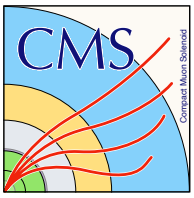
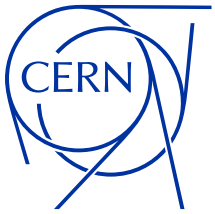




DISSERTATION

Firmware and emulator development of cut-based filter algorithms for the CMS Level-1 Global Trigger at the High-Luminosity LHC

zur Erlangung des akademischen Grades

Dr. der techn. Wissenschaften

im Rahmen des Studiums

Technische Physik

eingereicht von

Dipl.-Ing. Benjamin Huber BSc
Matrikelnummer: 01326034

ausgeführt am CERN mit Unterstützung
der Fakultät für Physik der Technischen Universität Wien

Betreuung
Betreuer: Privatdoz. Mag.phil. Mag. Dr.rer.nat. Manfred Jeitler
Mitwirkung: Dipl.-Ing. Dr.techn. Hannes Sakulin

Wien, 07.09.2025

(Unterschrift Verfasser/in) (Unterschrift Betreuer/in)

Abstract

The Large Hadron Collider (LHC) has already achieved significant milestones, most notably the discovery of the Higgs boson. Its journey continues with the upcoming High-Luminosity LHC upgrade, which will drastically enhance its luminosity and, consequently, its statistical reach. This upgrade poses considerable challenges to detector systems, particularly the CMS Level-1 Trigger, which must be redesigned to handle the increased data rate. The FPGA-based Level-1 Trigger is responsible for filtering events, thereby reducing the 40 MHz bunch crossing rate to a manageable 750 kHz for the CMS data acquisition system. The final element in this new Level-1 Trigger chain, which ultimately decides whether an event is to be completely read out, is the Global Trigger. Its readout decision is made by applying complex constraints to a wide range of reconstructed trigger objects, leveraging for the first time data from the silicon tracker and the new High-Granularity Calorimeter. This thesis presents a crucial element of the new Level-1 Global Trigger: the implementation of cut-based algorithms in both FPGA firmware and the corresponding emulation software. While machine learning techniques such as neural networks are gaining traction in event filtering, cut-based algorithms remain the robust backbone of the Global Trigger. The presented cut-based algorithm implementation is capable of applying constraints to individual reconstructed objects as well as topological correlations — such as invariant mass and angular distance — all while remaining within the strict latency budget of 1 μs. It reliably handles the 40 MHz bunch crossing rate with a compact resource footprint, fitting currently up to 1,000 algorithms into just 3 VU13P FPGAs, and retains full flexibility to constrain and correlate any combination of trigger objects. Both software and firmware components were extensively validated, demonstrating excellent agreement among each other and with the expectations set by extrapolating from the current LHC Run 3 system. Unavoidable errors in the calculations performed on FPGA fabric were investigated and found to be negligible when compared with the intrinsic resolution of the CMS detector. Beyond the technical implementation, this work addresses the operational challenge of managing up to 1,000 cut-based algorithms during High-Luminosity operation. An automated procedure is proposed to tailor cut-based algorithms to specific physics signatures. The procedure minimises an achievement scalarizing function combining the two competing objectives of maximising trigger efficiency and minimising trigger rate according to some selected trade-off preference. The effectiveness of this approach is illustrated through case studies targeting a B meson decay to two muons, a Higgs boson decay to two tau leptons, and a tau lepton decay to three muons.

Kurzfassung

Der Large Hadron Collider (LHC) hat bereits bedeutende Meilensteine erreicht, insbesondere die Entdeckung des Higgs-Bosons. Mit dem bevorstehenden Upgrade zum High-Luminosity LHC wird seine Reise fortgesetzt, wobei die Luminosität und damit die statistische Reichweite erheblich gesteigert wird. Dieses Upgrade stellt die Detektorsysteme vor große Herausforderungen, speziell den Level-1-Trigger des CMS-Experiments, der neu gestaltet werden muss, um die erhöhte Datenrate zu bewältigen. Der FPGA-basierte Level-1-Trigger ist dafür verantwortlich, Ereignisse zu filtern und so die Bunch-Crossing-Rate von 40 MHz auf eine für das CMS-Datenerfassungssystem handhabbare Rate von 750 kHz zu reduzieren. Das letzte Glied in dieser neuen Level-1-Trigger-Kette, das letztlich entscheidet, ob ein Ereignis vollständig ausgelesen wird, ist der Global Trigger. Seine Entscheidungsfindung basiert auf der Anwendung komplexer Bedingungen auf eine Vielzahl rekonstruierter Triggerobjekte, wobei erstmals Daten sowohl vom Silizium-Tracker als auch vom neuen High-Granularity-Kalorimeter genutzt werden. Diese Arbeit stellt ein zentrales Element des neuen Level-1 Global Triggers vor: die Implementierung schnittbasierter (cut-based) Algorithmen sowohl in der FPGA-Firmware als auch in der zugehörigen Emulationssoftware. Obwohl maschinelle Lernverfahren wie neuronale Netze zunehmend an Bedeutung bei der Ereignisfilterung gewinnen, bleiben schnittbasierte Algorithmen das robuste Rückgrat des Global Triggers. Die vorgestellte Implementierung dieser Algorithmen ist in der Lage, sowohl Bedingungen auf einzelne rekonstruierte Objekte als auch auf topologische Korrelationen — wie invariante Masse und Winkelabstand — anzuwenden, und das unter Einhaltung des strengen Latenzbudgets von 1 μs. Sie ermöglicht die zuverlässig Verarbeitung mit der Bunch-Crossing-Rate von 40 MHz bei kompaktem Ressourcenbedarf, der aktuell die Implementierung von bis zu 1.000 Algorithmen auf nur drei VU13P-FPGAs erlaubt. Dabei bleibt die volle Flexibilität erhalten, beliebige Kombinationen von Triggerobjekten zu beschränken und zu korrelieren. Sowohl die Software- als auch die Firmware-Komponenten wurden umfassend validiert und zeigen eine hervorragende Übereinstimmung miteinander sowie mit den aus dem aktuellen LHC-Betrieb (Run 3) extrapolierten Erwartungen. Unvermeidbare Rechenfehler durch die FPGA-Hardware wurden untersucht und als im Vergleich zur intrinsischen Auflösung des CMS-Detektors vernachlässigbar eingestuft. Über die technische Umsetzung hinaus behandelt diese Arbeit auch die operationelle Herausforderung, während des HL-LHC-Betriebs bis zu 1.000 schnittbasierte Algorithmen zu verwalten. Dazu wurde ein automatisiertes Verfahren vorgeschlagen, das diese Algorithmen gezielt auf spezifische physikalische Signaturen optimiert. Das Verfahren minimiert eine Ziel-Funktion, die die beiden konkurrierenden Ziele, Maximierung der Triggereffizienz und Minimierung der Triggerrate, gemäß einer gewählten Priorisierung vereint. Die Wirksamkeit des Ansatzes wird anhand von Fallstudien demonstriert, die den Zerfall eines B-Mesons in zwei Myonen, den Zerfall eines Higgs-Bosons in zwei Tau-Leptonen sowie den Zerfall eines Tau-Leptons in drei Myonen adressieren.

Contents

1 Introduction

> *We must not believe those, who today, with philosophical bearing and deliberative tone, prophesy the fall of culture and accept the ignorabimus. For us there is no ignorabimus, and in my opinion none whatever in natural science. In opposition to the foolish ignorabimus our slogan shall be:*
>
> ***We must know — we will know!***
>
> David Hilbert

While the Standard Model has been remarkably successful in describing the known elementary particles and their behaviours, it is far from complete. A range of unanswered questions points to deeper truths about the nature of reality, many of which remain elusive despite decades of research.

One of the most profound mysteries is the nature of dark matter. Astronomical and cosmological observations — including galactic rotation curves, gravitational lensing in colliding galaxy clusters, and anisotropies in the cosmic microwave background — provide compelling evidence that approximately 85% of the matter in the universe is non-luminous and non-baryonic [1, 2]. Yet, the particle nature of dark matter remains unknown. Particle colliders such as the Large Hadron Collider (LHC) offer a unique opportunity to produce potential dark-matter candidates under controlled conditions or to constrain their interactions through missing energy signatures and other indirect effects.

A second key puzzle is the observed asymmetry between matter and antimatter. The idea that antimatter could be sequestered in entire galaxies or clusters is unlikely, as we would expect intense gamma radiation from annihilations at the boundaries between matter and antimatter regions. However, the Standard Model and General Relativity treat matter and antimatter almost identically. Explaining how a small excess of matter emerged and persisted through the early universe's evolution is a question that cannot be answered within the current framework.

Big Bang nucleosynthesis calculations and cosmic microwave background observations constrain the baryon-to-entropy ratio today to be [3]

$$\eta_B \equiv \frac{n_b - n_{\bar{b}}}{n_\gamma} \approx 6 \times 10^{-10} \quad , \tag{1.1}$$

where n_b, $n_{\bar{b}}$ and n_γ are the baryon, anti-baryon, and photon number densities, respectively. This tiny, yet crucial, imbalance is far greater than would be expected if the early universe had been perfectly symmetric.

In 1967, Sakharov [4, 5] identified three necessary conditions, known as the Sakharov conditions, for generating this asymmetry dynamically via a mechanism called baryogenesis:

1. Baryon number violation.

2. Violation of charge conjugation symmetry (C) and charge-parity symmetry (CP).
3. Departure from thermal equilibrium.

Violation of baryon number allows the net baryon count to change, while C and CP violation are required to prevent the production of baryons and anti-baryons in equal amounts. The third condition, departure from thermal equilibrium, requires that the expansion rate of the universe outpace the rate of the asymmetry-generating interactions. Otherwise, the interactions would halt once a thermal equilibrium due to pair-annihilation processes is reached.

Although the Standard Model includes all three ingredients in principle, it fails to meet the required conditions in practice. The amount of CP violation it provides is too small, and the electroweak phase transition is a smooth crossover rather than the strongly first-order transition needed for sufficient electroweak baryogenesis.

The discovery of the Higgs boson in 2012 resolved some questions regarding the mechanism of mass generation, but it also raised new ones. When the Higgs field settles into its vacuum expectation value at roughly 250 GeV, it gives the W and Z bosons their masses of order 100 GeV. However, this electroweak scale is vastly smaller than the Planck scale ($\sim 10^{19}$ GeV), where gravitational interactions become strong. Why these scales are separated by 17 orders of magnitude is a central question, known as the hierarchy problem [6].

This disparity is not just aesthetically troubling. In quantum field theory, particles coupled to the Higgs field contribute quantum loop corrections that naturally push the Higgs mass toward the Planck scale. Without an extraordinary fine-tuning of parameters, the Higgs mass would not remain at its observed value of $\sim$125 GeV. This apparent need for delicate cancellations suggests the existence of a mechanism that stabilises the Higgs mass.

Several theoretical frameworks propose solutions. Supersymmetry (SUSY) introduces a superpartner for every known particle, whose contributions to Higgs mass corrections cancel those of the Standard Model particles. Extra-dimensional models (such as those predicting compactified spatial dimensions) dilute gravity's apparent weakness by allowing it to propagate in more dimensions than the Standard Model fields do. Composite Higgs theories postulate that the Higgs is not fundamental but a bound state of more basic constituents, with its mass stabilised by the strong dynamics binding them. Each scenario predicts new particles or phenomena that could be detected or constrained at the LHC [7].

Beyond the hierarchy problem, Higgs also raises questions about the stability of the vacuum. Specifically, is the vacuum state of our universe truly the lowest-energy configuration, or just a metastable state destined to decay? At tree level, the Standard Model Higgs potential

$$V(\Phi) = \mu^2 \Phi^* \Phi + \lambda (\Phi \Phi^*)^2 \tag{1.2}$$

has a minimum at $|\Phi| = v/\sqrt{2} \approx 174$ GeV. However, quantum corrections, particularly from the heavy top quark, cause the quartic coupling λ to run with energy [8]. At high

scales, λ may turn negative, introducing a new lower-energy minimum in the potential.

If λ remains positive up to the Planck scale, the vacuum is stable. If it becomes negative at some intermediate energy, our vacuum is metastable, with a finite lifetime governed by quantum tunnelling. Current measurements of the Higgs and top quark masses place us near the critical boundary between stability and metastability, an intriguing and potentially precarious position.

Precision measurements of Higgs and top quark properties, along with searches for new particles that could alter the running of λ, are crucial to determining the true nature of our vacuum.

Another major unresolved issue is the accelerated expansion of our universe, attributed to dark energy — often referred to as the cosmological constant. This constant acts as a uniform, repulsive pressure with an incredibly small energy density by particle physics standards (roughly 7×10^{-30} g/cm^3). However, naïve estimates of the vacuum energy in quantum field theory exceed the observed value by 46 to 120 orders of magnitude, depending on the chosen energy cutoff: 46 orders correspond to the proton mass scale, while 120 orders correspond to the Planck scale [9, 10]. The enormous mismatch is known as the cosmological constant problem.

Although dark energy is usually discussed in astrophysical or gravitational terms, particle colliders like the LHC can play an important, if indirect, role in probing ideas for its origin, such as:

- **Vacuum-energy contributions from new particles:** All quantum fields contribute to the vacuum energy. New particles — if they exist — could shift this value. Searching for such particles at the LHC helps constrain models that aim to address the cosmological constant.
- **Extra dimensions and modified gravity:** Some models propose that gravity — and consequently vacuum energy — can propagate into extra spatial dimensions, effectively diluting its apparent strength in our four-dimensional spacetime. These scenarios often predict Kaluza-Klein excitations (quantised modes arising from momentum in the compactified extra dimensions) which could be probed in high-energy collisions through their characteristic decay products [11].
- **Symmetry-based cancellations:** Certain symmetries might enforce a near-zero vacuum energy. For instance, unbroken supersymmetry would cancel contributions from bosons and fermions exactly. Although SUSY has not been observed, the LHC continues to test its viability, narrowing the parameter space where such cancellations could occur.

Despite the LHC's profound achievements, the questions that remain highlight the need for deeper exploration of the universe's fundamental structure. Some answers may lie just beyond our current experimental reach and could emerge with the enhanced statistical sensitivity provided by the LHC's upcoming High-Luminosity phase.

2 The Large Hadron Collider (LHC)

2.1 Introduction

The Large Hadron Collider (LHC) project was conceived in the early 1980s as a next-generation particle accelerator to replace the, by then, not yet operational Large Electron-Positron Collider (LEP). Construction was approved in December 1994 [12] and carried out by an international collaboration involving over 10,000 scientists and engineers from more than 100 countries. The LHC was completed in 2008, with the first beam established in September of the same year. As of today, it is the world's largest and most powerful particle accelerator, located at CERN (the European Organisation for Nuclear Research) near Geneva, Switzerland. The LHC is housed in a circular tunnel with a circumference of 26.7 kilometres, buried underground at a depth ranging from 50 to 175 metres. The tunnel, originally constructed for the LEP, was repurposed and upgraded to accommodate the LHC. The colossal machine was designed to accelerate protons (and occasionally heavy ions) and collide them at centre of mass energies of up to 14 TeV to study the resulting interactions.

The physics goals of the LHC included the finding or exclusion of the Higgs boson. Three clues existed from LEP and the Tevatron, a proton-antiproton collider at the Fermi National Accelerator Laboratory (Fermilab), that promised conclusive results with the energies and luminosities of the LHC [13]. They were:

- Indirect searches via precision measurements of the mass ratio of the W and Z boson, which is influenced by the Higgs mass through loop corrections. This study yielded $m_H = 94^{+29}_{-24}$ GeV.
- Direct searches of Higgs decays at LEP yielded a lower bound of $m_H > 114$ GeV.
- Direct searches of Higgs decays at Tevatron excluded the region 147 GeV $< m_H <$ 179 GeV.

Beyond these experimental clues, requiring a finite Higgs coupling strength up to the reduced Planck scale ($\kappa_P = 2.4 \times 10^{18}$ GeV) gives a theoretical upper limit of $m_H < 147$ GeV [14]. The Higgs boson was finally found in 2012 at the LHC with a mass of approximately 125 GeV, a clear excess was seen in the decay channels $H \to ZZ^* \to l^+l^-l^+l^-$ and $H \to \gamma\gamma$ [15, 16].

In order to reach the higher energies required for the discovery, just upgrading LEP with better magnets was not a viable path forward for CERN. The dominant limitation came from synchrotron radiation losses, which scale as $\propto \gamma^4$. For light, highly relativistic particles such as electrons and positrons, these losses grow so rapidly with energy that the accelerating radio frequency (RF) cavities of LEP, already operating at their design limits, could no longer replenish the energy lost per turn. This favoured the facilitation of heavier stable hadrons. The lightest such particle, the proton, is almost 2,000 times heavier than the electrons and positrons employed at LEP. Consequently, as the Lorentz factor of protons, given by

$$\gamma = \frac{E}{mc^2} \quad , \tag{2.1}$$

is 2,000 times smaller than that of electrons and positrons, resulting in synchrotron radiation about 13 orders of magnitude smaller than that of an electron or positron beam of similar energy. Nevertheless, using hadrons rather than leptons comes with additional challenges for the detector systems, which will be discussed in Section 3.2.

The choice of a collider was motivated by the much higher collision energies when compared to a fixed target setup. A fixed target setup, with the four-momentum of the beam $(E, p_x, p_y, p_z)^T$ and fixed target $(m, 0, 0, 0)^T$, can reach collision energies

$$\sqrt{s} = \sqrt{\begin{pmatrix} E+m \\ p_x \\ p_y \\ p_z \end{pmatrix} \cdot \begin{pmatrix} E+m \\ -p_x \\ -p_y \\ -p_z \end{pmatrix}} = \sqrt{E^2 + 2Em + m^2 - p^2} \quad . \tag{2.2}$$

With $p^2 = E^2 - m^2$ this can be written as

$$\sqrt{s} = \sqrt{2Em + 2m^2} \simeq \sqrt{2Em} \quad . \tag{2.3}$$

A collider experiment, with two beams in opposite direction $(E, p_x, p_y, p_z)^T$ and $(E, -p_x, -p_y, -p_z)^T$, on the other hand can reach

$$\sqrt{s} = \sqrt{\begin{pmatrix} E+E \\ p_x - p_x \\ p_y - p_y \\ p_z - p_z \end{pmatrix} \cdot \begin{pmatrix} E+E \\ -(p_x - p_x) \\ -(p_y - p_y) \\ -(p_z - p_z) \end{pmatrix}} = \sqrt{4E^2} = 2E \quad . \tag{2.4}$$

Hence, while in collider experiments the centre-of-mass energy $\sqrt{s}$ increases linearly with the beam energy, in a fixed-target setup it only grows with $\sqrt{E}$. Achieving comparable values of $\sqrt{s}$ in a fixed-target setup would therefore require a far more powerful accelerator for comparable collision energies $\sqrt{s}$.

2.2 Accelerator complex

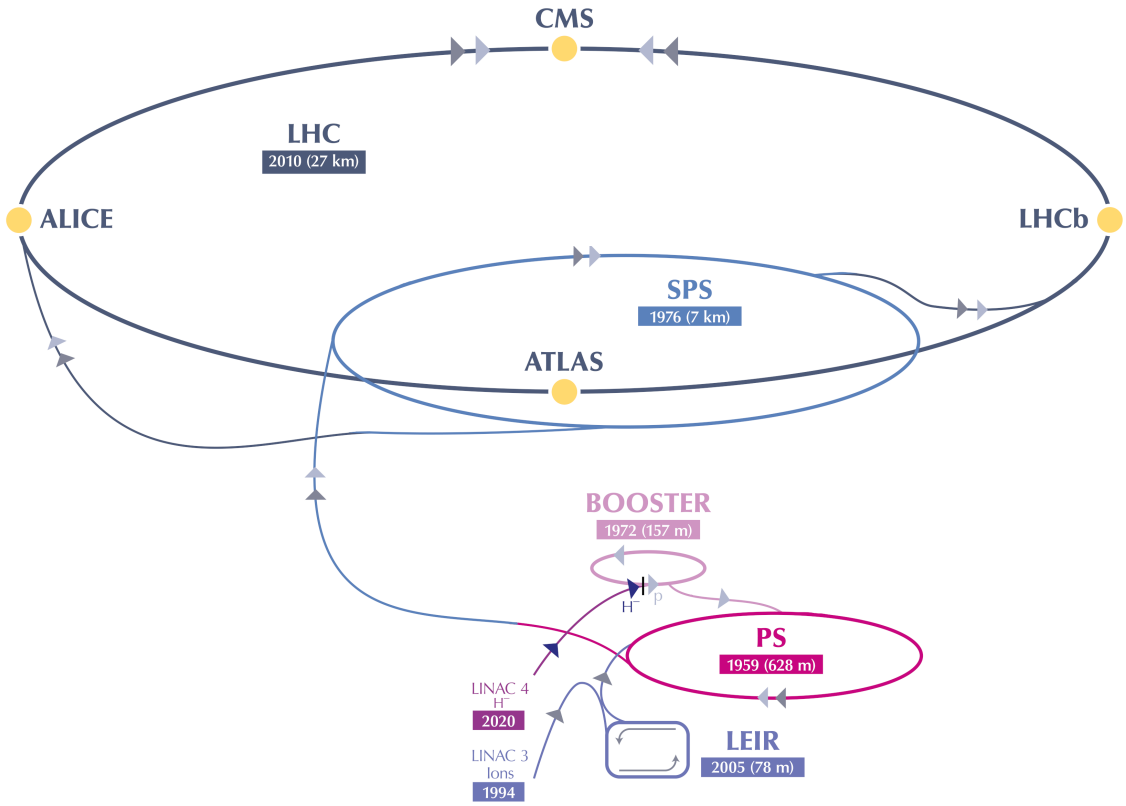


Figure 1: CERN's accelerator complex [17].

The whole accelerator complex at CERN (Fig. 1) consists of a series of machines that incrementally boost the energy of the particle beam [18]. Each machine increases the beam's energy before passing it on to the next one in the sequence. The final stage of this process is the LHC, where particle beams are accelerated to a record energy of up to 7 TeV per beam.

In 2020, Linear Accelerator 4 (Linac4) became the source of proton beams for the CERN accelerator complex. Linac4 accelerates negative hydrogen ions to 160 MeV. These ions are then injected into the Proton Synchrotron Booster (PSB), where their electrons are stripped away, leaving only protons. The PSB further accelerates the protons to 2 GeV before they are transferred to the Proton Synchrotron (PS), which increases their energy to 26 GeV. The protons then move to the Super Proton Synchrotron (SPS), where they are accelerated to 450 GeV.

Finally, the protons are injected into the two beam pipes of the LHC, with one beam circulating clockwise and the other counterclockwise. In the ideal case, filling a single LHC ring takes 4 minutes and 20 seconds, or 8 minutes and 40 seconds for both rings combined. In practice, however, injector setup, interleaving of the two beams, and pauses for operational checks extend the injection process to an average of about 90 minutes [19]. Once the injection is complete, the protons require an additional 20 minutes to be accelerated to their peak energy of up to 7 TeV.

Under normal operating conditions, the beams circulate for many hours inside the LHC beam pipes. The beams are made to collide inside four detectors — A Large Ion Collider Experiment (ALICE), A Toroidal LHC Apparatus (ATLAS), Compact Muon Solenoid (CMS), and LHC-beauty (LHCb) — where the total collision energy reaches up to 14 TeV.

In addition to protons, the LHC also collides heavy ions, most commonly lead. These ions are produced in a plasma source at Linac3, then linearly accelerated before being collected and further accelerated in the Low Energy Ion Ring (LEIR). From there, they follow the same path as the protons to reach their maximum energy.

2.3 Magnets

The LHC operates as a synchrotron, requiring each increase in beam energy to be accompanied by an increase in magnetic field in order to keep the beam on its circular path. At the peak energy of 7 TeV, a magnetic field of

$$B = \frac{p}{qR} \approx \frac{E}{cqR} = \frac{7\ \text{TeV}}{c\ 4.3\ \text{km}} = 5.4\ \text{T} \quad , \tag{2.5}$$

is required, assuming that the whole LHC beam pipe is covered with dipole magnets. This is, of course, not the case, requiring the actual magnetic field strength of the LHC dipole magnets to be slightly higher than estimated, with a maximum field of 8.3 T.

In addition to the 1,232 bending dipole magnets, each 15 metres long, the LHC employs 474 quadrupole magnets [20] to focus the beam. Each quadrupole magnet is capable of squeezing the beam in one dimension (Fig. 2). Combining two complementary quadrupole magnets, one squeezing the beam in the x-direction while the other squeezing it in the y-direction, results in an overall focusing of the beam. The focusing effect of the quadrupoles causes protons to oscillate around the ideal centre of the beam pipe. The number of oscillations per turn of the machine is referred to as the Q or the tune.

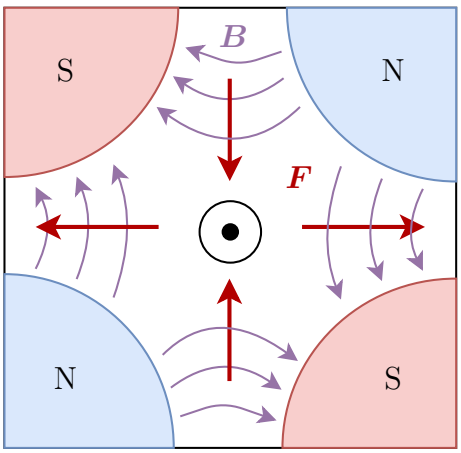


Figure 2: Schematic drawing of a quadrupole magnet. The magnetic field B squeezes the beam in one dimension via the resulting Lorentz force F.

Beyond the primary dipole and quadrupole magnets, the LHC employs higher-order multipole magnets to refine beam dynamics and compensate for various undesired effects. Sextupole magnets correct chromatic aberrations, while octupole magnets enhance beam stability by providing Landau damping through amplitude-dependent tune shifts, thereby suppressing coherent instabilities. In addition, decapole magnets are used to correct residual field errors in the main magnets [21].

2.4 Radio frequency cavities

Particles get injected as bunches into the LHC from the previous accelerators. The task of radio frequency (RF) cavities is keeping the particle bunches longitudinally close together while simultaneously increasing their energy incrementally with each revolution. One such cavity is depicted in Fig. 3 with its characteristic pill shape.

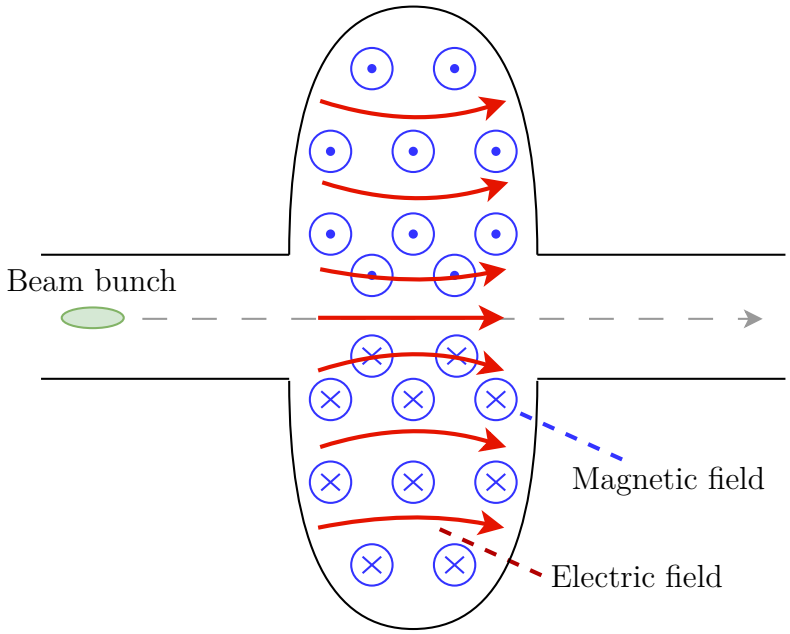


Figure 3: Schematic drawing of a RF cavity. The electric and magnetic fields are depicted in their fundamental accelerating mode TM_{010}.

In the lower energy linear accelerators, which CERN uses as pre-accelerators for the PS, drift tubes are employed. These tubes, essentially cylindrical plate capacitors, are positioned along the beamline at specific intervals. Their polarity alternates in sync with the beam particles, attracting them as they approach and repelling them as they depart. However, at higher energies, the use of drift tubes becomes impractical. As particle velocity increases, maintaining synchronisation would require either lengthening the tubes or increasing the frequency. The former is infeasible, while the latter would lead to significant energy loss through radiation.

The solution to this problem is to enclose the electromagnetic field in a so-called radio frequency cavity. The boundary conditions at the cavity walls, assuming perfect conductivity $R = 0$, require the parallel component of the electric field to vanish, which

results in the formation of standing waves in specific modes inside the cavity. In the LHC, superconducting RF cavities are used, featuring copper walls coated with a thin layer of niobium. These cavities can deliver up to 2 MV and are cooled to 4.5 K by being immersed in helium, ensuring the niobium remains below its critical temperature [22]. The LHC incorporates a total of 16 RF cavities, eight for each beam, housed within four cylindrical cryomodules.

Similar to drift tubes, the oscillating field in RF cavities must remain synchronised with the approaching beam particles. With the cavities operating at a frequency of 400 MHz, there are a total of

$$N = \frac{f_{RF}}{f_{rev}} = \frac{2\pi R f_{RF}}{c} \approx 35{,}600 \quad , \tag{2.6}$$

so-called "buckets" in the LHC. These are virtual positions along the storage ring where the conditions allow particles to stay in sync with the electromagnetic field generated by the cavities. The positions themselves do not need to be filled. In fact, the PS delivers bunches with a 25 ns spacing (corresponding to a frequency of 40 MHz), which would allow for up to 3,557 bunches in the LHC. However, additional constraints, such as the requirement for an abort gap — a gap that allows the magnets time to activate and redirect bunches to the dump — reduce the number of filled buckets to 2,808.

Besides acceleration, the cavities also provide a compacting effect on the bunch length. A particle perfectly synchronised with the fields of the cavities is referred to as a synchronous particle. Now, consider a particle with energy higher than that of the synchronous particle. According to Eq. 2.5, this particle will follow an orbit with a larger radius R, which increases the distance it travels per revolution, causing it to arrive later after completing one revolution. Because of this delay, the particle is no longer perfectly in sync with the field and is effectively slowed down compared to the synchronous particle. In the following revolution, the particle moves along an orbit with a smaller radius R, causing it to arrive earlier than the synchronous particle and receive an additional energy boost from the electric field. Overall, this effect leads to longitudinal oscillations around the synchronous particle and prevents the bunch from drifting apart.

2.5 High-Luminosity operation

A key figure of merit for any particle accelerator is the so-called instantaneous luminosity L. It is the key figure in describing the achievable number of collisions dN in a certain time interval dt as

$$\frac{dN}{dt} = \sigma_p \, L \quad , \tag{2.7}$$

where σ_p denotes the cross-section of colliding particles. While the cross-section represents a physical property of the colliding particles, independent of the accelerator and detector, luminosity is a characteristic of the accelerator, unaffected by the particles' physical properties. It is determined by the number of particles in each bunch N_1, N_2, their overlap cross-sectional area A, the number of bunches N_b, and the revolution

frequency f_{rev} via

$$L = \frac{N_1 N_2 f_{rev} N_b}{A} \quad . \tag{2.8}$$

If one assumes Gaussian-shaped bunches, the instantaneous luminosity Eq. 2.8 takes the form

$$L = \frac{N_1 N_2 f_{rev} N_b}{4\pi\sigma_x\sigma_y} \quad , \tag{2.9}$$

with σ_x, σ_y the tangential beam size [23]. The beam size is often expressed in accelerator physics in terms of the beam emittance ϵ, a quantity that refers to the area in the position-momentum phase space (x, x') occupied by the particles of the beam. The area is encapsulated by an ellipsis (Fig. 4) described by the equation

$$\gamma_{cs}x^2 + 2\alpha_{cs}xx' + \beta_{cs}x'^2 = \epsilon \quad , \tag{2.10}$$

where γ_{cs}, α_{cs} and β_{cs} denote the Courant-Snyder parameters, sometimes also referred to as Twiss parameters. It is worth noting that x' here is given as the angle between the particle's momentum p_x and the z-axis

$$\frac{p_x}{p_z} = \tan x' \approx \frac{dx}{dz} = x' \quad . \tag{2.11}$$

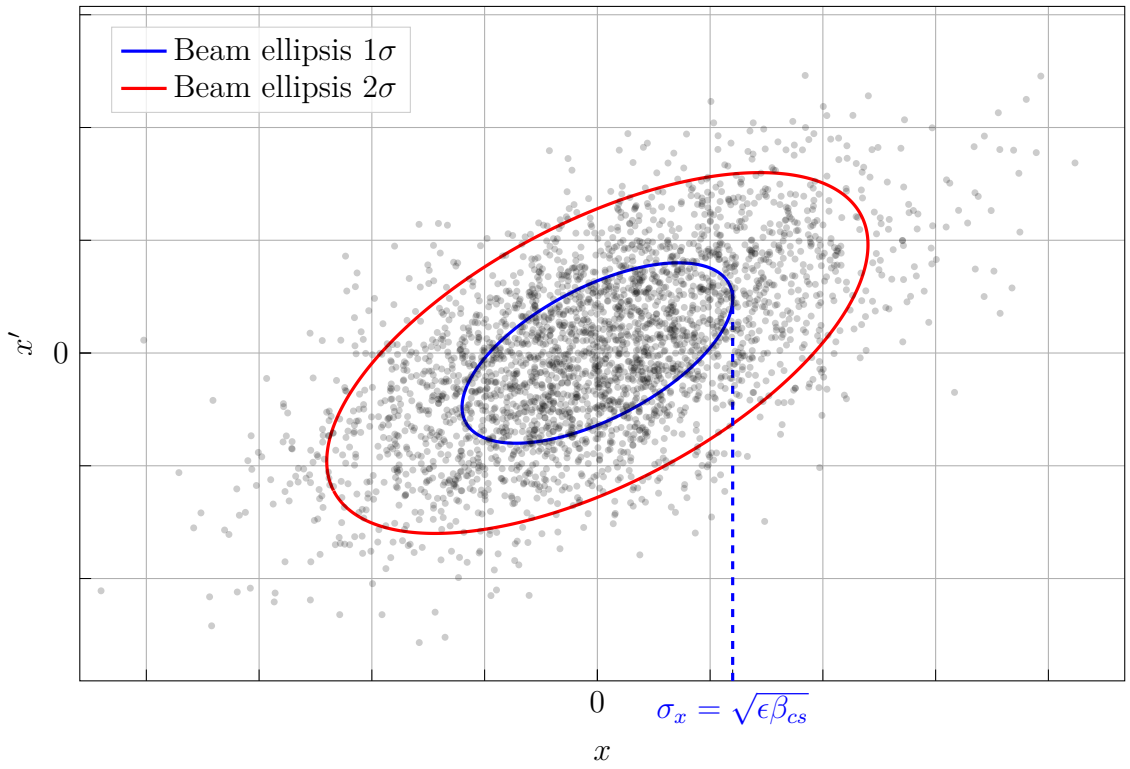


Figure 4: Phase space plot in the (x, x') plane, each dot represents a particle of the beam. Simulated by drawing from a bivariate normal distribution.

According to Liouville's theorem, the phase-space distribution, and consequently the emittance ϵ, remains constant when subjected to conservative forces. However, this is no longer the case during beam acceleration. To enable comparison of beam quality at different energies, the so-called normalised emittance

$$\epsilon_n = \gamma_z \beta_z \epsilon \quad , \tag{2.12}$$

is introduced, where γ_z is the Lorentz factor in z-direction and $\beta_z = v_z/c$. The normalised emittance ϵ_n remains invariant under acceleration, enabling the assertion of beam degradation. In the case of a storage ring, the Courant-Snyder parameter β_{cs} becomes a function of the position s in the ring, which depends in quite a complex way on the configuration of the focusing quadrupole magnets. The setup is designed so that $\beta_{cs}(s)$ reaches a minimum, referred to as β^*, at the interaction point in order to minimise the beam size $\sigma_x \sigma_y$.

In the LHC, another factor influences the luminosity. The bunches cross at a small angle, meaning they must first be diverted slightly from their original circular path to collide with opposing bunches. This angle alters the shape of the collision overlap, as shown in Fig. 5, resulting in a collision that is no longer perfectly head-on.

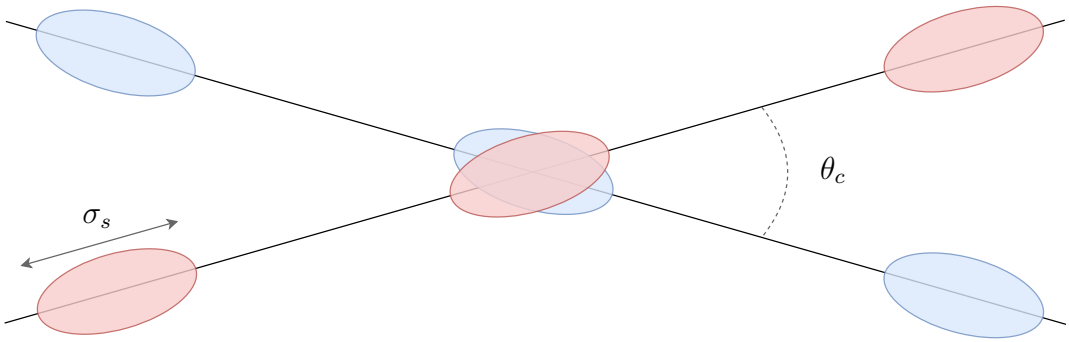


Figure 5: Schematic drawing of the angled LHC bunch crossing, crossing angle exaggerated.

The effect can be accounted for by a so-called luminosity reduction factor

$$S = \frac{1}{\sqrt{1 + \left(\frac{\sigma_x}{\sigma_s} \tan\frac{\theta_c}{2}\right)^2}} \frac{1}{\sqrt{1 + \left(\frac{\sigma_s}{\sigma_x} \tan\frac{\theta_c}{2}\right)^2}} \quad , \tag{2.13}$$

where σ_s denotes the bunch length and θ_c the crossing angle. The crossing angle effect is quite sizeable, leading to a luminosity reduction of roughly 20% [23], which is why so-called "crab cavities" are planned to be installed for the High-Luminosity upgrade [24]. They slightly tilt the bunches for a more perfect head-on collision at the interaction point.

Additional terms, accounting for slight transverse beam offsets and the hourglass effect, the variation in $\beta_{cs}(s)$, and consequently the beam density around the interaction

point (β^*), also contribute to a slightly lower luminosity than predicted by Eq. 2.9. It is important to note that transverse beam offsets are also intentionally introduced to regulate the luminosity delivered to a detector, a process known as luminosity levelling [25].

For the LHC we can estimate the instantaneous luminosity during Run 3 using a set of machine parameters $E = 6.8$ TeV, $\epsilon_n = 1.8$ μm, $\beta^* = 1.2$ m, $\theta_c/2 = 160$ μrad, $\sigma_s = 9$ cm, $N_1 = N_2 = 1.8 \times 10^{11}$ and $N_b = 2{,}736$ [26]. The luminosity can then be calculated, yielding

$$\sigma_x = \sigma_y = \sqrt{\frac{\epsilon_n \beta^*}{\beta_z \gamma_z}} \approx \sqrt{\frac{\epsilon_n \beta^*}{\gamma_z}} = \sqrt{\frac{\epsilon_n \beta^* m c^2}{E}} = 17.3 \ \mu\text{m} \quad , \tag{2.14}$$

$$S \approx \frac{1}{\sqrt{1 + \left(\frac{\sigma_s}{\sigma_x} \frac{\theta_c}{2}\right)^2}} = 0.77 \quad , \tag{2.15}$$

$$L = S \cdot \frac{N_1 N_2 f_{rev} N_b}{4\pi \sigma_x \sigma_y} = 2.0 \times 10^{34} \ \frac{1}{\text{cm}^2 \, \text{s}} \quad , \tag{2.16}$$

which is indeed the preferred luminosity by ATLAS and CMS during Run 3 for an average number of collision vertices per bunch crossing close to the limit of 60.

Finally, after having examined the current state of the LHC, we can take a look at which parameters are subject to change for High-Luminosity operation. While the bunch population is expected to only slightly increase to $N_1 = N_2 = 2.2 \times 10^{11}$, the main improvement stems from the advancement in beam optics around the interaction points of ATLAS and CMS, which should allow for a β^* of only 15 cm through the instalment of powerful inner triplet quadrupole magnets. These magnets are made of a novel superconducting compound, based on niobium and tin (Nb_3Sn) enabling a field of up to 12 T. Together with the other machine parameters $E = 7$ TeV, $\epsilon_n = 2.5$ μm, $\theta_c/2 = 250$ μrad, $\sigma_s = 9$ cm and $N_b = 2{,}760$ [27] we can estimate

$$\sigma_x = \sigma_y \approx \sqrt{\frac{\epsilon_n \beta^* m c^2}{E}} = 7.1 \ \mu\text{m} \quad , \tag{2.17}$$

$$S \approx \frac{1}{\sqrt{1 + \left(\frac{\sigma_s}{\sigma_x} \frac{\theta_c}{2}\right)^2}} = 0.30 \quad . \tag{2.18}$$

The more accurate value for the luminosity reduction factor, provided in the design report [27], is $S = 0.34$, underscoring the necessity of using "crab cavities" for High-Luminosity operation. With "crab cavities", the luminosity reduction factor improves

to $S = 0.716$, and the achievable instantaneous luminosity becomes

$$L_{peak} = S \cdot \frac{N_1 N_2 f_{rev} N_b}{4\pi\sigma_x\sigma_y} = 17.0 \times 10^{34} \ \frac{1}{\mathrm{cm}^2\,\mathrm{s}} \quad . \tag{2.19}$$

Both the energy deposition of collision debris in the inner triplet quadrupole magnets at the interaction points and the necessity to limit the number of collisions per bunch crossing require the peak luminosity to be levelled down. Initially, a continuous instantaneous luminosity of $5.0 \times 10^{34}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ is foreseen, but later stages of High-Luminosity operation should allow for $L = 7.5 \times 10^{34}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The levelling is expected to either be achieved by the variation of β^* or the variation of "crab cavity" voltage. The large peak luminosity L_{peak} compared to the target luminosity is essential in delivering a constant luminosity to the experiments, allowing the levelling to counteract the natural beam degradation during an ongoing run.

