks merged (pink), Global Muons (purple), Muon ID (orange). The top plot shows the current implementation, the bottom left plot shows the L3 Inside-Out reconstruction first, and the bottom right displays the L3 Outside-In reconstruction first.*

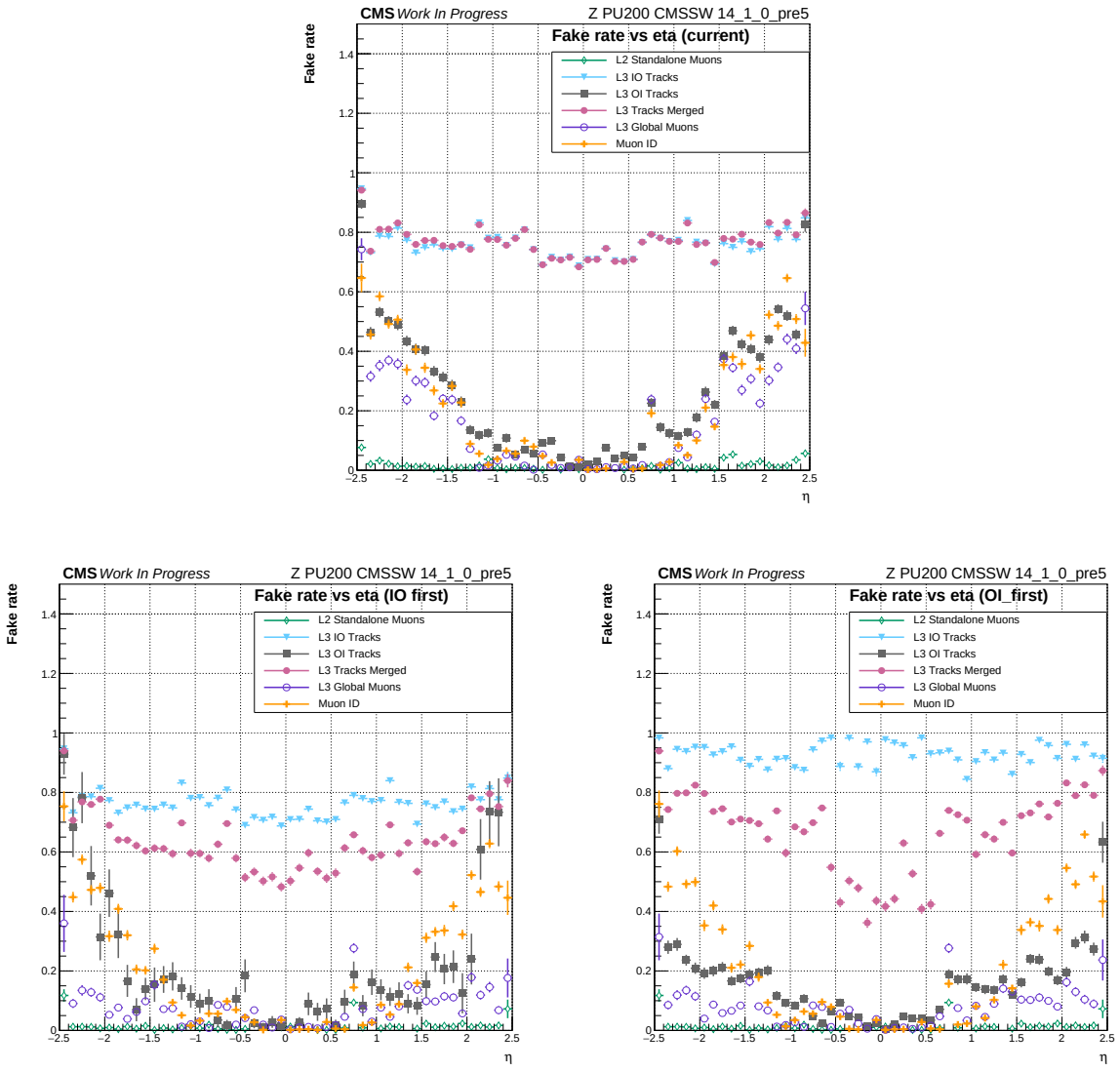


Figure 4.7: *Fake rate as a function of η for multiple reconstructed objects: Standalone muons (green), L3 Tracker tracks (IO) (light blue), L3 Muon tracks (OI) (grey), L3 tracks merged (pink), Global Muons (purple), Muon ID (orange). The top plot shows the current implementation, the bottom left plot shows the L3 Inside-Out reconstruction first, and the bottom right displays the L3 Outside-In reconstruction first.*

Standalone Muons and Muons seeds

Changes applied to the Standalone Muon seeding naturally propagate to the reconstructed Standalone Muons as well. Therefore, to verify the validity of the changes, both the seeds and the reconstructed muons have continually been monitored. To verify the performance of the seeds, the validation code extrapolates them to tracks using information from the simulated data and is thus able to assign them a χ^2, efficiency and all other parameters usually associated with reconstructed objects. Figure 4.8 shows the comparison between the current seeds and the modified seeds implemented for this work in efficiency as a function of η. The plot demonstrates similar performance at the seed level. In particular, the new implementation experiences some losses in the endcaps, specifically close to the limit of the acceptance. The new workflow also recovers some efficiency in the barrel region, probably because the matching of L1 information with segments in the muon chambers manages to recover some segments that would have been lost using the current extrapolation method.

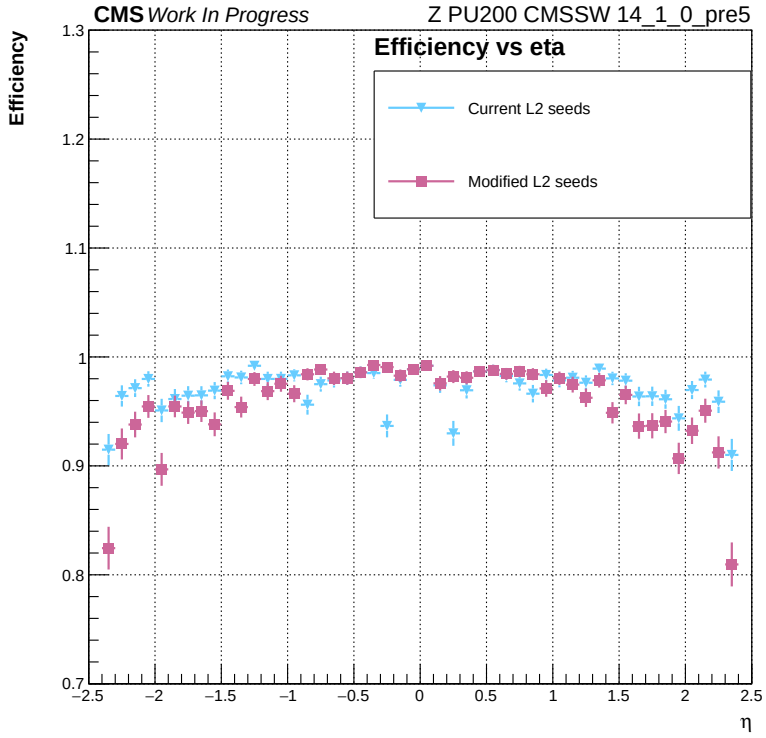
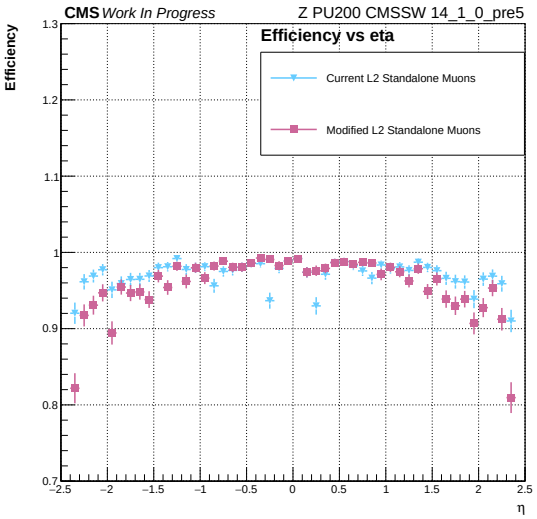