The amount of data collected by a particle physics experiment is typically expressed in terms of the integrated luminosity

$$L_{\mathrm{int}} = \int_0^T L(t)\,dt \quad . \tag{2.20}$$

The integrated luminosity represents the final figure of merit, which crucially also takes beam degradation and downtime into account. It relates to the total number of observed events through

$$L_{\mathrm{int}} \cdot \sigma_p = N \quad . \tag{2.21}$$

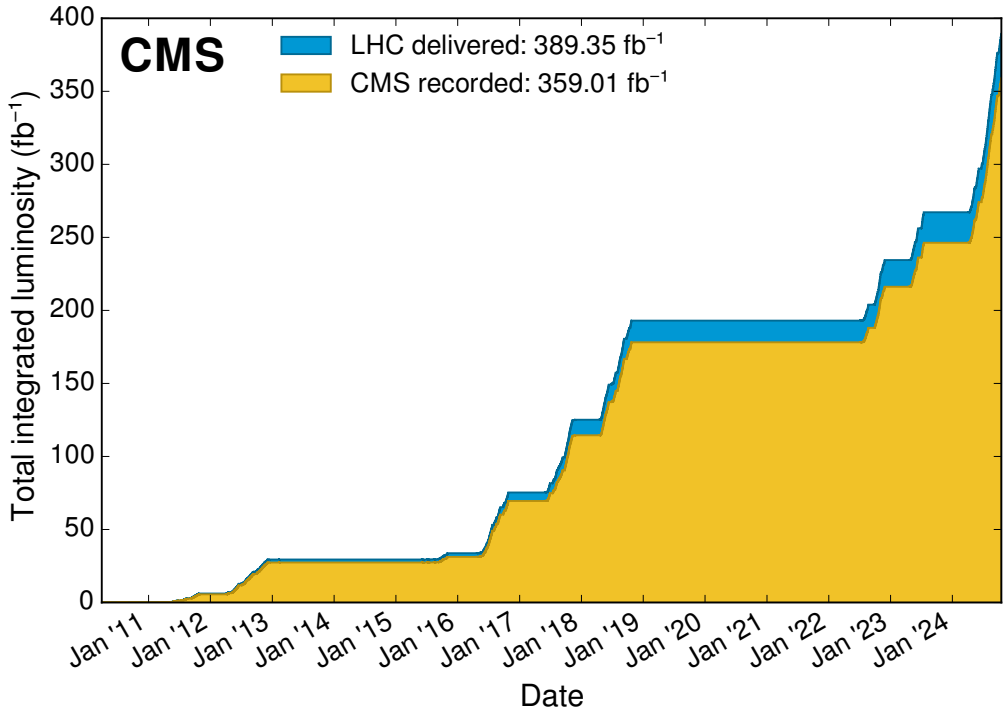


Figure 6: Plot of the integrated luminosity of proton-proton collisions delivered by the LHC compared to the recorded integrated luminosity by CMS [28]. Note that $1\ \text{fb} = 10^{-39}\ \text{cm}^2$.

From Fig. 6 one can conclude that with the total integrated luminosity of 359 fb^{-1} and a proton-proton cross-section σ_p of $\sim 80 \times 10^{-27}\ \text{cm}^2$ roughly 3×10^{16} collision events have been recorded. However, only a tiny fraction of these events were stored on disk, as most were immediately discarded by the trigger system (c.f. Section 4).

High-luminosity operation targets an annual integrated luminosity of 250 fb^{-1} [27], which should result in 3,000 fb^{-1} in the following 12 years after the upgrade, exceeding the current figure by almost a factor of 9.

3 The Compact Muon Solenoid (CMS) experiment

3.1 Introduction

The Compact Muon Solenoid (CMS) is a general-purpose particle detector at the Large Hadron Collider (LHC) positioned underground, near Cessy, France. Its broad physics program encompasses not only the precision measurements of the Standard Model particles and interactions, including the Higgs boson, but also searches for phenomena beyond the Standard Model, such as extra dimensions, potential dark matter particles and Supersymmetry. While CMS shares the same scientific objectives as the ATLAS experiment, it employs distinct technical approaches and utilises a different magnet system in its design.

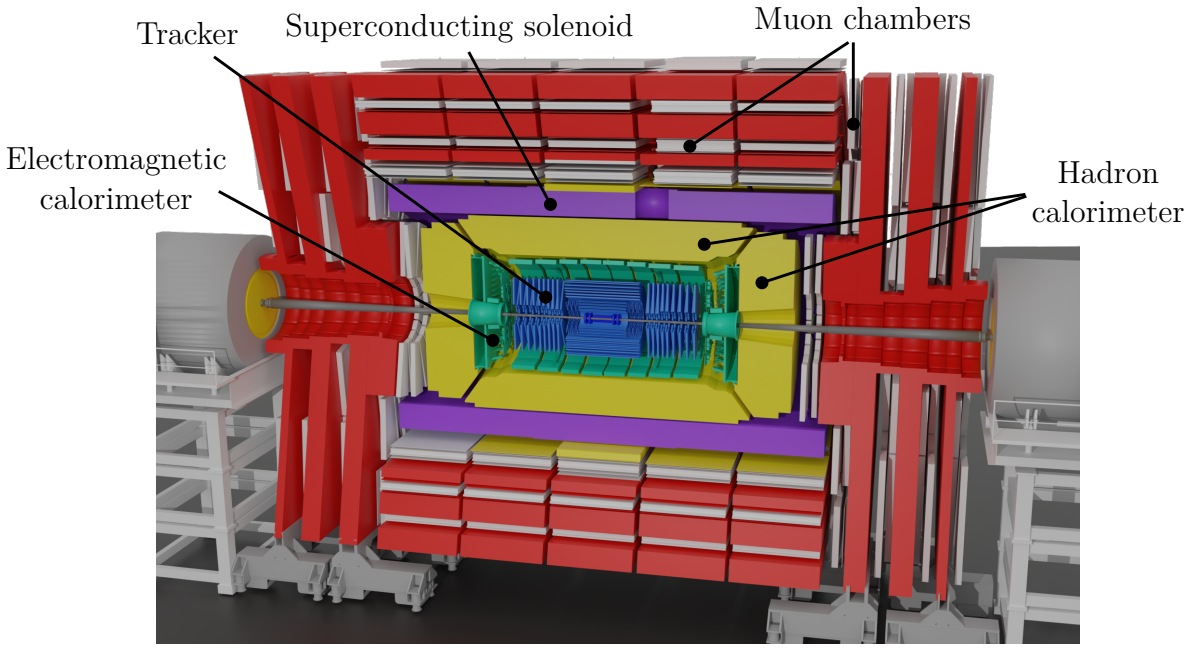


Figure 7: Rendered image of a 3D model [29] of the CMS detector during Run 3 sliced in half. Each colour represents a different detector subsystem.

The CMS detector measures 22 metres in length, 15 metres in diameter, and weighs 14,000 tons. Arranged around the interaction point 5 (IP5) of the LHC, it is designed to trigger on and identify various particles, including electrons, muons, photons, and both charged and neutral hadrons. The CMS's central component is a superconducting solenoid that measures 6 metres in internal diameter and 12.5 metres in length, capable of producing a magnetic field of 3.8 T. Within the solenoid's magnetic field are several key subsystems: a silicon pixel and strip tracker, an Electromagnetic Calorimeter (ECAL) made of lead tungstate crystals, and a Hadron Calorimeter (HCAL) composed of brass interspersed with plastic scintillating tiles. Both the ECAL and HCAL consist of barrel and endcap sections. Forward calorimeters extend the coverage beyond that of the barrel and endcap detectors. Muon detection is achieved using gas-ionisation detectors, which are placed within the steel flux-return yoke located outside the solenoid. An illustration

of the CMS detector with the mentioned subsystems is depicted in Fig. 7. The CMS experiment and its detector subsystems were designed with the key objectives of the LHC physics program in mind, which can be summarised as follows [30]:

- Accurate muon identification and momentum resolution across a broad range of momenta and angles, with good dimuon mass resolution at around 1% at 100 GeV and the ability to determine muon charge unambiguously for momenta smaller than 1 TeV.
- Excellent charged-particle momentum resolution and reconstruction efficiency in the inner tracker, along with efficient τ and b-jet tagging both in triggering and offline analysis, requiring precision pixel tracking detectors close to the interaction point.
- High electromagnetic energy resolution, along with good diphoton and dielectron mass resolution with about 1% at 100 GeV, broad geometric coverage, π^0 rejection, and effective photon and lepton isolation at high luminosities.
- Accurate missing transverse energy and dijet mass resolution, requiring hadron calorimeters with large hermetic coverage and fine lateral segmentation.

The goals above drove the technology choices made for the detector subsystems, described in the following sections.

3.2 Pileup (PU)

The total proton-proton cross-section, σ_p, at $\sqrt{s} = 14$ TeV was estimated to be approximately 80×10^{-27} cm^2 [31]. Using Eq. 2.7 and an instantaneous luminosity of $L = 5.0 \times 10^{34}$ cm^{-2} s^{-1} during High-Luminosity operation, this leads to around 4×10^9 inelastic collisions per second [32]. Given the LHC revolution frequency $f_{rev} = 11,246$ Hz and a bunch population of $N_b = 2{,}760$, this corresponds to an average of roughly 130 simultaneous collisions per bunch crossing[1]. Assuming the worst case scenario for the extrapolated total proton-proton cross-section σ_p, the average number of simultaneous collisions at the initial luminosity of 5.0×10^{34} cm^{-2} s^{-1} is expected to be 140. As the luminosity increases to 7.5×10^{34} cm^{-2} s^{-1}, this number could rise to 200.

The aim of achieving a large number of simultaneous collisions stems from a crucial difference between hadron and lepton colliders. In lepton colliders — such as LEP — collisions that result in the annihilation of a lepton-anti-lepton pair can produce new particles with the full centre-of-mass energy, $\sqrt{s}$. In contrast, hadron collisions are far more complex due to the internal structure of hadrons.

Most hadron collisions are classified as "soft", with energies too low to yield discoveries of new physics. A smaller fraction, termed "hard" collisions, involve high transverse momentum particles, potentially originating from the decay of massive objects. However, during High-Luminosity operation up to an average of 140 or 200 "soft" pileup (PU)

[1] Each collision event occurs independently of previous ones, causing the number of collisions per bunch crossing to fluctuate according to a Poisson distribution.

collisions can overlap with these interesting "hard" scatterings, which can impair the performance of the trigger and offline reconstruction systems. This pileup scenario represents a significant increase from current Run 3 conditions, which involve up to an average of 60 "soft" collisions, driving the need for upgrades in electronics, trigger systems, and reconstruction algorithms.

As a final remark, it should be noted that the fraction of "hard" collisions is so low that most bunch crossings only contain "soft" collisions. These events are often referred to as "Minimum Bias" events due to the absence of most of the bias from the otherwise more stringent trigger selection.

3.3 Coordinate system

In hadron collider experiments, only some constituents (partons) of the hadron interact to form new particles. The fraction of interacting constituents, however, is a priori unknown. Thus, the momenta of newly formed particles along the beamline (z-axis) are also unknown. It therefore becomes convenient to use a quantity whose differences are invariant under Lorentz boosts along the z-axis, counteracting the unknown momentum along the beamline. A quantity that fulfils this requirement is the relativistic rapidity,

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right) \quad . \tag{3.1}$$

Note in Eq. 3.1 we have used natural units ($c = 1$), a convention that will be continuously followed when describing kinematics. With a Lorentz boost in z-direction

$$E' = E \cosh \gamma - p_z \sinh \gamma \tag{3.2}$$
$$p'_z = p_z \cosh \gamma - E \sinh \gamma \quad , \tag{3.3}$$

the boosted rapidity becomes

$$y' = \frac{1}{2} \ln \left(\frac{(E + p_z)\,(\cosh \gamma - \sinh \gamma)}{(E - p_z)\,(\cosh \gamma + \sinh \gamma)} \right) = y + \frac{1}{2} \ln e^{-2\gamma} = y - \gamma \quad . \tag{3.4}$$

Hence, $\Delta y' = \Delta y$. We would now like to have a relation between the measurable polar angle θ and the rapidity. To achieve this we express p_z and E in terms of p, m and θ,

$$y = \frac{1}{2} \ln \left(\frac{\sqrt{p^2 + m^2} + p \cos \theta}{\sqrt{p^2 + m^2} - p \cos \theta} \right) \quad . \tag{3.5}$$

From Eq. 3.5 we can conclude that in general, y is a complicated function of p, m, and θ, where p and m are not directly measurable. Neglecting the rest mass m, however,

leads to the far simpler relation,

$$
\begin{aligned}
y \simeq \frac{1}{2} \ln \left[\frac{p + p\cos\theta}{p - p\cos\theta} \right] &= \ln \left[\left(\frac{1+\cos\theta}{1-\cos\theta} \right)^{\frac{1}{2}} \right] = \ln \left[\left(\frac{\cos^2\left(\frac{\theta}{2}\right)}{\sin^2\left(\frac{\theta}{2}\right)} \right)^{\frac{1}{2}} \right] \\
&= \ln \left[\frac{\cos\left(\frac{\theta}{2}\right)}{\sin\left(\frac{\theta}{2}\right)} \right] = \ln \left[\frac{1}{\tan\left(\frac{\theta}{2}\right)} \right] = -\ln \left[\tan\left(\frac{\theta}{2}\right) \right] \quad .
\end{aligned}
\tag{3.6}
$$

It therefore becomes convenient to formally define pseudorapidity,

$$
\eta \equiv -\ln \left[\tan\left(\frac{\theta}{2}\right) \right] \quad . \tag{3.7}
$$

Most directly measurable particles in the CMS detector have a low rest mass compared to their momentum, meaning they are ultrarelativistic. It is thus numerically sufficient to use pseudorapidity η instead of the rapidity y to retain the desired property of invariance of distances in η under Lorentz boosts along the beam axis.

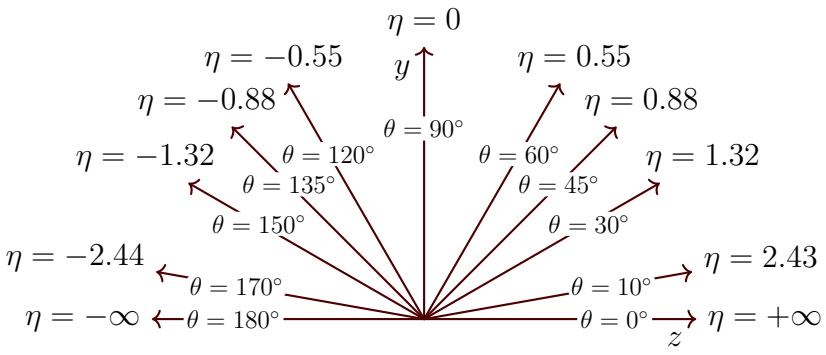


Figure 8: Pseudorapidity η values for various polar angles θ [33].

Equipped with this alteration of the polar angle, CMS's coordinate system (Fig. 9), the coordinate system of momentum vectors of detected particles can now be expressed in terms of the transverse momentum p_T, the pseudorapidity η and the azimuthal angle ϕ. The detector can be divided into a forward region ($\eta > 0$) and a negative forward region ($\eta < 0$).

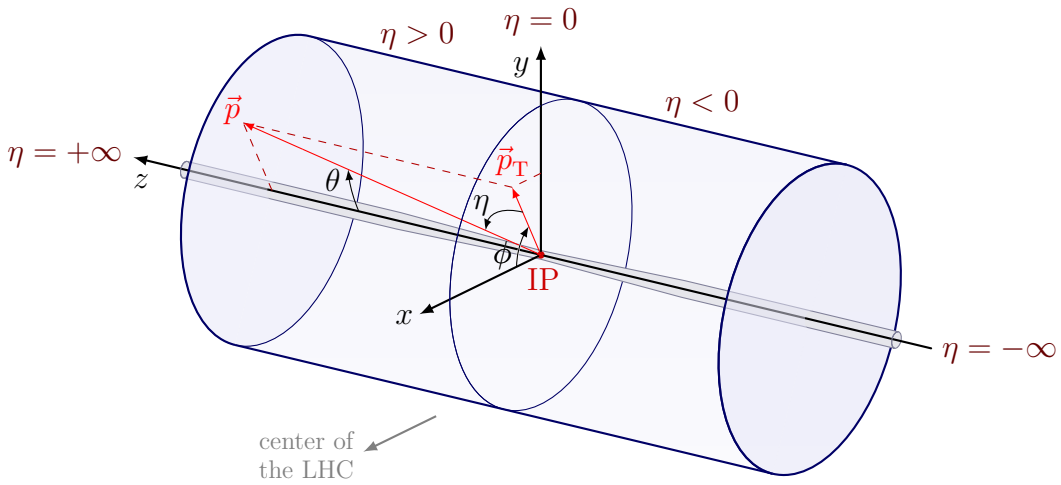


Figure 9: Schematic drawing of the CMS detector's coordinate system $(p_T, \eta, \phi)^T$. Modified from [33].

A conversion to cartesian coordinates can be found by expressing $\tan\theta$ in

$$p_z = \frac{p_T}{\tan\theta} \quad , \tag{3.8}$$

using the identity of the half angle, $\tan\theta = \frac{2\tan\left(\frac{\theta}{2}\right)}{1-\tan^2\left(\frac{\theta}{2}\right)}$. Eq. 3.8 then becomes

$$p_z = p_T \, \frac{1-\tan^2\left(\frac{\theta}{2}\right)}{2\tan\left(\frac{\theta}{2}\right)} \quad . \tag{3.9}$$

Plugging in $\tan\left(\frac{\theta}{2}\right)$ from Eq. 3.7 gives

$$p_z = p_T \, \frac{1-e^{-2\eta}}{2e^{-\eta}} = p_T \sinh\eta \quad . \tag{3.10}$$

The other two components are simple projections onto the x- and y-axis with,

$$p_x = p_T \cos\phi \tag{3.11}$$
$$p_y = p_T \sin\phi \quad . \tag{3.12}$$

Hence, the magnitude can be expressed as

$$|\boldsymbol{p}| = p_T \cosh\eta \quad . \tag{3.13}$$

3.4 Silicon tracker

The tracking system constitutes the innermost part of the CMS detector. It is tasked with tracking the trajectories of charged particles, enabling the reconstruction of their paths from the originating vertex to an energy deposition site in the calorimeters or muon systems. To avoid impairing the calorimeter's energy measurements, tracking detectors are typically designed with a low material budget. Consequently, these systems aim to measure ionisation signals in some active detector material while limiting the material required for power delivery, cooling, and electronics. In the case of the CMS detector, the active detector material is silicon pixels and strips that detect hits by measuring the electric charge collected from the formation and drift of electron-hole pairs.

Designed to operate at the nominal LHC luminosity of 1×10^{34} cm^{-2} s^{-1} during Run 1-3 the tracking system requires a complete replacement to cope with the conditions at the High-Luminosity LHC. The new system will feature increased radiation hardness, higher granularity and the ability to handle higher data rates and a longer trigger latency. These improvements aim to maintain, at a minimum, the current detector's performance levels in tracking and vertex reconstruction capabilities [34].

The layout for High-Luminosity operation consists of two parts, an Inner Pixel Tracker and an Outer Tracker, which comprises different detector modules.

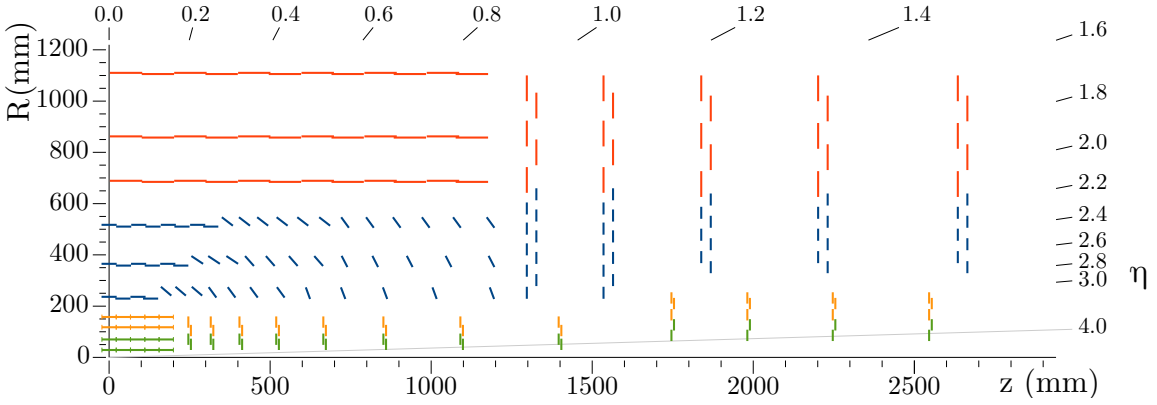


Figure 10: Schematic drawing of one-quarter of the High-Luminosity CMS tracker layout with coloured lines representing various detector modules. The Inner Tracker system comprises pixel modules, which are shown in orange for quad-chip modules and green for double-chip modules. The Outer Tracker consists of two types of modules, represented by blue lines for pixel-strip modules and red lines for two-strip modules [35].

The Inner Pixel Tracker will consist of two different modules, double-chip (1×2) modules with high read-out rate capability for the innermost layers and rings, and quad-chip (2×2) modules for the outer layers and disks. The innermost modules will primarily use 3D pixel sensors, which offer greater radiation resistance and lower power consumption. However, since 3D pixel sensors are complex to produce, the outer layers of the Pixel Tracker will instead use simpler thin planar pixel sensors. A new pixel readout chip is under development by RD53, a joint ATLAS-CMS collaboration. This chip will

handle digitisation, readout, and serial powering of the sensors — an essential feature for High-Luminosity operation that minimizes cabling material in the inner detector layers, thereby improving measurements performed by other detector subsystems.

The Outer Tracker will also comprise two different modules, one housing a macro-pixel sensor as well as a silicon strip sensor, while the other consists of two silicon strip sensors (Fig. 10). A crucial feature of the new Outer Tracker is the ability to provide tracking information at the collision rate of 40 MHz to the Level-1 trigger system (c.f. Section 4). This development is possible by leveraging the fact that mostly high transverse momentum (p_T) particles are relevant for physics reconstruction. Cleverly selecting pairs of tracker hits with the two types of modules — collectively referred to as p_T-modules — that adhere to a low bend trajectory (implying high p_T) in CMS's magnetic field allows for an effective reduction in required bandwidth (Fig. 11). The technique is performed on the dedicated readout chip of the p_T-modules to ensure that only passing pairs occupy valuable bandwidth on the optical links to the Level-1 trigger system. Only in the case of the Level-1 trigger system accepting an event will the whole Outer Tracker be read out.

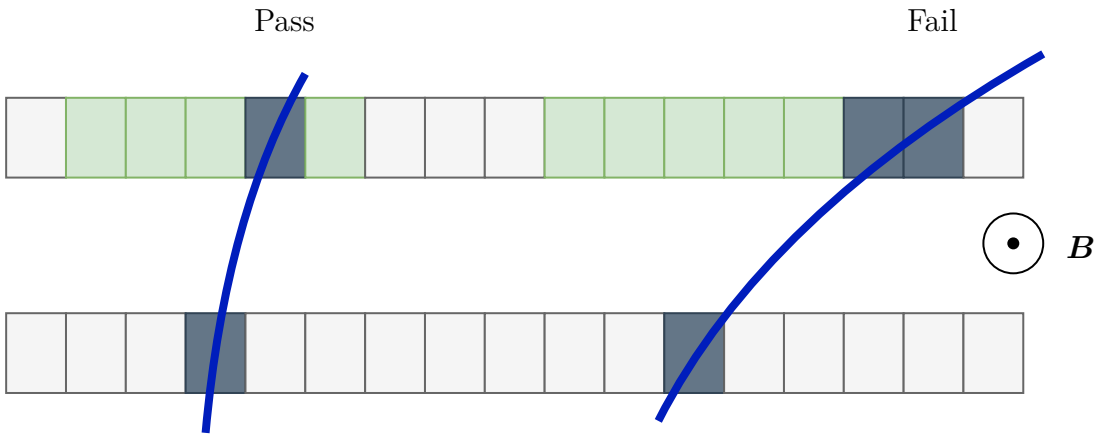


Figure 11: Schematic drawing of a p_T-module, only passing pairs of tracker hits are sent out to the trigger system as tracker "stubs".

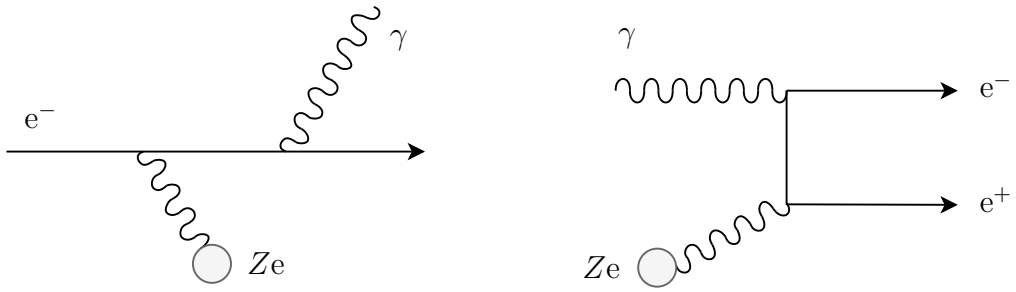
Overall, the design of the Outer Tracker covers up to a pseudorapidity $|\eta|$ of 2.4 and ensures that six layers of modules have to be traversed, enabling good tracking performance.

3.5 Calorimeters

The calorimeters are tasked with stopping and simultaneously measuring the energy of electrons, positrons, photons and hadrons. They can be separated into an inner Electromagnetic Calorimeter (ECAL) specialising in the energy deposition of electrons, positrons and photons and an outer Hadron Calorimeter (HCAL) for the energy deposition of hadrons.

The energy deposition of high-energy electrons, positrons, and photons happens through the formation of particle showers. The two effects responsible for these electromagnetic

cascades are bremsstrahlung and pair production. While bremsstrahlung affects electrons and positrons and causes them to radiate away a photon, pair production affects photons, causing them to convert into an electron-positron pair.



(a) Bremsstrahlung (b) Pair production

Figure 12: Illustration of bremsstrahlung and pair production, the two important processes for the formation of electromagnetic particle showers.

For both effects, the cross-section σ is proportional to Z^2, favouring the utilisation of materials such as lead, tungsten or uranium [36]. With increasing energy, the secondary particles for both effects are predominantly produced in the forward direction

$$\theta \propto \frac{1}{\gamma} = \frac{m_e c^2}{E} \quad . \tag{3.14}$$

These principles can be incorporated into a simplified model in which, on average, the number of particles doubles after a certain characteristic length X_0 due to bremsstrahlung and pair production. While the number of particles doubles after each step, their energy decreases. Below a certain critical energy E_c, energy loss through ionisation and excitation becomes dominant for electrons and positrons, preventing further photon generation and halting the cascade. An empirical equation for the shower development was found by Longo and Sestili [37]

$$\frac{dE}{dt} = E_0 \frac{b^a}{\Gamma(a)} t^{a-1} e^{-bt} \quad , \tag{3.15}$$

where Γ denotes the gamma function, $t = x/X_0$ the cascade step number and a and b are parameters that depend on E_0 and Z. The maximum of Eq. 3.15 occurs at

$$t_{max} = \frac{a-1}{b} = \ln \frac{E_0}{E_c} + C_j \qquad j = e, \gamma, \tag{3.16}$$

where $C_e = -0.5$ and $C_\gamma = 0.5$. The parameter b is very close to 0.5 for heavy elements with little dependence on the energy. The parameter a can then be determined via Eq. 3.16 or by fitting Eq. 3.15 to simulation or measurement data.

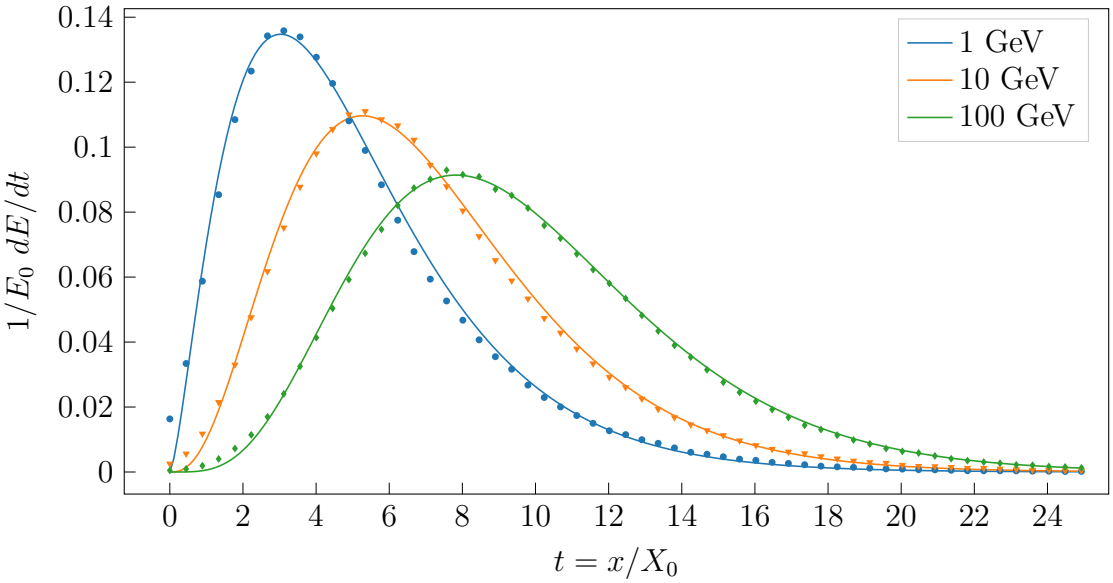


Figure 13: Plot of the longitudinal shower profile in $PbWO_4$ produced by incident electrons with various energies. The marks represent values obtained via a Geant4 [38, 39, 40] simulation, while the lines represent fits of Eq. 3.15 to the simulation data.

The dominant effects for the transversal profile of electromagnetic showers are multiple scattering for low-energy charged particles and Compton scattering for photons. A measure for the shower width is given by the Molière radius

$$R_M = \frac{m_e c^2}{E_c} \sqrt{\frac{4\pi}{\alpha}}\, X_0 \quad , \tag{3.17}$$

where α denotes the fine-structure constant. Approximately 99% of particles are within $3.5 \times R_M$.

Overall, we can conclude that the profile of the electromagnetic particle shower depends solely on the primary particle's energy and the properties of the absorber material. As a result, analysing the shower profile alone does not allow one to distinguish between electrons, positrons, and photons. However, exploiting tracking information with electrons bending in ϕ-direction, positrons in negative ϕ-direction and photons not bending at all in the magnetic field of the solenoid allows making this crucial distinction.

High-energy hadrons similarly produce cascades when traversing a medium, yet the responsible processes are far more numerous and complex [36]. They include:

- **Spallation and intra-nuclear cascades**, where the incident hadron scatters on a single nucleon of the nucleus and subsequent elastic scatterings inside the nucleus cause fragments to be ejected.
- **Evaporation**, where after spallation, the nucleus remains in an excited state resulting after about 10^{-18} s in the evaporation of nucleons, with kinetic energies of a few MeV.

- **Electromagnetic processes**, where the production of electromagnetically decaying short-lived mesons, mostly neutral pions, results in electromagnetic subshowers.
- **Weak processes**, where the production of weakly decaying mesons, mostly charged pions and kaons, results in the generation of neutrinos with undetectable kinetic energy.

The combination of hadronic, electromagnetic and weak processes leads to far greater fluctuations in the shower profile, resulting in a generally worse energy resolution when compared to purely electromagnetic showers. Furthermore, some of the incident energy can be lost in nuclear reactions as target nuclei absorb energy to increase their binding energy — an effect that is typically undetectable. This requires the use of various compensation techniques to account for these energy losses. Moreover, due to the complexity of the underlying models, hadronic showers are also far more challenging to simulate accurately with Monte Carlo methods.

A characteristic length for hadronic processes is the nuclear interaction length

$$\lambda_I = \frac{A}{N_A \, \rho \, \sigma_{inel}} \quad , \tag{3.18}$$

where A denotes the atomic mass number, N_A the Avogadro constant, ρ the density and σ_{inel} the inelastic cross-section for hadronic processes. The nuclear interaction length λ_I is typically much larger than the radiation length X_0, leading to far bigger hadron calorimeters than electromagnetic ones.

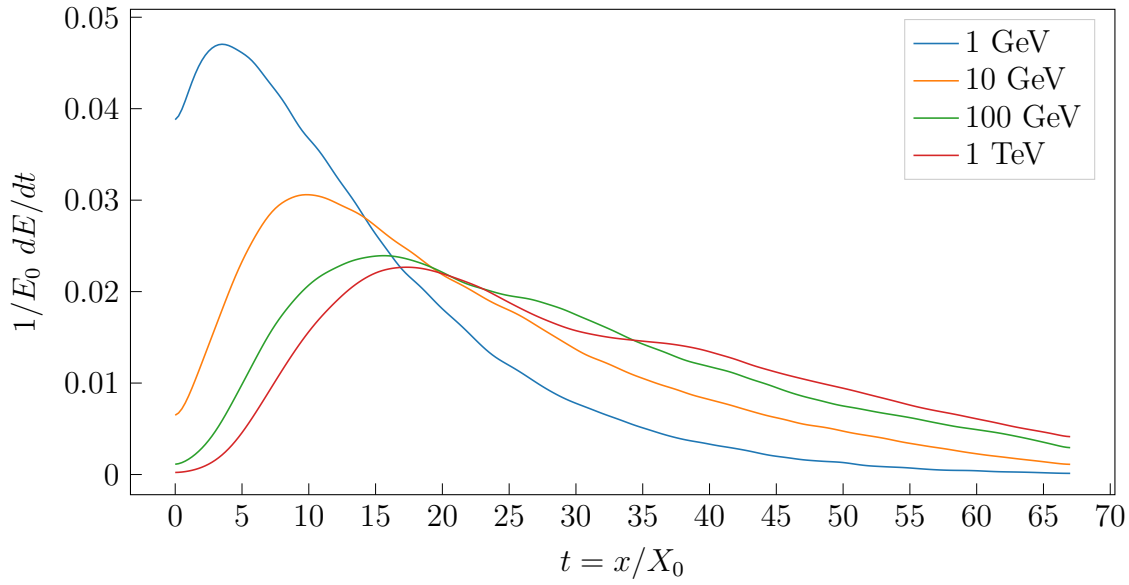


Figure 14: Plot of the averaged longitudinal shower profile in brass produced by incident π^- with various energies. With the radiation length $X_0 = 1.49$ cm and the nuclear interaction length $\lambda_I = 16.42$ cm, $\lambda_I \simeq 11 X_0$ of brass. Obtained via a Geant4 simulation (FTFP_BERT) [38, 39, 40]. Note that the shower profile extends much further into the material compared to the purely electromagnetic shower in Fig. 13.

Calorimeters are generally designed in one of two configurations. In a **sampling calorimeter**, the absorber and active detector materials are physically separated, typically arranged in a sandwich structure with alternating layers of absorber and active medium. This allows for a wider range of absorber materials, usually requiring low λ_I and X_0, to be used. The active medium may be a scintillator, a gas, silicon, or a Cherenkov radiator, which tend to have higher λ_I and X_0. In a **homogeneous calorimeter**, on the other hand, the whole volume contributes to the signal. These are often constructed from heavy scintillating crystals or non-scintillating Cherenkov radiators like lead fluoride or lead glass. However, due to significant differences in signal responses between hadronic and electromagnetic processes, along with the complexity of achieving effective 3D segmentation, homogeneous designs have so far been limited to electromagnetic calorimeters.

The energy resolution of a calorimeter can be expressed as

$$\frac{\sigma_E}{E} = \sqrt{\frac{A^2}{E} + \frac{B^2}{E^2} + C^2} = \frac{A}{\sqrt{E}} \oplus \frac{B}{E} \oplus C \quad , \tag{3.19}$$

where A summarises Poisson distributed stochastic fluctuations, B the electronic or energy-independent noise and C calibration errors and other irregularities such as shower leakage.

3.5.1 Electromagnetic calorimeter (ECAL)

In the CMS detector, the electromagnetic calorimeter is separated into a barrel covering $|\eta| < 1.48$ and two endcap sections extending the coverage to $|\eta| < 3.0$. In the current Run 3 setup, both sections are hermetic, homogeneous and made of 75,848 lead-tungstate scintillating crystals ($PbWO_4$), emitting blue-green (420-430 nm) light. The 61,200 barrel crystals are read out via Avalanche Photo-Diodes (APD), while the two endcaps with 7,324 crystals each use Vacuum Phototriodes (VPT). The short radiation length X_0 of 0.89 cm together with the small Molière radius R_M of 2.2 cm enable a compact calorimeter with a fine granularity [30].

For the upcoming High-Luminosity operation, the barrel section will undergo an electronics upgrade to accommodate the higher rate and latency requirements. The front-end system will feature upgraded, faster analogue electronics. The sampling rate will be increased from 40 MHz to 160 MHz, with 12-bit resolution, to achieve improved time resolution. This enhancement will aid in mitigating pile-up effects. Additionally, the trigger logic processing will be moved via high-speed optical links to off-detector electronics, allowing the implementation of more advanced trigger algorithms. Lastly, the operating temperature will be reduced from the current 18°C to 9°C to counteract the rise in radiation-induced dark current in the Avalanche Photo-Diodes. The overall barrel energy resolution for electrons has been measured in beam tests to be [41]

$$\frac{\sigma_E}{E} = \frac{2.8\%}{\sqrt{E}} \oplus \frac{12\%}{E} \oplus 0.3\% \quad , \tag{3.20}$$

with E in GeV.

The endcap sections, on the other hand, require a complete replacement due to the accumulated radiation damage and the expected further degradation, accelerated by the conditions at the High-Luminosity LHC. The replacement will be part of the new High-Granularity Calorimeter (HGCAL), which will be discussed in a subsequent section.

Currently, the endcaps are equipped with a pre-shower detector, consisting of two layers of lead with a total thickness of $3\,X_0$, where each layer is followed by a plane of silicon strip sensors. This detector enhances the separation of two-photon hits, compensating for the limited granularity of the ECAL during Run 3, particularly when distinguishing photons originating from a Higgs decay versus those from a neutral pion decay. However, due to anticipated radiation damage and the need to accommodate larger buffers for increased Level-1 Trigger latency during High-Luminosity operation, the ECAL pre-shower detector will be completely removed after Run 3. Its functionality will be assumed by the new High-Granularity Calorimeter (HGCAL).

3.5.2 Hadron calorimeter (HCAL)

The hadron calorimeter consists of a barrel section (HB), two endcap sections (HE), two forward sections (HF) and an outer shower tail catcher section (HO) for the barrel. While the barrel section (HB) covers up to $|\eta| < 1.392$ the endcaps (HE) extend this coverage to $|\eta| < 3.0$, with a small overlap between $1.305 < |\eta| < 1.392$ (Fig. 15). Both sections utilise a brass absorber interspersed with plastic scintillators as the active medium during Run 3. The barrel section has an additional front and backplate made from steel. The layout is illustrated in Fig. 15.

The blue light generated in the scintillating tiles is shifted to green via embedded wavelength-shifting fibres. The segmentation follows a tower structure in $\eta - \phi$ space of size 0.087×0.087 ($\Delta\eta \times \Delta\phi$). Towers are composed of 16 layers in the barrel and 17 layers in the endcap, with up to 7 layers being optically summed and read out by a single silicon photomultiplier (SiPM).

The tail catcher section (HO) is located outside the superconducting solenoid and is tasked with enhancing the energy resolution of the barrel section ($|\eta| < 1.3$) by adequately sampling the tail. High-energy incident hadrons can produce showers with a long tail not fully recorded by the barrel calorimeter, which has a depth ranging from 5.82 λ_I at $|\eta| = 0$ to 10.6 λ_I at $|\eta| = 1.3$. The outer calorimeter (HO) utilises the magnetic solenoid together with the iron return yoke as an absorber and samples the signal with plastic scintillator tiles. The scintillation light is read out via silicon photomultiplier (SiPM), similar to the barrel section.

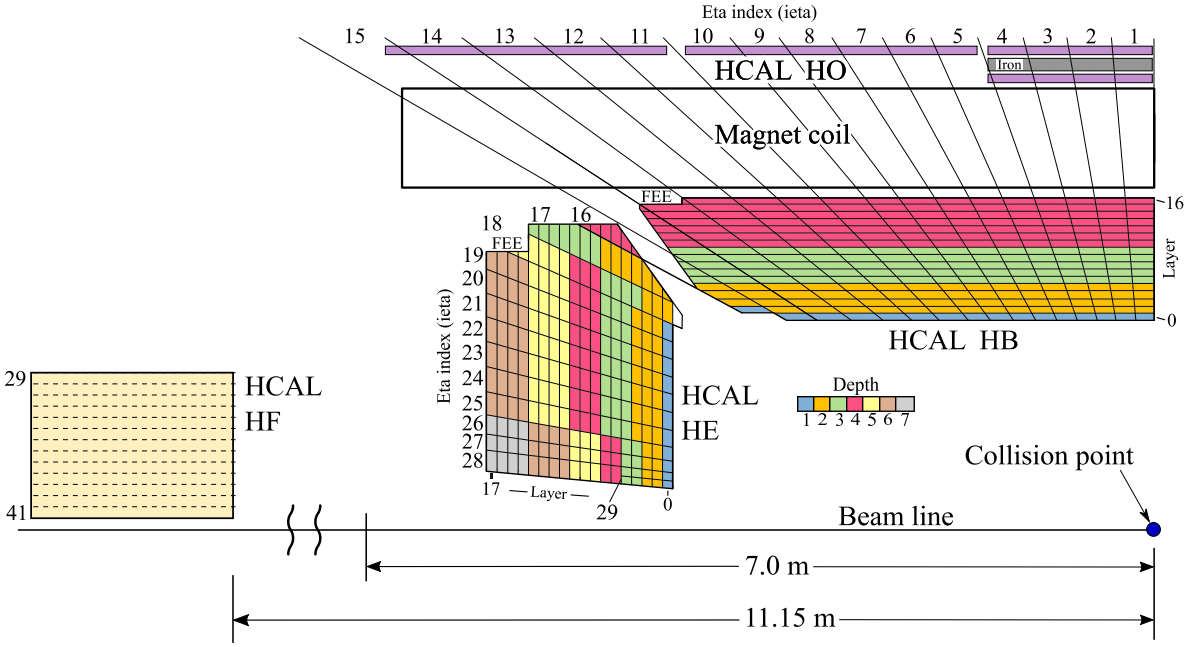


Figure 15: Schematic drawing of a quarter of the HCAL and its segmentation during Run 3 [42, 43]. Within a tower, layers of the same colour are connected to the same silicon photomultiplier (SiPM). The position of the front-end electronics is denoted by the acronym FEE.