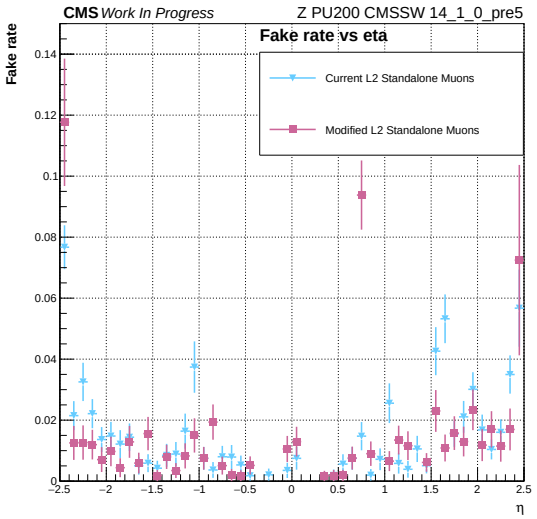
Figure 4.8: *Efficiency as a function of η for Standalone Muon seeds as are produced currently (light blue) and after the changes for this work (pink).*

As far as L2 Standalone Muons are concerned, the similarity in performance resembles

what happens with the seeds. Figures 4.9a and 4.9b show the comparisons for efficiency and fake rate for Standalone Muons before and after the changes. The comparison in efficiency shows largely compatible results, except the very edges of the η acceptance region which is still subject to changes since the same inefficiencies have been found for the new seeds and the current track finding does not fully exploit the new detectors installed for Phase-2 and the propagation of track information might be sub-optimal. Although Standalone Muons have always been clean reconstructed objects, characterized by low amounts of fakes, the changes implemented manage to further reduce the fake rate in practically all η regions which can result in faster reconstruction, especially in high pile-up conditions.



(a) *Efficiency as a function of η for Standalone Muons as they are currently (light blue line) and after the changes (pink line).*



(b) *Fake rate as a function of η for Standalone Muons as they are currently (light blue line) and after the changes (pink line).*

A more detailed view of the efficiency in the barrel, overlap, and endcap regions is presented in table 4.2. The performance comparison shows that the efficiencies of the current and new implementations are within the margin of error in both the barrel and overlap regions, while the new implementation experiences a loss of about 3% in the endcap region which needs to be investigated further.

Standalone Muons efficiency			
Reconstruction workflow	Barrel	Overlap	Endcap
Current	0.972 ± 0.008	0.98 ± 0.03	0.963 ± 0.014
New	0.967 ± 0.008	0.98 ± 0.03	0.933 ± 0.014

Table 4.2: *Efficiencies in the barrel, overlap, and endcap regions for Standalone Muons before and after the changes implemented for this work. The efficiency (ϵ) is calculated as the ratio between the number of reconstructed objects associated with their corresponding simulated track (i.e. at least 75% of the hits used in the reconstruction were produced by the correct simulated object) and the total number of simulated tracks. The ratios are computed separately for each region. The uncertainty is estimated as $1/\sqrt{N}$ where N is the numerator of the efficiency calculation (overestimating the statistically correct $\epsilon/\sqrt{N}$ since $\epsilon < 1$).*

Global and Tracker Muons

Global and Tracker Muons optimization (L3 OI and L3 IO tracks, respectively) was mostly aimed at reducing redundancies in reconstruction. Therefore, the results shown here compare the performance of L3 IO and OI tracks before and after the changes, taking into account both workflows Inside-Out first and Outside-In first. The main performance metrics are track efficiency as a function of η or p_T as well as the fake rate as a function of η. The optimization aimed at keeping the efficiency as close as possible, while significantly reducing the fake rate.

Figure 4.10 shows efficiency and fake rate as functions of η for L3 Tracker tracks (L3 IO) in the left and right plots, respectively. Efficiency for L3 IO tracks is perfectly superimposed for the current and Inside-Out reconstructions, as expected since they are produced using the same algorithms and from the same L1 Tracker Muons collection. In the Outside-In first case, L3 IO tracks naturally show much lower efficiency having to only fill gaps left by the first reconstruction pass. Comparing current and Inside-Out first fake rates, there is a significant reduction throughout almost the entire η acceptance, with the best improvement being in the barrel region. The Outside-In first reconstruction shows a higher fake rate, once again tied to the fact that it is executed as a second pass, thus producing much fewer candidates.

Table 4.3 reports the L3 IO tracks efficiency separated for barrel, overlap, and endcap regions for all the reconstruction workflows under analysis. The Inside-Out first reconstruction achieves a performance matching that of the current implementation, with the Outside-In first workflow showing low overall efficiency as expected and previously discussed.

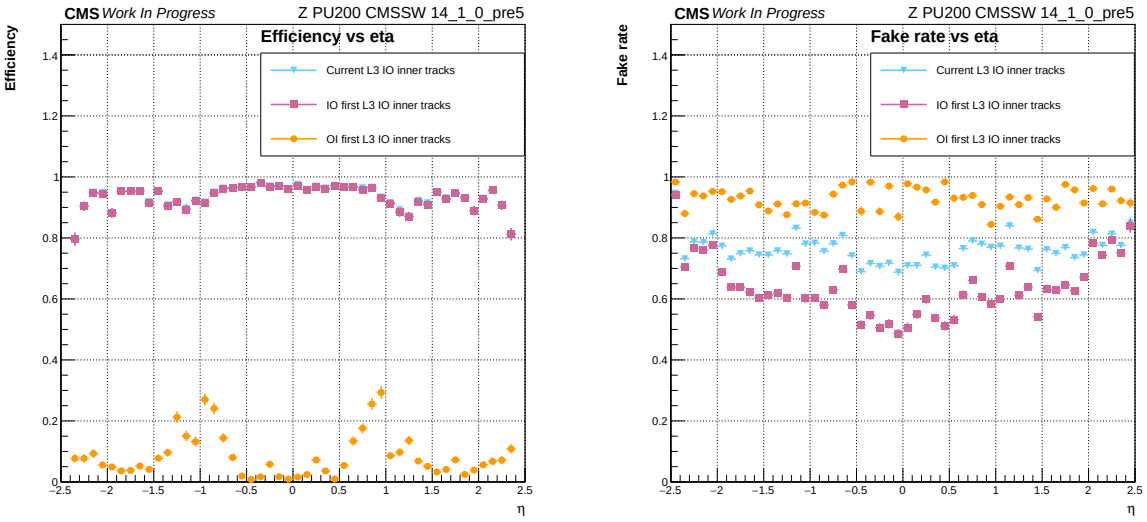


Figure 4.10: *L3 Tracker tracks (L3 IO) comparison between current implementation (light blue), Inside-Out first reconstruction (pink), and Outside-In first reconstruction (orange). Efficiency as a function of η is shown on the left, while fake rate as a function of η is shown in the right plot.*

L3 Tracker tracks (IO) efficiency			
Reconstruction workflow	Barrel	Overlap	Endcap
Current	0.948 ± 0.008	0.91 ± 0.03	0.914 ± 0.014
Inside-Out first	0.946 ± 0.008	0.91 ± 0.03	0.912 ± 0.014
Outside-In first	0.09 ± 0.03	0.16 ± 0.07	0.06 ± 0.05

Table 4.3: *Efficiencies in the barrel, overlap, and endcap regions for L3 IO tracks in the current, Inside-Out first, and Outside-In first reconstruction workflows. The efficiency (ϵ) is calculated as the ratio between the number of reconstructed objects associated with their corresponding simulated track (i.e. at least 75% of the hits used in the reconstruction were produced by the correct simulated object) and the total number of simulated tracks. The ratios are computed separately for each region. The uncertainty is estimated as $1/\sqrt{N}$ where N is the numerator of the efficiency calculation (overestimating the statistically correct $\epsilon/\sqrt{N}$ since $\epsilon < 1$).*

Figure 4.11 shows efficiency and fake rate as functions of η for L3 Muon tracks (L3 OI) in the left and right plots, respectively. Efficiency is mostly compatible between the current implementation and the Outside-In first reconstruction with only a few spots where one is more efficient than the other. L3 Muon tracks produced by the Inside-Out first reconstruction show efficiency compatible with the gaps left by Tracker tracks. The fake rate comparison shows a large improvement when comparing the current implementation and the Outside-In first reconstruction. Although Muon tracks (L3 OI) are usually less prone to producing fakes compared to Tracker tracks (L3 IO), the filter implemented for this thesis work allows a significant reduction of the fake rate. This is especially notable in the endcaps where the probability of random matching is higher due to the much greater occupancy. When Inside-Out reconstruction is executed first, much fewer Muon tracks are produced, leading to less statistics and thus larger error bars. However, the fake rate of Muon tracks (L3 OI) produced in this case is still largely compatible or lower than the one of the current implementation.

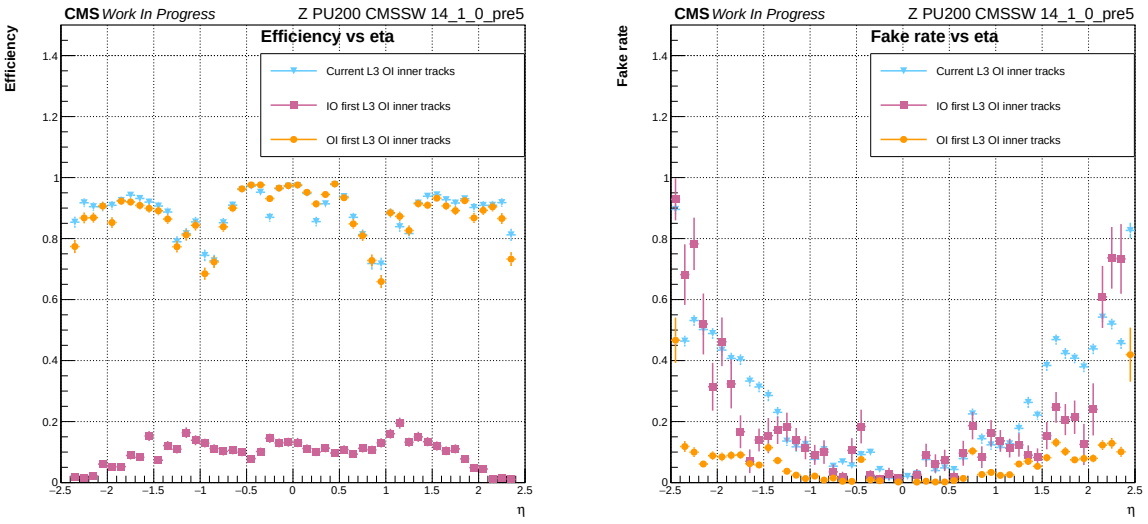


Figure 4.11: *L3 Muon tracks (L3 OI) comparison between current implementation (light blue), Inside-Out first reconstruction (pink), and Outside-In first reconstruction (orange). Efficiency as a function of η is shown on the left, while fake rate as a function of η is shown in the right plot.*

The detailed efficiency comparison for L3 OI tracks is reported in table 4.4. Once again, the first reconstruction pass achieves an efficiency comparable to the current, redundant, implementation in both the barrel and the overlap regions. The efficiency in the endcap shows a slight 2% drop tied to the worse performance of the new Standalone Muon seeds in this region since they are used as the starting point to produce L3 OI tracks. The objects reconstructed during the second pass achieve much lower efficiency

as they are only needed when the first pass fails or does not produce a track with good-enough quality.

L3 Global tracks (OI) efficiency			
Reconstruction workflow	Barrel	Overlap	Endcap
Current	0.891 ± 0.008	0.81 ± 0.03	0.904 ± 0.014
Inside-Out first	0.10 ± 0.02	0.16 ± 0.06	0.08 ± 0.05
Outside-In first	0.884 ± 0.008	0.81 ± 0.03	0.883 ± 0.015

Table 4.4: *Efficiencies in the barrel, overlap, and endcap regions for L3 OI tracks in the current, Inside-Out first, and Outside-In first reconstruction workflows. The efficiency (ϵ)is calculated as the ratio between the number of reconstructed objects associated with their corresponding simulated track (i.e. at least 75% of the hits used in the reconstruction were produced by the correct simulated object) and the total number of simulated tracks. The ratios are computed separately for each region. The uncertainty is estimated as $1/\sqrt{N}$ where N is the numerator of the efficiency calculation (overestimating the statistically correct $\epsilon/\sqrt{N}$ since $\epsilon < 1$).*

Figure 4.12 shows efficiency and fake rate as functions of η for L3 merged tracks (L3 IO + L3 OI) in the left and right plots, respectively. The efficiency plot shows great agreement among all analysed reconstruction workflows. This confirms that reducing the redundancy present in the current schema does not particularly affect physics performance when taking advantage of the new information available after the Phase-2 upgrade. The fake rate is also reduced in both the Inside-Out and Outside-In first approaches when compared to the current implementation. The Inside-Out first reconstruction performs better in the endcaps, while the Outside-In first approach manages to further reduce the fake rate in the barrel region.

A detailed efficiency comparison for L3 merged tracks is shown in table 4.5. Both the newly implemented reconstruction workflows achieve performance compatible with the current one. The endcap region shows the worst performance withing the margin of the statistical error as a result of the lower efficiency of the new Standalone Muon reconstruction in that region.

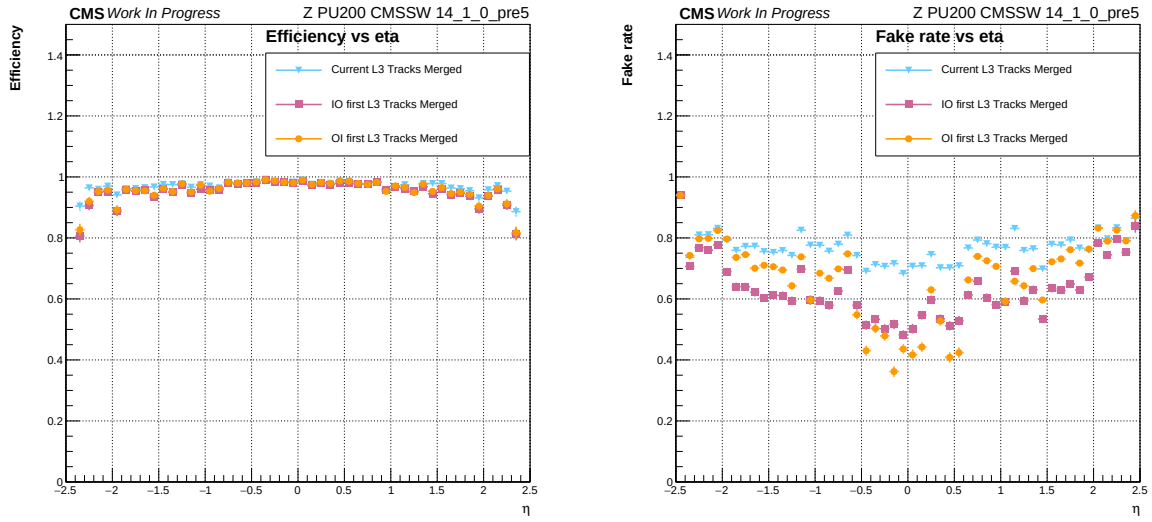


Figure 4.12: *L3 merged tracks (L3 IO + L3 OI) comparison between current implementation (light blue), Inside-Out first reconstruction (pink), and Outside-In first reconstruction (orange). Efficiency as a function of η is shown on the left, while fake rate as a function of η is shown in the right plot.*

L3 Global tracks (OI) efficiency			
Reconstruction workflow	Barrel	Overlap	Endcap
Current	0.973 ± 0.008	0.97 ± 0.03	0.957 ± 0.014
Inside-Out first	0.962 ± 0.008	0.96 ± 0.03	0.932 ± 0.014
Outside-In first	0.966 ± 0.008	0.963 ± 0.03	0.937 ± 0.014

Table 4.5: *Efficiencies in the barrel, overlap, and endcap regions for L3 merged tracks (L3 IO + L3 OI) in the current, Inside-Out first, and Outside-In first reconstruction workflows. The efficiency (ϵ) is calculated as the ratio between the number of reconstructed objects associated with their corresponding simulated track (i.e. at least 75% of the hits used in the reconstruction were produced by the correct simulated object) and the total number of simulated tracks. The ratios are computed separately for each region. The uncertainty is estimated as $1/\sqrt{N}$ where N is the numerator of the efficiency calculation (overestimating the statistically correct $\epsilon/\sqrt{N}$ since $\epsilon < 1$).*

Finally, figure 4.13 compares the performance of the final muon reconstructed objects after identification. The top plots show efficiency and fake rate as functions of η, while the bottom plot shows efficiency as a function of transverse momentum p_T. The results of this comparison are the same for all measured metrics: both the Inside-Out and Outside-In first workflows implemented for this thesis work achieve performance comparable with the current implementation while reducing complexity and unnecessary redundancies.