Since particles detected by the hadron calorimeter must also traverse the electromagnetic calorimeter, the overall energy resolution is a combined result of the performance of both the HCAL and the ECAL. The charged pion energy resolution was measured in beam test [44] and is given by

$$\frac{\sigma_E}{E} = \frac{84.7\%}{\sqrt{E}} \oplus 7.6\% \quad , \tag{3.21}$$

with E in GeV. Intercalibrations during Run 2 using isolated tracks from the tracker have refined the energy resolution, leading to a constant term C from calibration uncertainties of below 3% and an energy-independent term B below 2% [45]. The energy resolution of isolated pions with track momenta between 40 and 60 GeV showering in the HCAL was measured to be 19.4%, 18.8%, and 23.6% in the barrel, endcap, and transition regions, respectively.

The very forward region $3.0 < |\eta| < 5.2$ is covered by the two forward sections (HF) made from steel with embedded quartz fibres, running parallel to the beamline. The fibres constitute the active medium, which captures the Cherenkov light of charged shower particles. The design is thus most sensitive to the electromagnetic component of the shower. To differentiate electrons and photons from hadrons, the fibres are of two lengths: 1.65 m "long fibres", which measure the full signal, and 1.40 m "short fibres", which capture energy deposition only after the initial 22 cm of the steel absorber. The two types of fibres are separated and bundled into towers of 0.175×0.175 ($\Delta\eta \times \Delta\phi$). Light guides direct the Cherenkov light through radiation shielding to the read-out box,

where it is picked up by silicon photomultipliers (SiPMs). The system is designed to be highly radiation-hard, capable of withstanding the expected dose of up to 10 MGy by the end of High-Luminosity operation. The energy resolution of the forward sections has been determined in test beams with incident electrons, pions and muons [46]. It is given by

$$\frac{\sigma_E}{E} = \frac{280\%}{\sqrt{E}} \oplus 11\% \quad , \tag{3.22}$$

with E in GeV.

Extrapolating the observed reduction in light yield of the plastic scintillating tiles to an integrated luminosity of 1,000 fb^{-1} during High-Luminosity operation revealed that the anticipated degradation in the endcap sections (HE) would significantly impair the physics performance. Consequently, a complete replacement of the endcaps is being implemented as part of the new High-Granularity Calorimeter (HGCAL).

3.5.3 High-Granularity Calorimeter (HGCAL)

As mentioned in the previous sections, both the electromagnetic and hadronic endcap calorimeters require a complete replacement. Beyond the stringent requirement for the new High-Granularity Calorimeter (HGCAL) to withstand the much higher integrated radiation levels arising from the conditions at the High-Luminosity LHC with an integrated luminosity of 3,000 fb^{-1}, the new system needs to fulfil the following additional criteria:

- **Dense calorimeter:** To limit the lateral shower extent and ensure adequate sampling.
- **Fine lateral granularity:** For narrow two-jet separation, accurate pileup rejection, and to allow a high signal-to-noise ratio for minimum ionizing particle (MIP) calibration.
- **Fine longitudinal granularity:** To enable fine shower sampling and consequently good electromagnetic energy resolution, which also aids in the minimisation of pileup contributions.
- **Precision timing of high-energy showers:** To measure the timing with high precision of each cell containing a significant energy deposition, further supporting pileup mitigation and the identification of the triggering interaction's vertex.
- **Provide primitives to the Level-1 Trigger system:** Allowing the HGCAL to contribute to the trigger decision.

The system will comprise two compartments, an electromagnetic (CE-E) and a hadronic one (CE-H), housed within a thermally insulated volume maintained at -30°C and cooled using a two-phase CO_2 system.

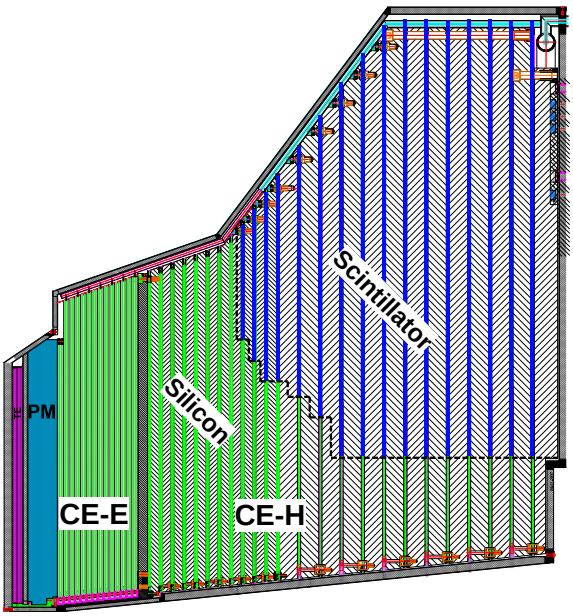


Figure 16: Schematic drawing of the upper half of one HGCAL endcap [47]. The electromagnetic compartment (CE-E) is followed by the hadronic compartment (CE-H). Silicon sensors are depicted in green, while scintillator tiles are shown in blue.

The electromagnetic compartment (CE-E) includes 28 sampling layers with a total thickness of 34 cm, corresponding to approximately $25\,X_0$ and $1.4\,\lambda_I$. Each active detector unit is a 163 mm wide hexagonal silicon sensor sandwiched between a 1.4 mm thick tungsten-copper (WCu, 75%/25%) baseplate and a printed circuit board (PCB) housing the front-end electronics. The sandwich structure forms a silicon module. Sensors with three sensitive thicknesses — 300, 200, and 120 μm — are deployed based on the radiation fluence in different regions. Modules are arranged on either side of a 6 mm thick copper cooling plate, which, along with the two WCu baseplates, constitutes one absorber layer. Alternate absorber layers use two 2.1 mm thick lead planes clad with 0.3 mm stainless steel sheets placed on either side of the module-cooling plate assembly. This structure is subdivided into 60°units or cassettes, and 14 such cassette layers form the complete 28 sampling layers [47].

The electromagnetic compartment (CE-E) will thus no longer function as a homogeneous calorimeter due to the requirement for radiation hardness. Its fine longitudinal segmentation is anticipated to result in a stochastic term of $\sim 25\%/\sqrt{E}$, which is relatively large when compared to the current homogeneous design (Eq. 3.20) but still acceptable for the High-Luminosity LHC's physics program. The energy resolution in a CE-E prototype has been measured in beam tests with incident positrons and was determined to be [48]

$$\frac{\sigma_E}{E} = \frac{22\%}{\sqrt{E}} \oplus 0.6\% \quad , \tag{3.23}$$

with E in GeV.

The absorber in the hadronic compartment (CE-H) consists of 12 layers of 35 mm thick stainless steel plates, followed by 12 layers of 68 mm thick stainless steel plates.

Silicon modules and scintillator tileboards are mounted on 6 mm thick copper cooling plates between these absorber layers, forming 30°cassettes similar to those in the CE-E. However, in the CE-H, sensors are mounted only on one side of the cooling plate, and absorbers are part of a separate mechanical structure. The total thickness of the calorimeter, including the CE-E and a polythene neutron moderator (PM) layer at the front, is $10.7\,\lambda_I$. The neutron moderator, with a thickness of 120 mm, reduces the neutron flux in the tracker.

All layers contribute to energy measurements, but only alternate layers in CE-E and all layers in CE-H are used to generate Level-1 trigger primitives.

In the CE-H, the transition radius from silicon sensors to plastic scintillators varies with the layer, as shown in Fig. 16. This transition is determined by radiation levels, ensuring that the scintillator's light loss due to radiation-induced damage remains below 50%, and neutron fluence stays under $8 \times 10^{13}\,\mathrm{neq/cm^2}$. This minimises the impact of increased silicon photomultiplier leakage current and light loss, maintaining low energy equivalent electronics noise, which is crucial for accurate minimum ionizing particle (MIP) response measurement.

The hadronic energy resolution of a prototype of the combined CE-E and CE-H system has been measured in charged pion beam tests and was found to be [49]

$$\frac{\sigma_E}{E} = \frac{129.1\%}{\sqrt{E}} \oplus 8.6\% \quad , \tag{3.24}$$

with E in GeV.

3.6 Muon system

3.6.1 Introduction

The mean energy loss or stopping power of high-energy muons can be described as

$$\left\langle -\frac{dE}{dx} \right\rangle = a(E) + b(E)E \quad , \tag{3.25}$$

where $a(E)$ is the electronic stopping power and $b(E)$ summarizes the radiative energy losses, direct e^+e^- pair production, bremsstrahlung and photonuclear interactions,

$$b \equiv b_{\mathrm{pair}} + b_{\mathrm{brems}} + b_{\mathrm{nucl}} \quad . \tag{3.26}$$

Muons are neither as short-lived as tau leptons, which decay within the tracker, nor light enough to experience substantial deflections by the nuclei in the calorimeters, which would otherwise result in significant cross-sections for bremsstrahlung and direct e^+e^- pair production. Additionally, muons do not interact via the strong force, limiting the cross-section for nuclear interactions to photonuclear processes. A comparison of mean energy losses of muons, electrons and charged pions is depicted in Fig. 17. The low stopping

power of muons leads to a large penetration depth, enabling them to pass through the inner layers of the CMS detector — tracker, electromagnetic calorimeter, hadron calorimeter and magnetic solenoid — with relatively little energy loss. Consequently, the muon system is positioned as the outermost component of the CMS detector, where it is responsible for measuring the trajectories of muons and inferring their momenta from the measurement.

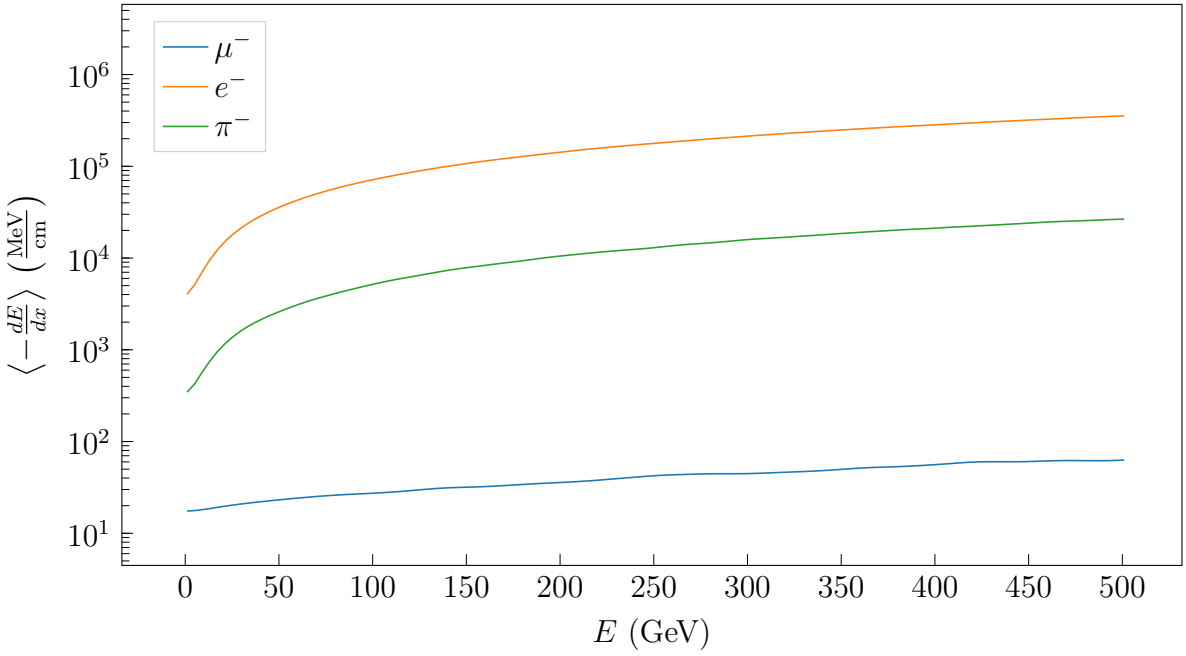


Figure 17: Mean stopping power of muons, electrons, and charged pions in brass with energies of 1-500 GeV. Obtained via a Geant4 simulation (FTFP_BERT) [38, 39, 40].

The muon's trajectory is determined by fitting a curve to the hits recorded in the four muon stations that form the muon system (Fig. 18). The particle's path is tracked across multiple active layers in each station. To enhance precision, these data are combined with measurements from the silicon tracker. The solenoid's magnetic field bends the trajectories of muons due to the Lorentz force. The bending radius is inversely proportional to the muon's momentum, meaning muons with low transverse momentum (p_T) experience greater curvature than those with high p_T. The muon stations are located in the gaps between the plates of the iron return yoke, which serves to confine and return the magnetic flux of the solenoid, ensuring a homogeneous magnetic field in the region outside the solenoid.

Various different types of gas detectors are employed in different sections of the muon system depicted in Fig. 18. These gas detectors provide a more cost-effective solution compared to silicon-based systems for covering the large outer volume of the detector.

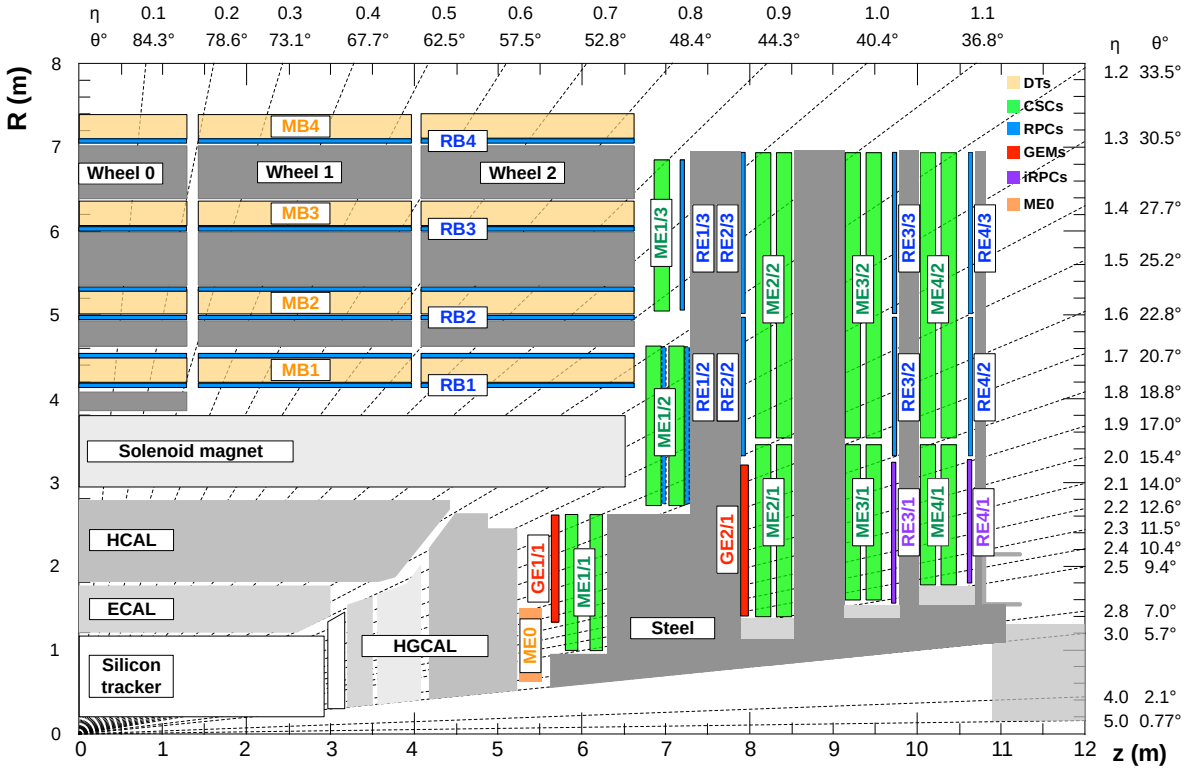


Figure 18: Schematic illustration of one quadrant of the muon system [50]. The different colours represent different muon detector subsystems, drift tubes (DTs) are marked in light orange, cathode strip chambers (CSCs) in green, resistive plate chambers (RPCs) in blue, gas electron multipliers (GEMs) in red, improved resistive plate chambers (iRPCs) in violet and the innermost gas electron multiplier station ME0 in orange. The design follows a layered structure, where the layers are referred to as stations. The first number in the labels indicates the station, starting with the innermost endcap muon detector ME0 and the innermost barrel muon detector MB1.

3.6.2 Resistive plate chambers (RPCs)

The CMS detector utilises resistive plate chambers (RPCs) in all barrel muon stations and most endcap muon modules. While they offer good spatial resolution, their standout feature is an extremely fast response time, significantly shorter than the 25 ns gap between consecutive bunch crossings. This response time allows resistive plate chambers (RPCs) to reliably assign the correct bunch crossing number to ionisation events occurring in the drift tubes (DTs) and cathode strip chambers (CSCs) while simultaneously contributing to the measurement of the muon's trajectory.

The CMS detector employs double-gap modules, which consist of two single-gap modules separated by readout strips, as illustrated in Fig. 19. A high voltage ($\sim$ 10 kV [30]) applied between the graphite electrodes generates electron avalanches when muons traversing the 2 mm wide gas gap produce ionisation clusters. If these clusters form at a sufficient distance from the anode, the avalanche can grow large enough for the induced signal to be detected by the readout strips. The high resistivity of the bakelite layers ($2\text{-}5\times10^{10}$ Ωcm), with a thickness of 2 mm each, effectively stops discharges between

electrodes after an avalanche has formed. The resistivity of the graphite electrodes ($\sim 10^5$ Ωcm) provides enough decoupling from the high-voltage source to enable the readout strips to capture the signal. To enhance detection efficiency and reduce voltage requirements, CMS's RPC modules feature two gas gaps that simultaneously influence the readout strips from both sides.

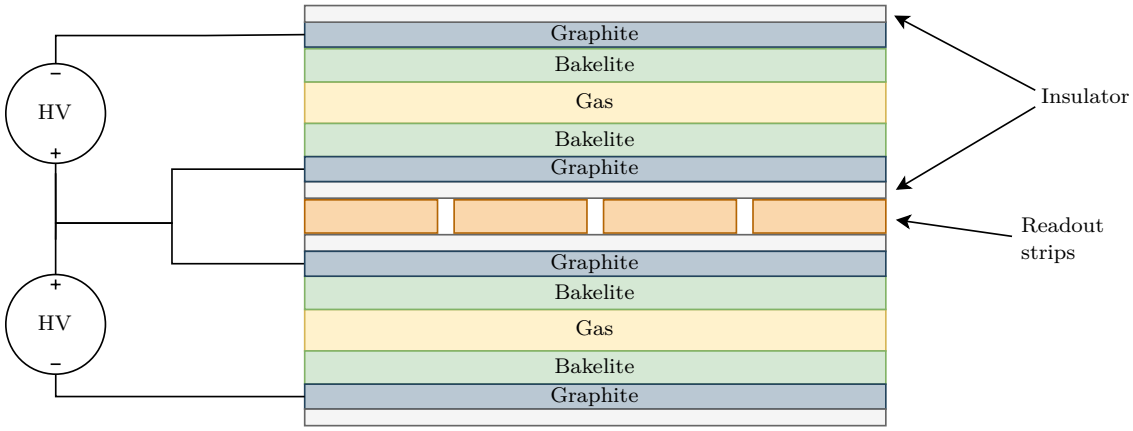


Figure 19: Schematic drawing of a resistive plate chamber (RPC) module.

The gas gaps are filled with the 3-component non-flammable mixture of 96.2% R134a ($C_2H_2F_4$), 3.5% iC_4H_{10} and 0.3% SF_6, with added water vapour to maintain a relative humidity of approximately 45% [30]. The main component R134a is responsible for providing a high density of ionisation clusters, while the additives iC_4H_{10} and SF_6 serve to capture UV photons and electrons, respectively, effectively limiting avalanche growth and quenching discharges.

In the barrel region, the RPC readout strips are oriented parallel to the beamline, with the modules distributed across five cylindrical wheels. In the endcap region, the modules have a trapezoidal shape, with their strips arranged radially and segmented into up to three radial sections (Fig. 18).

The improved resistive plate chambers (iRPCs) for the High-Luminosity operation of the LHC will be installed in the two outer muon endcap stations (Fig. 18) in the region $|\eta| > 1.9$, where the highest hit rates are expected. Their design largely follows that of their predecessor. Enhancements, including adjustments to the gas gap width, modifications to the thickness and materials of the high-resistivity layers and a decrease in readout strip pitch, enable improved timing and spatial resolution to handle the high hit rates close to the beam line.

3.6.3 Cathode strip chamber (CSC) system

The endcap regions ($0.9 < |\eta| < 2.4$) are instrumented with cathode strip chambers (CSCs), which are essentially multiwire proportional chambers with finely segmented cathode strip readouts. Each CSC module consists of seven parallel plates, six of which have one side coated with copper and are segmented into radially oriented

strips, functioning as cathodes. Anode wires are strung in the azimuthal (ϕ) direction, perpendicular to the cathode strips, within the six gaps formed between these plates. A single such layer is depicted in Fig. 20. The gaps are filled with a gas mixture of 40% argon, 50% CO_2, and 10% CF_4. Argon serves as the working gas, responsible for generating ionisation clusters, while CO_2 and CF_4 act as quenchers, absorbing secondary UV photons produced during electron avalanches.

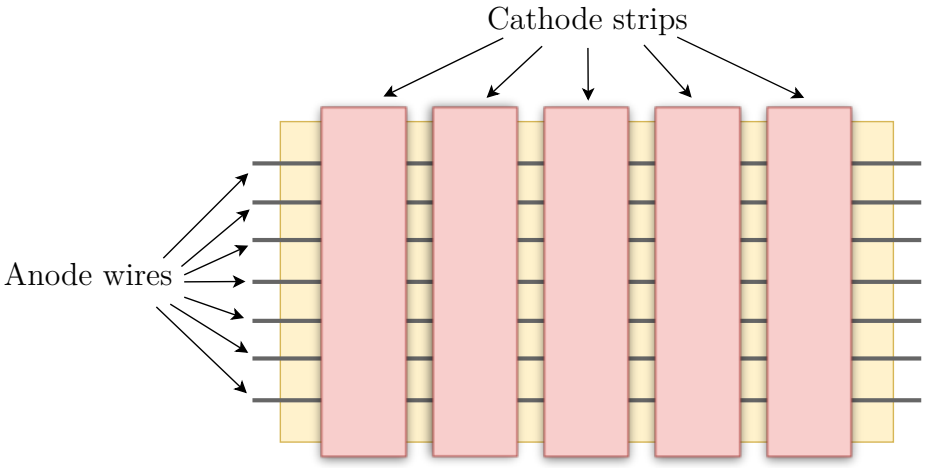


Figure 20: Schematic illustration of a cathode strip chamber (CSC) module. Note: In the innermost CSC modules, the wires are slightly tilted with respect to the axis perpendicular to the strips in order to counteract the Lorentz force due to the magnetic field.

The working principle is similar to that of an RPC. As a muon traverses the gas gap (7 or 9.5 mm), it produces ionisation clusters. A high voltage (2.9 or 3.6 kV), typically lower than in an RPC, is applied between the cathode strips and anode wires, creating a strong electric field across the gap. Near the thin anode wires, the field strength increases rapidly with the inverse of the radius ($E \propto 1/r$), triggering the ionisation clusters to develop into electron avalanches. When the avalanches reach the anode, a signal is detected in the form of collected electric charge, while the cathode strips capture the opposing positive charge. Measuring both opposing charges allows for accurately determining the hit position in the R-ϕ plane. Using multiple gas gaps further refines the trajectory measurement.

For High-Luminosity operation, the cathode strip chambers (CSCs) underwent an electronics upgrade to accommodate the increased Level-1 trigger latency and rate. This update was completed during the long shutdown prior to Run 3.

3.6.4 Gas electron multiplier (GEM) system

The high hit rate of above 5 kHz/cm^2 expected during High-Luminosity operation in the innermost muon layers close to the beamline ($|\eta| > 1.55$) exceeds the capabilities of traditional gas detectors — cathode strip chambers, resistive plate chambers and drift tubes. Silicon-based sensors, on the other hand, would be impractically expensive for covering the large muon detection area. The emergence of a new generation of

micro-pattern gaseous detectors (MPGDs), capable of operating at significantly higher rates, has opened up new opportunities for developing cost-effective muon detectors tailored for use in the most forward region of the CMS detector.

A gas electron multiplier (GEM) is a thin polymer foil with metal cladding, featuring a high density of chemically etched microscopic holes. The foil's bulk material is a 50 μm thick polyimide such as Kapton. It is coated with a 5 μm layer of copper on both sides. The holes in the GEM have a truncated double-cone shape, with outer diameters of approximately 70 μm, inner diameters around 50 μm, and are arranged in a hexagonal pattern with a pitch of 140 μm.

In the GEM modules used by CMS, three GEM layers are stacked with gas gaps of a few millimetres thickness in between, all enclosed between a cathode layer and a readout PCB with anode strip traces, as shown in Fig. 21. A progressively increasing electric potential is applied across the seven electrodes, from the readout PCB to the cathode. Electrons produced in the drift region are accelerated by the electric field toward the anode. As an electron passes through a GEM hole, it triggers an avalanche, generating additional secondary electrons and effectively amplifying the signal. By using three such GEM layers in succession, the signal undergoes further amplification, achieving a typical gain of around 15,000 relative to the initial signal. Once the electron avalanche reaches the anode, it is detected as deposited charge. The main purpose of the GEM foils is the confined electric field — and thus amplification — within their holes, thereby limiting amplification in the drift region and mitigating potentially destructive discharges that would otherwise result from the higher voltages required across that region.

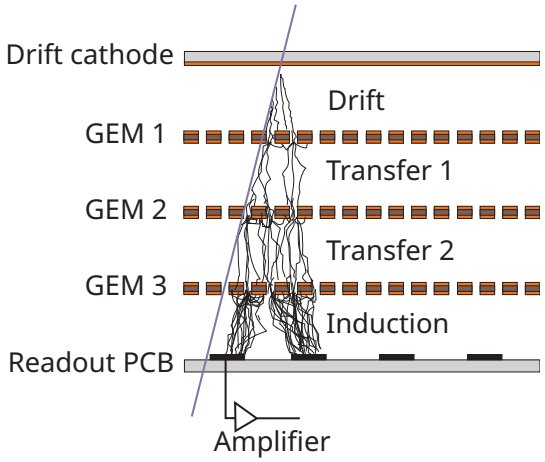


Figure 21: Schematic drawing of a triple-GEM module employed by CMS [51].

The readout strips on the board of each trapezoidal module are radially oriented and separated into eight η-sectors by a gap on the PCB. The innermost muon station, ME0, consists of six consecutive modules stacked together, totalling 36 stacks across both endcaps. In contrast, the two systems further outwards, ME1/1 and ME2/1, each consist of 72 chambers, with each chamber comprising four modules. Both systems feature a "front" layer followed by a "back" layer. The gas gaps in each module are filled with a mixture of 70% argon and 30% CO_2 [51].

The GEM modules play a crucial role in the physics program of the High-Luminosity LHC by enabling the precise measurement of low p_T muons close to the beamline. A feature especially of interest for the lepton flavour violating decay $\tau \to 3\mu$ [51].

3.6.5 Drift tube (DT) system

In the barrel section a total of 250 drift tubes are employed to cover the pseudorapidity region $|\eta| < 1.2$. This provides a more cost-effective solution compared to cathode strip chambers (CSCs), which are reserved for regions with higher hit rates.

Unlike cathode strip chambers, where multiple anode wires are housed within a single gas chamber, drift tubes are divided into multiple smaller gas volumes, each containing a single anode wire. Within a drift tube, most of the volume serves as the drift region, while near the anode wire, the electric field lines converge, enabling the formation of electron avalanches that amplify the signal. When a muon passes through the gas volume, ionisation clusters form and the free electrons drift toward the anode wire. Assuming the amplification region is small compared to the drift region, the muon track's position relative to the wire can be precisely determined by measuring the drift time. In CMS's drift tubes (Fig. 22), the cathodes and electrode strips (top and bottom) are set to voltages of -1,200 V and 1,800 V, respectively, while the anode wires operate at voltages between 3,500 and 3,600 V. This electric field configuration establishes a linear relationship between the drift time and the distance from the anode wire.

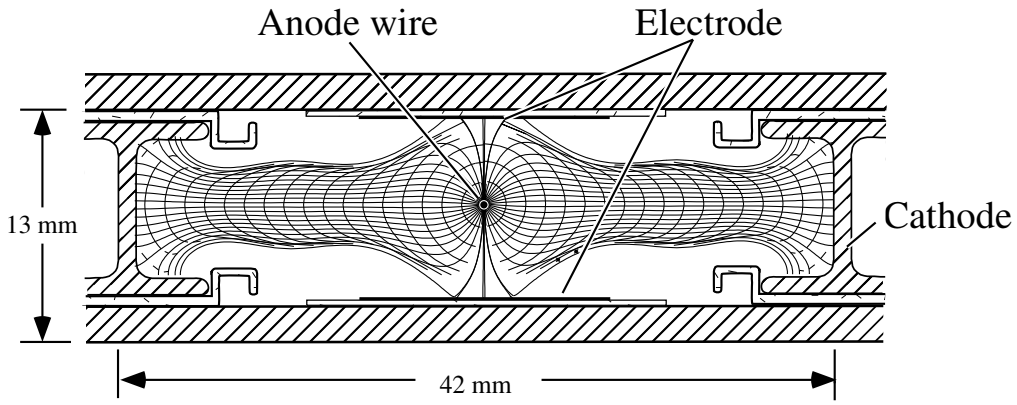


Figure 22: Illustration of a CMS drift tube [52], showing the electric field configuration with isochrone lines that represent equal drift times to the anode along each line.

The maximum drift path of 21 mm together with the operating gas mixture of 85% argon and 15% CO_2 corresponds to a maximum drift time of 380 ns [30]. A compromise that results in a negligible occupancy of multi-hits while keeping the number of active channels to an affordable value.

The individual drift cells are organised into layers, where four such layers constitute a super layer (SL). Up to three super layers are stacked with a spacer in between SL2 and SL3 to make up a chamber (Fig. 23). Two super layers track the muon in the R-ϕ plane while the super layer in the middle provides a measurement in the R-z plane. The chambers are mounted to one of five wheels constituting the barrel muon system,

where each of the twelve ϕ-sections contains four chambers (MB1, MB2, MB3, MB4), as illustrated in Fig. 18.

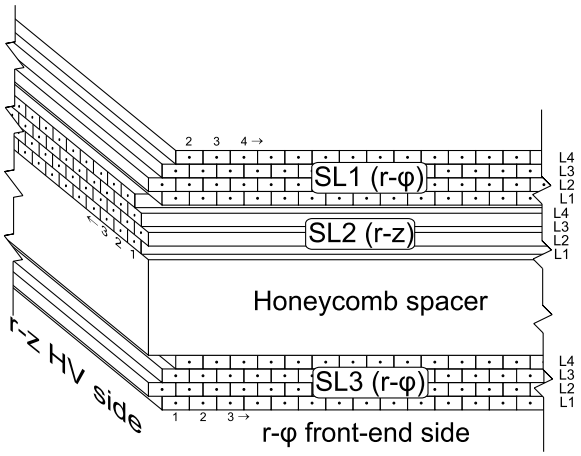


Figure 23: Schematic drawing of three drift tube super layers (SLs), each comprising four layers of drift tubes [53].

The use of four drift cell layers per super layer (SL) is essential for accurately determining the hit position. Although only three layers are needed in principle, the fourth layer enhances reliability, providing robustness in cases where a hit is missing or incorrect.

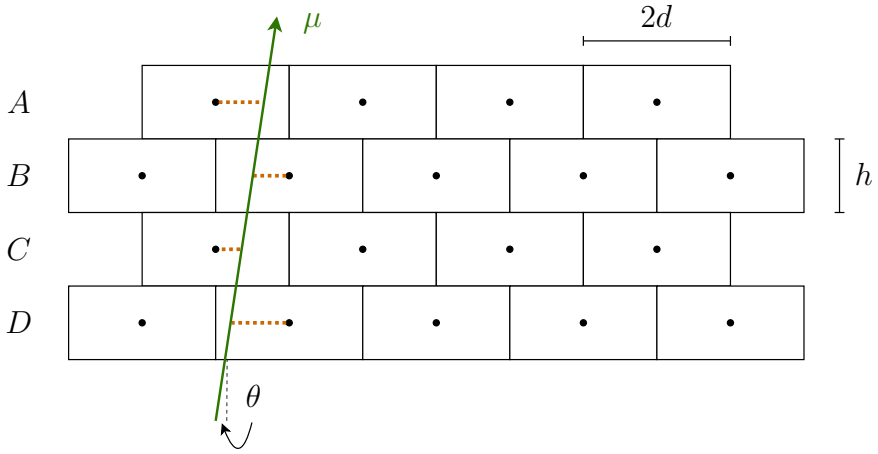


Figure 24: Schematic illustration of using drift time coincidence across drift tube layers to identify tracks.

By using basic trigonometry, we find for any three adjacent layers

$$d_A = d_C + 2h\tan\theta \tag{3.27}$$

$$d_B = d - h\tan\theta - d_C \quad . \tag{3.28}$$

Hence,

$$\frac{d_A + 2d_B + d_C}{2} = d \quad . \tag{3.29}$$

For non-adjacent layers, we can similarly find

$$d_A = d - d_D + 3h \tan\theta \tag{3.30}$$
$$d_B = d_D - 2h \tan\theta. \tag{3.31}$$

Hence,

$$\frac{2d_A + 3d_B - d_D}{2} = d \quad . \tag{3.32}$$

With similar equations for hits in the layers BCD and ACD. Since the drift time depends linearly on the distance, that is

$$T_i \propto d_i \qquad i = A, B, C, D \quad , \tag{3.33}$$

we can register the signal in a shift register. After each clock cycle (80 MHz during Run 3), the signal is shifted to the next register. Using Eq. 3.29 and Eq. 3.32 we can search for signal coincidences among the shift registers, meaning we find a coincidence in ABC when the distance of the signal in shift register A plus twice the distance in shift register B plus the distance in shift register C equals twice the constant distance d. Detecting such a coincidence not only allows us to determine the drift start time and, consequently, the muon position but also provides a trigger to identify the corresponding bunch crossing. This approach is known as the Mean-Timer technique [54], as it leverages the constant drift time $T_{\mathrm{MAX}} \propto d$ from either the left or right wall of the drift tube cell to the anode [55].

Since this process is inherently noisy and requires tolerances due to drift times not being perfectly aligned with the clock, the data from multiple super layers (SLs) within a chamber are correlated, significantly improving the tracking precision. The measurement and bunch crossing identification can be further refined by combining it with information from the RPCs.

While for High-Luminosity operation the drift tubes themselves will remain, an extensive electronics upgrade is planned to be carried out in the long shutdown after Run 3, accommodating the increased Level-1 trigger latency and rate.

3.7 Kinematics

3.7.1 Angular distance ΔR

The angular distance in hadron colliders is typically defined in a way that is invariant under Lorentz boosts along the beam axis. As discussed in Section 3.3 this is fulfilled by leveraging the relativistic rapidity y, which for negligible rest masses can be approximated by the pseudorapidity η. Hence, the angular distance in hadron colliders takes the form,

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \quad . \tag{3.34}$$

With $\Delta\eta = |\eta_1 - \eta_2|$ and $\Delta\phi$ the smallest azimuthal angle difference, taking into account that ϕ wraps around i.e. $\phi = -\pi = \pi$.

3.7.2 Invariant mass

The invariant mass represents the portion of a system's mass that is independent of its motion. More specifically, it is a Lorentz-invariant quantity derived from the system's energy and momentum. In the centre-of-momentum reference frame, if such a frame exists, the invariant mass equals the system's total mass. Using relativistic kinematics and four-vector notation, one can express the invariant mass via a system's total four-momentum as

$$M^2 = P_\mu P^\mu \quad . \tag{3.35}$$

Where P_μ is the covariant total four-momentum of the system and P^μ it's contravariant, $P^\mu = g^{\mu\nu} P_\nu$. In the following, we adopt the mostly-minus metric convention

$$g^{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad . \tag{3.36}$$

For a system of N particles, the total four-momentum is given by

$$P_\mu = \sum_{i=1}^{N} p_\mu^{(i)} \quad . \tag{3.37}$$

3.7.3 Two-body invariant mass

The two-body invariant mass in CMS's coordinate system $(p_T, \eta, \phi)^T$ can be expressed as,

$$\begin{aligned} M^2 = &\begin{pmatrix} \sqrt{p_{T_1}^2 \cosh^2(\eta_1) + m_1^2} + \sqrt{p_{T_2}^2 \cosh^2(\eta_2) + m_2^2} \\ p_{T_1}\cos(\phi_1) + p_{T_2}\cos(\phi_2) \\ p_{T_1}\sin(\phi_1) + p_{T_2}\sin(\phi_2) \\ p_{T_1}\sinh(\eta_1) + p_{T_2}\sinh(\eta_2) \end{pmatrix} \cdot \\ &\begin{pmatrix} \sqrt{p_{T_1}^2 \cosh^2(\eta_1) + m_1^2} + \sqrt{p_{T_2}^2 \cosh^2(\eta_2) + m_2^2} \\ -p_{T_1}\cos(\phi_1) - p_{T_2}\cos(\phi_2) \\ -p_{T_1}\sin(\phi_1) - p_{T_2}\sin(\phi_2) \\ -p_{T_1}\sinh(\eta_1) - p_{T_2}\sinh(\eta_2) \end{pmatrix} \quad . \end{aligned} \tag{3.38}$$

Multiplying out the dot product and using the identities $\sin^2(x) + \cos^2(x) = 1$ and $\cosh^2(x) - \sinh^2(x) = 1$ to cancel out $p_{T_1}^2$ and $p_{T_2}^2$ yields

$$\begin{aligned} M^2 = m_1^2 + m_2^2 - 2p_{T_1}p_{T_2}\sinh(\eta_1)\sinh(\eta_2) + \\ 2\sqrt{\left(p_{T_1}^2\cosh^2(\eta_1) + m_1^2\right)\left(p_{T_2}^2\cosh^2(\eta_2) + m_2^2\right)} - \\ 2p_{T_1}p_{T_2}\left(\cos(\phi_1)\cos(\phi_2) + \sin(\phi_1)\sin(\phi_2)\right) \quad . \end{aligned} \tag{3.39}$$

We can neglect the rest masses of the constituents m_1 and m_2 in Eq. 3.39 due to them being generally much smaller than their kinetic energy, which gives

$$\begin{aligned} M^2 = 2p_{T_1}p_{T_2}\left(\cosh(\eta_1)\cosh(\eta_2) - \sinh(\eta_1)\sinh(\eta_2)\right) - \\ 2p_{T_1}p_{T_2}\left(\cos(\phi_1)\cos(\phi_2) + \sin(\phi_1)\sin(\phi_2)\right) \quad . \end{aligned} \tag{3.40}$$

By further using the identities $\cosh(x-y) = \cosh x \cosh y - \sinh x \sinh y$ and $\cos(x-y) = \cos x \cos y + \sin x \sin y$, we arrive at the well-known equation in high-energy physics,

$$M^2 = 2p_{T_1}p_{T_2}\left(\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2)\right) \quad . \tag{3.41}$$

From Eq. 3.41 we can see that the invariant mass depends only on the differences in azimuthal angles $\Delta\phi = \phi_1 - \phi_2$ and in pseudorapidities $\Delta\eta = \eta_1 - \eta_2$ and is scaled by the product of transverse momenta p_{T_1}, p_{T_2}. Lastly, to get a better understanding of Eq. 3.41 we'll examine one of the edge cases. Imagine that the constituents are flying in the same direction i.e. $\eta_1 = \eta_2$, $\phi_1 = \phi_2$ or $\Delta R = 0$, in this case, the term $\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2) = 1 - 1$ vanishes and the invariant mass becomes 0. This means that the two constituents cannot come from the decay of a hypothetical mother particle, as they cannot "carry away" the rest mass of the mother particle without violating energy-momentum conservation. It should be noted that highly boosted mother particles will similarly exhibit small azimuthal angle and pseudorapidity differences.

Furthermore, one can define related quantities like $\frac{M}{\Delta R}$ to better separate heavy boosted mother particles from lighter unboosted ones.