The detailed comparison of HLT Muon ID efficiencies is reported in table 4.6. The efficiencies in the barrel and overlap regions are largely compatible among all versions of the reconstruction with only minor differences. The endcap region suffers the most with an approximate 2% loss in efficiency: a result of the propagation of the efficiency loss in the new Standalone Muon seeds.

Muon ID efficiency			
Reconstruction workflow	Barrel	Overlap	Endcap
Current	0.966 ± 0.008	0.95 ± 0.03	0.954 ± 0.014
Inside-Out first	0.956 ± 0.008	0.95 ± 0.03	0.930 ± 0.014
Outside-In first	0.960 ± 0.008	0.95 ± 0.03	0.936 ± 0.014

Table 4.6: *Efficiencies in the barrel, overlap, and endcap regions for HLT Muon ID in the current, Inside-Out first, and Outside-In first reconstruction workflows. The efficiency (ϵ) is calculated as the ratio between the number of reconstructed objects associated with their corresponding simulated track (i.e. at least 75% of the hits used in the reconstruction were produced by the correct simulated object) and the total number of simulated tracks. The ratios are computed separately for each region. The uncertainty is estimated as $1/\sqrt{N}$ where N is the numerator of the efficiency calculation (overestimating the statistically correct $\epsilon/\sqrt{N}$ since $\epsilon < 1$).*

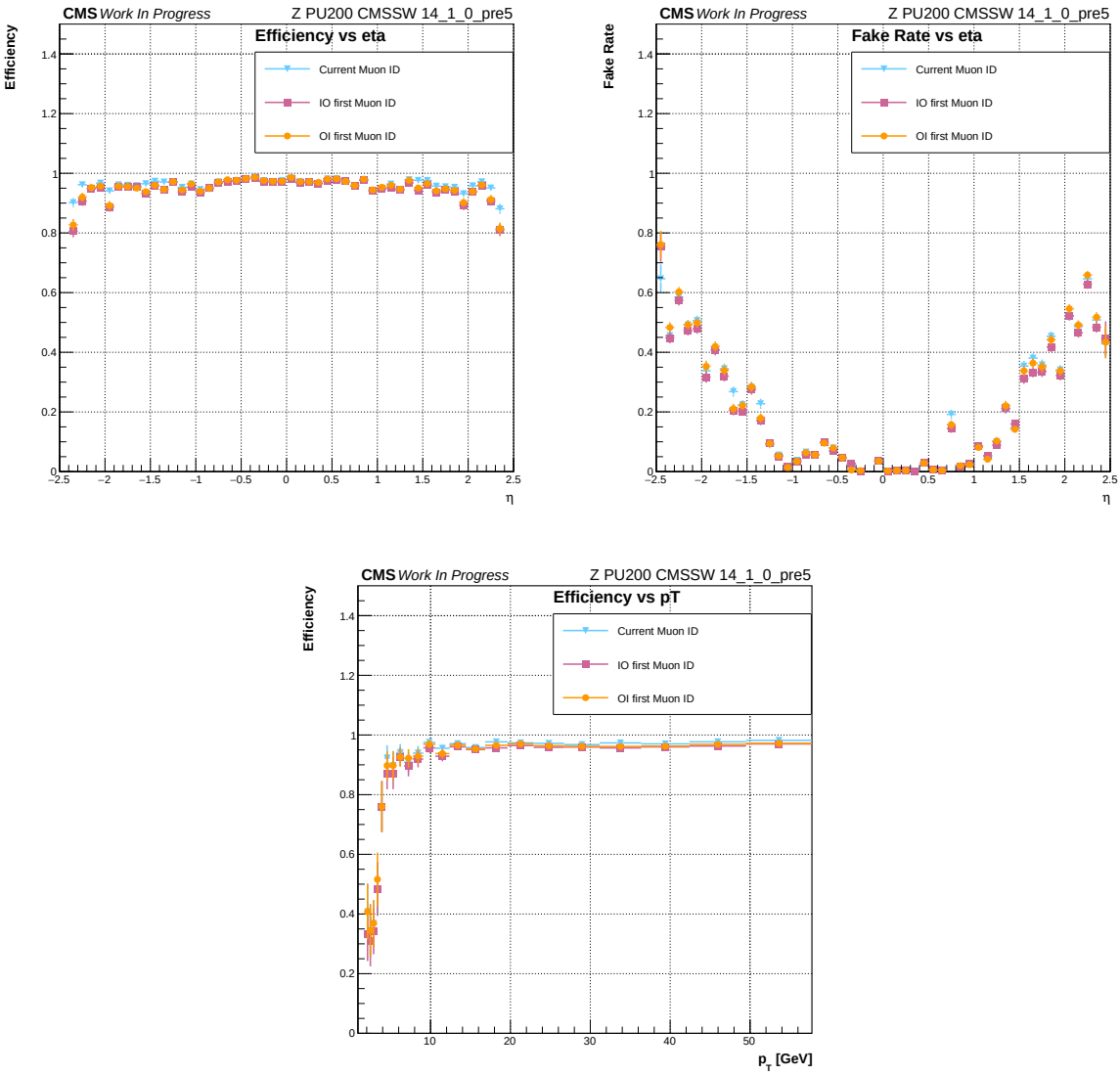
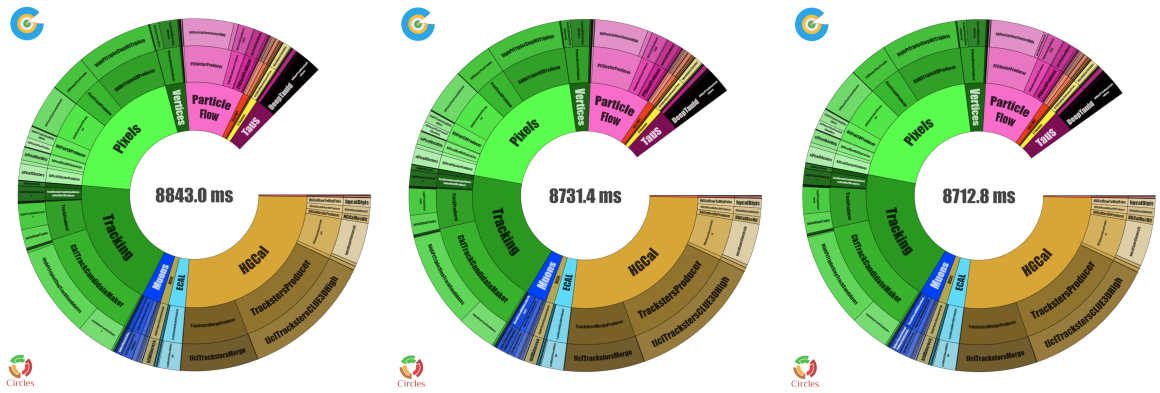


Figure 4.13: *Final reconstructed muons (Muon ID) comparison between current implementation (light blue), Inside-Out first reconstruction (pink), and Outside-In first reconstruction (orange). Efficiency and fake rate as a function of η are shown in the top left and top right plots, respectively. The bottom plot shows the efficiency as a function of the transverse momentum p_T.*

4.3.2 Timing and computing performance

As previously discussed, the current reconstruction approach cannot effectively scale up to Phase-2 conditions for multiple reasons, including timing and computing resources utilisation. Therefore, together with the physics performance, the timing of the new modules implemented for this thesis work has also been measured. In particular, this section takes into account the measurements of the total execution time of the HLT Online reconstruction. The highest-level result is represented by a full HLT Online reconstruction timing, similar to the one shown in figure 4.1, but for Phase-2 conditions. Because of this, a sample of about 7000 $t\bar{t}$ at $\sqrt{s} = 14$ TeV with pile-up 200 (Phase-2 conditions) was used to time the current, Inside-Out first, and Outside-In first reconstructions. The results are shown in figure 4.14 and are generally compatible with what was shown in the simulations for the Phase-2 Technical Design Report [35].