3.7.4 Three-body invariant mass

The derivation of the three-body invariant mass is analogous to the two-body version. In this case, we'll directly neglect the rest masses m_1, m_2 and m_3. Analogously to Eq. 3.38 we have

$$\begin{aligned} M^2 = &\begin{pmatrix} p_{T_1}\cosh(\eta_1) + p_{T_2}\cosh(\eta_2) + p_{T_3}\cosh(\eta_3) \\ p_{T_1}\cos(\phi_1) + p_{T_2}\cos(\phi_2) + p_{T_3}\cos(\phi_3) \\ p_{T_1}\sin(\phi_1) + p_{T_2}\sin(\phi_2) + p_{T_3}\sin(\phi_3) \\ p_{T_1}\sinh(\eta_1) + p_{T_2}\sinh(\eta_2) + p_{T_3}\sinh(\eta_3) \end{pmatrix} \cdot \\ &\begin{pmatrix} p_{T_1}\cosh(\eta_1) + p_{T_2}\cosh(\eta_2) + p_{T_3}\cosh(\eta_3) \\ -p_{T_1}\cos(\phi_1) - p_{T_2}\cos(\phi_2) - p_{T_3}\cos(\phi_3) \\ -p_{T_1}\sin(\phi_1) - p_{T_2}\sin(\phi_2) - p_{T_3}\sin(\phi_3) \\ -p_{T_1}\sinh(\eta_1) - p_{T_2}\sinh(\eta_2) - p_{T_3}\sinh(\eta_3) \end{pmatrix} \quad . \end{aligned} \tag{3.42}$$

The calculation is a bit more cumbersome. We again have terms $p_{T_1}^2$, $p_{T_2}^2$, $p_{T_3}^2$ from the identity $\cosh^2(x) - \sinh^2(x) = 1$ and terms $-p_{T_1}^2$, $-p_{T_2}^2$, $-p_{T_3}^2$ from the identity

$\sin^2(x) + \cos^2(x) = 1$ that cancel. The remaining mix terms are

$$\begin{aligned} M^2 = {} & 2p_{T_1}p_{T_2}\left(\cosh(\eta_1)\cosh(\eta_2) - \sinh(\eta_1)\sinh(\eta_2)\right) - \\ & 2p_{T_1}p_{T_2}\left(\cos(\phi_1)\cos(\phi_2) + \sin(\phi_1)\sin(\phi_2)\right) + \\ & 2p_{T_1}p_{T_3}\left(\cosh(\eta_1)\cosh(\eta_3) - \sinh(\eta_1)\sinh(\eta_3)\right) - \\ & 2p_{T_1}p_{T_3}\left(\cos(\phi_1)\cos(\phi_3) + \sin(\phi_1)\sin(\phi_3)\right) + \\ & 2p_{T_2}p_{T_3}\left(\cosh(\eta_2)\cosh(\eta_3) - \sinh(\eta_2)\sinh(\eta_3)\right) - \\ & 2p_{T_2}p_{T_3}\left(\cos(\phi_2)\cos(\phi_3) + \sin(\phi_2)\sin(\phi_3)\right) \quad . \end{aligned} \tag{3.43}$$

By again using the argument addition identities of cosh and cos we can write this as

$$\begin{aligned} M^2 = {} & 2p_{T_1}p_{T_2}\left(\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2)\right) + \\ & 2p_{T_1}p_{T_3}\left(\cosh(\eta_1 - \eta_3) - \cos(\phi_1 - \phi_3)\right) + \\ & 2p_{T_2}p_{T_3}\left(\cosh(\eta_2 - \eta_3) - \cos(\phi_2 - \phi_3)\right) \quad . \end{aligned} \tag{3.44}$$

Hence, Eq. 3.44 is simply the sum of all permutations of Eq. 3.41,

$$M^2 = M_{123}^2 = M_{12}^2 + M_{13}^2 + M_{23}^2 \quad . \tag{3.45}$$

3.7.5 Transverse mass

In particle physics, the term transverse mass can have two meanings. For single particles, it refers to an invariant under Lorentz boosts in z-direction, which, analogously to the invariant mass ($m^2 = E^2 - p^2$) may be defined as

$$m_T^2 \equiv E^2 - p_z^2 = m^2 + p_x^2 + p_y^2 = m^2 + p_T^2 \quad . \tag{3.46}$$

Yet, in this section, we will focus on the correlational transverse mass, whose significance stems from the fact that it is possible to infer the transverse component of the missing momentum in hadron collider detectors, commonly called the missing transverse momentum. It arises from one or more particles that cannot be directly detected, but whose transverse momentum p_T can be deduced from the visible momenta through the principle of conservation of momentum (Fig. 25).

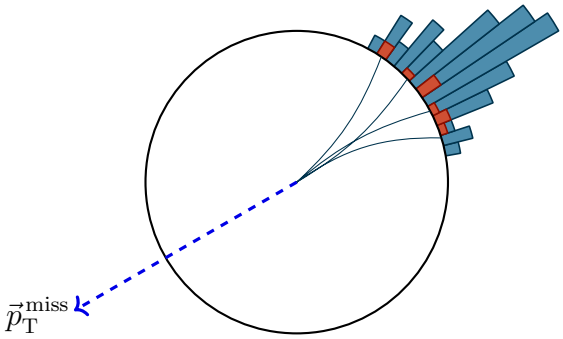


Figure 25: Illustration of jet constituents in the tracker, electromagnetic calorimeter and hadron calorimeter. Balancing the transverse momentum of the jet requires the presence of undetectable missing transverse momentum $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$. Modified from [33].

Due to the generally unknown longitudinal momentum of generated particles (c.f. Section 3.3), it is not possible to infer the longitudinal component of the missing momentum. As a result, calculating an invariant mass from the missing momentum is not feasible, requiring alternative quantities to extract meaningful constraints. One such quantity is obtained by assuming the absence of a longitudinal component altogether (c.f. Eq. 3.38), leading to the correlational transverse mass. For two final-state particles, it is defined as

$$M_T^2 = \begin{pmatrix} \sqrt{p_{T_1}^2 + m_1^2} + \sqrt{p_{T_2}^2 + m_2^2} \\ p_{T_1}\cos(\phi_1) + p_{T_2}\cos(\phi_2) \\ p_{T_1}\sin(\phi_1) + p_{T_2}\sin(\phi_2) \\ 0 \end{pmatrix} \cdot \begin{pmatrix} \sqrt{p_{T_1}^2 + m_1^2} + \sqrt{p_{T_2}^2 + m_2^2} \\ -p_{T_1}\cos(\phi_1) - p_{T_2}\cos(\phi_2) \\ -p_{T_1}\sin(\phi_1) - p_{T_2}\sin(\phi_2) \\ 0 \end{pmatrix} \quad . \tag{3.47}$$

Neglecting the rest masses m_1, m_2 in Eq. 3.47 and multiplying out the dot product leads to

$$M_T^2 = 2p_{T_1}p_{T_2}\left(1 - \cos(\phi_1 - \phi_2)\right) \quad . \tag{3.48}$$

As $\cosh(\eta_1 - \eta_2) \geq 1$ in Eq. 3.41 we can conclude that the transverse mass M_T provides a lower bound for the invariant mass M. Evidently, in the case where $\cosh(\eta_1 - \eta_2) = 1$ the transverse mass and invariant mass coincide. This fact can be leveraged when measuring the transverse mass spectrum of a decay. The peak at the higher end of the transverse mass spectrum corresponds to the invariant mass, while the tail towards the higher end can be used to estimate the decay width. This was in fact where the transverse mass saw its first application during the study of the process $W \to l\nu$ [56].

Analogously to the three-body invariant mass M, the three-body transverse mass M_T can be derived from

$$M_T^2 = \begin{pmatrix} \sqrt{p_{T_1}^2 + m_1^2} + \sqrt{p_{T_2}^2 + m_2^2} + \sqrt{p_{T_3}^2 + m_3^2} \\ p_{T_1}\cos(\phi_1) + p_{T_2}\cos(\phi_2) + p_{T_3}\cos(\phi_3) \\ p_{T_1}\sin(\phi_1) + p_{T_2}\sin(\phi_2) + p_{T_3}\sin(\phi_3) \\ 0 \end{pmatrix} \cdot \begin{pmatrix} \sqrt{p_{T_1}^2 + m_1^2} + \sqrt{p_{T_2}^2 + m_2^2} + \sqrt{p_{T_3}^2 + m_3^2} \\ -p_{T_1}\cos(\phi_1) - p_{T_2}\cos(\phi_2) - p_{T_3}\cos(\phi_3) \\ -p_{T_1}\sin(\phi_1) - p_{T_2}\sin(\phi_2) - p_{T_3}\cos(\phi_3) \\ 0 \end{pmatrix} \quad . \tag{3.49}$$

Which when neglecting the rest masses m_1, m_2 and m_3 leads to

$$\begin{aligned} M_T^2 = {} & 2p_{T_1}p_{T_2}\left(1 - \cos(\phi_1 - \phi_2)\right) + \\ & 2p_{T_1}p_{T_3}\left(1 - \cos(\phi_1 - \phi_3)\right) + \\ & 2p_{T_2}p_{T_3}\left(1 - \cos(\phi_2 - \phi_3)\right) \quad . \end{aligned} \tag{3.50}$$

As is the case for the three-body invariant mass, this is equivalent to the sum of two-body permutations

$$M_T^2 = M_{T_{123}}^2 = M_{T_{12}}^2 + M_{T_{13}}^2 + M_{T_{23}}^2 \quad . \tag{3.51}$$

3.7.6 Combined two-particle transverse momentum P_T

The combined two-particle transverse momentum P_T is the transverse momentum of a hypothetical mother particle decaying into two daughters. It can be easily calculated via vector addition in the xy-plane,

$$P_x = p_{T_1} \cos(\phi_1) + p_{T_2} \cos(\phi_2) \tag{3.52}$$
$$P_y = p_{T_1} \sin(\phi_1) + p_{T_2} \sin(\phi_2) \quad . \tag{3.53}$$

The combined transverse momentum, then obviously, is

$$P_T^2 = P_x^2 + P_y^2 \quad . \tag{3.54}$$

Plugging Eq. 3.52 and Eq. 3.53 into Eq. 3.54 leads to

$$P_T^2 = p_{T_1}^2 + p_{T_2}^2 + 2 p_{T_1} p_{T_2} \cos(\phi_1 - \phi_2) \quad . \tag{3.55}$$

In general, the combined two-particle transverse momentum can be used to differentiate between highly boosted and low-boosted particles, providing clues about potential decay chains. For instance, various beyond the Standard Model theories predict very heavy particles that decay into highly boosted intermediate decay products.

4 The Level-1 Trigger system

4.1 Introduction

As discussed in Section 3.2, the substantial increase in superimposed proton-proton collisions during High-Luminosity operation demands a more advanced trigger architecture to efficiently separate interesting physics events from the overwhelming background. Although the latency budget will be slightly extended to 12.5 μs (compared to 3.8 μs in Run 3), the stringent low-latency requirements still constrain the design to application-specific integrated circuits (ASICs) in the detector front-end and high-performance field-programmable gate arrays (FPGAs) in the back-end trigger stages, located outside the radiation environment of the CMS detector cavern.

The new system adopts a local-to-global pipeline, progressively combining spatial and subdetector-specific information across successive trigger stages (Fig. 26). High-speed optical fibres (up to 25 Gb/s) facilitate communication between boards across consecutive stages [50]. The final stage of this pipeline, the Global Trigger, is responsible for aggregating and correlating all reconstructed trigger objects. It ultimately decides whether an event should be accepted by the Level-1 Trigger and forwarded to the software-based High-Level Trigger (HLT) [57]. The upgraded data acquisition system will enable an increase in the allowed acceptance rate to 750 kHz, up from 100 kHz in Run 3, permitting approximately every 42$^{\text{nd}}$ event to be accepted.

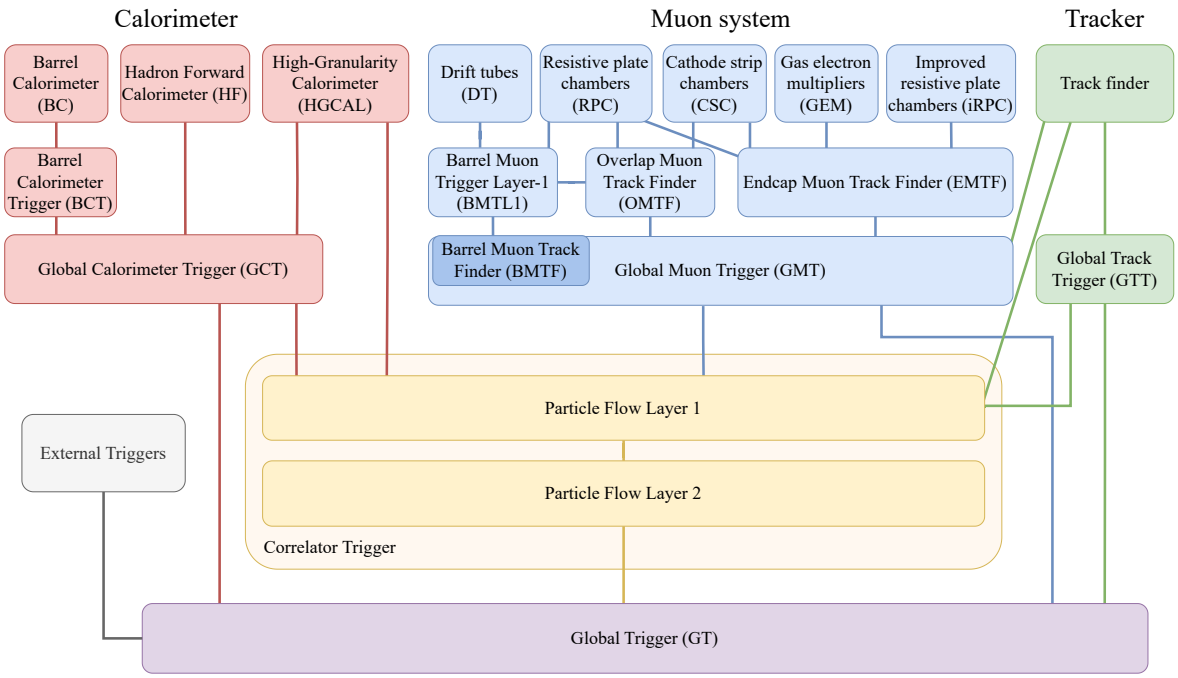


Figure 26: Schematic overview of the Level-1 Trigger system during High-Luminosity operation, where the first layer from the subdetector front-ends provides so-called Trigger primitives as inputs to subsequent trigger stages.

A central element of the new system is the inclusion of tracking information enabled by the novel p_T-modules described in Section 3.4. This data will be utilised by the redesigned Global Muon Trigger (GMT) to improve the muon reconstruction efficiency and by the new Correlator Trigger, which employs particle-flow reconstruction to mitigate pileup contributions and reconstruct higher-quality trigger objects [58]. Additionally, the track information will be consolidated by the Global Track Trigger (GTT) into trigger objects and propagated to the Global Trigger, where correlations with other objects can be performed.

4.2 Calorimeter Trigger

4.2.1 Barrel Calorimeter (BC)

Trigger primitives for the electromagnetic barrel calorimeter are generated in the detector back-end by barrel calorimeter processor boards (BCPs) using data transmitted from twelve front-end cards, each covering a 5×5 array of scintillating crystals. Similarly, hadron barrel calorimeter primitives are produced by BCPs in the detector back-end, with each BCP handling all η-segments and 4 ϕ-segments, resulting in a total of 18 barrel calorimeter processors [59]. Both electromagnetic and hadronic Barrel Calorimeter (BC) primitives are processed by one of 36 Barrel Calorimeter Triggers (BCTs), also known as regional calorimeter triggers (RCTs) [60].

Each Barrel Calorimeter Trigger (BCT) covers a region of 1.479×0.348 ($\Delta\eta \times \Delta\phi$) and is tasked with generating electron/photon clusters and hadron calorimeter towers. The reconstruction is achieved in a two-step process starting from a seed crystal with $p_T > 1$ GeV where clusters in the Electromagnetic Calorimeter (ECAL) are formed as 3×5 ($\Delta\eta \times \Delta\phi = 0.0525 \times 0.0875$) arrays of crystals around the seed. The precise position is determined via barycentric energy-weighting and then fixed to the closest crystal position. Bremsstrahlung can be accounted for by extending the cluster in ϕ direction with another 3×5 window, either above or below the initial cluster.

Each cluster is assigned an isolation value, calculated as the sum of the individual p_T of all crystals within a 27×27 region around the seed crystal, excluding crystals belonging to the core cluster. This region corresponds to an isolation cone size of approximately $\Delta R = 0.3 - 0.4$. The isolation calculation relies solely on ECAL information, and with a minimum p_T threshold of 0.5 GeV per crystal, it does not require pileup corrections.

4.2.2 Hadron Forward Calorimeter (HF)

Hadron Forward Calorimeter (HF) trigger primitives retain their current Run 3 definition, using signals from long and short fibres sampled at 40 MHz to determine tower energy. The transverse energy E_T reconstruction algorithm suppresses anomalous signals from charged particles hitting photomultiplier tube windows. Two feature bits are included: one identifies electromagnetic shower signatures based on the long-to-short fibre energy ratio, and the other is an ADC-over-threshold indicator for minimum-bias triggers [50].

4.2.3 High-Granularity Calorimeter (HGCAL)

The calorimeter trigger primitive generator processes raw input data from the High-Granularity Calorimeter (HGCAL) silicon and scintillator sections. In the silicon section, energies are summed into trigger cells (~ 4 cm^2), while in the scintillator section, trigger cells are groups of tiles covering 2.5 degrees azimuthally (4-10 cm dimensions). The charge in trigger cells is compressed to 7 bits using a floating-point format without timing information due to bandwidth constraints. Only trigger cells above an energy threshold (1-2 MIP/$\sin\theta$) are transmitted, with unselected channels summed within each HGCAL module to compensate for energy loss.

Data reduction, including thresholding and summation, is handled by custom ASICs, which send compressed data to back-end electronics over ~6,000 10 Gb/s optical links. The back-end system consists of a two-stage FPGA architecture. Stage 1 FPGAs perform data repacking and calibration, and send the processed data on 16 Gb/s links to Stage 2 (Fig. 27). Each Stage 2 board covers one-third of an endcap (120 degrees) and processes one out of 18 bunch crossings using time-multiplexed inputs. Boundary data are duplicated for seamless processing. The design enables large-area data processing and the implementation of 3D clustering algorithms.

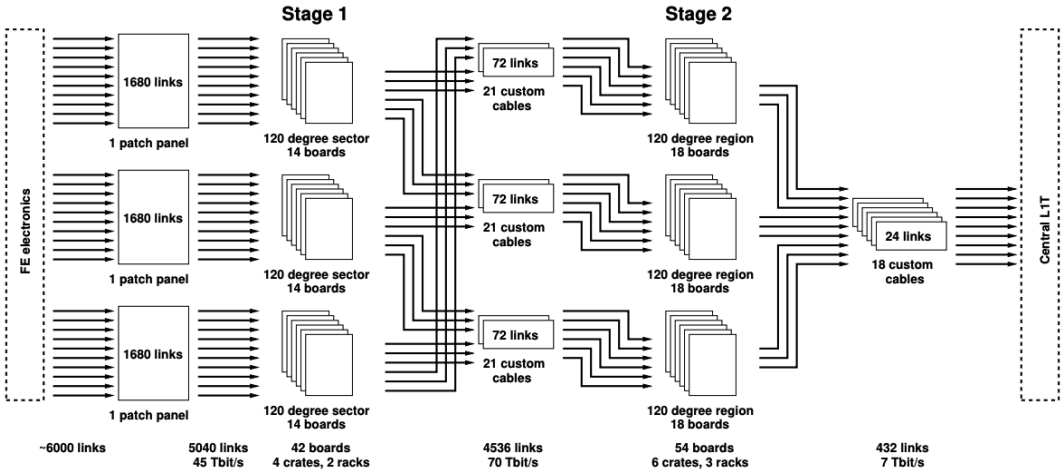


Figure 27: Schematic drawing of the HGCAL trigger primitive generator system for each endcap. The setup comprises two processing stages, with an identical copy for the other endcap. Stage 1 collects trigger data from on-detector electronics and sends it to Stage 2 via time-multiplexed connections. Stage 2 interfaces with the Central Level-1 Trigger (L1T) [61].

In terms of reconstruction, primitives are generated from seeds with energies exceeding a threshold of $E = 10$ minimum ionizing particles (MIPs) per $\sin\theta$. The position of each seed is calculated using a barycentric energy-weighting of all trigger cells. Surrounding trigger cells within a distance in the $x/z - y/z$ plane — ranging from 0.015 in the initial layers to 0.050 in the final layers — are added to the seed energy in a clustering process [50].

4.2.4 Global

The Global Calorimeter Trigger (GCT) combines primitives from the Barrel Calorimeter Trigger (BCT), Hadron Forward Calorimeter (HF), and High-Granularity Calorimeter (HGCAL). By accessing calorimeter data over the full η-ϕ space from all subdetectors, it applies a variety of trigger algorithms to reconstruct objects such as hadronically decaying tau leptons (τ_h), jets, and electrons/photons (e/γ) — which are indistinguishable in the calorimeter and thus reconstructed as a single object — as well as missing transverse energy (MET and MHT) [60].

Jet reconstruction begins with electromagnetic towers, hadronic towers and e/γ clusters that match the size of hadronic towers. The process first identifies a seed tower with the highest p_T. After constructing the jet and excluding associated towers, the reconstruction iterates with the remaining towers until no towers with $p_T > 2.5$ GeV are left. Each reconstructed jet consists of the seed tower surrounded by a cone of towers within an angular distance of approximately $\Delta R = 0.4$. The resulting pattern resembles a 7×7 square with the towers at the corners removed. To reduce pileup contributions, only towers within the cone with $p_T > 0.5$ GeV are included in the reconstructed jet. Additionally, various energy corrections are applied to improve the accuracy of the reconstructed jet's transverse momentum p_T, taking into account pileup effects, pseudorapidity (η), and the specific subdetector used for measurement.

Hadronically decaying tau leptons (τ_h) can be reconstructed in a manner similar to jets. A key distinguishing feature is that the decay products of τ_h are generally more collimated than jets produced in soft collisions. In the endcap region covered by the High-Granularity Calorimeter (HGCAL), depth information and fine segmentation provide additional valuable insights that can be leveraged to distinguish background-induced jets from hadronically decaying taus (τ_h).

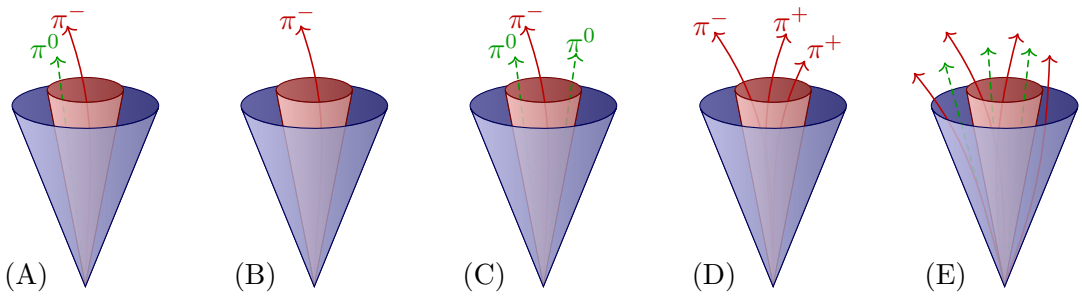


Figure 28: Schematic drawing of jets and their constituents originating from the hadronic tau decays with the highest branching ratios (A)-(D) compared to a quark/gluon jet (E). For tau-induced jets, the constituents are confined within a much narrower cone, highlighted in red. Modified from ref [33].

Two variants of **missing transverse energy** are calculated: the total missing transverse energy (MET), which accounts for all energy depositions, and the hadronic missing transverse energy (MHT), determined solely from hadronic objects like jets and taus. Both are obtained as the negative of the vector sum of the transverse components

of the reconstructed trigger-object energies — a detector-level convention that, in the ultrarelativistic limit ($E \approx p$) approximates the true missing transverse momentum $\boldsymbol{P}_{\boldsymbol{T}}^{\text{miss}}$.

Objects reconstructed at the Global Calorimeter Trigger (GCT) are forwarded to both the Correlator Trigger and the Global Trigger (GT) for further processing and correlation with other trigger objects.

4.3 Muon Trigger

The muon trigger is divided into three regions: the barrel ($|\eta| < 0.83$), the overlap region ($0.83 < |\eta| < 1.24$), and the endcap ($1.24 < |\eta| < 2.8$), with the overlap region leveraging trigger primitives from both the barrel and endcap section.

4.3.1 Barrel

The drift tubes (DTs) in the barrel region are read out at the detector front-end by the new On-detector Board for Drift Tubes (OBDT). Unlike the current Run 3 system, the OBDT front-end electronics only perform time measurements and do not apply bunch crossing identification using the Mean-Timer technique (see Section 3.6.5). The captured signal is time-to-digital converted and transmitted to the Barrel Muon Trigger Layer-1 (BMTL1) in the detector back-end for processing [62].

Resistive plate chamber (RPC) data is read out at a frequency of 640 MHz, which is 16 times the bunch crossing frequency of 40 MHz used during Run 3. This enables a sub-bunch crossing time resolution with a granularity of one-sixteenth of a bunch crossing, which is made available to the Barrel Muon Trigger Layer-1 (BMTL1) in the detector back-end.

The Barrel Muon Trigger Layer-1 (BMTL1) is responsible for generating the individual drift tube and resistive plate chamber primitives and subsequently merging them into super primitives. For the drift tube primitives, a so-called analytical method is employed, where cells of drift tube super layers are first analysed via the mean-timer technique (see Section 3.6.5) to yield tracks across either three or four layers. In the second step, the two $R-\phi$ super layers (Fig. 23) are then correlated. A match is found if compatible tracks are found within a ± 25 ns window, resulting in a recalculation of associated track parameters by combining both hits. In case no match is found, all per super layer primitives are forwarded to the next step. Resistive plate chamber (RPC) primitives are generated by clustering individual RPC hits within the BMTL1.

To merge the two primitive types, the coordinates of RPC clusters are first transformed into the drift tube convention. Matching is then performed based on the azimuthal angle within a ± 3 bunch crossing window centred around the DT bunch crossing information. For matched super primitives, the timing information from the RPCs is preserved. To provide a more extensive picture to the subsequent track-finding step, super primitives formed from clusters of two RPC layers are allowed in the first two muon stations. Additionally, unmatched RPC clusters and DT segments are also passed on to the track-finding step.

Track finding happens in the Global Muon Trigger (GMT) as part of the Barrel Muon Track Finder (BMTF) algorithm, which employs a Kalman filter [63]. The process starts from the state vector of super primitives in the outermost muon station $\boldsymbol{x_n} = (k, \phi, \phi_b)^T$, where $k = q/p_T$ denotes the signed curvature ($q = \pm 1$), ϕ the azimuth angle and ϕ_b the bending angle in the azimuth plane. The state vector is gradually propagated inwards via

$$\boldsymbol{x_{n+1}} = F\boldsymbol{x_n} \qquad \text{or} \qquad \begin{pmatrix} k \\ \phi \\ \phi_b \end{pmatrix}_{n+1} = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & b \\ c & 0 & 1-b \end{pmatrix} \begin{pmatrix} k \\ \phi \\ \phi_b \end{pmatrix}_{n} \quad , \tag{4.1}$$

with F denoting the propagation matrix whose parameters a, b and c are determined from the detector geometry and simulations. An additional 3×3 covariance matrix P expresses the uncertainties of the state and follows a similar propagation relation via

$$P_{n+1} = FP_nF^T + Q \quad , \tag{4.2}$$

where the covariance matrix Q represents external uncertainties arising from multiple scattering in the iron return yoke. After each propagation step, the closest stub to the propagated state is selected, resulting in an update of the track parameters with the measurement data and uncertainties of the stub. The measurement is represented by $\boldsymbol{z_n} = \left(\phi^{\text{stub}}, \phi_b^{\text{stub}}\right)^T$ and corresponding measurement uncertainties via the covariance matrix R. The state update is facilitated by the so-called Kalman gain matrix,

$$K = HPH^T \left(HPH^T + R\right)^{-1} \quad , \tag{4.3}$$

with H a matrix mapping the state vector to the measurement vector. Hence, the residual between the predicted and measured state becomes

$$\boldsymbol{r_n} = \boldsymbol{z_n} - H\boldsymbol{x_n} = \begin{pmatrix} \phi^{\text{stub}} \\ \phi_b^{\text{stub}} \end{pmatrix}_n - \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} k \\ \phi \\ \phi_b \end{pmatrix}_n \quad . \tag{4.4}$$

Yielding an updated state and covariance matrix given by

$$\boldsymbol{x_n^{\text{upd}}} = \boldsymbol{x_n} + H^T K \boldsymbol{r_n} \tag{4.5}$$

$$P_n^{\text{upd}} = P_n - H^T K H P_n \quad . \tag{4.6}$$

Each track candidate is propagated through all four muon stations, with the state updated at each station using the measured stub data. Two types of muon candidates can be reconstructed: prompt muons, which undergo a final propagation step to the beamspot[1], and displaced muons, which cannot be reliably propagated to the beamspot. Displaced muons arise from the decay of heavier mother particles at a displaced or secondary vertex sufficiently far away from the beamspot.

[1]The beamspot is an estimate of the expected proton collision region, determined from the distribution of collision vertices.

4.3.2 Endcap

Endcap RPCs operate similarly to their barrel counterparts, featuring a readout frequency of 640 MHz and sub-bunch crossing time resolution during High-Luminosity operation. Data is transmitted to the detector back-end via optical links, where hit clustering is performed. Each chamber is expected to transmit up to four hits per bunch crossing.

With the improved resistive plate chambers (iRPCs), clustering is performed using signals from both ends of the strips within a 1.5 ns time window, accounting for potential variations in signal rise time. Clusters from both ends are then merged, allowing for the possibility of a missing signal from one edge. A final filtering step is applied to separate clusters composed of hits with incompatible radial distances.

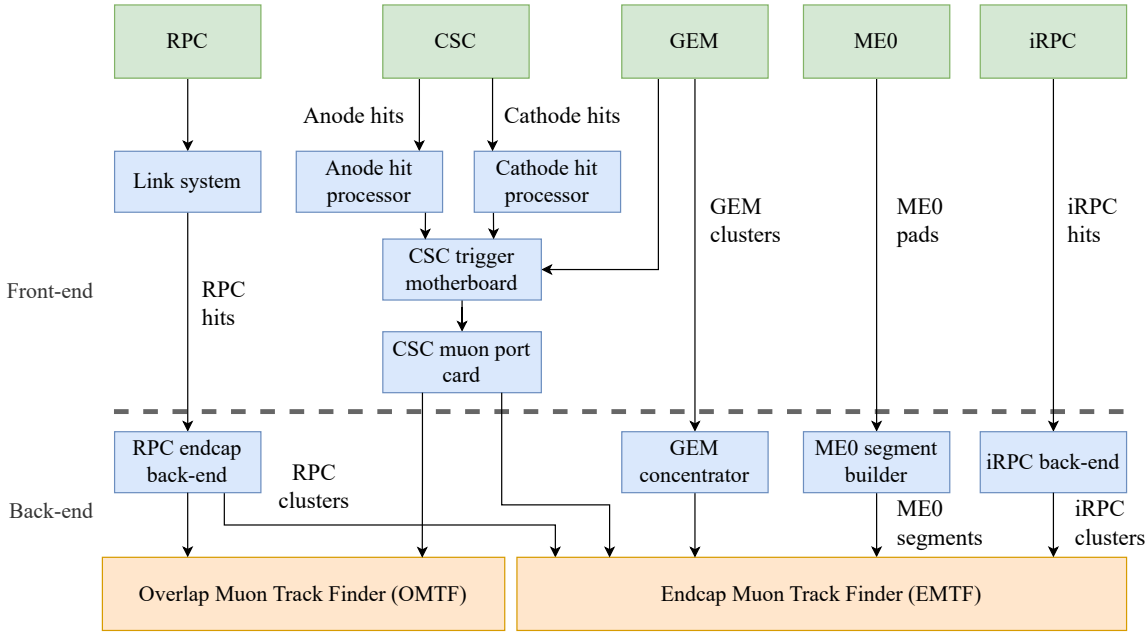


Figure 29: Architecture of the endcap and overlap primitive generation. The dashed line separates the detector front-end, located within the radiation environment of the CMS experimental cavern, from the back-end, situated outside of it.

The endcap cathode strip chambers (CSCs) achieve a hit accuracy of half a strip by performing analogue comparisons of charge deposition on each strip and its neighbouring strips. The CSC anode electronics front-end employs constant-fraction discriminators, which register muon hits with minimal timing variation, ensuring precise bunch crossing identification. The CSC trigger motherboard identifies local charged tracks based on straight-line coincidence patterns observed between the anode and cathode. Each module, consisting of six gas gaps (cf. Section 3.6.3), requires four such coincidences to construct a local charged track. In the $R-z$ plane, a single anode pattern is allowed, indicating that the muon originated near the CMS interaction point, while nine cathode patterns exist covering the $\phi-z$ hyperplane and accounting for the bending effect caused

by the CMS's magnetic solenoid. In muon stations equipped with both CSC and gas electron multipliers (GEMs), the resolution of the bending trajectory can be improved by combining data from both subdetectors within the CSC trigger motherboard. The obtained primitives are forwarded to the track finders via the CSC muon port card.

The gas electron multiplier (GEM) chambers installed in the very forward region of the muon system generate trigger pads using an on-chamber ASIC called VFAT3. These pads are formed by logically OR'ing two neighbouring strips, resulting in an angular resolution of 0.9 mrad in ϕ [50]. An opto-hybrid board collects pads from 12 VFAT3 ASICs [64] and assembles clusters from up to eight adjacent trigger pads within the chamber with a latency of 1.5 bunch crossings. These clusters are transmitted to both the CSC trigger motherboard and the GEM concentrator at the detector back-end (see Fig. 29). From there, the data are forwarded to the Endcap Muon Track Finder (EMTF).

Track finding in the endcap is carried out using a three-step process known as the EMTF++ algorithm, implemented on FPGAs in the detector back-end [65]:

1. Primitives of all subdetectors for each of the four muon stations, plus ME0 are separately analysed. Patterns consistent with a traversing muon in that station are kept as so-called muon stubs.

2. Tracks are constructed from the stubs identified in the previous step. In cases of ambiguity, the stubs are prioritised based on $\Delta\eta$ and $\Delta\phi$ compatibility between stations. Lower priority track candidates are discarded as ghosts.

3. Muon candidates get assigned a p_T value computed from the available track information, including $\Delta\eta$, $\Delta\phi$, bending and η. In the current Run-3 system, this computation is performed using a Boosted Decision Tree (BDT). For High-Luminosity operation, the implementation of a neural network that incorporates all subdetector features as inputs is planned.

The new EMTF++ algorithm is also designed to reconstruct displaced muons, whose trajectories do not originate from a collision vertex. These muons are produced from the decay of other particles at a displaced or secondary vertex. The reconstruction process largely mirrors that of prompt muons described earlier, but excludes beamspot vertex constraints on stub patterns and tracks. Additionally, the neural network not only assigns a p_T value but also determines the impact parameter d_0, defined as the distance of closest approach of the track candidate to the z-axis (Fig. 30) [65].

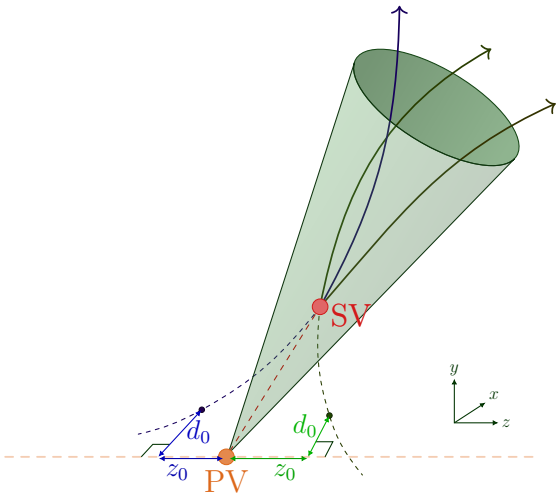


Figure 30: An illustration of a jet initiated by a bottom quark originating from a collision or primary vertex (PV). The bottom quark decays at a displaced or secondary vertex (SV), producing three muons. The longitudinal and transverse impact parameters, z_0 and d_0, respectively, of the muon tracks are depicted. Modified from ref [33].

4.3.3 Overlap region

The overlap region employs a dedicated Overlap Muon Track Finder (OMTF) algorithm utilising trigger primitives from both the barrel and endcap in the delicate overlap region. The algorithm devised for the current Run 3 system and ported to the one for High-Luminosity operation employs a naive Bayes classifier, enabling classification into 52 p_T^k bins, with 26 bins for each charge $(q \pm 1)$ [66]. It is based on the conditional probability for each class k, given the observed detector hits x_i, as

$$P\left(p_T^k \mid x_1, \ldots, x_L\right) = \frac{P\left(x_1, \ldots, x_L \mid p_T^k\right) P\left(p_T^k\right)}{P\left(x_1, \ldots, x_L\right)} \quad . \tag{4.7}$$

This is where the core concept of a naive Bayes classifier comes into play. By assuming that the individual x_i are uncorrelated, $P\left(x_1, \ldots, x_L \mid p_T^k\right)$ can be expressed as $\prod_{i=1}^{L} P\left(x_i \mid p_T^k\right)$. However, in our case, the absolute muon hit positions ϕ_i across different layers are highly correlated, following a bend trajectory in CMS's magnetic field. To account for this, the distance between the absolute hit position ϕ_i and a chosen reference hit ϕ^{ref} is used,

$$\phi_i^{\text{dist}} = \phi_i - \phi^{\text{ref}} - \Delta\phi_i^{\text{mean}}\left(p_T^k\right) \quad , \tag{4.8}$$

where $\Delta\phi_i^{\text{mean}}\left(p_T^k\right)$ denotes the average ϕ distance for a given p_T^k in layer i (Fig. 31).

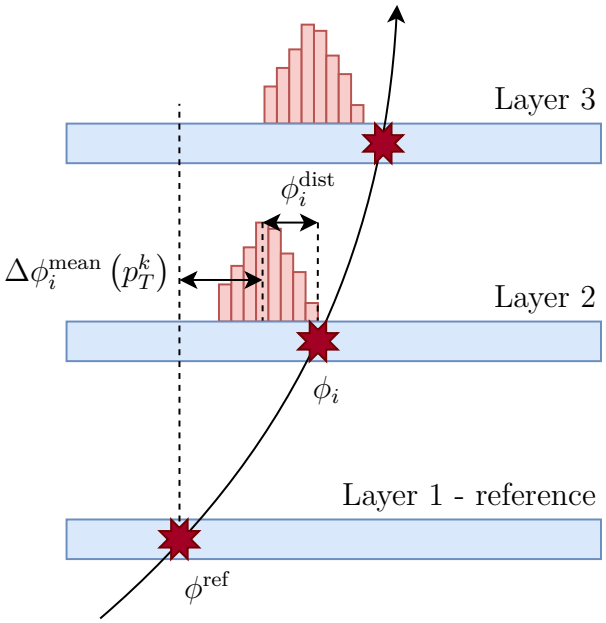


Figure 31: Illustration of the ϕ coordinate definition in the OMTF algorithm. The stub (red star) in the first layer serves as the reference hit.

By substituting x_i with ϕ_i^{dist} in Eq. 4.7 and selecting the p_T^k with the highest probability, we obtain

$$p_T = \underset{k}{\text{argmax}} \prod_{i=1}^{L} P\left(\phi_i^{\text{dist}} \mid p_T^k\right) P\left(p_T^k\right) \quad , \tag{4.9}$$

where the denominator of Eq. 4.7 was dropped because it is a constant for all p_T^k bins and thus does not affect the selection of the maximum. When multiple compatible hits are detected, the hit with the smallest ϕ_i^{dist} is selected. If stubs are missing in certain layers, those layers are excluded, and muon candidates with fewer than three stubs are discarded. To maintain robustness when a hit is absent in the reference layer, the algorithm is executed four times, each with a different reference layer. Any duplicates are subsequently removed based on the number of stubs associated with each muon candidate.

In practice, the distributions for $P\left(\phi_i^{\text{dist}} \mid p_T^k\right)$ are obtained from Monte Carlo simulations and their logarithms are stored in so-called look-up tables (LUTs) on the FPGA running the OMTF algorithm. The computation of Eq. 4.9 then simplifies to a sum of logarithms.

To facilitate the reconstruction of displaced muons during High-Luminosity operation, which in the baseline version of the algorithm mimic low transverse momentum muons, an additional angle ϕ_B relative to the muon chamber is measured by the first drift tube (DT) chambers. The measured angle ϕ_B — effectively incorporating the additional angle seen due to displacement — can be used to correct Eq. 4.8 through extrapolation to other muon stations [67].

In recent years, neural networks have gained significant popularity, and their application for the Overlap Muon Track Finder (OMTF) is actively investigated [68].

4.3.4 Global

Although the energy loss of muons in the calorimeters is generally small compared to other particles (see Fig. 17), it can become significant for low-momentum muons, potentially impairing their reconstruction. Such low-momentum muons may only leave hits in one or two of the innermost muon stations, which is insufficient to assign track parameters like p_T and charge or to differentiate them from spurious tracks. Reconstruction performed by the High-Level Trigger (HLT) [69] and during offline analysis has shown that incorporating tracking information from the silicon tracker into the muon reconstruction leads to significant improvements. This approach enables notably better transverse momentum determination, particularly for low-momentum muons and in the challenging overlap region between the two fiducial regions.

During High-Luminosity operation, the Global Muon Trigger (GMT) will receive tracks from the Global Track Trigger (GTT), reconstructed using silicon tracker stubs provided by the novel p_T-modules (see Section 3.4). Matching tracks to muon candidates can be performed either at the level of muon stubs or using reconstructed tracks from the BMTF, OMTF, and EMTF++ algorithm. While matching at the muon stub level provides greater acceptance for muons that might otherwise be missed, it is also more prone to noise and punch-through hadrons, leading to a higher overall rate of misidentified muons.

The current baseline **tracker plus stubs** (TPS) algorithm [50,70] starts from reconstructed silicon tracker tracks and propagates them to the muon stations. The propagation of the azimuthal angle ϕ_j and bending angle $\phi_{b,j}$ to the j-th muon station is approximated as

$$\phi_j^{\text{prop}} = \phi + c_j k_j \quad , \tag{4.10}$$

$$\phi_{b,j}^{\text{prop}} = c_{b,j} k_j \quad , \tag{4.11}$$

where $k = q/p_T$ denotes the signed curvature ($q = \pm 1$), ϕ the azimuthal angle at the interaction point and c_j and $c_{b,j}$ the propagation coefficients for the ϕ_j and the bending angle $\phi_{b,j}$, respectively. To account for ionisation losses, curvatures are corrected using

$$p_{T,j} = p_T - \epsilon_j \rightarrow k_j = \frac{k}{1 - \epsilon_j |k|} = k + \text{sgn}(k)\epsilon_j k^2 + \mathcal{O}(k^3) \quad . \tag{4.12}$$

Hence, Eq. 4.10 and Eq. 4.11 become

$$\phi_j^{\text{prop}} = \phi + \text{sgn}(k)(\underbrace{c_j}_{\text{LUT}} |k| + \underbrace{c_j \epsilon_j}_{\text{LUT}} |k|^2) \quad , \tag{4.13}$$

$$\phi_{b,j}^{\text{prop}} = \text{sgn}(k)(\underbrace{c_{b,j}}_{\text{LUT}} |k| + \underbrace{c_{b,j} \epsilon_j}_{\text{LUT}} |k|^2) \quad . \tag{4.14}$$

These computations are efficiently implemented in FPGA firmware using precomputed LUTs derived from simulations. A muon stub is matched to the propagated track variables if it satisfies certain limits a_j, b_j via

$$\frac{|\phi_j^{\text{stub}} - \phi_j^{\text{prop}}|}{\sigma_j} < a_j \quad , \tag{4.15}$$

$$\frac{|\phi_{b,j}^{\text{stub}} - \phi_{b,j}^{\text{prop}}|}{\sigma_j} < b_j \quad . \tag{4.16}$$

where $\sigma_j = \sqrt{\alpha_j k^2 + \beta_j} \approx \alpha_j'|k| + \beta_j'$ represents the resolution, accounting for multiple scattering (α) and measurement uncertainties (β). Once all tracks are matched to stubs, a cleaning step is performed to eliminate stubs used in multiple track-matched muon candidates. Preference is given to muon candidates with a higher number of associated stubs or, if equal, those with stubs possessing smaller deviations from the propagated track. The remaining muon candidates are p_T-sorted and are assigned an additional isolation variable defined as $P_T^{\text{isol}} = \sum p_T$, the p_T-sum of tracks within a cone of $\Delta R = 0.3$ (see Section 3.7.1) around the matched track. Finally, up to twelve track-matched muons are forwarded as a collection to the Global Trigger (GT).