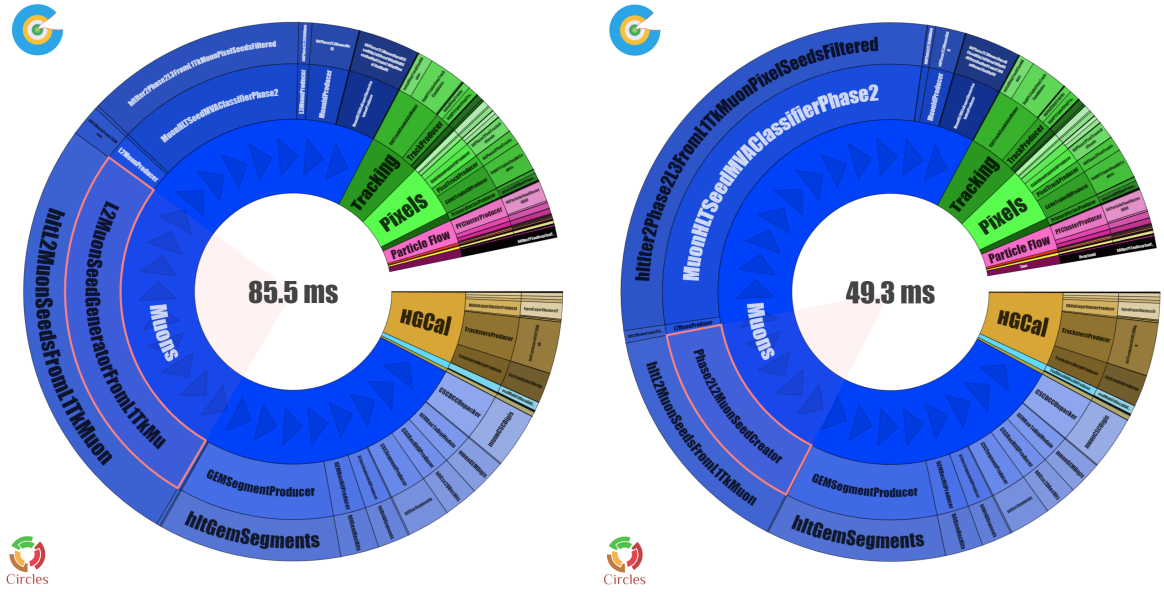


(a) Current. *(b) Inside-Out first.* *(c) Outside-In first.*

Figure 4.14: *Timing of the full CMS HLT Online reconstruction on a sample of about 7000 $t\bar{t}$ events in Phase-2 conditions. The innermost rings show the reconstruction macro-areas, while moving towards the outside the rings become more and more specific until the C++ reconstruction module is shown in the outermost ring. From left to right the performance refers to: (a) the current muon reconstruction, (b) the optimized reconstruction with the Inside-Out pass done first (see section 4.2.1), and (c) the optimized reconstruction with the Outside-In pass done first (see section 4.2.2). Both optimized reconstruction workflows benefit from the changes to Standalone Muons seeding. Total execution time is shown at the centre of each circle, the size of each module is proportional to the percentage of execution time it takes up.*

These measurements take into account all HLT subsystems and are thus largely dominated by the Pixel, Tracker, and HGCAL reconstructions, especially when run entirely on CPUs as in this case. Figure 4.15 shows only the expanded muon sector, with a particular focus on the Standalone Muon seeding for the current reconstruction (left) and after the changes for this work (right). As it has already been shown that the new

seeding approach produces seeds with the same quality as the ones presently produced, while reducing complexity and fake rate, the timing plot shows a $\approx$ 42% increase in performance at virtually no cost.



(a) *(b)*

Figure 4.15: *Zoomed timing chart from figure 4.14 to focus on the module responsible for the Standalone Muon seeding. (a) shows the timing of the current implementation (only Standalone seeds matching with L1 Tracker Muon, Offline seeds creation time is negligible in this context). (b) shows the timing of the improved Standalone Muon seeding module as described in section 4.1. The comparison between the two shows a* $\approx$ 42% *improvement due to the changes implemented.*

Measuring the computing performance difference between the current implementation and the Inside-Out first and Outside-In first reconstructions could also benefit from a complementary approach. This is mostly because the module responsible for filtering the tracks produced in the first pass has a negligible impact on timing and the resulting track finding performance is only slightly impacted. However, this does not mean that the changes have no impact on the reconstruction as a whole since the timing improvement can be directly related to the reduction of fakes. Table 4.7 reports the number of each type of reconstructed object when executing the current, Inside-Out first, and Outside-In first reconstructions. The data was obtained from the same sample used for physics performance measurements: about 15000 $Z \to \mu\mu$ events at $\sqrt{s} = 14$ TeV with 200 pile-up. In this case, reconstructed and fake objects are not defined using the matching by hit with simulated information, but a less accurate separation criterion is employed. An object is defined as "Signal" when it is found within a $\Delta R = \sqrt{\phi^2 + \eta^2} < 0.01$ (0.05

for Standalone Muons) with respect to a simulated Muon. Any other object is marked as "Fake", including muons not originating from the main $Z \to \mu\mu$ event. From the table, it is clear that the number of Reconstructed objects is mostly consistent among all different categories, while the number of Fakes can vary widely, with the worst offenders being the L3 Tracker Muons tracks (L3 IO). This was expected since those tracks originate from seeds produced in the environment with the highest occupancy and, therefore with the largest probability of having a poor quality. Moreover, to retain efficiency at extremely low p_T (< 5 GeV) the loosest Muon ID criteria are used, introducing a large number of random matches and duplicates. The large amount of Tracker Muon tracks produced carries over to the merged tracks, increasing substantially the relatively low number of Fakes produced by L3 Muon tracks (L3 OI). Most of the fakes present in the merged collection are subsequently pruned by the Muon ID which processes all previously reconstructed objects.

The Inside-Out first reconstruction filters the L3 IO tracks requiring them to be of good-enough quality. This reduces the number of Fakes produced by $\approx 42\%$ with respect to the current reconstruction while retaining comparable physics performance. The same approach produces very few L3 OI tracks, reducing the timing of their associated module by more than 90%. The tracks thus produced are extremely clean, resulting in a merged collection with significantly fewer Fakes than in the current implementation. This eases the pressure on the Muon ID which performs about 30% faster while maintaining the same physics performance.

In the Outside-In first case, an extremely clean collection of L3 Muon tracks (L3 OI) is produced. This results in very few L3 Tracker Muon tracks (L3 IO) being required specifically for cases where the Outside-In reconstruction failed. The combination of these two factors results in the overwhelming majority of the L3 IO tracks produced being identified as Fakes. However, the new merged tracks collection is still cleaner than what is currently produced, with a reduction of about 30% in the number of Fakes. This results in a similar timing performance improvement for the Muon ID as the one observed in the Inside-Out first case. In the end, the three final collections of muons that passed the Muon ID show a similar Fake/Reco ratio, but the current implementation has to process a much larger number of candidates, pruning most of them. Furthermore, removing the redundancy of the L3 reconstruction eliminates the need to perform the Muon ID on two reconstructed candidates that correspond to the same muon. Both the Inside-Out first and the Outside-In first approaches provide a significant advantage since they reduce the complexity of the identification, a feature especially relevant for high pile-up and occupancy environments. The final Muon candidates that passed the identification show a slight variation in number between the various reconstructions, with the Inside-Out first approach producing a slightly cleaner collection. The overall efficiency is compatible within 2% for all the reconstruction workflows and coarse $|\eta|$ bins under consideration, in agreement with what was discussed in the previous section and reported in table 4.6.

Reconstruction	Object	Total	Signal	Fakes	Fake/Signal
Common	L1 Tracker Muons	49324	21857	27467	1.26
Current	L2 Standalone Muons	33881	21543	12339	0.573
	L3 IO tracks	101257	22189	79068	3.56
	L3 OI tracks	30273	20218	10055	0.497
	L3 tracks merged	109982	22984	86998	3.79
	Muon ID	32999	21922	11077	0.505
Inside-Out first	L2 Standalone Muons	25487	21437	4051	0.189
	L3 IO tracks	101254	22189	79065	3.56
	L3 IO track filtered	67115	21564	45551	2.11
	Unmatched L2 Standalone Muons	3067	2502	565	0.226
	L3 OI tracks	2793	2301	492	0.214
	L3 tracks merged	67967	22023	45944	2.09
	Muon ID	31412	21606	9806	0.454
Outside-In first	L2 Standalone Muons	25487	21437	4051	0.189
	L3 OI tracks	24258	20193	4065	0.201
	L3 OI tracks filtered	21593	19970	1623	0.0812
	Unmatched L1 Tracker Muons	28817	2003	26814	13.4
	L3 IO tracks	60321	2059	58262	28.3
	L3 tracks merged	81786	21930	59856	2.73
	Muon ID	32653	21696	10957	0.505

Table 4.7: *Number of various physics objects produced throughout different reconstruction workflows: current (see figure 3.9, Inside-Out first (see figure 4.3, and Outside-In first (see figure 4.4). Signal objects are defined as the ones found within* $\Delta R = \sqrt{\phi^2 + \eta^2} < 0.01$ *with respect to a simulated muon in global coordinates. L2 Standalone Muons have a slightly less strict requirement* $\Delta R < 0.05$*. Anything which is not marked as Signal is automatically tagged as Fake, possibly including muons not coming from the leading* $Z \to \mu\mu$ *simulated event.*

Finally, detailed timing information of the various reconstruction modules is reported in table 4.8. In this case, the data was obtained from a sample of about 7000 simulated $t\bar{t}$ events in Phase-2 conditions. The time reported refers to the total execution time of the full reconstruction sequence responsible for the creation of a specific type of object (e.g. Standalone Muons timing includes both the seeding and the track finding step: from the seed creation until a Standalone Muon track is produced). The "MVA classifier"(*) refers to a module used to prune the list of Muon seeds produced in the inner tracker and is further discussed in the following section. These timing results largely corroborate the previous discussion and show a noticeable improvement in both the Standalone Muon reconstruction and the second L3 tracks to be reconstructed. Muon ID shows the largest improvement in the Inside-Out first reconstruction since, in this case, most of the events require a single pass therefore the resulting merged collection is significantly cleaner than both the current and Outside-In first reconstructions. In comparison, the Outside-In first reconstruction requires both reconstruction passes in most cases, leading to a larger number of fakes being present in the merged collection, mostly coming from the unfiltered L3 IO tracks produced during the second pass. Therefore, the Muon ID processes far fewer tracks in the Inside-Out first case than both in the Outside-In first and the current workflows, with the Outside-In first still requiring less time than the current implementation.

Reconstruction	Sequence	Total execution time (ms)
Current	Standalone Muons (L2)	91.57 ± 0.02
	L3 Tracker Muon tracks (IO)	22.55 ± 0.01
	L3 Global Muon tracks (OI)	19.17 ± 0.01
	MVA classifier *	42.46 ± 0.02
	Muon ID	9.07 ± 0.01
Inside-Out first	Standalone Muons (L2)	51.87 ± 0.02
	L3 Tracker Muon tracks (IO)	23.21 ± 0.01
	L3 Global Muon tracks (OI)	1.80 ± 0.01
	MVA classifier *	101.06 ± 0.02
	Muon ID	6.15 ± 0.01
Outside-In first	Standalone Muons (L2)	52.85 ± 0.02
	L3 Tracker Muon tracks (IO)	19.76 ± 0.01
	L3 Global Muon tracks (OI)	10.26 ± 0.01
	MVA classifier *	87.64 ± 0.02
	Muon ID	8.20 ± 0.01

Table 4.8: *Detailed timing measurements for Muon reconstruction sequences in the current, Inside-Out first, and Outside-In first reconstruction workflows. A sequence includes all modules responsible for the creation of a specific reconstructed object.*

The Muon seeds MVA module

Currently, all Muon reconstruction workflows employ a module designed to filter the muon seeds produced in the pixel detector by using information from the L1 trigger. This module takes advantage of a machine learning approach using Multivariate Analysis (MVA) to assign a score to each seed, and then pick a configurable number of seeds with the highest score, resulting in a filtered collection of high-quality seeds. Although the module's functionality is beyond the original scope of this thesis, an issue with its execution time was noted throughout the work presented. In both the Inside-Out first and Outside-In first approaches the MVA module performs noticeably worse than in the current reconstruction (see table 4.8), even if there is no obvious correlation between the changes implemented and the module itself. To improve the timing of this module, some light code optimization was performed, taking advantage of modern coding patterns and refactoring some of the code to improve its performance. This led to a roughly 30% improvement in the timing of the module itself: bringing the timing of the module, in the current reconstruction, from about 42.5 ms to about 30.1 ms. These metrics were obtained by comparing the performance of the same reconstruction workflow, having modified only the MVA module. This optimisation was however not enough to regain the performance lost by the new reconstruction approaches. Therefore, further investigation is needed to assess the performance of the model used by the MVA module, possibly retraining it on a new dataset containing L1 Tracker Muons information, or rewriting the module from scratch.

4.4 Future work

This work lays the foundation for a larger redesign of the Online Muon reconstruction strategy. Having reduced the redundancy in both the Standalone seeding and Tracker Muon reconstruction, the focus will shift to displaced muons, track finding improvements and computing performance optimisation, possibly looking into taking advantage of heterogenous solutions exploiting GPUs.

Firstly, the slight loss in efficiency in the endcap region measured with the new seeding module should be further investigated. In addition, the performance after the changes might be tested for close-by and displaced muons. The former is to assess the capability of the new module to discern tracks in difficult conditions since having access to a more precise momentum measurement at the seed level should translate into smaller extrapolation windows. Therefore, this would improve the efficiency of the reconstruction for muons produced close one to the other. The latter represents one of the most promising signatures for Beyond the Standard Model physics at colliders. As previously discussed, the new Standalone seeding approach allows tagging prompt and displaced candidates in a single pass. This is not true for the current muon track reconstruction, which must

be executed twice to retain efficiency for both prompt and displaced muons. Currently, after Offline Muon seeds are produced using only information from the muon chambers, they are matched with L1 Tracker Muons producing a collection of seeds for prompt muons in this first pass. Then, a second pass is needed to tag seeds that were not previously matched and might thus be associated with displaced muons. On the other hand, the new module foregoes the matching step, producing a single collection of seeds and can thus be used to produce seeds both for prompt muons by matching a L1 Tracker muon with segments in the muon chambers and for displaced muons by using only the segments in the muon chambers when no suitable L1 Tracker Muon is found. Therefore, one short-term change would be to implement displaced reconstruction using the new seeding approach. This would also require further tuning of the parameters, specifically for the displaced approach incorporating, for example, the improved p_T measurement of the seed.

Another avenue to pursue would be the re-evaluation of the MVA module used to filter Muon seeds produced in the pixel either by re-training the same model taking advantage of the new detector information and considering the changes implemented in the reconstruction, or by rewriting a module with the same purpose from scratch.

Since the L3 reconstruction is the most computationally demanding task of the Muon HLT, solutions exploiting heterogeneous computing might be investigated. Similarly to what has been done with the pixel and tracker reconstruction [45], this approach could significantly improve timing and efficiency while reducing computing resources usage. In this case, the reconstruction approach would be to borrow from the heterogeneous inner tracker reconstruction whenever possible, introducing new modules with original algorithms where necessary.

Finally, the validation of the new reconstruction workflows will be performed with the intent of fully integrating into the CMS reconstruction software. The analysis of the integration, considering different physics topologies, will result in a physics-driven decision on what and how to optimize next.