The Global Muon Trigger also consolidates muon candidates reconstructed without tracking information during earlier steps using the BMTF, OMTF, and EMTF++ algorithms. Overlaps are removed, and the remaining muon candidates are organised into two collections of up to twelve objects each: the standalone muon collection and the standalone displaced muon collection. Both collections are p_T-sorted and sent to the Global Trigger (GT).

4.4 Track Trigger

As described in Section 3.4, the newly installed p_T-modules in the Outer Tracker are designed to correlate two tracker hits that exhibit a low bend trajectory. The correlation pattern is configurable, allowing the selection of a p_T threshold, with the current baseline set at $p_T > 2$ GeV. Together with the requirement $|\eta| < 2.5$, this approach achieves effective data reduction by approximately an order of magnitude, which is essential for transmitting data to the detector back-end at the bunch crossing rate of 40 MHz.

4.4.1 Track finder

At the back-end, state-of-the-art FPGAs are utilised in a two-step process. The first step employs the "tracklet" pattern matching technique to identify seeding track candidates [35, 71]. This process begins by identifying two compatible seeding stubs in adjacent tracking layers that are consistent with an origin at the beamspot. The "tracklet" is then projected both inward and outward to define a search window. Stubs within this search window are added to the track candidate (see Fig. 32). If multiple matching stubs are found in a specific layer, the stub closest to the projected trajectory is selected. Between four and six stubs are required to construct a complete track candidate. To

minimise duplication, track candidates sharing stubs are merged before proceeding to the second tracking step.

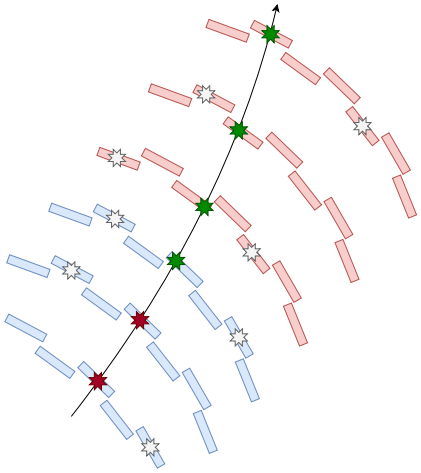


Figure 32: Illustration of a seeding stub pair highlighted in red forming a "tracklet". The "tracklet" is then projected onto other layers, and stubs near the projected trajectory (green) are matched.

The second step employs a Kalman filter [71, 72], propagating the state vector

$$\boldsymbol{x_n} = \left(\frac{1}{2R}, \phi, \cot\theta, z\right)^T_n \quad , \tag{4.17}$$

where R denotes the curvature radius, defined as $R = p_T/(qB)$, θ is the polar angle relating to the pseudorapidity η via Eq. 3.7, ϕ is the azimuthal angle and z represents the longitudinal impact parameter[1] extended to the distance r from the beamline. The state update is performed via

$$\boldsymbol{x_{n+1}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \Delta r & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \Delta r & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{2R} \\ \phi \\ \cot\theta \\ z \end{pmatrix}_n \quad , \tag{4.18}$$

where Δr is the radial separation between hits in adjacent tracker layers. A 4×4 covariance matrix, P_n, describes the uncertainties of the state. The process follows the description of the Kalman filter for the Barrel Muon Track Finder (BMTF) outlined in Section 4.3.1, where each prediction step is followed by a state update. To distinguish true tracks from randomly aligned stubs (false tracks), a χ^2-test is applied to the reconstructed track. The value of χ^2 can be computed iteratively across Kalman filter

[1] The longitudinal impact parameter z_0 refers to the distance on the beamline from the interaction point to the point of closest approach of the track, see Fig. 30.

steps, accumulating the squared deviations weighted by their variances

$$\chi^2 = \sum_n \boldsymbol{r_n}^T \left(HPH^T + R\right)^{-1} \boldsymbol{r_n} \quad , \tag{4.19}$$

where $\boldsymbol{r_n}$ denotes the residual (Eq. 4.4), H the state to measurement mapping matrix and R the covariance of the matrix of the measurement (c.f. Section 4.3.1). Tracks are initially filtered based on the number of stubs, favouring those with more, followed by those with a lower χ^2 value. Additional selection criteria are applied based on the track's p_T and z value. To optimise resource usage, all filtering steps are conducted after each Kalman filter iteration, allowing early elimination of low-quality track candidates. A final filtering step is performed once the track is fully reconstructed, after which it is evaluated by a Boosted Decision Tree (BDT) to assess track quality [71].

To enable the reconstruction of displaced tracks that do not originate from the beamspot, "tracklet" seeds consisting of three compatible hits are used to initiate track finding [73].

4.4.2 Vertex reconstruction

The baseline for reconstructing collision vertices or primary vertices (PVs) is a histogram-based method. This approach utilises a weighted histogram with 256 equal z_0 bins (Fig. 33), currently spanning from -20.47 cm to 20.47 cm, which are populated with the p_T values of reconstructed tracks provided by the track finder. To identify the primary vertex, a sliding window covering three consecutive z_0 bins moves across the histogram, determining the position of the maximum scalar p_T-sum within the window. The primary vertex position is then calculated as the weighted average z_0 position within the window, forming a trigger object along with the corresponding scalar p_T-sum. This process can be repeated to identify the primary vertex (PV) with the next largest scalar p_T-sum.

To mitigate degradation caused by poor z_0 resolution in high η tracks and the influence of high p_T fake tracks in the previously described method, a new neural network-based approach has been developed [74]. In this approach, the input weights for the histogram are generated by a deep neural network. As a result, the histogram weights no longer directly correspond to the scalar p_T-sum. To address this, a second deep neural network, leveraging track features and the track's distance from the primary vertex (PV), assigns a scalar p_T-sum to the PV.

The Global Track Trigger (GTT) reconstructs up to ten primary vertices (PVs) per event and transmits them to the Global Trigger (GT). Additionally, plans include the Global Track Trigger (GTT) providing up to two displaced vertices per event to the Global Trigger (GT). Recent studies on displaced vertex reconstruction have explored the use of gradient-boosted decision trees to rank intersections observed in the tracking data [75].

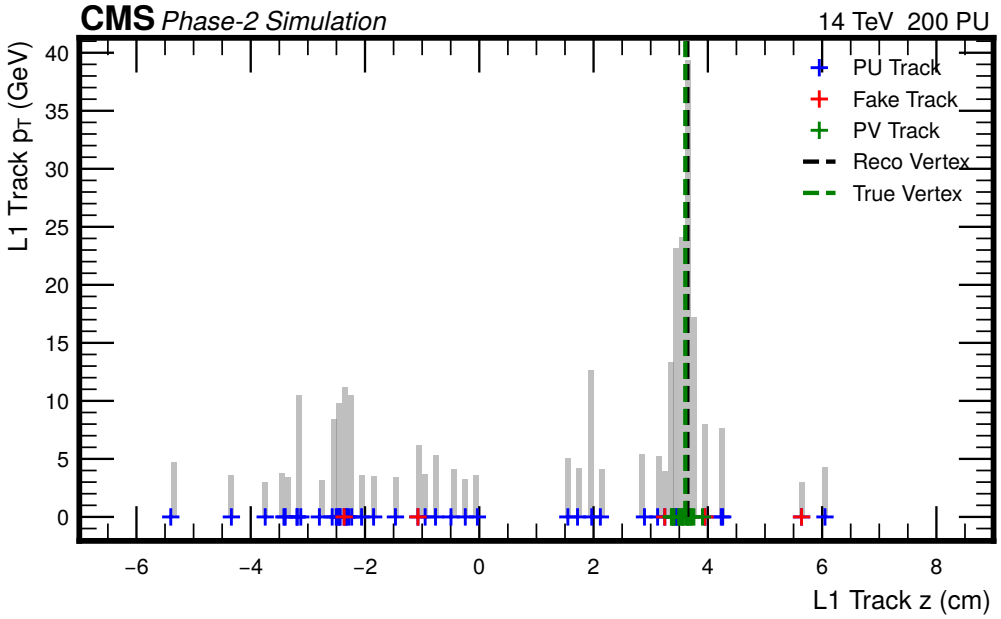


Figure 33: A p_T-weighted histogram representing the track positions along the z-axis. The grey markers denote the scalar sum of track p_T within each bin, while the coloured crosses represent the actual positions of individual tracks. The colours indicate the track origins: blue for tracks from pileup interactions, red for fake tracks, and green for tracks from the primary vertex. The black dashed line marks the vertex position identified by the algorithm used during offline reconstruction, while the green dashed line represents the true vertex position. From ref [50].

4.4.3 Jet reconstruction

The current baseline algorithm segments the η-ϕ plane along with the z_0-axis into bins or cells. The z_0 bins provide an overlapping double coverage, offset by half a z_0 bin, enabling more precise z_0 assignment. Each z_0 bin is assigned an η-ϕ clustering grid, where tracks that meet certain purity criteria are included. Optionally, tracks can be filtered based on the distance from the primary vertex (PV), which is the current default of the emulator within Compact Muon Solenoid Software (CMSSW) [70].

The clustering process begins with a search for local peaks in η within each ϕ bin, incorporating neighbouring η bins into the clustering seeds. In the next step, the resulting clusters are merged with adjacent clusters in ϕ. The final clustered jet candidates are assigned a scalar p_T-sum of included tracks, along with ϕ, η, z_0, and their track count.

To identify displaced jets, a separate clustering process is performed with displaced tracks.

4.4.4 Missing transverse energy

Similar to the Global Calorimeter Trigger (GCT), the Global Track Trigger (GTT) reconstructs two kinds of missing transverse energy: the overall MET and the hadronic-only MHT.

In MET reconstruction, the default configuration in CMSSW first filters the reconstructed tracks using a stricter χ^2 criterion than that applied during the initial track reconstruction. Particular attention is given to assessing the bending reconstruction quality for tracks via a dedicated χ^2_{bend} criterion, limiting the impact of poorly reconstructed track momenta. The remaining tracks are then associated with the primary vertex (PV), and their transverse momentum components are summed vectorially as $\boldsymbol{P_T} = \sum_i (p_x, p_y)_i^T$ [70]. Finally, the components are inverted to obtain the missing momentum $\boldsymbol{P_T^{\text{miss}}} = -\boldsymbol{P_T}$ and then converted back into the transverse momentum magnitude (P_T) and azimuthal angle (ϕ) representation.

MHT is computed in a similar manner using the vector sum of transverse momenta but considers only reconstructed jets in the summation. Alongside the vector sum, the algorithm also determines the scalar sum of reconstructed jets. Two versions exist: a prompt MHT, derived from tracks associated with the primary vertex (PV), and a displaced MHT, based on tracks originating from secondary or displaced vertices.

4.4.5 Other trigger objects

In addition to the described trigger objects, the Global Track Trigger (GTT) is also expected to exploit tracking information to reconstruct hadronically decaying τ leptons, light mesons such as ϕ-, ρ- and B_s-mesons, as well as the rare $W^{\pm} \to \pi^{\pm}\pi^{\mp}\pi^{\pm}$ decay [70]. Furthermore, it is planned to forward up to 12 prompt and 12 displaced tracks that meet specific isolation and quality criteria to the Global Trigger (GT).

4.5 Correlator Trigger

The Correlator constitutes an intermediate system between the subdetector-specific triggers and the Global Trigger (GT) leveraging and combining all subdetector information. It is separated into two particle-flow layers. Particle-flow has been an essential tool in both offline and High-Level Trigger (HLT) event reconstruction [57, 58, 69] and is now also possible at the Level-1 Trigger during High-Luminosity operation [76, 77].

4.5.1 Layer-1

To efficiently process the large number of input objects, the total detection area is divided into smaller regions in η and ϕ, enabling parallel processing across multiple boards. Each region is expanded into a subregion with an additional 0.25 overlap in η and ϕ with neighbouring regions. This overlap ensures accurate association of input objects at region boundaries. In total, there are 12×9 ($\eta \times \phi$) subregions, as implemented in the current CMSSW emulators [70]. These are distributed as follows: 6×9 for the barrel, 4×9 for both endcaps, and 2×9 for the two forward calorimeter sections (HF), as shown in Fig. 34.

In the initial stage, the "regionizer" sorts all inputs into their respective subregions. These inputs include tracks from the track finders, standalone muons from the Global Muon Trigger (GMT), ECAL barrel electromagnetic (EM) clusters, HCAL barrel hadronic towers, as well as HGCAL and HF clusters, as described in the previous sections.

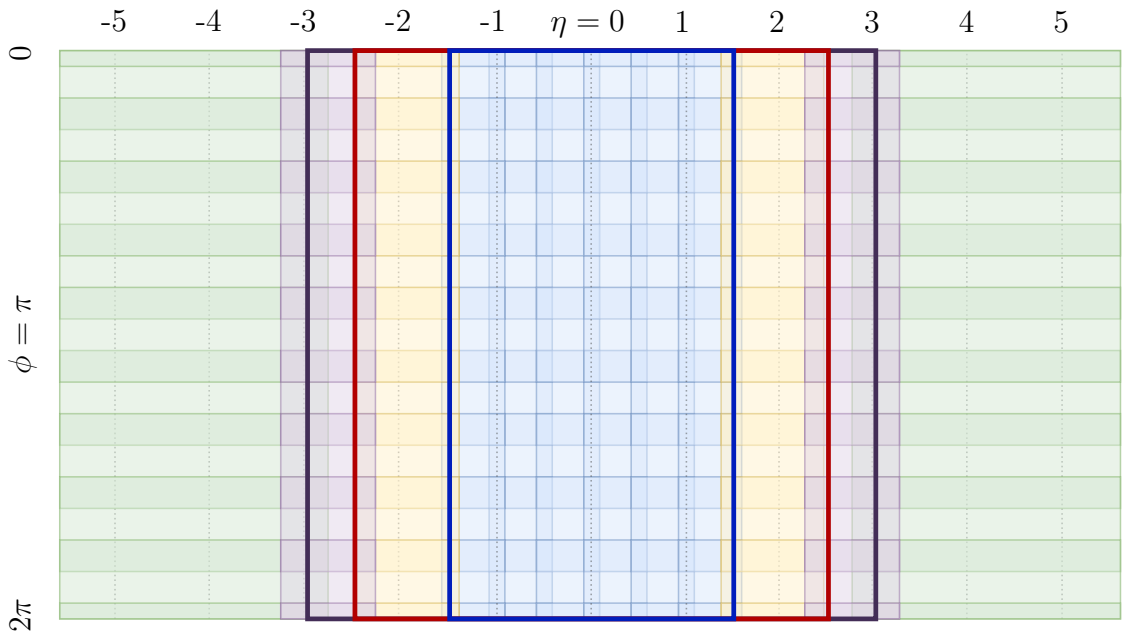


Figure 34: An illustration depicting the division of particle-flow regions across the barrel calorimeter, endcap calorimeters (HGCAL), tracker, and forward calorimeters (HF). Subregions, including overlaps, are drawn as transparent rectangles. Subdetector boundaries are highlighted in blue for the barrel calorimeter, red for the tracker, and deep purple for the endcap.

Once the "regionizer" assigns inputs to specific subregions, each subregion is processed independently to reconstruct particle-flow candidates (Fig. 35):

- **Track-Muon association:** Tracks are matched to the nearest muon candidates based on their ΔR and p_T. Once linked, these tracks are excluded from further association steps.
- **Track-EM cluster association:** Tracks within $\Delta R < 0.04$ of an electromagnetic cluster are linked, and their total transverse momentum ($\sum p_T^{\text{track}}$) is computed. Clusters without track associations are classified as photons. Clusters satisfying $p_T^{\text{cluster}} \geq \sum p_T^{\text{track}} - 2\sigma$[1] are identified as electrons. A significant excess of p_T^{cluster} over $\sum p_T^{\text{track}}$ leads to an additional photon candidate. If $p_T^{\text{cluster}} < \sum p_T^{\text{track}} - 2\sigma$, the cluster is disregarded as an electron or photon, as it likely originates from a hadronic shower starting in the electromagnetic calorimeter.

[1] $\sigma = \max\left[\sigma\left(p_T^{\text{cluster}}, |\eta|\right), \sigma\left(\sum p_T^{\text{track}}, |\eta|\right)\right]$

- **Hadronic cluster association:** Combined ECAL-HCAL 3×3 tower clusters are matched to the nearest electromagnetic clusters. The energy of linked hadronic clusters is corrected by subtracting the energy of associated EM clusters. If the corrected energy falls below 10% of the original, the hadronic cluster is discarded.
- **Track-Hadronic cluster association:** Remaining tracks are linked to the nearest hadronic clusters if $\Delta R < 0.15$, provided they satisfy $p_T^{\text{had}} > \sum p_T^{\text{track}} - 2\sigma\left(p_T^{\text{had}}\right)$ to reject fake high-p_T tracks. If multiple clusters meet the criteria, the best match is determined by minimising the combined angular and p_T distance

$$q^2 = \left(\frac{\Delta R}{0.15}\right)^2 + \left(\frac{\max\left(p_T^{\text{track}} - p_T^{\text{had}}, 0\right)}{\sigma\left(p_T^{\text{track}}\right)}\right)^2 \quad . \tag{4.20}$$

  Linked tracks are labelled as charged hadrons. Unlinked tracks may also be classified as charged hadrons if they satisfy $p_T > 10(20)$ GeV under a loose (tight) quality requirement. If $p_T^{\text{had}} \gg \sum p_T^{\text{track}}$, the track-cluster combination is instead classified as a neutral hadron or photon.
- **Classification of out-of-tracker-coverage clusters:** Clusters beyond the tracker's $|\eta|$ range are categorised as photons or neutral hadrons based on their relative energy depositions in the electromagnetic and hadronic calorimeters only.

With HGCAL inputs, a single track association step is performed using preidentified hadron and electromagnetic clusters, analogous to the Track-EM cluster and Track-Hadronic cluster associations described above.

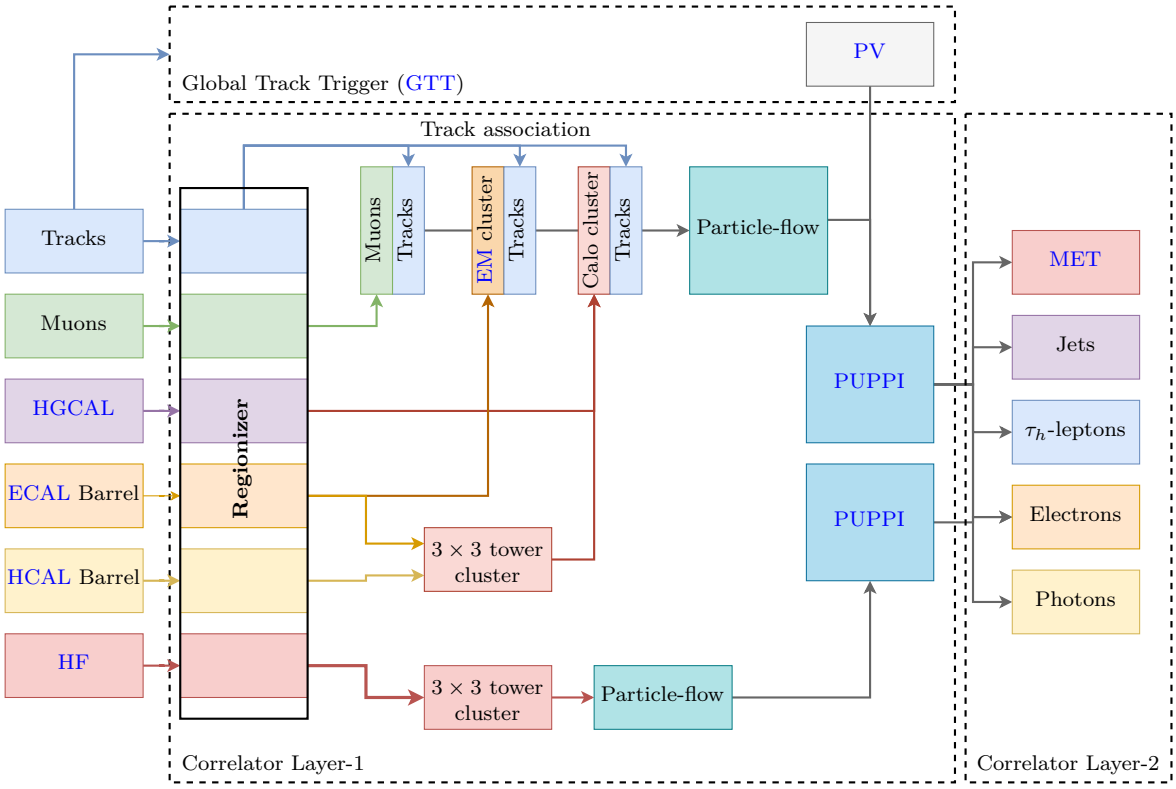


Figure 35: Illustration of the Correlator Trigger data flow through both layers.

The particle-flow algorithm reconstructs candidates regardless of whether they originate from a "hard" scatter or pileup. A subsequent pileup per particle identification (PUPPI) step is responsible for mitigating and removing pileup contributions [78]. A key aspect of the PUPPI algorithm is assigning particle-flow candidates to a primary vertex (PV), which is reconstructed within the Global Track Trigger (GTT) (see Section 4.4.2). Charged candidates can be reliably traced back to the PV via the tracks' longitudinal impact parameter distance Δz_0 from the PV, allowing the removal of incompatible particle-flow candidates. For neutral particle-flow candidates, however, direct backtracking to the PV is not possible. Instead, they are assigned a weight or probability ω, which reflects the likelihood that they originate from the PV. This is facilitated through a metric a_C (C stands for central) defined as

$$a_C = \ln \sum_{i \in \mathrm{PV},\, \Delta R < 0.3} \left[\frac{\min\left(p_T^i, p_T^{\max}\right)}{\max\left(\Delta R, \Delta R^{\min}\right)} \right]^2 \quad , \tag{4.21}$$

where the summation is over all tracks originating from the PV within an angular distance $\Delta R < 0.3$ from the neutral particle-flow candidate. The constant $p_T^{\max} = 50$ GeV is used, with $\Delta R^{\min}$ set to 0.07 in the barrel and 0.04 in the endcap. In the forward region, beyond the tracker's coverage, a slightly modified metric α_F is introduced

$$\alpha_F = \ln \sum_{i \in \mathrm{PF},\, \Delta R < 0.3} \left[\frac{\min\left(p_T^i, p_T^{\max}\right)}{\max\left(\Delta R, \Delta R^{\min}\right)} \right]^2 \quad , \tag{4.22}$$

where the summation now considers particle-flow candidates instead of tracks. While the constants $p_T^{\max}$ and $\Delta R^{\min}$ remain unchanged in the HGCAL region, they are increased to $p_T^{\max} = 100$ GeV and $\Delta R^{\min} = 0.1$ in the HF region to account for its different granularity. These α parameters can be interpreted as collimation metrics: they take on large values for narrowly clustered, high-p_T constituents and smaller values for softer, more widely separated ones that are more likely to originate from pileup. Using the corresponding α (α_C or α_F) parameter, the weight ω_i applied to neutral particle-flow candidates is calculated using a sigmoid function

$$\omega_i = \frac{1}{1 + e^{-\left(x_\alpha + x_{p_T} - x_{\mathrm{PU}}\right)}} \quad , \tag{4.23}$$

where

$$x_\alpha = \min\left[\max\left(C_\alpha\left(\alpha - \alpha_{\mathrm{PU}}\right), -x_\alpha^{\max}\right), x_\alpha^{\max}\right] \quad , \tag{4.24}$$

$$x_{p_T} = C_{p_T}\left(p_T - p_{T,\mathrm{PU}}\right) \quad , \tag{4.25}$$

$$x_{\mathrm{PU}} = \ln\left(\frac{N_{\mathrm{PU}}}{200}\right) + C_0 \quad . \tag{4.26}$$

The parameters α_{PU} and $p_{T,\mathrm{PU}}$ represent typical values of α and p_T for particles originating from pileup. The other constants, $x_\alpha^{\max}$, C_α, C_{p_T} and C_0 are optimised to ensure

precise assignment of PUPPI-corrected momenta using $p'_T = \omega_i \cdot p_T$. Candidates with $\omega_i < 1\%$ or a corrected momentum p'_T below an *eta*-dependent threshold are classified as pileup and discarded. The remaining now PUPPI candidates are forwarded to the second layer of the Correlator.

4.5.2 Layer-2

The Correlator Layer-2 receives PUPPI candidates from all individually processed subregions and first processes them via a so-called "Deregionizer", which merges subregions into a global event topology. Multiple boards receive identical inputs from Layer-1, with each board processing a different one of six consecutive bunch crossing events, creating a temporal processing distribution known as time multiplexing (TMUX). Following this initial merging step, advanced techniques leveraging particle-flow and PUPPI objects are used to reconstruct trigger objects such as jets, hadronically decaying taus, missing transverse energy, electrons, and photons (see Fig. 35).

Missing transverse energy (MET): Similar to track- and calorimeter-based MET calculations, two variants exist. The hadronic MHT is derived from the vector sum of the transverse momentum of reconstructed jets at the Correlator, while the total MET is obtained using all PUPPI candidates (see Section 4.4.4 for details). In addition to the reconstructed hadronic missing transverse momentum MHT, the scalar p_T-sum of reconstructed jets is also calculated. The particle-flow preselection and PUPPI correction of components in the calculation help suppress the impact of pileup contributions.

Jets: The Seeded-Cone algorithm [79, 80] serves as the current baseline for jet reconstruction at the Correlator. The algorithm starts by selecting the highest p_T PUPPI candidate as the clustering seed, adding other PUPPI candidates that fall within a cone of ΔR around the seed. The jet axis is determined as the p_T-weighted barycentre of its constituents in η and ϕ, while the jet's transverse momentum (p_T) is calculated as the sum of its constituents' p_T, with additional corrections derived from simulations. Once the jet clustering is completed, its constituents are removed from the event topology, and the process repeats. The algorithm has two variants, using cone sizes of $\Delta R = 0.4$ and $\Delta R = 0.8$, with the $\Delta R = 0.4$ version serving as the baseline and the $\Delta R = 0.8$ version aimed at future trigger developments. Additionally, jet flavour tagging has been implemented on top of the baseline variant using a convolutional neural network to assign a score indicating the likelihood that a bottom quark is among the jet's constituents [81].

Hadronically decaying tau leptons (τ_h): The current baseline for identifying hadronically decaying tau leptons at the Correlator begins by selecting the highest p_T PUPPI candidate as a seed. Tau constituents are then clustered from PUPPI candidates within a cone of $\Delta R = 0.1$ around the seed. Additionally, PUPPI candidates within a larger cone of $\Delta R = 0.4$ are gathered as inputs to a neural network, which assigns a score indicating the likelihood that the tau cluster originates from a real hadronically decaying tau. The neural network also provides a p_T correction factor C_{NN}, allowing the tau's transverse momentum to be determined as $p_T = C_{\mathrm{NN}} \cdot p_T^{\mathrm{seed}}$. The seed, along with all constituents within the larger $\Delta R = 0.4$ cone, is then removed from the event topology, and the process is repeated up to four more times [70].

Electrons and photons (e/γ): As outlined in Section 4.5.1, the first layer of the Correlator associates tracks with calorimeter information, identifying electromagnetic clusters with associated tracks as electrons, while those without tracks are classified as photons. The second layer is responsible for aggregating the tagged candidates, sorting them, and forwarding them to the Global Trigger (GT) in two separate collections. Additionally, it calculates an isolation variable based on other PUPPI candidates within an annular region around the tagged candidate, defined by an inner radius ΔR_{min} and an outer radius ΔR_{max}, as

$$P_T^{\text{isol}} = \sum_{\Delta R_{\text{min}} < \Delta R < \Delta R_{\text{max}}} p_T \quad , \tag{4.27}$$

where the parameters are set to $\Delta R_{\text{min}} = 0.03$ and $\Delta R_{\text{max}} = 0.2$ for electrons, and $\Delta R_{\text{min}} = 0.07$ and $\Delta R_{\text{max}} = 0.3$ for photons [70].

4.6 External Triggers

Additional systems are expected to supply information to the Global Trigger (GT). The beam pickup timing for experiments (BPTX) sends a "BEAM1" and "BEAM2" signal, indicating the presence of bunches in the two colliding beams during a specific bunch crossing [50]. The BPTX system is part of the beam radiation, instrumentation, and luminosity detectors (BRIL) [82], whose detectors may provide further, yet unspecified, data to the Global Trigger (GT) during High-Luminosity operation. Beyond the BRIL detectors, inputs from the precision proton spectrometers (PPS), a system positioned around CMS's interaction point that measures the proton-proton interaction cross-section, are foreseen.

5 The newly developed Level-1 Global Trigger

5.1 Introduction

The Global Trigger (GT) serves as the final stage in the Level-1 Trigger chain, receiving at least 66 links from upstream Trigger systems (Fig. 36) to determine whether a given bunch crossing event should be fully read out by the data acquisition (DAQ) [83]. This readout initiates a more detailed processing at the High-Level Trigger (HLT) [57], which utilises detector data at full resolution to decide whether the event should be stored for offline analysis. During High-Luminosity operation, the GT will consist of up to thirteen boards: twelve dedicated to running cut-based and machine learning-based algorithms, and one Final-OR board reserved for producing a logical "OR" of selected algorithms and forwarding the final trigger decision to the Trigger Control and Distribution System (TCDS), which is responsible for relaying the readout decision.

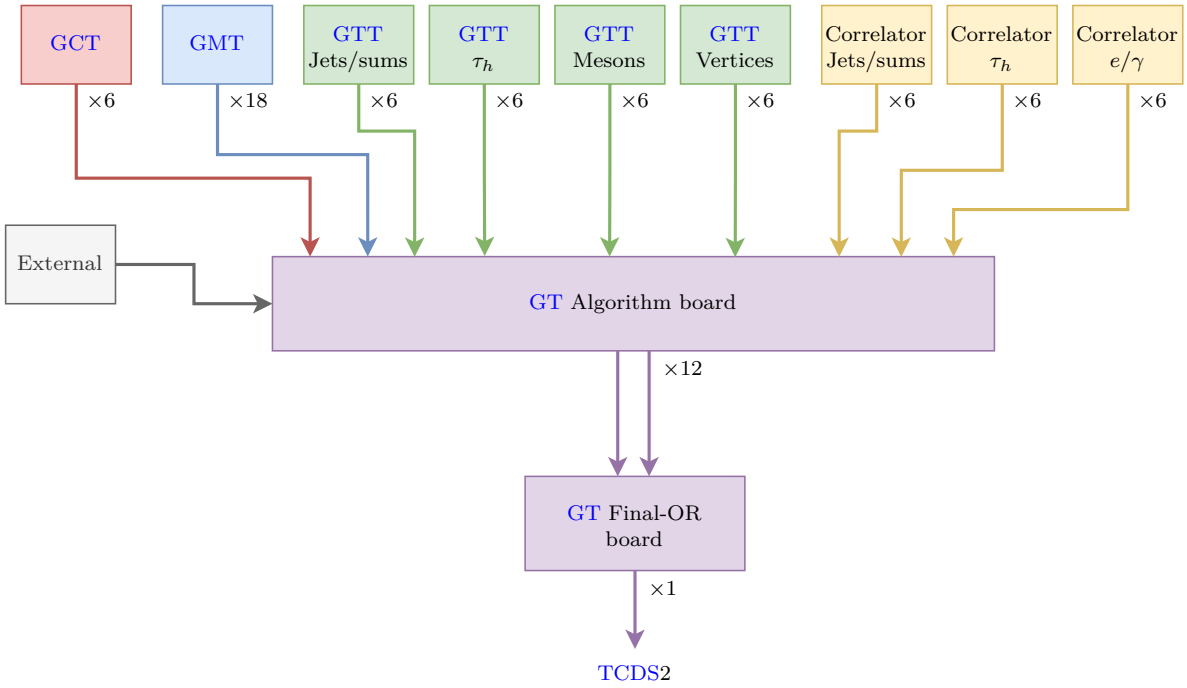


Figure 36: Illustration of the Global Trigger (GT) architecture and input links from upstream Trigger systems. The factor depicted next to each link represents the number of times the corresponding processing unit and link are replicated. For upstream Trigger systems (top), this reflects the TMUX factor, which describes the distribution of consecutive bunch crossing events across multiple units. For GT Algorithm boards, on the other hand, this factor denotes the number of boards available, where each board processes a subset of the full trigger menu.

5.1.1 The Serenity platform

All thirteen GT boards are expected to be Advanced Telecommunications Computing Architecture (ATCA) compatible Serenity boards [84], developed by the Serenity consortium and equipped with Xilinx UltraScale+ FPGAs. The Serenity platform overall comprises an ATCA carrier card, complemented by a firmware and software framework that includes EMP firmware [85], as well as SMASH and EMP software. Earlier designs used daughter cards to host up to two FPGAs, whereas newer versions integrate a single FPGA directly onto the carrier card. This integration reduces the overall board height and enhances cooling efficiency by enabling the use of longer cooling fins, while preserving ATCA-compatible dimensions [86].

The carrier card provides essential services such as power distribution, clocking, optical interfaces, electrical interconnection, intelligent platform management controller (IPMC) functionality, and an onboard CPU (an AMD Xilinx Kria System-on-Module in newer designs [87]) for control and management. The FPGA, by contrast, is dedicated to application-specific processing. In the earlier daughter card designs, a standardised footprint allowed compatibility with a range of FPGAs. The final design, however, integrates a single Xilinx Virtex UltraScale+ VU13P FPGA. In those earlier designs, each daughter card could connect to up to 16 Samtec Firefly modules [88], enabling optical transmission. These modules are available as unidirectional variants with 12 channels or as bidirectional variants with 4 transmit and 4 receive channels, each operating at 25.7 Gb/s. The final design features 10 pairs of unidirectional 12-channel Firefly modules, along with one additional 4×4 bidirectional module [87].

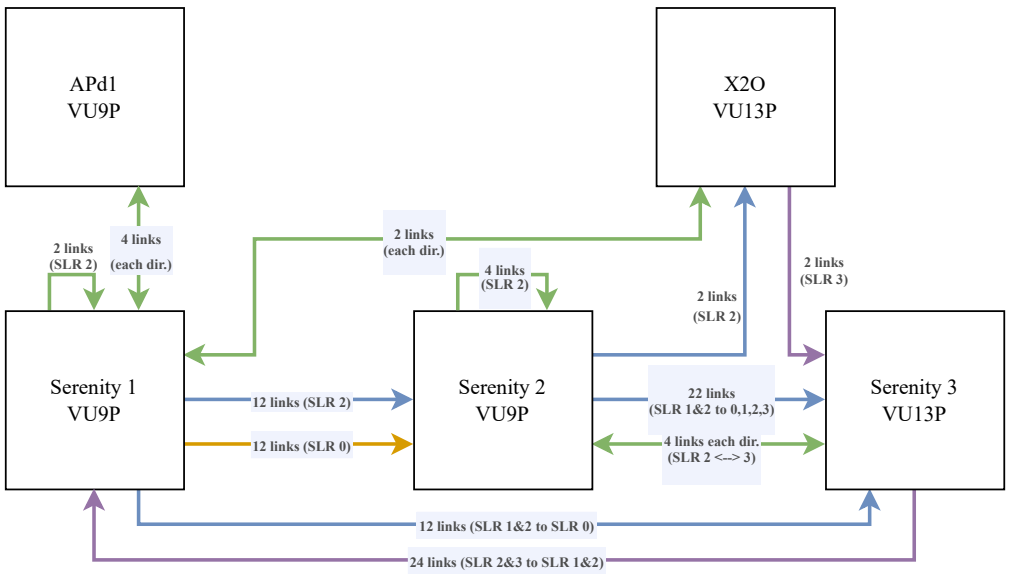


Figure 37: Illustration of available boards in the integration facility together with their interconnections at the time of writing [89]. APd1 and X2O are boards used by the calorimeter trigger and muon trigger, respectively.

The EMP firmware, together with the SMASH and EMP software frameworks, offers a modular architecture that supports multiple daughter cards and FPGAs. The EMP

firmware framework delivers essential common features such as the link protocol, clock generation, and an IPbus [90] configuration interface, while the software framework provides tools for board monitoring and configuration.

At the time of writing, a few Serenity boards are available for testing and evaluation (see Fig. 37). Three boards are specifically designated for integration tests (slice tests) involving optical interconnection between multiple boards. Serenity 1 houses two daughter cards, one equipped with a Xilinx Virtex UltraScale+ VU9P and the other with a Xilinx Kintex UltraScale+ KU15P. Serenity 2 contains a single daughter card with a VU9P, while Serenity 3 features a single daughter card equipped with the larger Xilinx Virtex UltraScale+ VU13P FPGA.

5.1.2 Algorithm boards

During High-Luminosity operation, up to twelve boards are expected to run a trigger menu consisting of up to approximately 1,000 algorithms [91]. Each algorithm of this menu targets one or more physics signatures. In cut-based algorithms, this is achieved by decomposing the algorithm into a logical combination of conditions, ranging from the simplest case of a single condition to more complex multi-condition structures. Each condition searches for between one and four distinct trigger objects that meet a specific set of kinematic constraints (see Fig. 38). These constraints, expressed as inequalities or equalities, can be applied to basic quantities such as the magnitude and direction of a trigger object's momentum vector or correlations between two or more objects. A detailed description of the condition modules and the available constraints or cuts, as well as their implementation, is provided in Section 5.2 and Section 5.3.

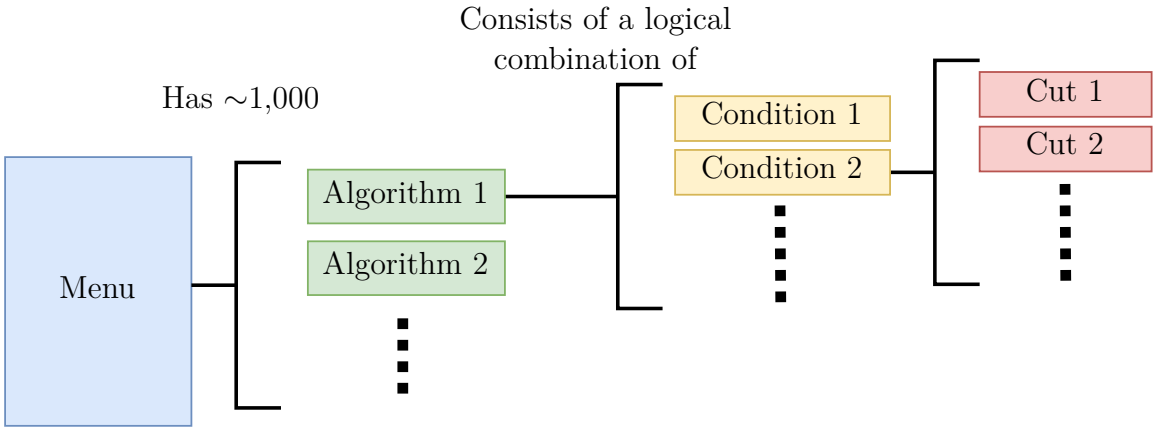


Figure 38: Illustration of the menu building blocks for cut-based algorithms: Conditions are combined using logical operators ("and", "or", "not") to create algorithms, with each condition consisting of a specified set of cuts applied to various quantities [92].

In addition to cut-based algorithms, the GT is also expected to run neural network-based algorithms during High-Luminosity operation [93, 94]. The output of a neural network-based classifier is typically a score between 0 and 1, representing the likelihood that a given bunch crossing event matches the physics signature (or signatures) the

neural network was trained to identify. An algorithm decision can be made by applying a threshold cut to this score.

5.1.3 Final-OR board

The "Final-OR" board [94], implemented on a VU13P FPGA, is tasked with applying configurable "bunch masks" and "pre-scales" to algorithm decisions. These features provide a mechanism to limit the acceptance rate of certain algorithms, enabling quick adaptation during an ongoing data-taking run. The "bunch mask" overrides algorithm decisions for specific bunch indices in the LHC, effectively rejecting selected bunch crossings. Once a "bunch mask" is applied, the decision may be further "pre-scaled". This process involves incrementing an algorithm-specific counter each time the algorithm accepts a bunch crossing. If the counter reaches a configured threshold, the decision is forwarded and the counter resets; otherwise, the decision is ignored. Updates to the "bunch masks" and "pre-scales" take effect at the start of the next luminosity section, defined as 2^{18} revolutions (orbits) in the LHC, or approximately 23.3 s.

In addition to these functions, the Final-OR board includes monitoring capabilities for detailed tracking of algorithm decisions. The monitoring logic consists of counters that increment with each algorithm accept. For every algorithm, four counters are instantiated: one for the initial decision, one for the decision after applying the "bunch mask", one for the decision after applying the "pre-scale", and one for the "pre-scale preview" (functionally identical to the "pre-scale" but with the decision not forwarded). At the end of each luminosity section (2^{18} orbits), counter values are stored in memory and can be read via IPbus [90], with counters resetting afterwards.

After passing "bunch masks" and "pre-scales", the algorithm decisions are grouped via logical "ORs" into up to eight trigger types and sent to TCDS2. Special algorithms designated as vetoes can override the final decision by masking all trigger types. A similar counter-based monitoring system tracks these grouped trigger decisions across the eight types.