Conclusions

The work presented in this thesis aims to revisit the Online High-Level Trigger (HLT) Muon reconstruction at the CMS experiment. This is made necessary by the upcoming HL-LHC upgrade, which would push the current reconstruction technical performance beyond its limits. Therefore, there is a need to optimise the reconstruction, reducing its computing resources usage while maintaining the remarkable physics performance demonstrated up to the present LHC run. The main focus of this work is thus split between Standalone Muon reconstruction (L2) and Tracker/Global Muon reconstruction (L3 Inside-Out and L3 Outside-In, respectively).

As far as the Standalone Muon reconstruction is concerned, the algorithms responsible for the production of L2 Muon seeds (i.e. the starting states for the Standalone track finding) was entirely rewritten, taking advantage of new information coming from the upgraded hardware trigger. This allows the new module to produce roughly a factor of 5 fewer seeds with respect to the current Offline seeds produced using only information from the muon system, by matching the hardware trigger information with hits and segments in the muon chambers. This results in about a 42% better timing performance while maintaining compatibility within a $< 0.5\%$ difference for efficiency in the barrel and overlap regions, and $<3\%$ difference for the endcap, and slightly reducing the fake rate.

The Tracker/Global Muon reconstruction optimisation is mainly aimed at reducing the current reconstruction redundancy. In fact, at present, all muon candidates are reconstructed twice: once starting from the tracker information around a L1 Tracker Muon identified by the Level-1 hardware trigger and matching segments in the muon chambers (Inside-Out), and once starting from a L2 Standalone Muon and building a track inwards until it is matched with information in the pixel detector (Outside-In). This allows the reconstruction to maintain optimal efficiency in the entire η acceptance region and a wide momentum range. However, this reconstruction is also the most expensive from the computational point of view. The new Muon reconstruction allows to choose which algorithm to execute first, Inside-Out or Outside-In, with the second one to be executed only looking for candidates that were not reconstructed during the first pass or did not meet specific quality criteria. This results in a substantial decrease of the fake rate for most of the intermediate reconstructed objects while the efficiency over the

full η coverage and at low p_T is preserved. In particular, the filter implemented to assess the quality of the tracks produced during the first pass reduces the number of fake tracks by more than 30%, decreasing the computational load on the HLT Muon identification module, responsible for creating the final reconstructed HLT Muon candidates. In the end, the final reconstructed objects in the Inside-Out (Outside-In) first approach achieve an efficiency of 95.6 (96.0)%, 95 (95)%, 93.0 (93.6)% in the barrel, overlap, and endcap regions, respectively, showing excellent compatibility with differences at $< 1\%$ level, closely resembling their compatibility with the current implementation.

Although the physics results presented are quite comprehensive for the limited samples analysed, the reconstruction of close-by muons (i.e. pairs of muons produced extremely close one to the other) has not been investigated and should be addressed in further developments, together with the physics performance on a more varied event topology.

Moreover, the new seeding module for Standalone Muons was designed to tag both prompt and displaced candidates in a single pass, but no specific workflow has been implemented to take advantage of this and no study has been carried out to assess the performance of the new seeding module for the reconstruction of displaced muons produced in the decays of long-lived, beyond the Standard Model, neutral particles. This is another point to be addressed in the future since displaced muons are one of the prime candidates for beyond the Standard Model physics at colliders.

Finally, the Tracker and Global Muon reconstruction remains the most computationally intensive task of the muon reconstruction chain. Having addressed the current redundancy, the natural next step would be to investigate heterogeneous solutions to further increase computational performance by taking advantage of GPUs. This would require a major rethinking of the reconstruction to make it parallel by design, with a focus on data structures and would thus represent a longer-term task.

Appendix A

Muon Triggering and Reconstruction Glossary

In the previous chapters, an explanation of the terminology commonly used when dealing with Muons was made when necessary. This Appendix offers a comprehensive list of the terms used to facilitate the reading of this thesis:

- **General terms**
  - LHC: Large Hadron Collider;
  - HL-LHC: High-Luminosity Large Hadron Collider. Substantial upgrade of the accelerator at CERN, targeting a luminosity increase of up to a factor 7.5;
  - MB: Muon Barrel. Ensemble of muon detectors in the $|\eta| < 1.2$ region of the muon chambers;
  - ME: Muon Endcap. Ensemble of muon detectors in the $\eta > 0.9$ region of the muon chambers;
  - Muon ID: Muon identification module. Responsible for creating the final muons used in the High-Level Trigger by merging all intermediate reconstructed objects, performing quality cuts, and attaching identification variables to all candidates.
- **Muon detectors**
  - DT: Drift Tube. Detector used in the barrel region of CMS;
  - CSC: Cathode Strip Chamber. Used in the endcap region of CMS;
  - RPC: Resistive Plate Chamber. Used in both the barrel and endcap regions of CMS;

  - GEM: Gas Electron Multiplier. Detector used in the forward region of the endcap, currently testing a single detector with an addition planned for the Phase-2 upgrade;
  - iRPC: improved Resistive Plate Chamber. Detector to be installed in the endcap region of CMS for the Phase-2 upgrade;

- **High-Level Trigger intermediate reconstructed muon objects**
  - L1 Standalone Muon: hardware trigger tracks built using only the trigger primitives in the muon chambers;
  - L1 Tracker Muon: hardware trigger track built by matching a Tracker trigger track with one or more trigger primitives in the muon chambers;
  - L2 Standalone Muon: track built using only information from the muon chambers.
  - L3 Tracker Muon (L3 IO track): muon object identified combining tracker tracks with one or more DT or CSC segments. A track reconstructed inside-out, using only information from the inner tracker is referred to as L3 IO track;
  - L3 Global (or combined) Muon (L3 OI tracks): muon object built by matching a L2 Standalone Muon with a tracker track reconstructed propagating from the outer tracker towards the pixel detector. This object contains a tracker track referred to as L3 OI track;
  - HLT Muon: the final type of reconstructed object produced by the HLT after reconstruction and identification combining the two kinds of L3 tracks and adding the Muon ID variables.

Appendix B

Code and data availability

All codes discussed, developed, and implemented during the thesis work are available in the following repositories on GitHub:

- **Changes to L2 Standalone Muon seeding**:
  GitHub repository forked from CMS-SW/CMSSW
  https://github.com/Parsifal-2045/cmssw/tree/L2Seeder;
- **Implementation of flexible module to execute L3 Muon reconstruction Inside-Out or Outside-In first**:
  GitHub repository forked from CMS-SW/CMSSW
  https://github.com/Parsifal-2045/cmssw/tree/L3_selector;
- **All changes implemented into the CMS reconstruction software during this thesis work**:
  GitHub repository forked from CMS-SW/CMSSW
  https://github.com/Parsifal-2045/cmssw/tree/Master_thesis
- **Analysis and plotting software**:
  GitHub repository clone Parsifal-2045/muon_analyzer
  https://github.com/Parsifal-2045/muon_analyzer

All the plots and data used to produce them are available (CERN login required): https://lferragi.web.cern.ch/plots/muon_hlt_phase2/thesis_plots/

Bibliography

[1] Stephen Myers and Emilio Picasso. "The LEP Collider". In: *Scientific American* 263.1 (1990), pp. 54–61. ISSN: 00368733, 19467087. URL: http://www.jstor.org/stable/24996863 (visited on 08/13/2024).

[2] *ATLAS: technical proposal for a general-purpose pp experiment at the Large Hadron Collider at CERN*. LHC technical proposal. Geneva: CERN, 1994. DOI: 10.17181/CERN.NR4P.BG9K. URL: https://cds.cern.ch/record/290968.

[3] CMS Collaboration. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08004.

[4] *ALICE: Technical proposal for a Large Ion collider Experiment at the CERN LHC*. LHC technical proposal. Geneva: CERN, 1995. URL: https://cds.cern.ch/record/293391.

[5] *LHCb : Technical Proposal*. Geneva: CERN, 1998. URL: https://cds.cern.ch/record/622031.

[6] Ewa Lopienska. "The CERN accelerator complex, layout in 2022. Complexe des accélérateurs du CERN en janvier 2022". In: (2022). General Photo. URL: https://cds.cern.ch/record/2800984.

[7] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 30–61. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.021. URL: https://www.sciencedirect.com/science/article/pii/S0370269312008581.