5.1.4 Latency budget

From the total latency budget of 12.5 μs allocated to the Level-1 Trigger system, 1 μs is assigned to the GT, corresponding to 40 bunch crossings (BX) at a 40 MHz clock rate. The latency budget is distributed as follows (see Fig. 39). Half of a link latency, amounting to 8 BX, is assigned to the input links and the other half to the output link connecting to TCDS2. One full link latency is allocated to the connections between the GT Algorithm boards and the GT Final-OR board. The GT Final-OR board itself incurs a latency of 4 BX.

For the GT Algorithm board, the latency distribution is broken down as follows. 6 BX are allocated to the demultiplexers (see Section 5.3.2), approximately 1 BX is used for distributing inputs across the entire chip, which is divided into so-called Super Logic Regions (SLRs). 3 BX are reserved for correlations across ± 3 BX, and roughly 1 BX is required to route the results to the output. This leaves approximately 9 BX available for processing algorithms.

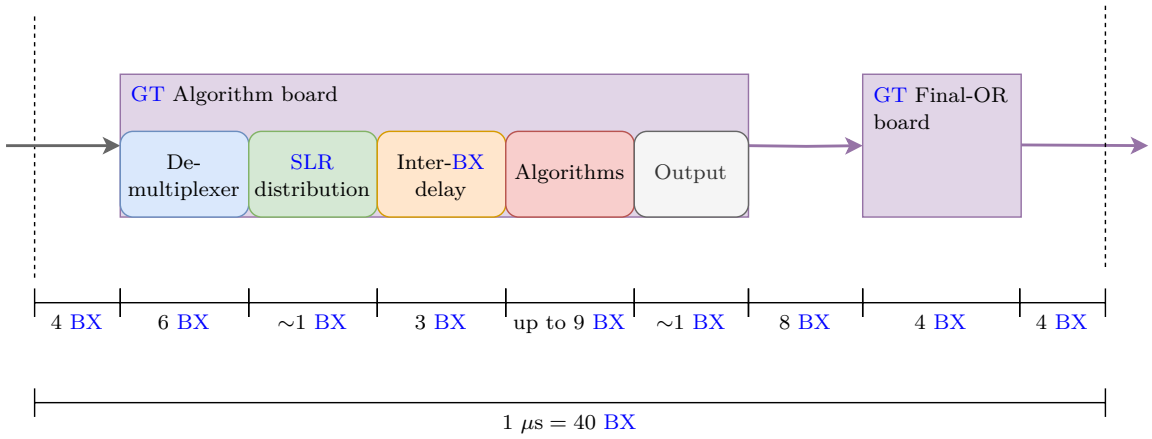


Figure 39: Illustration of the latency distribution across elements of the Global Trigger (GT) in units of the bunch crossing (BX) period, 25 ns. The elements' latencies are not drawn to scale.

5.2 Global Trigger emulation in CMSSW

5.2.1 Introduction: CMSSW

The software framework for CMS is known as Compact Muon Solenoid Software (CMSSW), which includes Monte Carlo event generation, detector and trigger simulations, offline data analysis, and online trigger processing of the High-Level Trigger (HLT). Structurally, it consists of a core framework and various application-specific packages and sub-packages. For instance, the Level-1 Trigger emulators are organised as sub-packages within the *L1Trigger* package.

CMSSW employs an Event Data Model (EDM), where different modules store products in a C++ type-safe event container for each recorded or simulated bunch crossing event [95, 96].

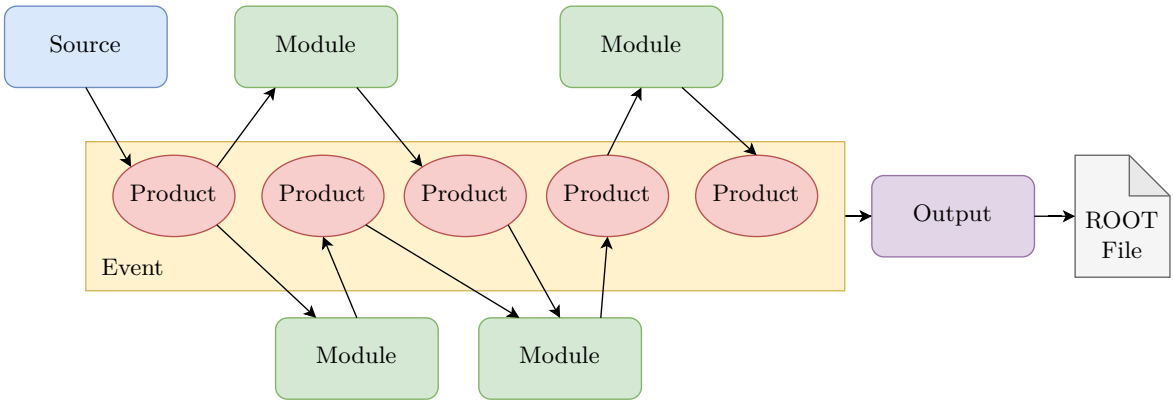


Figure 40: Illustration of CMSSW's Event Data Model (EDM).

The main types of modules are as follows:

- *Source*: Provides events populated with products from local or remote ROOT files, or empty events that can later be populated by an event generator.
- *EDProducer*: Creates new products by processing products from other modules.
- *EDFilter*: Similar to *EDProducer* but also produces a filter result, determining whether subsequent modules in the same *Path* should continue processing the event.
- *EDAnalyzer*: Processes products from other modules without creating new products or affecting subsequent processing.
- *OutputModule*: Writes selected event data to an output ROOT file.

Each module (or plugin) derived from these types declares the products it consumes and produces during construction. This declaration allows the framework to determine the appropriate execution order. Package developers are responsible for ensuring their modules are thread-safe, meaning that modules retaining a state across multiple events should ensure that potential concurrent access is free from unexpected behaviours.

Modules are configured using a *ParameterSet*, which is passed to the framework from the run configuration written in Python. A *ParameterSet* holds various name value pairs, allowing for fine control over a module's runtime behaviour. This flexibility enables adjustments without recompiling the module's package and allows the use of the same module for multiple purposes.

To efficiently utilise *EDFilter* modules, a path structure can be defined to control the execution flow. The result of *EDFilter* modules determines whether subsequent modules on the same *Path* are executed (see Fig. 41). *EDFilter* modules can also be assigned to multiple *Path*s, enabling simultaneous filtering across all of them with only a single *EDFilter* execution. There are two types of paths: standard *Path*s and *EndPath*s. *EndPath*s are executed after all other paths and are typically used for *OutputModule*s and, in some cases, *EDAnalyzer*s.

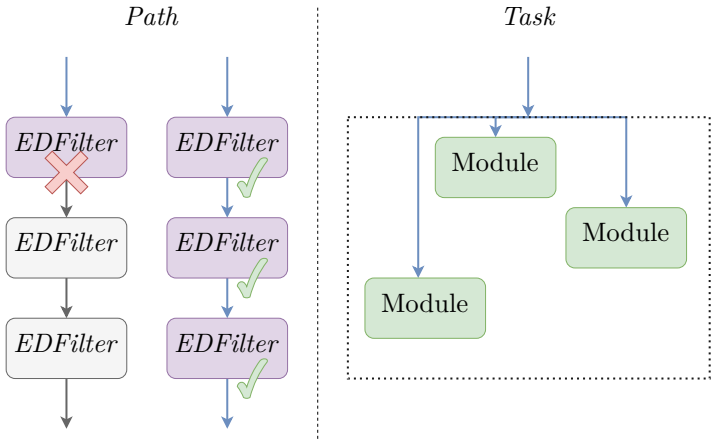


Figure 41: Left: Illustration of ordered execution within *Path*s, where a negative result from an *EDFilter* stops the execution of subsequent modules. **Right:** In a *Task*, all modules are executed, with the execution order determined by the dependencies between the products they consume and produce.

If the execution order is not important, modules can be placed in a *Task*, allowing the framework to manage the execution order automatically while disregarding any filter results.

To better organise reusable, ordered groups of modules, the *Sequence* construct exists. A *Sequence* behaves like a *Path* but is not automatically executed. Instead, it is designed to be included as a sub-entry within one or more *Path*s.

Modules added to *Path*s and *Task*s are executed as part of a process. These *Path*s and *Task*s can be run in either unscheduled mode — where all defined ones are executed — or by associating specific ones with a *Schedule*, limiting execution to the selected *Path*s and *Task*s.

The process can be configured to execute multiple streams, with each stream handling a different event on a separate thread to optimise resource usage on multicore machines. For cases where a single machine is insufficient to complete a simulation or analysis within a reasonable time, the CMS Remote Analysis Builder (CRAB) tool [97] can be used to distribute the workload across multiple machines in CERN's computing grid.

5.2.2 Objectives

The integration of the Global Trigger (GT) emulator within CMSSW serves three essential purposes, each critical to the success of the Level-1 Trigger project:

1. **Firmware evaluation:** To ensure the correctness of the firmware implementation, it must be validated against a proven reference. This is achieved by generating input and output buffer files using a software-based emulator whose implementation matches the firmware as closely as possible. The input buffer files are loaded into the board's buffers, driving the FPGA logic, which then fills the designated output buffers. Comparing the FPGA output buffers with the emulated output buffer files over a sufficiently large range of possible inputs allows for a thorough assessment of the firmware's correctness. This procedure can be applied not only to individual boards but also across multiple boards and emulators (see Fig. 42 and Fig. 43).

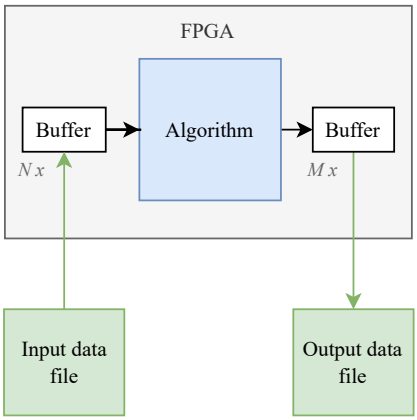


Figure 42: Schematic drawing of a single-board test setup using input and output buffers.

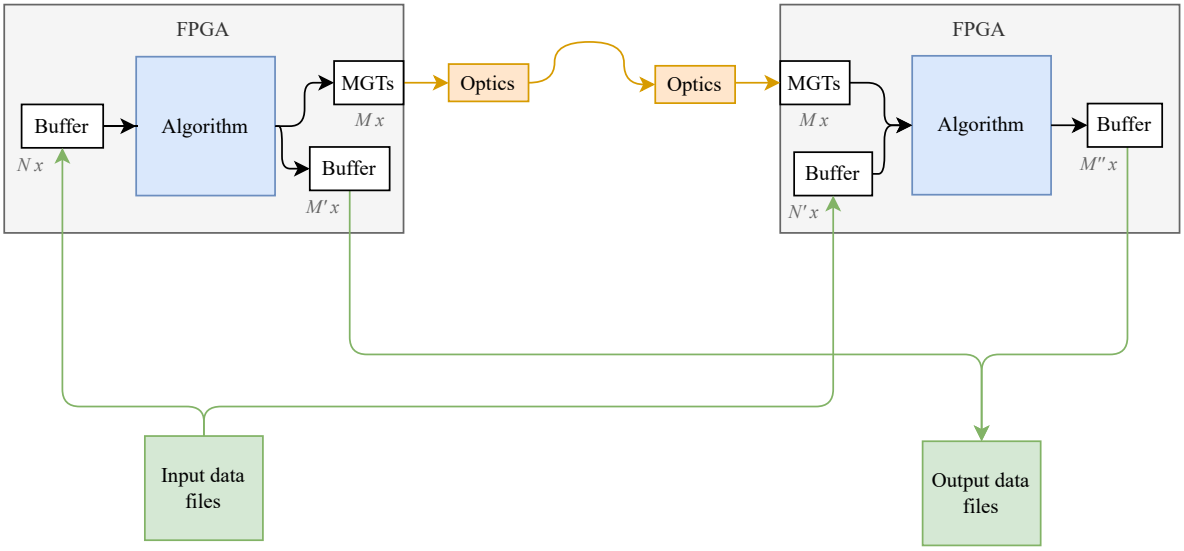


Figure 43: Schematic drawing of a two-board test setup, showing the use of input and output buffers along with the optical interconnection between the two boards. The validation process involves comparing both output buffer files with the corresponding emulated buffer files.

2. **Level-1 Trigger simulation studies:** To evaluate a potential Level-1 Trigger menu in terms of efficiency and trigger rate for integration into the real system, an accurate software representation is essential. This allows a thorough assessment using Monte Carlo simulated events. Efficiency refers to the fraction of accepted to total events containing the physics signature of interest, while the trigger rate represents the total number of accepted (including background) events per second. It is crucial to retain as many interesting events as possible while keeping the trigger rate below the threshold that the data acquisition (DAQ) can handle.

3. **Seeding the High-Level Trigger (HLT):** Every Level-1 Trigger accept decision results in further processing at the HLT, which greatly benefits from understanding how the Level-1 Trigger decision was made. Once an event is accepted, it is first processed by the Level-1 Trigger emulators to generate a detailed record describing which algorithms are responsible for the positive decision as well as the objects passing those algorithms, enabling the HLT to inspect those objects more closely. The seeding could, in principle, also be performed using a readout record generated directly by the FPGA firmware. However, this approach significantly increases firmware complexity, needs additional resources, and — due to its intricacy — is particularly prone to errors. These drawbacks outweigh the modest benefit of slightly reduced computing requirements when the GT emulator is not run as part of the HLT.

Beyond the above purposes, the emulator's Level-1 Trigger menu configuration serves as the baseline, which, after successful validation, can be translated to a VHDL representation [98, 99], ready to be embedded into the rest of the firmware code and then implemented into the target FPGA. This workflow allows quick adaptation to evolving run conditions and shifting physics goals at the CMS experiment.

5.2.3 Structure

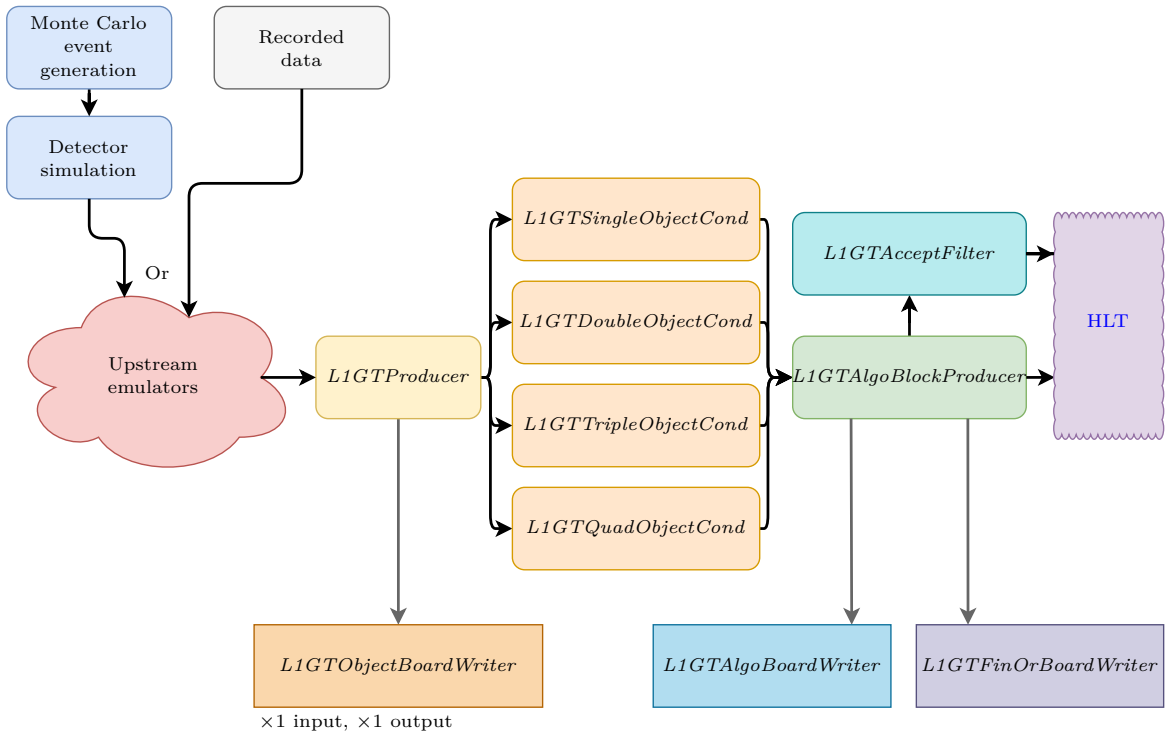


Figure 44: Illustration of the GT emulator modules and their relation to other components within CMSSW. Upstream emulators can process either simulated data (offline) or recorded data during an active data-taking run (online).

Structurally, the GT emulator is divided into multiple modules (or plugins):

- ***L1GTProducer*:** An *EDProducer* module responsible for converting all input trigger objects into an intermediate superset object (*P2GTCandidate*). It standardises variables into a common integer-based format to replicate the calculations performed within FPGAs, which primarily use integer arithmetic as these are more efficiently implementable. Additionally, using scale parameters from the configuration, it stores certain physical floating-point values for use at the HLT.
- ***L1GTSingleObjectCond*:** An *EDFilter* module that searches for a single object within a configurable collection that meets specific criteria (cuts). It utilises the *P2GTCandidate* objects created by *L1GTProducer*. If one or more matching objects are found, references to them are added to the event container, and the filter returns *true*.
- ***L1GTDoubleObjectCond*:** Similar to *L1GTSingleObjectCond*, this *EDFilter* module searches for two distinct objects within configurable collections that meet specific criteria. If matching pairs are found, their references are added to the event container, and the filter returns *true*.
- ***L1GTTripleObjectCond*:** This *EDFilter* module extends the functionality to

three distinct objects, searching for sets that meet the defined selection criteria. References to matching sets are added to the event container, and the filter returns *true*.

- ***L1GTQuadObjectCond*:** Like the previous condition modules, this *EDFilter* module searches for sets of four distinct objects that fulfil the specified selection criteria. References to matching sets are stored in the event container, and the filter returns *true*.

- ***L1GTAlgoBlockProducer*:** An *EDProducer* module configured with algorithm definitions. Each algorithm decision comprises a logical combination of condition *Path* results, including cases where an algorithm consists of only one condition *Path* result. Since standard *Path* logic does not support a logical "OR", each condition is assigned a separate *Path*, with the *L1GTAlgoBlockProducer* handling the logical combination of *Path* results. In addition to generating algorithm decisions, this module emulates the Final-OR board by applying a configurable bunch "mask" and "pre-scales", storing these decisions alongside the initial algorithm results. The final object added to the event container is a map consisting of pairs of algorithm names and *P2GTAlgoBlock* objects, where each pair represents one algorithm.

- ***L1GTAcceptFilter*:** An *EDFilter* module that determines the final Level-1 Trigger decision by logically combining all algorithm decisions using a logical "OR" operation for a specific trigger type while considering potential vetoing algorithms. These decisions are obtained from the map generated by the *L1GTAlgoBlockProducer*. This module enables the HLT to efficiently retrieve the final Level-1 Trigger decision.

- ***L1GTObjectBoardWriter***: An *EDAnalyzer* module that writes input and output buffer files of the trigger objects received by the GT. In input mode, it writes data following the TMUX structure, as illustrated in Fig. 36, where consecutive events are separated into different columns representing links within the file. This organisation mimics the event-wise workload division across multiple boards, where each board sends its reconstructed objects on a separate link to the GT. In output mode, the *L1GTObjectBoardWriter* writes the received objects in a demultiplexed format, where consecutive events are stored sequentially, distributing objects that exceed a single link's capacity across multiple links. Currently, these outputs are only used for validation and are not transmitted, with data being accessed by reading output buffers during testing.

- ***L1GTAlgoBoardWriter*:** An *EDAnalyzer* module that writes algorithm decisions into a file with two output channels (columns) for validating the GT Algorithm boards. An optional mask can be applied to force specific algorithms in the file to zero. This feature allows for the validation of multiple boards, each handling a different set of algorithms, enabling comparison against separate reference files.

- ***L1GTFinOrBoardWriter*:** An *EDAnalyzer* module that writes algorithm decisions, along with the decisions after applying the "bunch-mask" and "pre-scales", into a file. Additionally, the final trigger decision is grouped into trigger

types on a separate link or channel, yielding a total of seven channels or columns. These buffer files are intended for validating the Final-OR board.

By leveraging information from upstream emulators within CMSSW, these modules are executed once their respective input objects have been added to the event container. The *L1GTAlgoBlockProducer* generates a map of algorithm names paired with *P2GTAlgoBlock* objects, which is then utilised by the HLT, forming the processing chain illustrated in Fig. 44. The buffer file writer modules are designed to be placed on an *EndPath*, making them independent or orthogonal to any modules that follow their parent module in the execution sequence.

To minimise code duplication, the condition modules utilise three helper classes. The class *L1GTSingleCollectionCut* applies cuts to a single collection and is instantiated for each object to be constrained. *L1GTCorrelationalCut* handles correlational cuts between object permutations, allowing, for example, a triple-object condition to have correlational cuts for every two-object permutation — specifically, $(1, 2)$, $(1, 3)$, and $(2, 3)$. *L1GT3BodyCut* applies three-object correlational cuts, currently supporting three-body invariant mass (see Section 3.7.4) and three-body transverse mass (see Section 3.7.5). Similar to two-object correlations, three-object correlations can be instantiated for all possible permutations. In the quad-object condition, these permutations are $(1, 2, 3)$, $(1, 2, 4)$, $(1, 3, 4)$, and $(2, 3, 4)$.

The computation of quantities closely follows the firmware implementation (see Sections 5.3.3 to 5.3.7) by relying solely on integer arithmetic and avoiding complex operations such as division and square root calculations. Furthermore, the internal superset object (*P2GTCandidate*) provides access to its variables through "getter" functions that exclusively return *ap_int* types, ensuring integer bit lengths are constrained to match the firmware implementation. Functions like cosh and cos, required for computing invariant mass (Section 3.7.2), transverse mass (Section 3.7.5), two-object P_T (Section 3.7.6), and the $M/\Delta R$ quantity, are handled using precomputed look-up tables (LUTs) to follow the firmware implementation.

5.2.4 Configuration

In CMSSW, modules and their relationships with *Path*s and *Task*s are configured using Python syntax. When a run is initiated from a configuration, the Python interpreter first executes it, performing tasks such as variable substitutions and calculations. The finalised configuration is then passed to the framework, which instantiates modules based on the provided key-value pairs.

To minimise redundancy in common configuration parameters, we extensively leverage Python's capabilities. Parameters like conversion scales, used for translating between hardware and physical values, and the precomputed tables of LUT values are provided through baseline configurations for condition modules within "_cfi.py" files. The four baseline objects, which can be cloned and extended to reuse common parameters, are: *l1tGTSingleObjectCond*, *l1tGTDoubleObjectCond*, *l1tGTTripleObjectCond* and *l1tGTQuadObjectCond*.

A separate baseline configuration is available for the two producer modules, *L1GTPro-*

ducer and *L1GTAlgoBlockProducer*, which typically require modifications only in niche use cases.

The configuration of condition modules follows an ambiguity resolution scheme. When no ambiguities exist, parameters should be defined at the top level of the nested parameter set (*PSet*) structure. For conditions involving multiple objects, a sub-parameter set with the key *collectionX* must be specified to define constraints for each object X. Similarly, for conditions with multiple two- or three-object correlations, sub-parameter sets with keys *correlXY* or *correlXYZ* must be defined for each permutation (X, Y) or (X, Y, Z) requiring constraints. Examples illustrating this configuration scheme are shown in Figures 45 to 47.

```
process.SingleTkMuon22 = l1tGTSingleObjectCond.clone(
  tag = cms.InputTag("l1tGTProducer", "GMTTkMuons"),
  maxAbsEta = cms.double(2.4),
  regionsAbsEtaLowerBounds = cms.vdouble(0, 0.83, 1.24),
  regionsMinPt = cms.vdouble(22, 21, 20)
)
```

Figure 45: Example of the definition of a single-object condition, with the module name *SingleTkMuon22* highlighted in yellow. Since no ambiguities exist, all parameters are defined at the top level.

```
process.DoubleTkEle2512 = l1tGTDoubleObjectCond.clone(
  collection1 = cms.PSet(
    tag = cms.InputTag("l1tGTProducer", "CL2Electrons"),
    minPt = cms.double(20),
    maxAbsEta = cms.double(2.4),
    regionsAbsEtaLowerBounds = cms.vdouble(0, 1.479),
    regionsQualityFlags = cms.vuint32(0b0010, 0b0000)
  ),
  collection2 = cms.PSet(
    tag = cms.InputTag("l1tGTProducer", "CL2Electrons"),
    minPt = cms.double(9),
    maxAbsEta = cms.double(2.4),
    regionsAbsEtaLowerBounds = cms.vdouble(0, 1.479),
    regionsQualityFlags = cms.vuint32(0b0010, 0b0000)
  ),
  maxDz = cms.double(1),
)
```

Figure 46: Example of a double-object condition definition, with the module name *DoubleTkEle2512* highlighted in yellow and the keys for the sub-parameter sets defining constraints on the target objects highlighted in green. While different cuts can be applied to each object, correlations between them remain unambiguous.

```
process.TripleTkMuon533 = l1tGTTripleObjectCond.clone(
  collection1 = cms.PSet(
    tag = cms.InputTag("l1tGTProducer", "GMTTkMuons"),
    minPt = cms.double(5),
    maxAbsEta = cms.double(2.4),
    qualityFlags = cms.uint32(0b0001)
  ),
  collection2 = cms.PSet(
    tag = cms.InputTag("l1tGTProducer", "GMTTkMuons"),
    minPt = cms.double(3),
    maxAbsEta = cms.double(2.4),
    qualityFlags = cms.uint32(0b0001)
  ),
  collection3 = cms.PSet(
    tag = cms.InputTag("l1tGTProducer", "GMTTkMuons"),
    minPt = cms.double(3),
    maxAbsEta = cms.double(2.4),
    qualityFlags = cms.uint32(0b0001)
  ),
  correl12 = cms.PSet(maxDz = cms.double(1)),
)
```

Figure 47: An example of a triple-object condition definition, with the module name *TripleTkMuon533* highlighted in yellow. The keys for the sub-parameter sets defining constraints on the target objects are highlighted in green, while the additional sub-parameter set constraining the correlation permutation $(1, 2)$ is highlighted in red. In this case, both single-object constraints and two-object correlation constraints are ambiguous.

A comprehensive list of cuts along with their definitions can be found in Section A.1. These cuts can be assigned within their respective sections when defining a condition. Any cuts not explicitly specified in the configuration are automatically disabled. This is achieved by initialising the corresponding *std::optional* holding the cut definition with *std::nullopt* during module instantiation.

Since path logic lacks a built-in logical "OR", we opted for a structure where each condition module is on a separate *Path*. The *L1GTAlgoBlockProducer* is then configured with the logical *Path* expression that defines the algorithm, along with an optional algorithm name. If no name is specified, the logical *Path* expression itself is used as the algorithm name.

The algorithm configuration can also include optional parameters to support Final-OR emulation, such as a "pre-scale" factor, a "bunch-mask", a list of trigger types the algorithm belongs to, and whether it should function as a veto. Fig. 48 illustrates an example of an algorithm definition without the optional Final-OR emulation parameters.

```
from L1Trigger.Phase2L1GT.l1tGTAlgoBlockProducer_cff import algorithms

process.pSingleTkMuon22 = cms.Path(process.SingleTkMuon22)
process.pDoubleTkEle25_12 = cms.Path(process.DoubleTkEle2512)

algorithms.append(cms.PSet(
  name = cms.string("TkMuon22_TkEle25_12"),
  expression = cms.string("pSingleTkMuon22 or pDoubleTkEle25_12")
))
```

Figure 48: Example of an algorithm definition that includes a logical combination of *Path*s, with the referenced *Path* definitions highlighted in blue.

When evaluating the firmware, the buffer file writers described in Section 5.2.3 must also be configured. Default configurations are available for boards equipped with a VU9P and VU13P. However, depending on the specific tests, modifications may be required to account for the available links in the integration facility (see Fig. 37).

It is important to note that in multithreaded mode, limited control over the event execution order can lead to inconsistencies in event ordering across different buffer files when using multiple buffer file writers. A workaround that also ensures a consistent event order with generated output root files is producing buffer files in single-threaded mode only. Additionally, to maintain a uniform number of events across files from different systems, it is highly recommended to explicitly define the number of events per buffer file.

5.2.5 Conversion scales: Mapping physical to hardware values

The approach for handling the conversions from physical values to hardware integers was designed to address three key objectives:

1. **Bookkeeping:** When reading a recorded root file, it must be possible to retrieve the exact physical representation of converted hardware values.
2. **Single source of truth (SSOT):** The hardware conversion process should be identical whether running in CMSSW or generating VHDL module instantiations using the VHDL writer tool [98].
3. **Adaptability:** Newer versions of CMSSW should be executable with older scale parameter definitions without requiring package recompilation.

To meet all objectives, scale parameters are first defined within a Python configuration, serving as the single source of truth (SSOT). These parameters are automatically included in the condition and producer module configurations, allowing CMSSW's default configuration storage system to manage bookkeeping. Defining these parameters in Python also enables modifications without requiring recompilation, making it possible to use older scale parameters with newer CMSSW versions.

Additionally, the conversion functions, implemented in C++, are exposed to Python via

pybind11 [100]. This allows the VHDL writer tool [98] to utilise the same conversion functions as the CMSSW emulator when translating cut parameters.

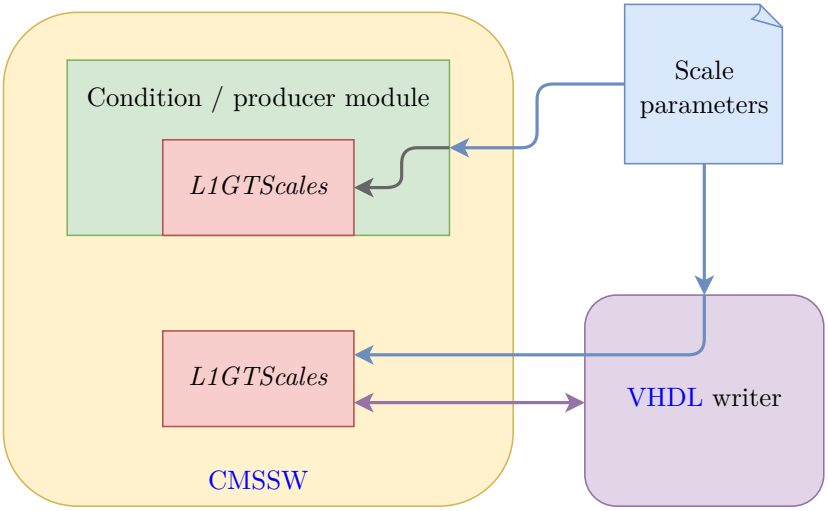


Figure 49: Illustration of accessing hardware conversion functions by instantiating the *L1GTScales* class within CMSSW and the VHDL writer tool, using identical scale parameters from a shared Python configuration file.

To ensure optimal alignment with a floating-point-based tool used for developing potential algorithms for High-Luminosity operation, the hardware conversions are designed to avoid rounding and truncation effects that could introduce inconsistencies.

These effects can be illustrated with a simple example. Consider the cut $\eta < 2.4$, which, when converted to hardware representation using an LSb corresponding to $\pi/2^{12}$ radians, results in

$$\eta_{\text{HW}} < 3129.1135 \quad . \tag{5.1}$$

If the cut value were simply rounded or truncated, inconsistencies with a floating-point-based emulator could arise when $\eta_{\text{HW}} = 3129$ (approximately 2.3999 radians).

To prevent such discrepancies, all "less than" cuts are converted using the ceil function, while all "greater than" cuts are translated with the floor function. Applying this method to the example above results in

$$\eta_{\text{HW}} < 3130 \quad , \tag{5.2}$$

which ensures that the boundary value $\eta_{\text{HW}} = 3129$ is correctly included.

Certain truncation effects in variable conversion are inevitable to some extent when using reasonable bit widths for calculations. This applies to cuts defined as a ratio of two variables, such as the relative isolation cut (Tab. 7). In this case, the smallest theoretical difference that can occur is

$$\Delta \frac{P_T^{\text{isol}}}{p_T} = \frac{0.25 \text{ GeV}}{2047.9375 \text{ GeV}} - \frac{0.25 \text{ GeV}}{2047.96875 \text{ GeV}} = 1.86 \times 10^{-9} \quad . \tag{5.3}$$

To limit truncation effects, we ensured a precision of 2^{-15} (approximately 3×10^{-5}) for the relative isolation cut. Validation with 14,000 $t\bar{t}$ events and 1.5 million "Minimum Bias" events showed no discrepancies compared to the floating-point-based tool. While this does not guarantee the complete absence of differences, it ensures that any potential deviations remain negligible when assessing algorithms in terms of trigger rate and efficiency.

5.3 Firmware implementation

5.3.1 Introduction: FPGAs

Field-programmable gate arrays (FPGAs) are programmable integrated circuits, with "field" referring to their ability to be configured by the user after manufacturing, "in the field". They consist of various discrete functional blocks positioned at fixed locations on a microchip, which can be configured (placed) and then interconnected (routed) using an underlying connection grid.

Common primitive components within an FPGA, which may exist as separate blocks or be combined into configurable logic blocks (CLBs), include [101]:

- **Look-up tables (LUTs):** LUTs are digital logic elements with a predefined number of inputs, whose output is determined by a configurable "truth table", which refers to a table stating for each combination of digital input signals the corresponding output signal. This enables the implementation of a wide range of logic operations using the same physical LUT primitive.
- **Flip-flops (registers):** Flip-flops are digital storage elements capable of maintaining two stable states. To prevent metastability, they are typically clocked, meaning state updates occur only on specific clock signal transitions (synchronous logic). Unclocked or asynchronous flip-flops, known as latches, are generally avoided. The most common type in FPGAs is the D flip-flop, which captures an input signal (D) and holds it at the output (Q). Many D flip-flops also include a set or reset input that asynchronously forces the Q output to one or zero, respectively.
- **Shift registers:** Used for delaying signals by a specific number of clock cycles, shift registers help optimise resource usage by reducing the need for individual flip-flops. They are often implemented as shift register look-up tables (SRLs), where the output delay can be dynamically adjusted via input signals.
- **Distributed RAMs:** Certain LUTs can be configured to store and retrieve data at addresses defined by input signals. In single-port mode, a single address input is used for both read and write operations. By allocating additional LUTs, dual-port, quad-port, or even octa-port configurations can be realised, each providing separate address inputs for reading. By combining multiple such single-bit-wide elements, a distributed RAM of the desired data width can be created.
- **Carry logic chains:** To preserve valuable LUT resources for common carry bit operations, certain FPGAs incorporate dedicated carry logic chains within some or all CLBs.

- **Block random-access memorys (BRAMs):** A BRAM is a fixed-size storage element (e.g., 36 Kbits on UltraScale+ FPGAs [102]) accessed via an address input. It supports configurable width (bit width of individual data elements) and depth (number of data elements). Single-port BRAM permits only one access (read or write) at a time. A simple dual-port BRAM supports both read and write operations, but not concurrently. In contrast, a true dual-port BRAM provides two fully independent ports, each with its own address, data input, data output, and clock signal, enabling simultaneous read-write, dual-read or dual-write operations. The simultaneous read-write capability is particularly useful for transferring large data blocks between different clock domains.
- **Digital signal processing (DSP) blocks:** DSPs are designed for efficient multiplications and often include additional features like pre-adders, accumulators, and pattern detectors. In UltraScale+ FPGAs, they support 27×18-bit signed multiplications with a 48-bit output [103], where larger operations can be performed by combining results from multiple DSP blocks.

Beyond these core elements, FPGAs often include specialised components for clock signal generation and distribution, as well as high-speed communication interfaces for data transmission and reception.

FPGA designs can be developed using either graphical tools or text-based languages. Among text-based languages, one can distinguish between the very descriptive hardware description languages such as VHDL and Verilog and "high-level" languages that typically relate strongly to common software programming languages such as C++. Text-based designs require an additional synthesis step, where FPGA vendor tools translate the described constructs into a netlist of primitive components and their interconnections.

The GT relies solely on VHDL for the implementation of its infrastructure and cut-based algorithms due to the language's high degree of control over latencies, resource utilisation, and logic distribution across registers. All of these features are crucial for implementing a highly adaptable menu of ∼1,000 cut-based algorithms while keeping the total latency within the 1 μs budget allocated to the GT.

5.3.2 Demultiplexers, SLR distribution and inter-bunch crossing delay

The demultiplexing architecture builds upon a previous design by Elias Leutgeb [99], with minor modifications to generalise the deserializer VHDL module. This adaptation allows it to support any link structure by passing the appropriate generics that define the logical link configuration during module instantiation.

The deserializer module operates as follows: Serial link data, received as a stream of 64-bit words within the 360 MHz clock domain, are assigned to buffers based on their index in the data stream, following a repeating link structure. Each trigger object collection, consisting of twelve objects packed into multiple 64-bit words, is placed into a separate buffer. Once all buffers for the link are filled, a signal that is synchronised with all other deserializers triggers the read-out process in the 480 MHz clock domain. The read-out process cycles through all twelve buffer addresses to sequentially retrieve trigger objects of 64, 96, or 128 bits from all stored collections simultaneously (see Fig. 50).

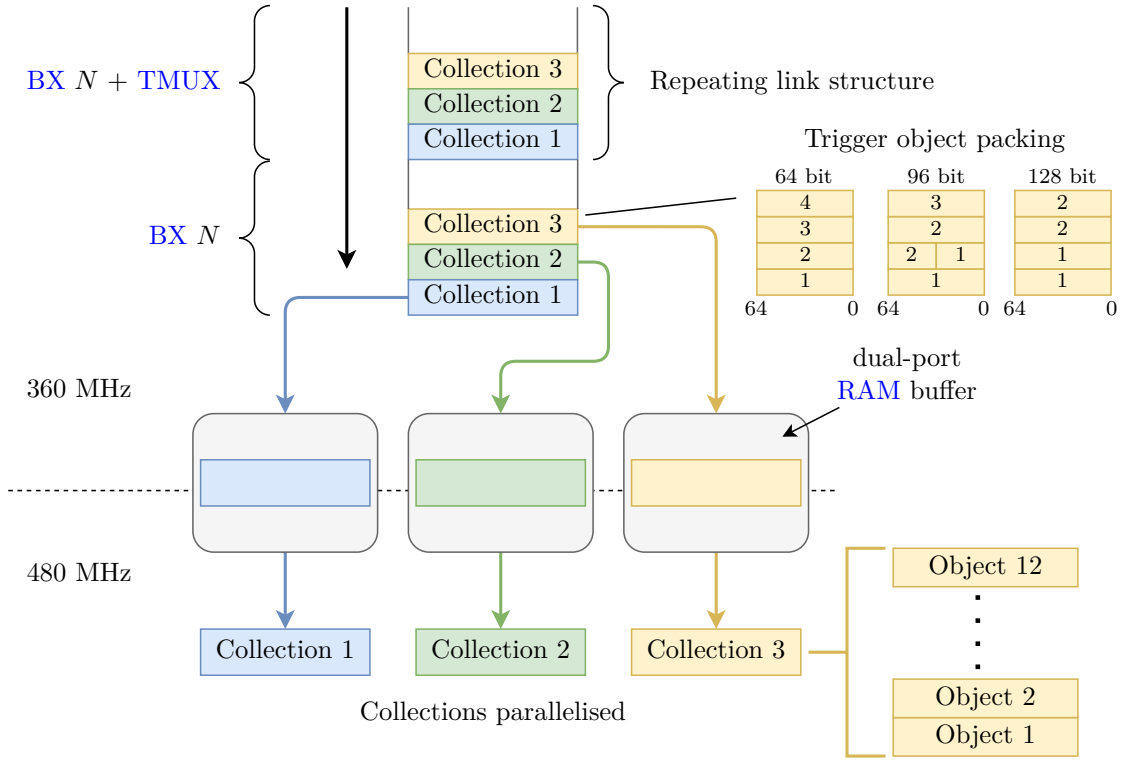


Figure 50: Diagram depicting the logic of the deserializer module, where trigger object collections received via a serial link structure are distributed across multiple dual-port RAM buffers. The upper right side illustrates how trigger objects of sizes 64, 96, and 128 bits are packed into 64-bit words on the link.

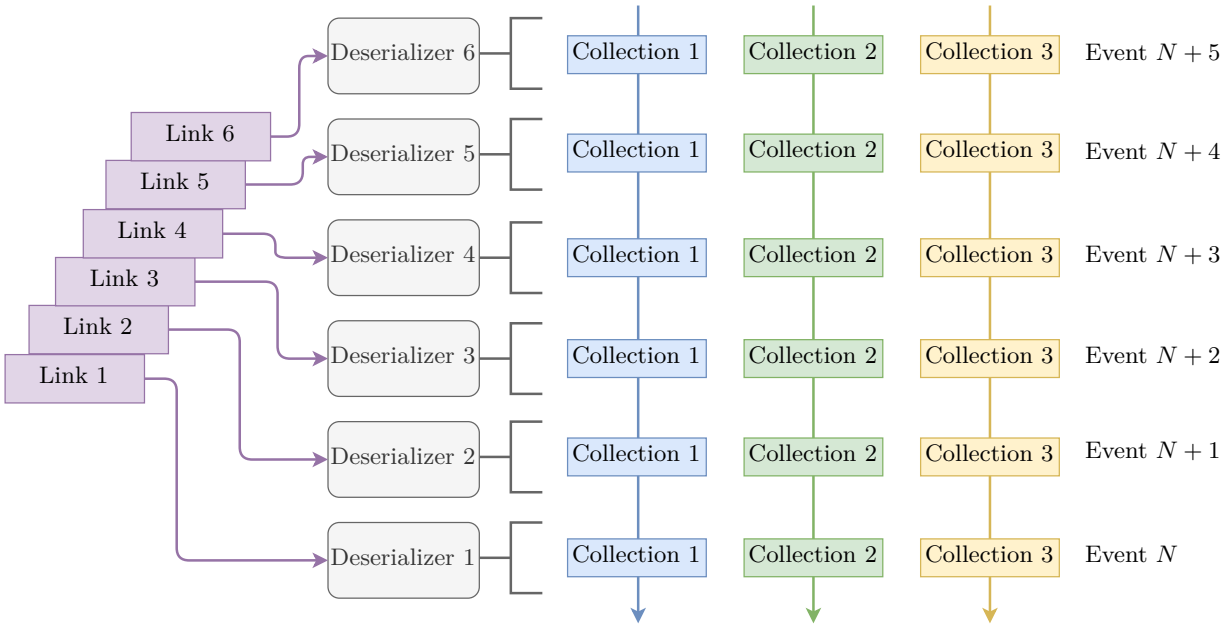


Figure 51: Diagram illustrating how multiple deserializers, each assigned to a separate link, work together to form the demultiplexed event chain.

Due to the TMUX structure, where data from different bunch crossings are processed by separate boards and arrive at different links, each deserializer module is instantiated either eighteen times (GMT) or six times (all other systems). As a result, each deserializer provides the collections of one event, followed by a gap of either five or seventeen bunch crossings. To properly demultiplex the data and fill these gaps with data from the remaining bunch crossings, all deserializers handling a specific link type dynamically inject their collections into the stream based on whether their buffers contain data or are still being filled. This approach effectively generates a continuous stream of trigger objects, separated into their respective collections (Fig. 51). The processing clock frequency of 480 MHz was deliberately chosen to be twelve times the bunch crossing frequency of 40 MHz, enabling the serial processing of an entire collection of twelve objects.

The large UltraScale+ FPGAs used in the GT consist of either three (VU9P) or four (VU13P) so-called Super Logic Regions (SLRs). Each SLR is essentially an independent silicon die, interconnected via stacked silicon interconnect (SSI) technology to form a single large FPGA. This technology employs a passive silicon layer with embedded traces to connect the SLR dies. However, connections between SLRs typically have higher propagation delays due to their increased length, requiring careful guidance to the vendor tools to infer SLR-crossing register chains without additional logic elements to achieve proper timing closure.

Since data arrives at different SLRs, and subsequent algorithms require access to the full event topology, it is essential to distribute the arriving data streams (currently spanning 29 collections) across all SLRs. This is achieved through register chains, where specific registers are constrained to individual SLRs, ensuring their placement in the designated SLR by vendor tools. For example, a set of collections originating in SLR 0 must pass through a register chain into SLR 1, then SLR 2, and finally into SLR 3. By the end of this process, each SLR contains an identical copy of the objects in the collections. Similar data paths are established for all collections to ensure the uniform distribution of objects across the entire FPGA.

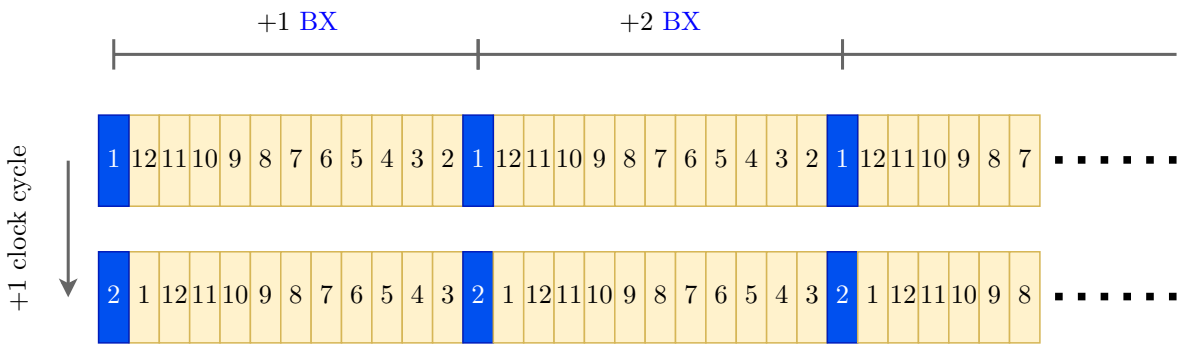


Figure 52: Illustration demonstrating the operating principle of the inter-bunch crossing delay module. The shift register connects to the output at intervals of twelve, with the corresponding register highlighted in dark blue.

Before being processed by the condition modules, each trigger object collection passes through a shift register with output connections at intervals of twelve (Fig. 52). This interval corresponds to the maximum number of trigger objects within a collection and represents the time between two consecutive bunch crossing events, given that a shift occurs on each rising edge of the 480 MHz clock. There are a total of seven output connections, providing access to event data shifted by 0 to 6 bunch crossings, with indexing ranging from -3 to 3. This delay mechanism enables correlations across different bunch crossings by selecting objects from a specific indexed bunch crossing, where 0 represents the default index.

The inter-bunch crossing delay module is instantiated for each collection within each SLR, ensuring that a set of bunch crossings to choose from is available for every trigger object collection. Condition modules then internally select a specific bunch crossing index according to their configuration. This functionality enables correlations across consecutive bunch crossings, a feature that is particularly intriguing for triggering on various proposed long-lived particle signatures [104, 105].

Beyond providing collections to the condition modules, an additional signal is generated to indicate the presence of at least one valid object in any collection. This signal is produced by performing an "OR" operation with all "valid" bits received as an extra control signal at the input links. Once generated, the signal is transferred to the 480 MHz clock domain, where it remains synchronised with the corresponding collections and follows the same distribution across SLRs, including the delay introduced by the inter-bunch crossing delay module. The signal enables object indexing within a collection by initiating a wrapping counter, which runs from 0 to 11, upon the signal's transition from zero to one. This indexing mechanism is extensively used throughout the design.

5.3.3 Cuts on simple single-object quantities

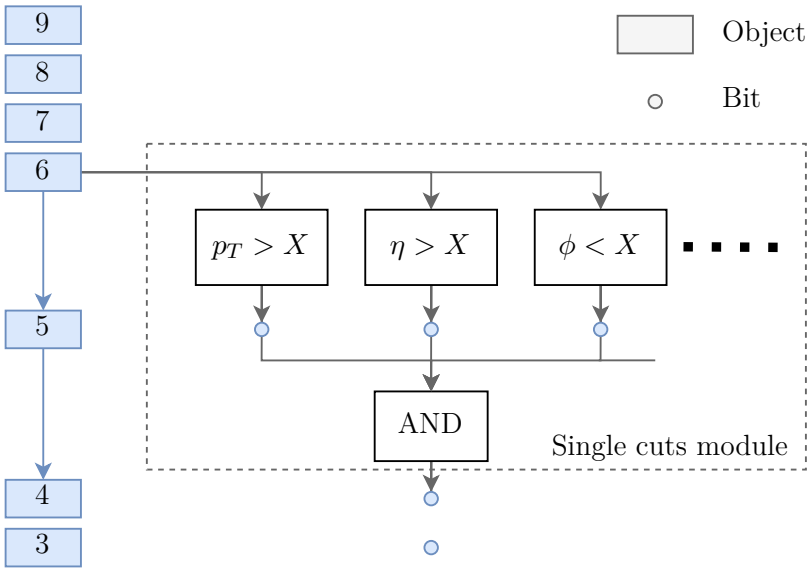


Figure 53: Diagram illustrating the data flow into and out of the simple single cuts module, where dots represent a registered bit and rectangles denote a registered trigger object.

The selected collections pass through a module that applies simple single-object cuts to various quantities, such as the components of the momentum vector $(p_T, \eta, \phi)^T$, impact parameters z_0 (and eventually d_0), various quality metrics, and the scalar sum of p_T. The original object stream within each collection, as produced by the demultiplexers, remains unchanged, with comparisons performed simultaneously across all quantities to constrain one object. The results of these parallel comparisons for each object are registered for one clock cycle before being logically combined using an "AND" operation. This yields a final bit indicating whether an object in the collection passes all defined simple single-object cuts.

5.3.4 Cuts on advanced single-object quantities

The more advanced single-object cuts follow the same principles as the simpler ones but involve computations with longer latencies. These include the η-regional cuts (Tab. 8), which allow for different thresholds in different detector sections, and the relative isolation cut, defined as $P_T^{\text{isol}}/p_T \lesseqgtr X$.

For simple η-regional cuts (p_T and quality flags), the module applies the cuts for all regions simultaneously, generating a list of results. Meanwhile, the η-region selection logic determines the appropriate index for a specific η-region by comparing the object's η value with the configured region boundaries. This index is then used to select the corresponding cut result from the list.

The relative isolation cut is computed using $P_T^{\text{isol}} \lesseqgtr X \cdot p_T$. This requires multiplying the cut value X by p_T, a task typically performed using a single DSP. However, large X values may necessitate the use of two DSPs. To ensure proper pipelining and utilise the internal registers of the DSP, this operation requires five clock cycles to complete. The computed result is then compared against the absolute isolation value P_T^{isol} (see Sections 4.3.4 and 4.5.2). The cut value X is either chosen dynamically from a predefined list, based on the η-region selection index or kept constant when applied uniformly across the entire η-range.

5.3.5 Cuts on two-object correlations

With a clock frequency of 480 MHz (twelve times the LHC clock of 40 MHz) two-object correlations, like single-object cuts, are designed for efficient reuse of FPGA components. This efficiency is achieved by first parallelising one of the two trigger object collections involved in the correlation computations. To do this, objects from one collection are first aggregated in a shift register over twelve 480 MHz clock cycles. They are then transferred to another register, where they remain for the next twelve cycles. With this collection now parallelised, the second collection is streamed past it, allowing each streamed object to form correlation pairs with the twelve parallel objects. A dedicated correlation module then handles the correlation computations for each object pair, aggregates all correlations, and ultimately combines them into a single result.

The key advantage of this design is that the correlation logic needs to be instantiated only twelve times to process all $12^2 = 144$ correlations. This represents a significant improvement over the current Run 3 system [106,107,108], which computes all correlations

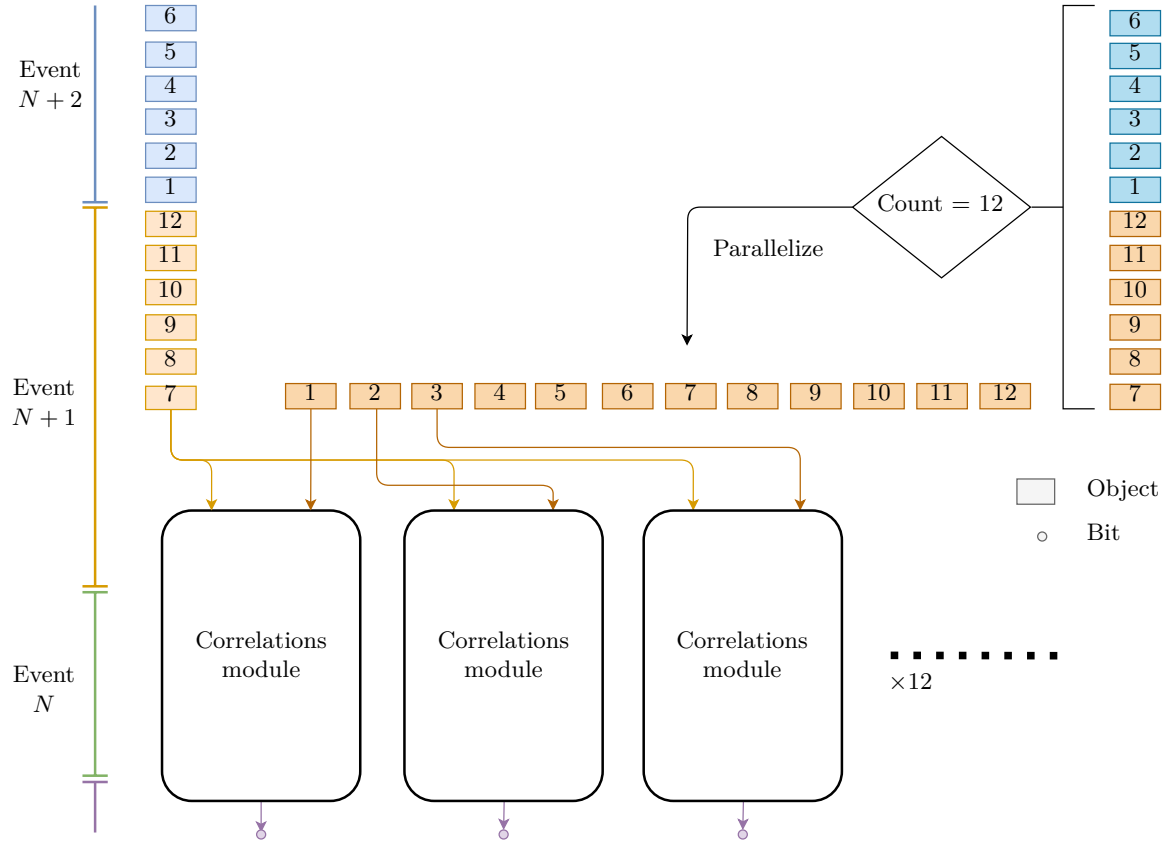


Figure 54: Diagram illustrating the data flow into the correlation module, which is instantiated twelve times and requires one collection to be parallelised first. Dots represent registered bits, while rectangles indicate registered trigger objects.

in parallel, albeit with fewer objects per collection.

A comprehensive list of implemented two-object correlations, along with their corresponding mathematical expressions, is provided in Tab. 9. Below is a brief overview of how various correlations are computed within the correlations module:

- $\Delta\eta$: Calculated as the absolute difference between the two pseudorapidities, $\Delta\eta = |\eta_1 - \eta_2|$.
- $\Delta\phi$: To ensure the smallest azimuthal angle difference within the range $\Delta\phi \in [0, \pi]$, both $\Delta\phi_{1,2} = |\phi_1 - \phi_2|$ and $\Delta\phi_{2,1} = |\phi_2 - \phi_1|$ are computed, with $\phi_{1|2} \in [-\pi, \pi)$. These values are stored using the same bit width as $\phi_{1|2}$, interpreted as unsigned values while ignoring possible overflows that reflect $\Delta\phi$ periodicity. The final $\Delta\phi$ is then determined as the unsigned minimum of the two, which by construction is confined to the interval $[0, \pi]$.
- ΔR^2: The angular distance ΔR (see Section 3.7.1) is computed as ΔR^2 using the results of the computation of $\Delta\eta$ and $\Delta\phi$. Two separate DSPs are employed to calculate both $\Delta\eta^2$ and $\Delta\phi^2$, which are then summed to obtain ΔR^2.
- Δz_0: The difference in longitudinal impact parameter z_0 is computed as a simple

absolute subtraction, $\Delta z_0 = |z_{0_1} - z_{0_2}|$.

- Combined two-particle transverse momentum P_T: The two-object P_T (see Section 3.7.6) is calculated as P_T^2 by first evaluating the products $p_{T_1}^2$, $p_{T_1}^2$ and $p_{T_1} p_{T_2}$ using a single DSP for each. Simultaneously, $\Delta\phi$ is used to retrieve the value of $\cos(\Delta\phi)$ from a precomputed LUT, where stored values are scaled by a factor C_s to avoid subsequent floating-point operations. The retrieved $\cos_{\text{LUT}}\,\Delta\phi$ value is multiplied by $p_{T_1} p_{T_2}$ using two DSPs. Meanwhile, the products $p_{T_1}^2$ and $p_{T_1}^2$ are scaled with the same factor C_s using two DSPs per scaling operation. The term $p_{T_1} p_{T_2} \cos_{\text{LUT}}(\Delta\phi)$ is left-shifted by one bit to account for the additional factor of two before all terms are added together to yield the final two-particle P_T.
- Invariant mass M: The two-object invariant mass (see Section 3.7.3) is computed as $M^2/2$. This involves calculating $p_{T_1} p_{T_2}$ using a single DSP while simultaneously retrieving $\cosh(\Delta\eta)$ and $\cos(\Delta\phi)$ from precomputed LUTs. To optimise resource utilisation and minimise latency, a single scaling factor C_s is applied to both LUTs. However, due to the vastly different ranges of $\cosh\Delta\eta \in [1, \sim 143{,}376)$ with $\Delta\eta \in [0, 4\pi)$ and $\cos\Delta\phi \in [-1, 1]$ with $\Delta\phi \in [0, \pi]$, two distinct $\Delta\eta$-regimes are introduced at $\Delta\eta = 2\pi$, each with its own scale factor (see Fig. 55). In the upper regime ($\Delta\eta \geq 2\pi$), the retrieved value for $\cos(\Delta\phi)$ is neglected.

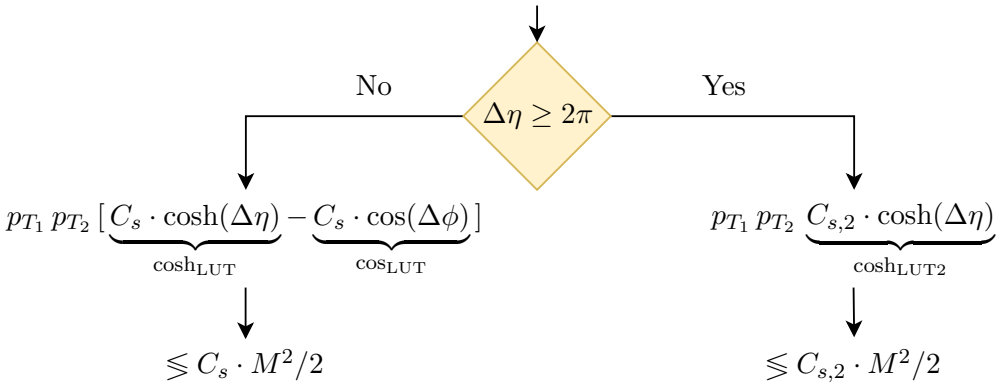


Figure 55: Diagram depicting the invariant mass calculation split at 2π ($\cosh(2\pi) \simeq 268$) [92].

After retrieving the appropriate values, $\cos_{\text{LUT}}(\Delta\phi)$ is subtracted where applicable, and the difference is multiplied by $p_{T_1} p_{T_2}$ using two DSPs. The result is then compared against the predefined cut value corresponding to the relevant $\Delta\eta$-regime. To conserve BRAM resources, the read-only $\cosh_{\text{LUT2}}$ is instantiated only if $\Delta\eta \geq 2\pi$ is not excluded by other η cuts, which typically applies to silicon tracker objects and muon candidates (see Sections 3.4 and 3.6). Furthermore, both cosh LUTs were optimised to fit within a single 36 kb BRAM each, utilising 11 address bits and 18 data bits, while the cos LUT fits within a single 18 kb BRAM with 11 address bits and 9 data bits.

- Transverse mass M_T: Similar to invariant mass, the transverse mass (see Section 3.7.5) is computed as $M_T^2/2$. The computation starts by evaluating $p_{T_1} p_{T_2}$

using a single DSP, while $\cos(\Delta\phi)$ is retrieved from a LUT. The difference $C_s - \cos_{\text{LUT}}(\Delta\phi)$ is then computed, where C_s represents the value 1 scaled to match the LUT scaling. This result is multiplied by $p_{T_1}p_{T_2}$ to obtain the final value $M_T^2/2$ before being compared against the predefined cut value.

- Invariant mass over angular distance $M/\Delta R$: To compute the ratio of invariant mass to angular distance, the following condition is evaluated

$$C_{s|s,2} \cdot M^2/2 \lesseqgtr X \cdot \Delta R^2 \quad . \tag{5.4}$$

Both $M^2/2$ and ΔR^2 are calculated as described earlier. The term $X \cdot \Delta R^2$ is computed using four DSPs, after which a comparison with $M^2/2$ is performed. Since this comparison is based on a ratio, fractional accuracy is achieved by left-shifting $C_{s|s,2} \cdot M^2/2$ to account for fractional bits within X. Similar to the invariant mass cut, two separate comparators are used to account for the different $\Delta\eta$-regimes, ensuring the appropriate scaling factor — C_s or $C_{s,2}$ — is applied.

The simplest form of correlations, the two charge correlations — "same sign" and "opposite sign" — are computed outside the correlations module through direct comparisons between the streamed object and all parallel ones.

5.3.6 Cuts on three-object correlations

The current three-object correlation cuts include the three-body invariant mass and the three-body transverse mass (see Sections 3.7.4 and 3.7.5). Since both are essentially sums over all two-body permutations, they directly utilise the results computed in the two-object correlations module. The difficulty in implementing this stems from the large number of possible three-body mass combinations, given by $12^3 = 1{,}728$.

To maintain a general implementation capable of handling correlations across different collection types and bunch crossings, we opted not to restrict the calculation to correlations within a single collection type. Such a restriction would have significantly reduced the number of possible three-body mass combinations to

$$\binom{12}{3} = \frac{12!}{3!(12-3)!} = 220 \quad . \tag{5.5}$$

Given the large number of combinations, the bit widths of input two-body masses are dynamically limited based on the cut value. This bit-width adjustment is configured during the instantiation of the three-object correlations module and ensures routability of the design by effectively dropping both most and least significant bits of the inputs.

- **Most significant bits** of the input two-body masses are dropped if they exceed the most significant bit set of the (lower) cut value. Since three-body mass calculations only involve positive terms, an early comparison of the cut value against each two-body mass can be performed. If any two-body mass already exceeds the threshold, a single bit is collected to override any subsequent sums that include this specific two-body mass.

- **Least significant bits** of the inputs are dropped based on the desired resolution relative to the cut value. The resolution of squared two-body masses is set to 2^{-15} of the (lower) cut value's most significant set bit, yielding

$$\frac{\Delta M_{123}}{M_{123}} = \frac{1}{2}\left[\frac{\Delta M_{12}^2}{M_{123}^2} + \frac{\Delta M_{13}^2}{M_{123}^2} + \frac{\Delta M_{23}^2}{M_{123}^2}\right] < 0.005\% \quad . \tag{5.6}$$

  This demonstrates that reducing the input resolution results in only a negligible loss of cut precision, even allowing for further tightening of the input resolution if improved routability and reduced resource consumption are required.

Due to the two different η-regimes resulting in two differently scaled LUTs used for the computation of the two-body invariant mass, all input two-body invariant masses must first be normalised, a step that is not necessary for the transverse mass. Since the scaling factors for the two cosh LUTs differ by an exact power of two, normalisation is performed via a simple left-shift applied to invariant masses in the lower η-regime.

Masses are then compared against the cut value to collect overriding bits before being constrained to the resolution of 16 bits. As described in the previous section, two-body masses are computed by keeping one collection in parallel while streaming the other past it to form all correlations. This results in the parallel computation of twelve masses over twelve clock cycles. The same approach is used for three-body mass calculations, but in this case, $12^2 = 144$ operations are performed in parallel over twelve clock cycles.

The computation proceeds as follows:

1. All 144 two-body masses from one of the three possible permutations ($M_{12}^2/2$) are aggregated and stored in a register for the next twelve clock cycles.
2. The other two permutations, each processed as a stream of twelve two-body masses, first compute $M_{13}^2/2 + M_{23}^2/2$ in 144 parallel operations.
3. In the following clock cycle, all $M_{12}^2/2$ values are added to the previous results, forming $M_{123}^2/2$.
4. One clock cycle later, all 144 results are compared against the cut value while accounting for any set overriding bits.

The computation of the three-body transverse mass follows the same approach but forgoes the added normalisation step.

5.3.7 Other cuts

Additional cuts include:

- **A z_0-distance from a selected primary vertex (PV):** Implemented by first selecting a PV Z_0 from the corresponding GTT collection based on a configured index. Once selected, trigger objects from the target collection are streamed past the PV Z_0, and absolute differences $|z_0 - Z_0|$ are computed, which are one clock cycle later compared against a predefined threshold [109].

- **A multiplicity threshold cut:** Implemented by first applying a threshold cut to specific quantities (currently only p_T). The number of instances where the threshold is exceeded within a single bunch crossing is then counted. If this count meets or exceeds a preconfigured value N, all objects within the event are accepted; otherwise, all are rejected.

- **A partial sum cut:** Implemented by summing specific object quantities (currently the quality score) within a bunch crossing. To simplify implementation, the number of contributing terms matches the number of condition target objects (e.g., two for a double-object condition, three for a triple-object condition, etc.). As with other cuts, one collection's quantities are streamed over twelve clock cycles, while up to three other collections' quantities are parallelised, computing 12^{N-1} sums in parallel. An optimisation is applied when summing from a single collection, reducing parallel operations to $\binom{12}{N-1}$ by skipping equivalent permutations. This ensures that the design remains routable when instantiated within the quad-object condition.

5.3.8 Combining cut results

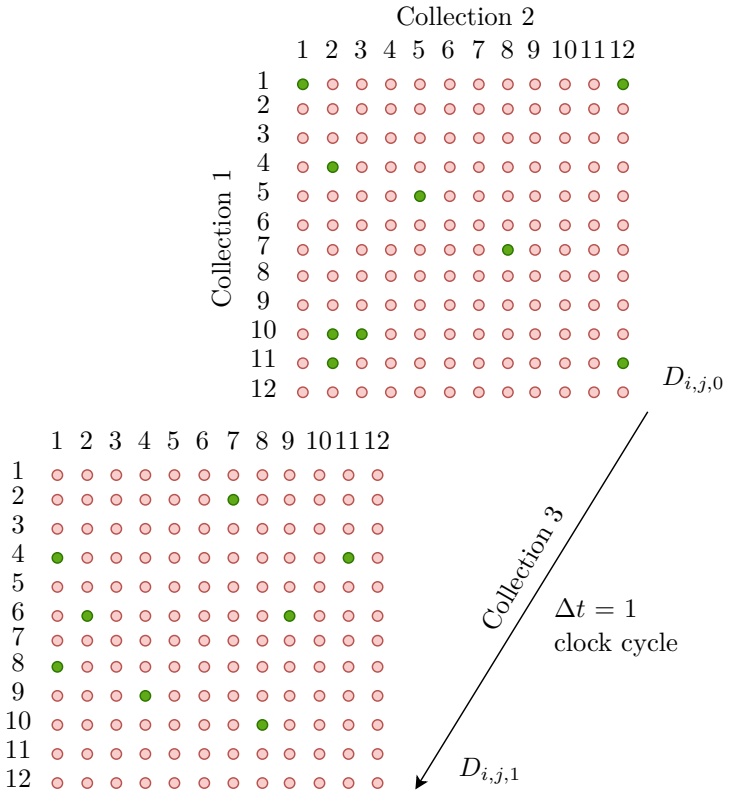


Figure 56: Illustration of the decision tensor $D_{i,j,k}$ for the triple-object condition. Each dot represents a decision bit, with green indicating one and red indicating zero. The last collection's dimension is always temporal, spanning twelve clock cycles, similar to the other dimensions.

Given the significant variation in latencies among the implemented cuts and the GT's overarching goal of collapsing information into single bits as early as possible to conserve register resources, the final decision tensor D, representing passing object combinations, is constructed incrementally throughout the condition module's pipeline. This tensor has order N, corresponding to the number of targeted objects of the condition, with each dimension being of length twelve, one of which temporal spanning twelve clock cycles (Fig. 56).

The initial tensor is constructed from the results of simple and advanced single-object cuts, charge correlations, the PV z_0-distance cut, as well as multiplicity and partial sum cuts. These decisions are first parallelised across all collections except the last one, whose decisions remain streamed.

In the triple- and quad-object conditions, the cut combination process occurs in stages. Initially, two or three vectors are generated by merging the results of each parallel collection with the streamed one. In the subsequent clock cycle, these vectors are combined to form either a matrix or a third-order tensor encompassing all spatial dimensions.

A critical step in constructing the decision tensor is eliminating diagonals when two input collections contain the same trigger objects. This prevents double-counting an object for different cuts. For instance, when identifying two muons satisfying $p_{T_1} > 18$ GeV and $p_{T_2} > 10$ GeV, the muon meeting the p_{T_1} requirement should not also be counted toward the p_{T_2} cut. Diagonal removal can apply to spatial dimensions (see Fig. 57) or the temporal one, using a counter-based indexing.

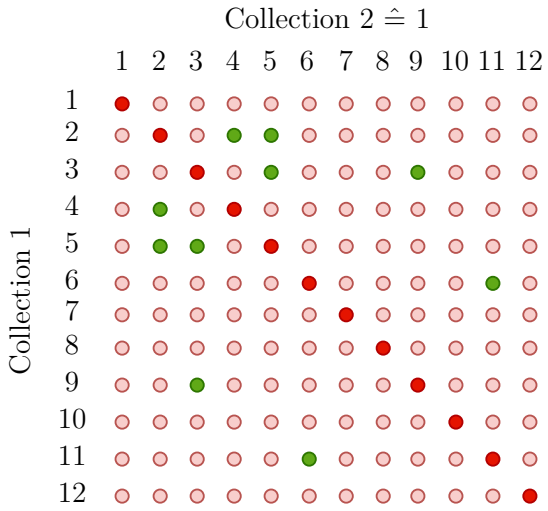


Figure 57: Illustration of the decision tensor $D_{i,j}$ where collections 1 and 2 contain the same objects, necessitating diagonal removal by forcing it to zero. In such cases, the tensor must also be symmetric under the exchange of identical collections, i.e., $D_{i,j} = D_{j,i}$.

Once the results from the two- and three-object correlation modules are available, they are merged with the previously constructed decision tensor using logical "AND"

operations. After integrating all cuts, the tensor undergoes temporal reduction via logical "OR" operations across twelve clock cycles, collapsing it into a single spatial sub-tensor. The spatial dimensions are then further reduced using "OR" operations to produce a single bit reflecting the final decision of the condition module. In the quad-object condition, the final spatial reduction is distributed over two clock cycles to minimise the number of logic blocks (LUTs) between registers, ensuring design routability.

For algorithms that involve logical combinations of conditions, condition decision bits can be merged outside the condition modules using a predefined logical expression to generate the final algorithm result.

5.3.9 Implemented demonstrator design

To evaluate the design's efficiency in terms of the number implementable algorithms, we created a dummy menu with complexity comparable to or greater than that of the first prototype menu for High-Luminosity operation (Section A.2). The design was successfully deployed on both the Serenity board with a VU9P and the one with a VU13P. The VU9P dummy menu contains 252 algorithms (Fig. 58 and 59), while the VU13P version accommodates 336 (Fig. 60 and 61).

Since the designated target FPGA is the VU13P, only three Serenity boards are needed to run a comparable menu of approximately 1,000 algorithms. This leaves nine boards available for advanced algorithms, such as those using machine learning classifiers.

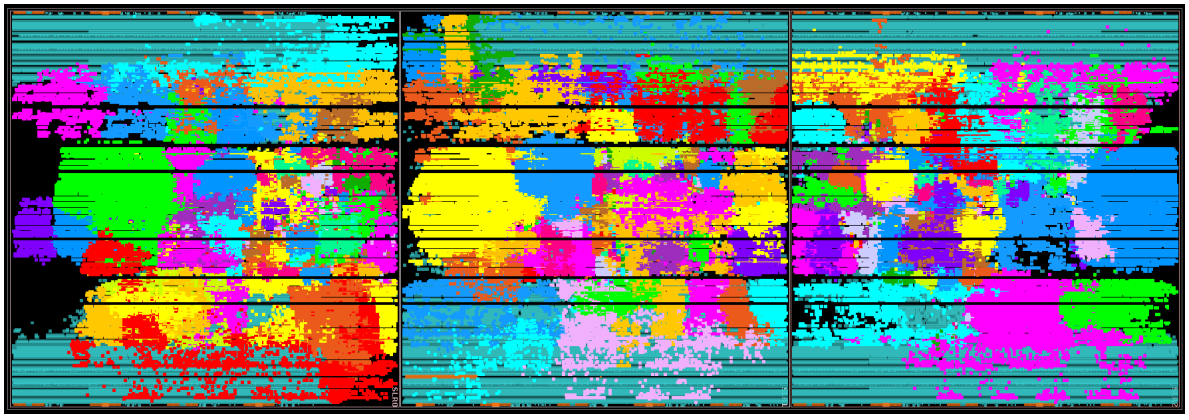


Figure 58: Floor-plan of the Global Trigger (GT) design featuring a dummy menu with 252 algorithms implemented on a Xilinx VU9P FPGA. The three distinct sections, represent the individual SLRs, with SLR 0 positioned on the left. The cyan blocks along the edges correspond to the EMP framework [85] and demultiplexers, while the various colours in the centre each represent one of the 252 implemented algorithms.

The resource utilisation plots (Fig. 59 and 61) further highlight that there is still available capacity to accommodate modifications to the EMP framework, link structure and additional GT functionality. This is particularly evident as the two most heavily utilised components, DSPs and BRAMs, are primarily used for specific correlational cuts (DSPs and BRAMs) and the EMP framework's link buffers (BRAMs). Both are expected to remain largely unaffected by potential future developments.

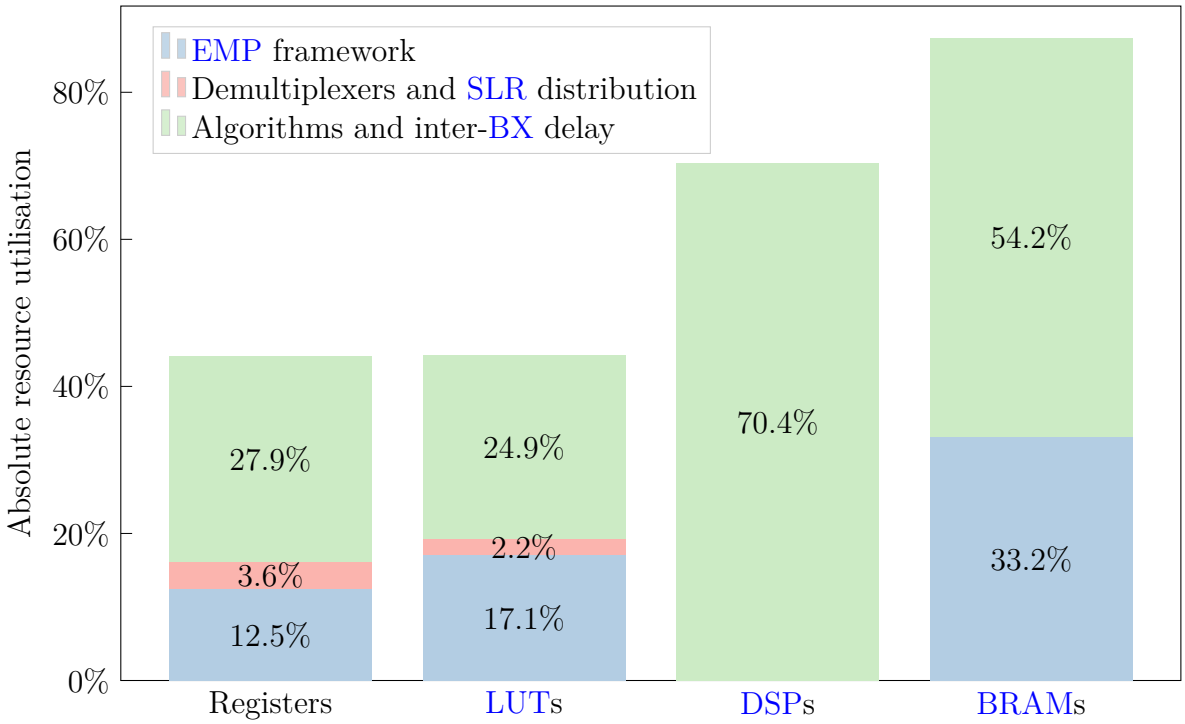


Figure 59: Plot depicting the resource utilisation of various FPGA components relative to the total available resources of the VU9P for the demonstrator design with 252 implemented algorithms. The resource utilisation is further broken down into the three functional blocks: EMP framework; demultiplexers and SLR distribution; and algorithms including the inter-BX delay.

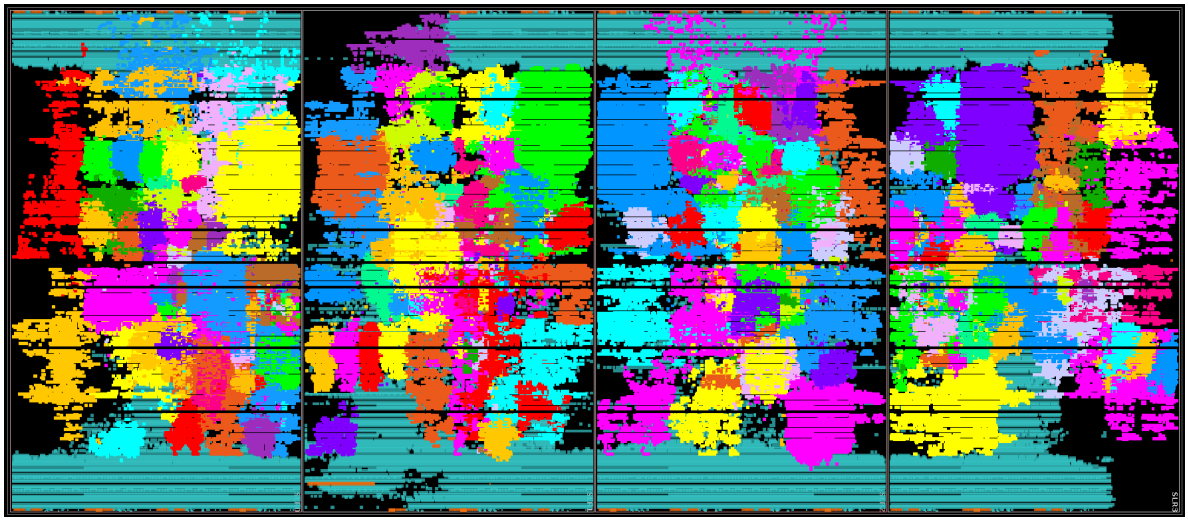


Figure 60: Floor-plan of the Global Trigger (GT) design with a dummy menu containing 336 algorithms implemented on a Xilinx VU13P FPGA. The design is divided into four sections, representing the SLRs, with SLR 0 positioned on the left. The cyan blocks at the edges correspond to the EMP framework and demultiplexers, while each colour in the centre represents one of the 336 implemented algorithms.

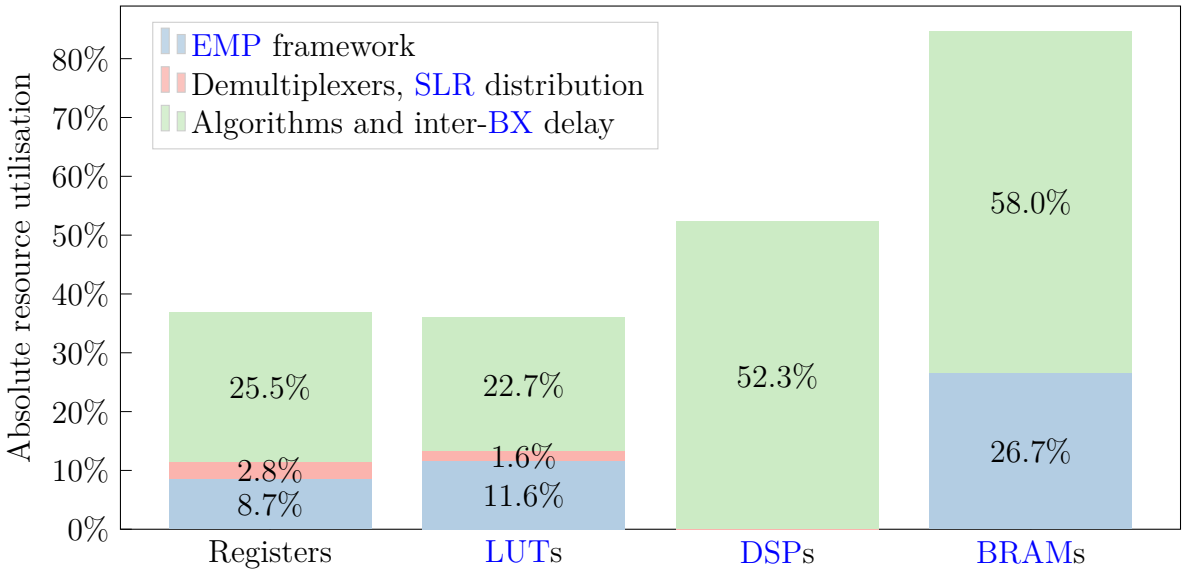


Figure 61: Plot illustrating the resource utilisation of various FPGA components relative to the total available resources of the VU13P in the demonstrator design with 336 implemented algorithms. Utilisation is further categorised into the three functional blocks: EMP framework; demultiplexers and SLR distribution; and algorithms including the inter-BX delay.

5.3.10 Validation results

To validate the firmware against the CMSSW reference implementation, a representative menu for High-Luminosity operation is required. The current prototype of this menu is detailed in Section A.2. The reference menu algorithms are emulated in CMSSW using a sample with sufficient variation in algorithm decisions to generate input and output buffer files. The VHDL writer tool [98] is then used to translate the reference menu into an equivalent VHDL representation. Once integrated into the firmware scaffold, the resulting design enables output generation through either simulation or hardware testing, allowing for direct comparison with the reference files from the emulation.

This validation approach was applied to a sample of 100,000 $t\bar{t}$ events, each including an average of 200 superimposed pileup collisions (see Section 3.2), selected for their expected high variability in algorithm decisions (see Fig. 62). No discrepancies were observed between the CMSSW emulator and the firmware across this large dataset, reinforcing our high confidence in the correctness of the firmware implementation.

In addition to the rigorous validation mentioned above, all implemented cuts (see Section A.1), including those not present in the prototype menu, were tested using both a dummy menu within the standalone tool (see also Section 5.3.9) and a separate dummy menu within CMSSW. The standalone tool validation generated trigger objects from random numbers, while the CMSSW validation employed emulated trigger objects using a sample of 10,000 $t\bar{t}$ events, see Fig. 63. In both cases, no discrepancies were observed between the emulator and the firmware.