[8] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 1–29. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.020. URL: https://www.sciencedirect.com/science/article/pii/S037026931200857X.

[9] CERN Engineering. *Pulling together: Superconducting electromagnets*. URL: https://home.web.cern.ch/science/engineering/pulling-together-superconducting-electromagnets.

[10] CERN Engineering. *Accelerating: Radiofrequency cavities*. URL: https://home.web.cern.ch/science/engineering/accelerating-radiofrequency-cavities.

[11] *The ALICE Experiment*. URL: https://home.cern/science/experiments/alice.

[12] *The LHCb Experiment*. URL: https://home.cern/science/experiments/lhcb.

[13] *The ATLAS Experiment*. URL: https://home.cern/science/experiments/atlas.

[14] *The CMS Experiment*. URL: https://home.cern/science/experiments/cms.

[15] The LHCf Collaboration et al. "The LHCf detector at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08006. DOI: 10.1088/1748-0221/3/08/S08006. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08006.

[16] *The LHCf Experiment*. URL: https://home.cern/science/experiments/lhcf.

[17] Cristiano Alpigiani et al. "A Letter of Intent for MATHUSLA: A Dedicated Displaced Vertex Detector above ATLAS or CMS." In: (July 2018). arXiv: 1811.00927 [physics.ins-det].

[18] *The MATHUSLA Experiment*. URL: https://mathusla-experiment.web.cern.ch/.

[19] Jae Hyeok Yoo. "The milliQan Experiment: Search for milli-charged Particles at the LHC". In: *PoS* ICHEP2018 (2019). proceeding for ICHEP 2018 SEOUL, International Conference on High Energy Physics, 4-11 July 2018, SEOUL, KOREA, p. 520. DOI: 10.22323/1.340.0520. arXiv: 1810.06733. URL: https://cds.cern.ch/record/2645863.

[20] James Pinfold et al. *Technical Design Report of the MoEDAL Experiment*. Tech. rep. CERN, 2009. URL: https://cds.cern.ch/record/1181486.

[21] *The MoEDAL-MAPP Experiment*. URL: https://www.home.cern/science/experiments/moedal-mapp.

[22] G. Anelli et al. "The TOTEM experiment at the CERN Large Hadron Collider". In: *JINST* 3 (2008), S08007. DOI: 10.1088/1748-0221/3/08/S08007.

[23] *The TOTEM Experiment*. URL: https://home.cern/science/experiments/totem.

[24] Henso Abreu et al. "The FASER detector". In: *Journal of Instrumentation* 19.05 (May 2024), P05066. DOI: 10.1088/1748-0221/19/05/P05066. URL: https://dx.doi.org/10.1088/1748-0221/19/05/P05066.

[25] *The FASER Experiment*. URL: https://home.cern/science/experiments/faser.

[26] G. Arduini et al. "High Luminosity LHC: challenges and plans". In: *Journal of Instrumentation* 11.12 (Dec. 2016), p. C12081. DOI: 10.1088/1748-0221/11/12/C12081. URL: https://dx.doi.org/10.1088/1748-0221/11/12/C12081.

[27] CMS Collaboration. "The CMS trigger system". In: *Journal of Instrumentation* 12.01 (Jan. 2017), P01020. DOI: 10.1088/1748-0221/12/01/P01020. URL: https://dx.doi.org/10.1088/1748-0221/12/01/P01020.

[28] CMS Collaboration. "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of Instrumentation* 15.10 (Oct. 2020), P10017. DOI: 10.1088/1748-0221/15/10/P10017. URL: https://dx.doi.org/10.1088/1748-0221/15/10/P10017.

[29] CMS Collaboration. "The CMS Phase-1 pixel detector upgrade". In: *Journal of Instrumentation* 16.02 (Feb. 2021), P02027. DOI: 10.1088/1748-0221/16/02/P02027. URL: https://dx.doi.org/10.1088/1748-0221/16/02/P02027.

[30] CMS Collaboration. "Silicon Strip Tracker Performance results 2018". In: (2018). URL: https://cds.cern.ch/record/2638062.

[31] R. BENETTA et al. "THE CMS ECAL READOUT ARCHITECTURE AND THE CLOCK AND CONTROL SYSTEM". In: *Calorimetry in Particle Physics*, pp. 162–172. DOI: 10.1142/9789812701978_0021. eprint: https://www.worldscientific.com/doi/pdf/10.1142/9789812701978_0021. URL: https://www.worldscientific.com/doi/abs/10.1142/9789812701978_0021.

[32] CMS Collaboration. "Construction and test of the final CMS Barrel Drift Tube Muon Chamber prototype". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 480.2 (2002), pp. 658–669. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(01)01227-X. URL: https://www.sciencedirect.com/science/article/pii/S016890020101227X.

[33] CMS Collaboration. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}$=13 TeV". In: *Journal of Instrumentation* 13.06 (June 2018), P06015. DOI: 10.1088/1748-0221/13/06/P06015. URL: https://dx.doi.org/10.1088/1748-0221/13/06/P06015.

[34] CMS Collaboration. "Development of the CMS detector for the CERN LHC Run 3". In: *Journal of Instrumentation* 19.05 (May 2024), P05064. DOI: 10.1088/1748-0221/19/05/P05064. URL: https://dx.doi.org/10.1088/1748-0221/19/05/P05064.

[35] CMS Collaboration. *Technical Proposal for the Phase-II Upgrade of the CMS Detector*. Tech. rep. CERN, June 2015. DOI: 10.17181/CERN.VU8I.D59J.

[36] CMS Collaboration. *A MIP Timing Detector for the CMS Phase-2 Upgrade*. Tech. rep. Geneva: CERN, 2019. URL: https://cds.cern.ch/record/2667167.

[37] CMS Collaboration. *The Phase-2 Upgrade of the CMS Endcap Calorimeter*. Tech. rep. Geneva: CERN, 2017. DOI: 10.17181/CERN.IV8M.1JY2. URL: https://cds.cern.ch/record/2293646.

[38] Thomas Hebbeker, Andrey Korytov, and CMS Collaboration. *The Phase-2 Upgrade of the CMS Muon Detectors*. Tech. rep. CERN, Sept. 2017.

[39] CMS Collaboration. *The Phase-2 Upgrade of the CMS Level-1 Trigger*. Tech. rep. Final version. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2714892.

[40] The CMS Collaboration. *The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger*. Tech. rep. This is the final version of the document, approved by the LHCC. Geneva: CERN, 2021. URL: https://cds.cern.ch/record/2759072.

[41] CMS Collaboration. "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV". In: *Journal of Instrumentation* 7.10 (Oct. 2012), P10002. DOI: 10.1088/1748-0221/7/10/P10002. URL: https://dx.doi.org/10.1088/1748-0221/7/10/P10002.

[42] CMS Collaboration. *Performance of the CMS TwinMux Algorithm in late 2016 pp collision runs*. Tech. rep. CERN, 2016. URL: https://cds.cern.ch/record/2239285.

[43] R. Frühwirth. "Application of Kalman filtering to track and vertex fitting". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 262.2 (1987), pp. 444–450. ISSN: 0168-9002. DOI: https://doi.org/10.1016/0168-9002(87)90887-4. URL: https://www.sciencedirect.com/science/article/pii/0168900287908874.

[44] CMS Collaboration. "Search for long-lived particles decaying to final states with a pair of muons in proton-proton collisions at $\sqrt{s} = 13.6$ TeV". In: *Journal of High Energy Physics* 47 (2024). ISSN: 1029-8479. DOI: 10.1007/JHEP05(2024)047. URL: https://doi.org/10.1007/JHEP05(2024)047.

[45] A. Bocci et al. "Heterogeneous Reconstruction of Tracks and Primary Vertices With the CMS Pixel Tracker". In: *Frontiers in Big Data* 3 (2020). ISSN: 2624-909X. DOI: 10.3389/fdata.2020.601728. URL: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2020.601728.

[46] Andrea Bocci. *CMS HLT Run-3 timing reference*. URL: https://fwyzard.web.cern.ch/circles/web/piechart.php?local=false&dataset=Run3_HLT_14.0.2%2FCPU%20only%20-%20reference&resource=time_real&colours=default&groups=hlt&show_labels=true&threshold=0.