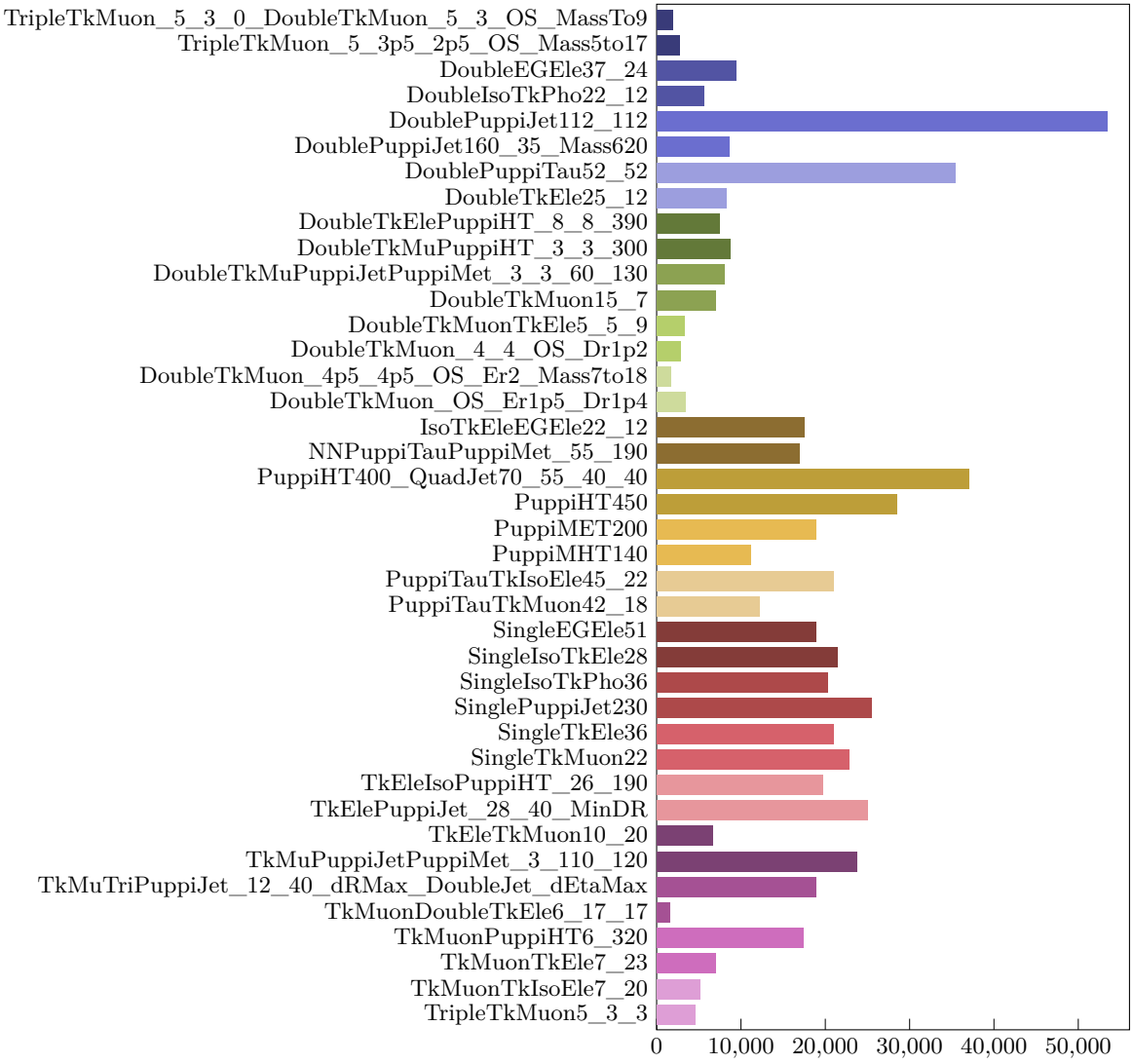


Figure 62: Vertical histogram plot depicting the number of accepted events per algorithm in a sample of 100,000 $t\bar{t}$ events. The algorithms correspond to those in the first prototype menu for High-Luminosity operation (see Section A.2).

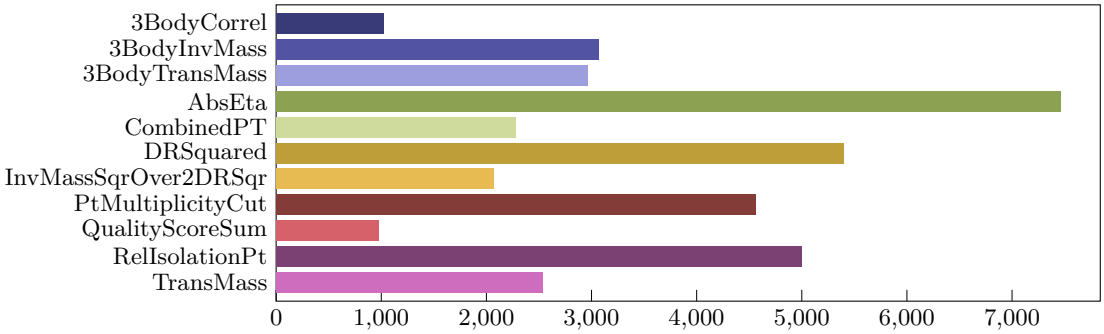


Figure 63: Vertical histogram plot showing the number of accepted events per algorithm in a sample of 10,000 $t\bar{t}$ events. The algorithms were designed to evaluate cuts which are not yet utilised in the prototype menu.

5.4 Computation accuracy

The LUTs for cosh and cos were pre-filled with scaled-up precomputed values. The cosh function was divided into two $\Delta\eta$-regimes at $\Delta\eta = 2\pi$ (see Section 5.3.5). The scaling factor C_s was chosen such that

$$\max_{\Delta\eta\in[0,2\pi),\Delta\phi\in[0,\pi]} [\cosh_{\mathrm{LUT}}(\Delta\eta) - \cos_{\mathrm{LUT}}(\Delta\phi)] \quad , \tag{5.7}$$

fits within the signed 18-bit input of a single DSP — equivalent to an unsigned 17-bit input — while maximizing resolution. Similarly, for $\Delta\eta \geq 2\pi$, the scale factor $C_{s,2}$ was chosen so that $\cosh_{\mathrm{LUT2}}(\Delta\eta)$ fits within the signed 18-bit DSP input, while ensuring that the ratio $C_s/C_{s,2}$ is an exact power of two. This ratio constraint simplifies the computation of the three-body invariant mass by making the normalisation operation trivially implementable in firmware.

However, enforcing this ratio requires capping some LUT outputs for $\Delta\eta > 12$ at the maximum positive value representable by an 18-bit signed integer. Since the maximum expected $\Delta\eta$ occurs when two objects are reconstructed in opposite HF sections, with each section covering $3.0 < |\eta| < 5.2$ (see Section 3.5.2), this capping does not compromise the accuracy of invariant mass calculations.

To conserve BRAM resources within the FPGA, the address width is also reduced by dropping the two least significant bits of $\Delta\eta$ and $\Delta\phi$ before using them as address inputs to the precomputed LUT.

The invariant mass error due to neglecting the $\cos(\Delta\phi)$ term in the upper $\Delta\eta$-regime can be evaluated by computing its maximum value, which is given by[1]

$$\frac{\Delta M}{M}(\Delta\eta = 2\pi, \Delta\phi = 0) = \frac{\sqrt{\cosh(2\pi)} - \sqrt{\cosh(2\pi) - \cos(0)}}{\sqrt{\cosh(2\pi) - \cos(0)}} = 0.2\,\% \quad . \tag{5.8}$$

This small error can be regarded as negligible, particularly when compared to the detector's intrinsic energy resolution, discussed below. Furthermore, it remains confined to the immediate vicinity above the transition.

The overall errors introduced by the LUTs in computing the invariant mass M, transverse mass M_T, two-object transverse momentum P_T, and the derived quantity $M/\Delta R$ are evaluated by calculating these variables from reconstructed trigger objects using a $t\bar{t}$ sample of 30,000 events.

[1] A previously reported calculation erroneously overestimated this error [92].

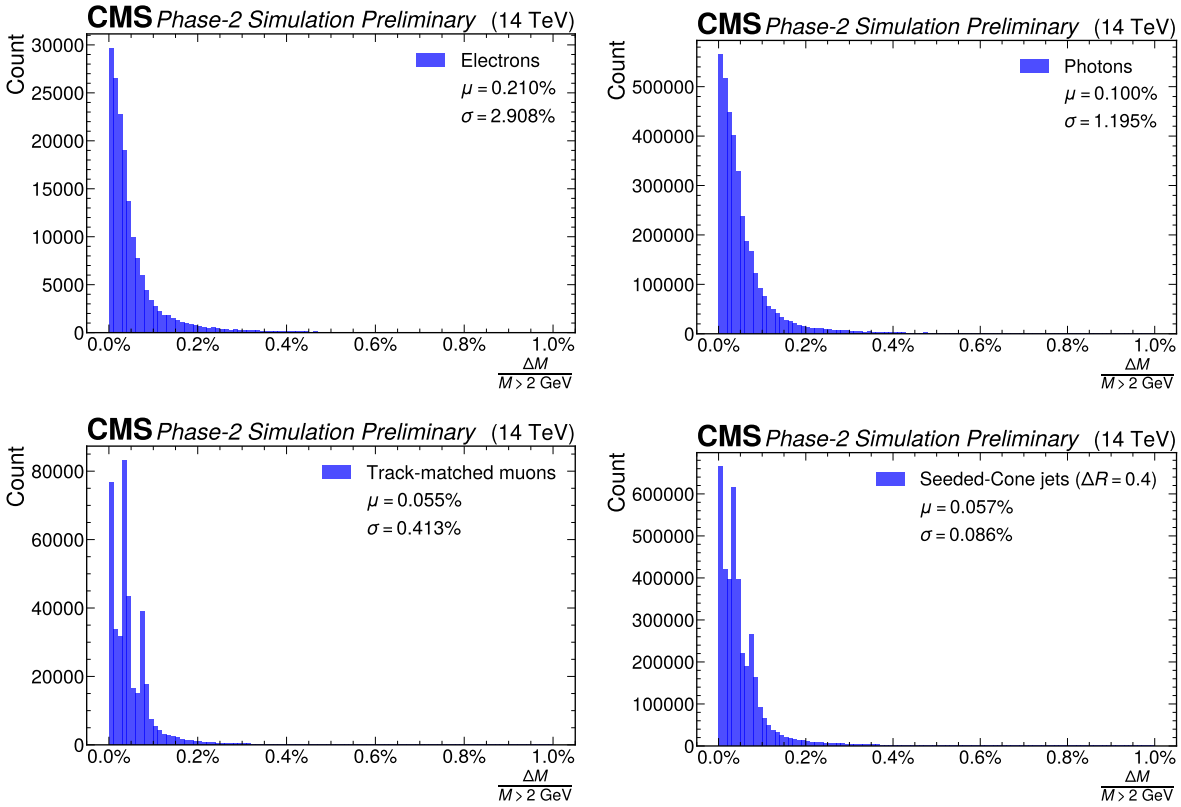


Figure 64: Histogram plots showing the distribution of relative invariant mass errors for various trigger objects, computed for all combinatorically possible invariant masses $M > 2$ GeV within a sample of 30,000 $t\bar{t}$ events.

Since the computation of ΔR does not rely on LUTs, the relative error of the derived quantity $M/\Delta R$ follows that of the invariant mass M,

$$\frac{\Delta\left(\frac{M}{\Delta R}\right)}{\frac{M}{\Delta R}} = \frac{\frac{\Delta M}{\Delta R} + \frac{M}{\Delta R^2}\cancel{\Delta(\Delta R)}}{\frac{M}{\Delta R}} = \frac{\Delta M}{M} \quad . \tag{5.9}$$

To compare the observed relative errors (Fig. 64 and 65) with the intrinsic energy resolution of the detector, we first express the relative invariant mass uncertainties in terms of the transverse momentum resolution using error propagation. This provides a lower bound on the relative invariant mass uncertainties, neglecting contributions from uncertainties in pseudorapidity η and azimuthal angle ϕ,

$$\frac{\Delta M}{M} = \frac{\Delta M_T}{M_T} = \frac{1}{2}\sqrt{\left(\frac{\Delta p_{T_1}}{p_{T_1}}\right)^2 + \left(\frac{\Delta p_{T_2}}{p_{T_2}}\right)^2} \quad . \tag{5.10}$$

In the ultrarelativistic limit, $\Delta p/p \approxeq \Delta E/E$ holds, which in the case of no uncertainties in pseudorapidity η can be expressed as $\Delta p/p = \Delta p_T/p_T$. Using the calorimeter with

the best resolution, the barrel ECAL, we can plot the invariant mass uncertainties as a function of p_{T_1} and p_{T_2} (Fig. 66), applying the resolution given in Eq. 3.20.

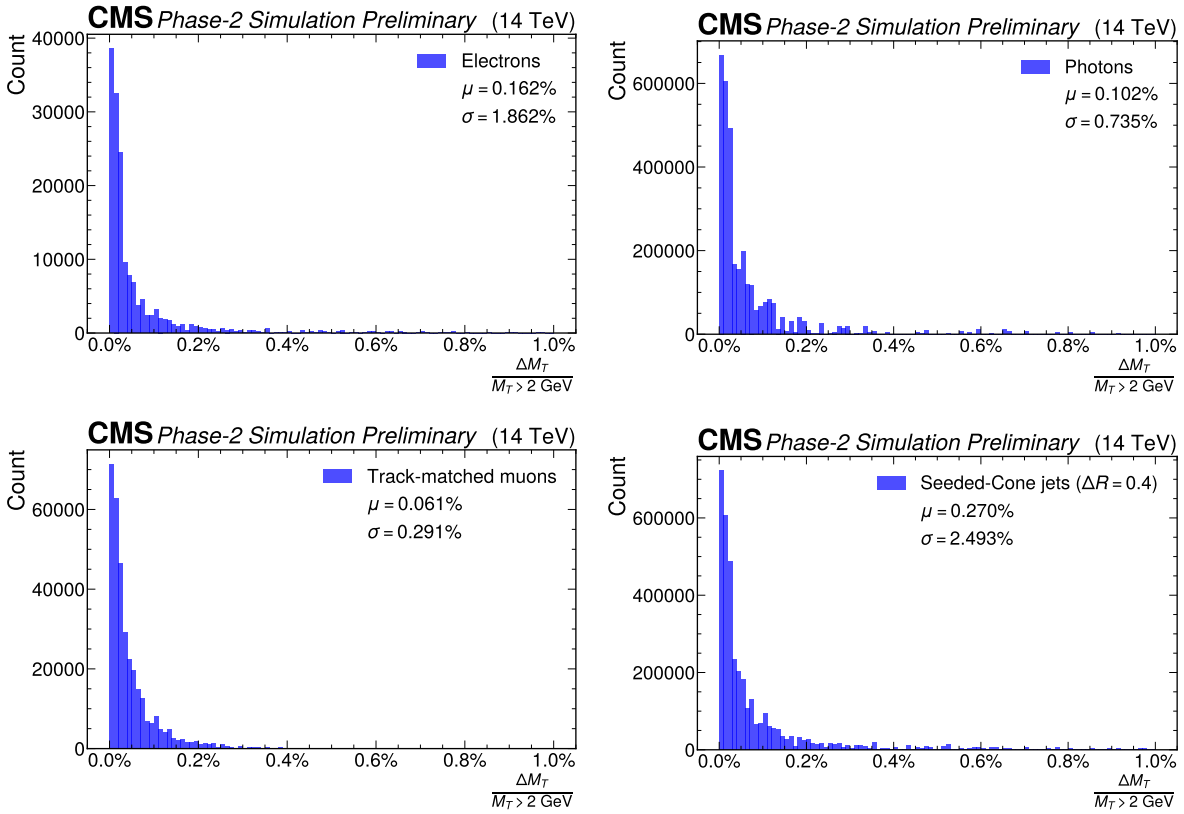


Figure 65: Histogram plots illustrating the distribution of relative transverse mass errors for various trigger objects, calculated for all possible transverse masses $M_T > 2$ GeV in a sample of 30,000 $t\bar{t}$ events.

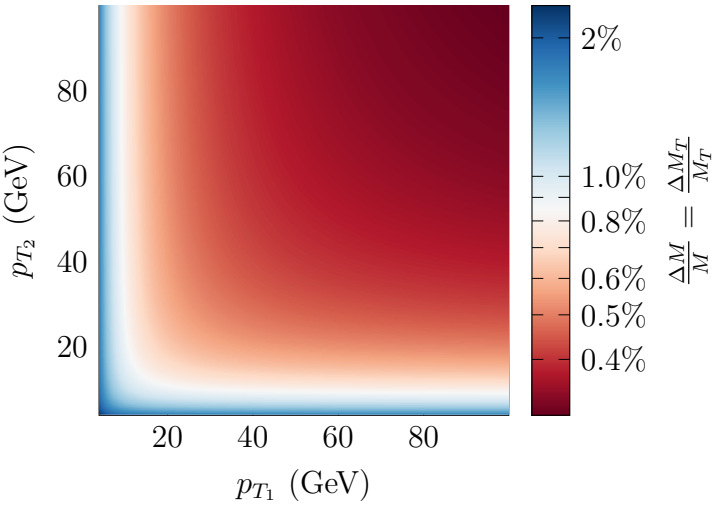


Figure 66: Estimation of the invariant mass and transverse mass resolution due to the intrinsic energy resolution of the ECAL barrel for two e/γ objects with transverse momenta (p_T) ranging from 4 GeV to 100 GeV.

Comparing Fig. 66 with Fig. 64 and 65, we conclude that even the highest-resolution calorimeter introduces uncertainties that are about twice as large as the computational errors. This underscores that the computational precision is more than sufficient for applying accurate cuts on invariant and transverse masses.

Note: The absence of the two peaks (at ∼0.04% and ∼0.08%) in the upper two plots of Fig. 64 arises from the η binning of Correlator electrons and photons. Unlike track-matched muons and Seeded-Cone jets, these objects do not use the two least significant bits (LSbs) to represent their η values in hardware and therefore avoid the characteristic truncation errors. The number of observed peaks corresponds to the distance from the next LUT address: two values lie at distance one from the closest address (producing the larger peak), while one value lies at distance two from the closest address (producing the smaller peak). A feature significant only in the cosh LUT, due to the exponentially increasing spacing between mapped values at larger addresses.

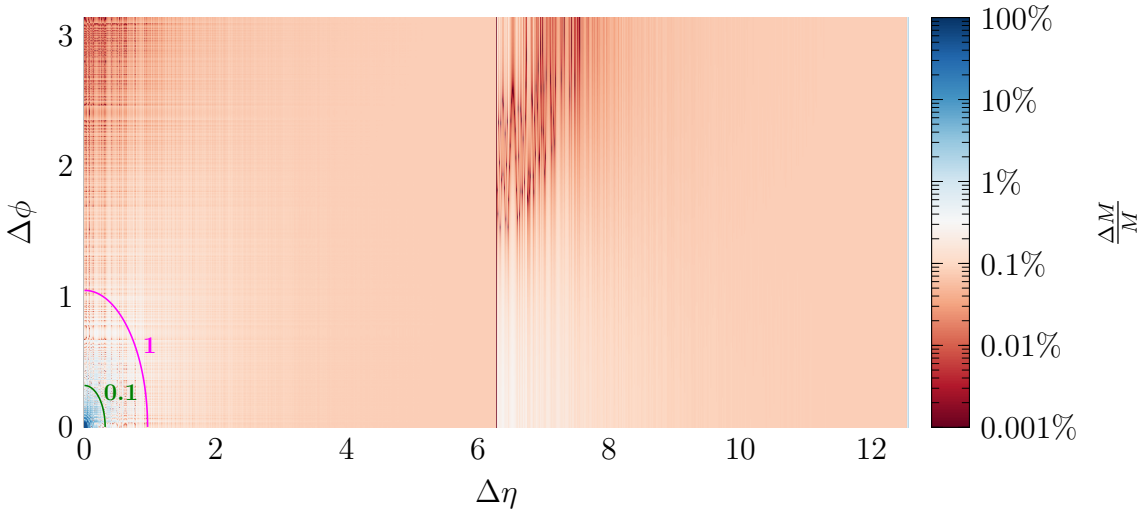


Figure 67: Plot of the relative invariant mass error as a function of $\Delta\eta$ and $\Delta\phi$. The colour scale represents the maximum error within each 4×4 batch in $\Delta\eta_{\mathrm{HW}} \times \Delta\phi_{\mathrm{HW}} = 4\pi/2^{12} \times 4\pi/2^{12}$, corresponding to a single LUT address, as the two least significant bits of $\Delta\eta_{\mathrm{HW}}$ and $\Delta\phi_{\mathrm{HW}}$ are omitted when used as address inputs. The magenta and green contour lines indicate the values of the denominator in Eq. 5.12.

For both masses, a more rigorous approach can be taken to assess the errors introduced by the LUTs by considering the propagated error,

$$\Delta M = \frac{1}{2}\sqrt{\frac{2p_{T_1}p_{T_2}}{\cosh(\Delta\eta)-\cos(\Delta\phi)}}\,\Delta\left[\cosh(\Delta\eta)-\cos(\Delta\phi)\right] \quad , \tag{5.11}$$

which leads to the relative error after dividing by M,

$$\frac{\Delta M}{M} = \frac{1}{2}\,\frac{\Delta\left[\cosh(\Delta\eta)-\cos(\Delta\phi)\right]}{\cosh(\Delta\eta)-\cos(\Delta\phi)} \quad . \tag{5.12}$$

Eq. 5.12 enables plotting of the relative invariant mass errors as a function of $\Delta\eta$ and $\Delta\phi$ (Fig. 67).

In Fig. 67, two previously discussed features of the LUT are evident. First, an increase in error is observed near $\Delta\eta \geq 2\pi$, primarily due to the omission of the cos LUT in the upper $\Delta\eta$ regime. The maximum error in this transition region, accounting for both the neglect of $\cos(\Delta\phi)$ and the omission of the two least significant bits in the LUT address, reaches 0.35%, a value slightly higher than that arising from the neglect of $\cos(\Delta\phi)$ alone, as given in Eq. 5.8.

A second notable feature is the thin band of higher errors just below the maximum $\Delta\eta$ ($\sim 4\pi$), which arises from enforcing the ratio between the two LUT scalings in different $\Delta\eta$-regimes to be an exact power of two. This constraint requires capping certain LUT values for large $\Delta\eta$ and consequently leads to larger relative errors.

The most striking feature in Fig. 67 is the large relative error as $\Delta\eta \to 0$ and $\Delta\phi \to 0$. This occurs because the denominator in Eq. 5.12 approaches zero, amplifying the unavoidable deviations caused by neglecting the two least significant bits of $\Delta\eta$ and $\Delta\phi$. This effect is further emphasised by the contour lines in Fig. 67, which serve to highlight the region where the denominator exhibits this behaviour. However, this effect is primarily limited to very small invariant masses and, when considering the full detector and reconstruction chain, is overshadowed by the two-object separation resolution. To mitigate this issue, the plots in Fig. 64 only consider masses above 2 GeV.

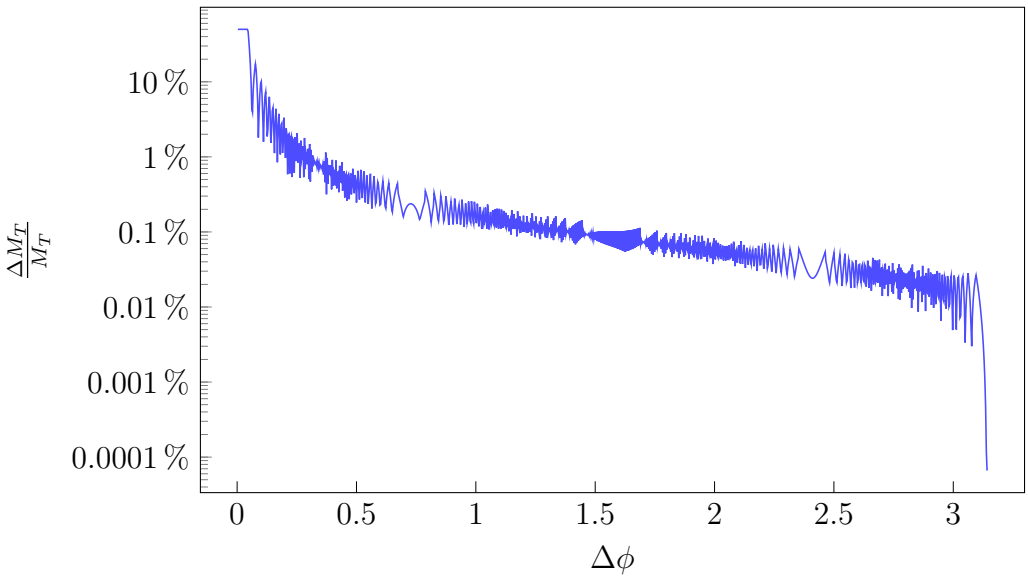


Figure 68: Plot of the relative invariant mass error as a function of $\Delta\phi$. The error is determined as the maximum over four consecutive $\Delta\phi_{\text{HW}}$ values corresponding to a single LUT address, arising from the omission of the two least significant bits.

The relative error of the transverse mass can be similarly evaluated using error propaga-

tion, leading to

$$\frac{\Delta M_T}{M_T} = \frac{1}{2} \left| \frac{\Delta \left[\cos(\Delta\phi)\right]}{1 - \cos(\Delta\phi)} \right| \quad . \tag{5.13}$$

Similar to the invariant mass error plot in Fig. 67, we observe a large relative error as $\Delta\phi \to 0$, stemming from the denominator of Eq. 5.13 approaching zero. To limit the error to at most 10%, a minimum separation of $\Delta\phi > 0.1$ is required, while $\Delta\phi > 0.2$ reduces the error to at most 2%. As with the invariant mass case, this error is, to some extent, unavoidable and occurs only for small transverse masses (M_T).

One notable remark regarding our estimates in Eq. 5.12 and 5.13 is that the use of error propagation via a first-order Taylor series expansion assumes small errors relative to the computed quantity. This is no longer fulfilled when the denominator of Eq. 5.12 and 5.13 approaches zero, requiring the inclusion of higher-order terms for a more accurate estimate. However, our primary interest in Fig. 67 and 68 lies in the well-behaved regions, while in the less well-behaved ones, we are mainly concerned with capturing the overall trend.

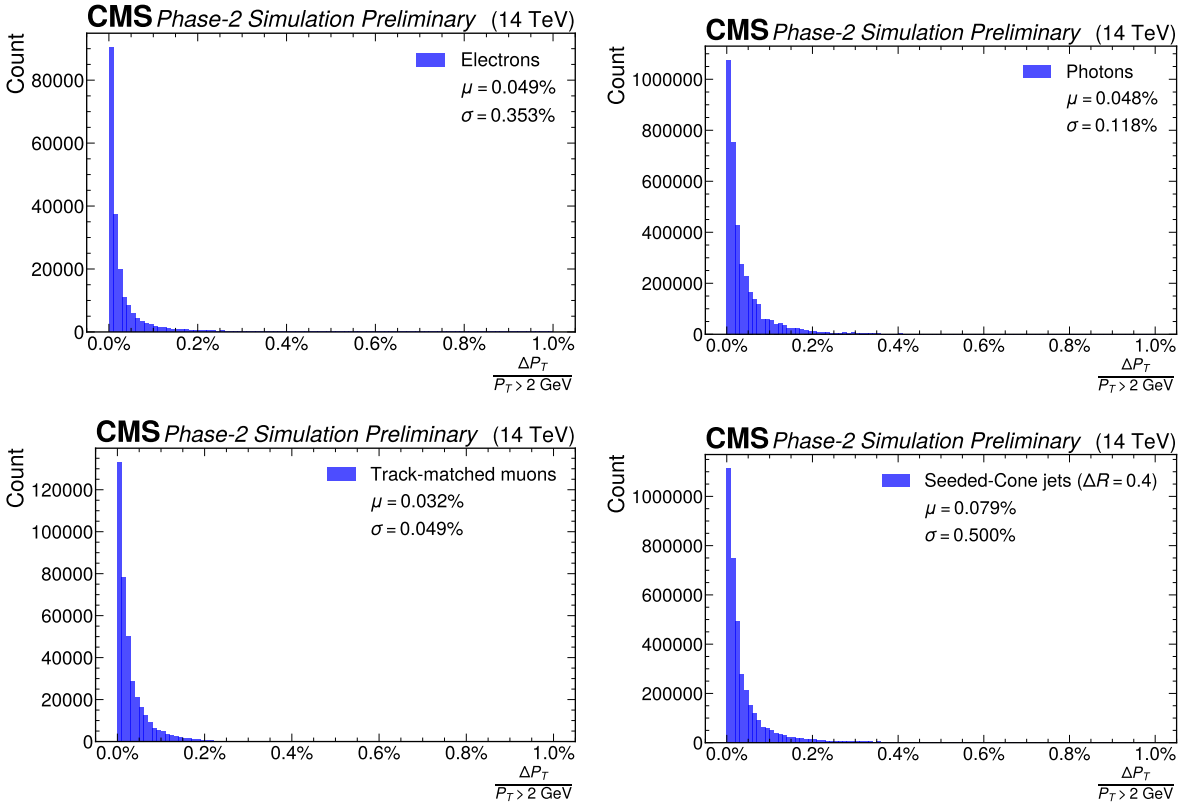


Figure 69: Histogram plots depicting the distribution of relative combined transverse momentum errors for various trigger objects, evaluated for all possible combined transverse momenta $P_T > 2$ GeV in a sample of 30,000 $t\bar{t}$ events.

From Fig. 69, we observe that the relative errors in the combined two-particle transverse momentum (P_T) are approximately half of those exhibited in Fig. 65. This can be

loosely explained by examining Eq. 3.55, where the terms $p_{T_1}^2$ and $p_{T_2}^2$, which are not affected by LUT errors, contribute to the total transverse momentum. As a result, the influence of the last term, $2p_{T_1}p_{T_2}\cos(\Delta\phi)$, which is affected by LUT errors, is effectively reduced by half.

Furthermore, we can again compare the relative errors introduced by the LUT to the intrinsic energy resolution of the barrel ECAL, the calorimeter with the highest energy resolution. By applying uncertainty propagation, we can use the intrinsic energy resolution Eq. 3.20 — which, under the assumption of no uncertainties in pseudorapidity η and in the ultrarelativistic limit, is equivalent to the relative transverse momentum resolution $\Delta p_T/p_T$ — to compute the uncertainties in the combined two-object transverse momentum P_T, yielding

$$\frac{\Delta P_T}{P_T} = \frac{\sqrt{\left(2p_{T_1}^2 + 2p_{T_1}p_{T_2}\cos(\Delta\phi)\right)^2\left(\frac{\Delta p_{T_1}}{p_{T_1}}\right)^2 + \left(2p_{T_2}^2 + 2p_{T_1}p_{T_2}\cos(\Delta\phi)\right)^2\left(\frac{\Delta p_{T_2}}{p_{T_2}}\right)^2}}{2\left(p_{T_1}^2 + p_{T_2}^2 + 2p_{T_1}p_{T_2}\cos(\Delta\phi)\right)} \,. \tag{5.14}$$

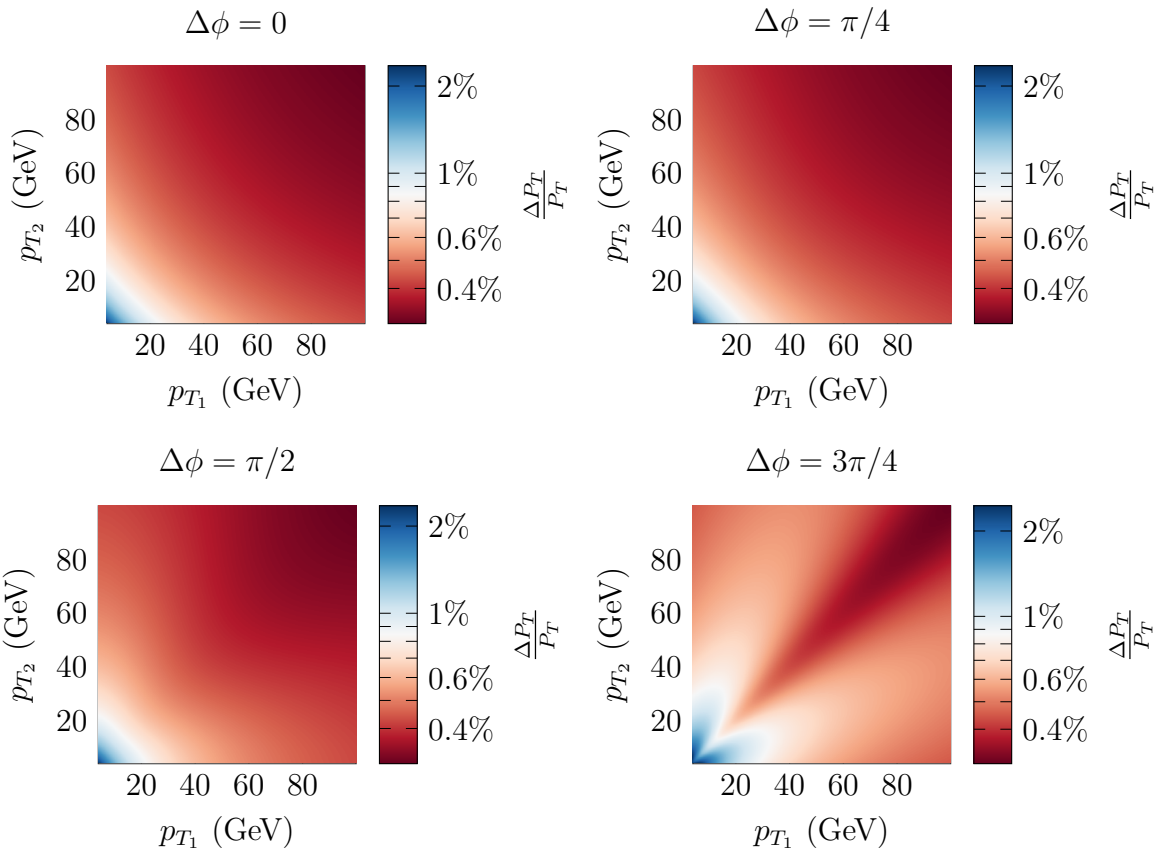


Figure 70: Estimation of the combined two-object P_T resolution as a function of the transverse momenta p_{T_1} and p_{T_2} of two e/γ objects for various azimuthal angle differences $\Delta\phi$, computed using Eq. 5.14. The transverse momenta of both objects range from 4 GeV to 100 GeV.

A comparison of Fig. 70 and Fig. 69 reveals that the computational errors are almost an order of magnitude smaller than the intrinsic uncertainties stemming from the energy resolution of the barrel ECAL. This highlights that the computational precision is more than adequate for implementing accurate cuts on the combined two-object transverse momentum P_T.

5.5 Heuristic cut optimisation

This chapter is largely based on previously published and presented work on the optimisation of cut-based algorithms [110].

5.5.1 A two-objective optimisation problem

The Level-1 Trigger menu for CMS is expected to grow significantly more complex during LHC's High-Luminosity phase, making the manual design of cut-based algorithms for specific physics signatures increasingly labour-intensive. To manage trigger rates, individual thresholds such as the transverse momentum (p_T) can be adjusted to tighten the selection criteria. This approach works reasonably well for simple event topologies. However, for more intricate physics signatures characterised by high object multiplicity and complex inter-object correlations, such a one-dimensional tuning strategy often fails to achieve optimal performance in terms of efficiency and trigger rate.

This challenge arises not only from the sheer number of trigger objects — with up to twelve per trigger collection — but also from the intricate, nonlinear dependencies among the variables that need to be constrained simultaneously. As a result, tuning such algorithms by hand becomes increasingly challenging with the conditions at the High-Luminosity LHC.

The task of designing an effective trigger algorithm constitutes a two-objective optimisation problem. On the one hand, the goal is to maximise the selection of events that contain the desired physics signature, referred to as trigger efficiency. On the other hand, this must be balanced against the need to keep the overall event rate within the bandwidth constraints of the CMS data acquisition system. If the selection criteria are too permissive, the resulting trigger rate could overwhelm the system; if they are too strict, valuable signal events may be lost.

This trade-off defines an optimisation landscape with no single global optimum. Instead, one obtains a Pareto front of solutions, ranging between the two trivial cases of accepting all events (maximal rate, maximal efficiency) and rejecting all events (zero rate, zero efficiency). Every point along this front represents a different optimal solution with an efficiency and rate.

To select a specific solution from this set, one can introduce a reference point that encodes a preference for a particular balance between the two objectives. Using the reference point, it becomes possible to define an achievement scalarizing function, turning the multi-objective problem into a scalar optimisation task. By shifting the reference point, different optimal solutions can be explored systematically [111, 112].

A commonly used form of an achievement scalarizing function is given by

$$s(\boldsymbol{f}(x)) = \max_{i=1,\dots,k} \left[\omega_i \left(f_i(x) - g_i\right)\right] + \rho \sum_{i=1}^{k} \omega_i \left(f_i(x) - g_i\right) \quad , \tag{5.15}$$

where ω_i are scaling weights, $f_i(x)$ denotes the individual components of the objective function, g_i represents the components of a user-defined reference point, and $\rho > 0$ is a small augmentation parameter.

The first term ensures weak Pareto optimality, meaning no alternative solution exists that improves all objectives simultaneously. The second term, often called the augmentation term, introduces a preference around the reference point. This encourages trade-offs in its vicinity and guarantees a properly Pareto optimal solution, allowing improvements in some objectives even if others are not simultaneously improved.

In this study, the two objective functions to be minimised are defined as

$$f_{\text{eff}} = 1 - \text{efficiency} = 1 - \frac{a_{\text{sig}}}{n_{\text{sig}}} \tag{5.16}$$

$$f_{\text{rate}} = \text{rate} = \frac{a}{n} \quad , \tag{5.17}$$

where a is the number of accepted events and n the total number of events. For optimisation, the scaling factors were set to $\omega_{\text{eff}} = 1$ and $\omega_{\text{rate}} = 1/g_{\text{rate}}$, with an augmentation parameter $\rho = 0.05$. Under these choices, Eq. 5.15 becomes

$$s(\boldsymbol{f}(x)) = \max\left[f_{\text{eff}}(x) - g_{\text{eff}},\ \frac{f_{\text{rate}}(x) - g_{\text{rate}}}{g_{\text{rate}}}\right] + 0.05\left(f_{\text{eff}}(x) + \frac{f_{\text{rate}}(x)}{g_{\text{rate}}}\right) + \text{const} \quad . \tag{5.18}$$

The specific values chosen for ω_{eff}, ω_{rate}, and ρ reflect a preference for minimising relative deviations from the reference point. In particular, changes in rate were weighted more heavily than changes in efficiency. Additionally, these parameters were selected to promote faster convergence of the optimisation procedure. It is worth noting that alternative optimisation strategies or different priorities in the efficiency-rate trade-off may require adjusting ω_{eff}, ω_{rate}, and ρ.

5.5.2 Optimisation technique

While various optimisation techniques exist, only gradient-free methods were applicable in our case, as the constraining inequalities (and equalities) of cut-based algorithms are not analytically differentiable. This limitation narrowed our options to heuristic approaches. We chose a variant of steepest-ascent hill climbing, which, due to our goal of minimising a loss function, effectively operated in reverse by following the path of steepest descent. This specific optimisation procedure used is known as “pattern search”, also referred to as “direct search” [113] and can be summarised as follows:

1. **Initialisation:** Cut parameters are initialised based on the mean and standard deviation of the signal distribution, assuming no correlations and single-object instances.
2. **Neighbourhood generation:** For each cut parameter, generate neighbouring configurations using a step size δ_i defined as a percentage of the variable's standard deviation.
3. **Performance estimation:** Evaluate each candidate configuration's trigger efficiency and rate using sufficiently large simulation samples.
4. **Loss calculation:** Compute the scalar loss for each configuration using the achievement scalarizing function defined in Eq. 5.18.
5. **Best neighbour selection:** For each cut, identify the neighbour with the greatest loss reduction and form a new candidate solution from all best neighbours.
6. **Adaptive step size update:** Adjust each δ_i based on the relative improvement observed.
7. **Iteration check:** If the scalar loss improves, repeat from step 2 using the new candidate.
8. **Refinement:** If no improvement is found and the minimum step sizes $\delta_{i,\min}$ remain above a threshold ϵ, halve the minimum step sizes and resume the search from step 2 to further refine the solution.

5.5.3 $B_s^0 \to \mu^+\mu^-$

The rare decay $B_s^0 \to \mu^+\mu^-$, mediated by a flavour-changing neutral current, serves as a highly sensitive probe for new physics beyond the Standard Model. In the Standard Model, this decay is forbidden at tree level and can only proceed through higher-order loop processes [114]. As a result, any deviation from the predicted branching ratio could indicate contributions from beyond Standard Model particles entering at tree level.

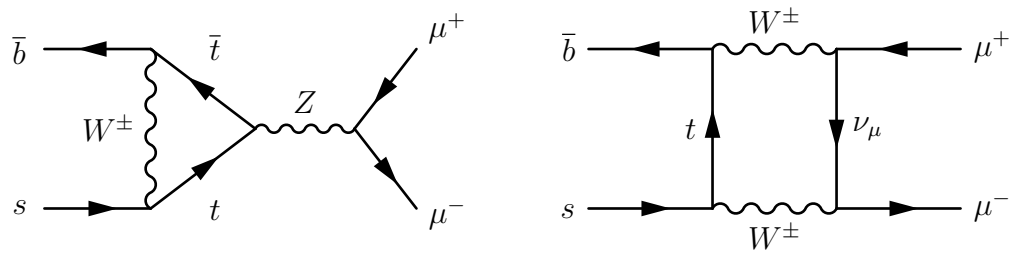


Figure 71: Leading-order Feynman diagrams of the $B_s^0 \to \mu^+\mu^-$ decay within the Standard Model.

To optimise an algorithm targeting this decay, three datasets were generated. The first consisted of a single proton-proton collision producing the $B_s^0 \to \mu^+\mu^-$ decay. The second added an average of 200 pileup collisions per event (denoted PU 200) to the same signal process. The third sample, referred to as "Minimum Bias" (see Section 3.2),

included only pileup events with no signal decay. All datasets were simulated using the CMSSW framework, employing Pythia 8 [115] as the event generator. During optimisation, the "Minimum Bias" sample was used to estimate trigger rates, while the signal-only sample (without pileup) was used to evaluate efficiency. The signal sample with PU 200 was reserved for a final validation of the obtained algorithm under real conditions.

As an initial hypothesis for a suitable trigger algorithm, we focused on events containing two muon candidates reconstructed via track matching in the GMT (see Section 4.3.4). Selection criteria were applied to each muon based on transverse momentum (p_T), pseudorapidity (η), and quality of reconstruction. Additional cuts were introduced on correlated muon properties, including angular separation ΔR (Section 3.7.1), invariant mass m (Section 3.7.2), charge correlations q, and the difference in longitudinal impact parameter Δz_0.

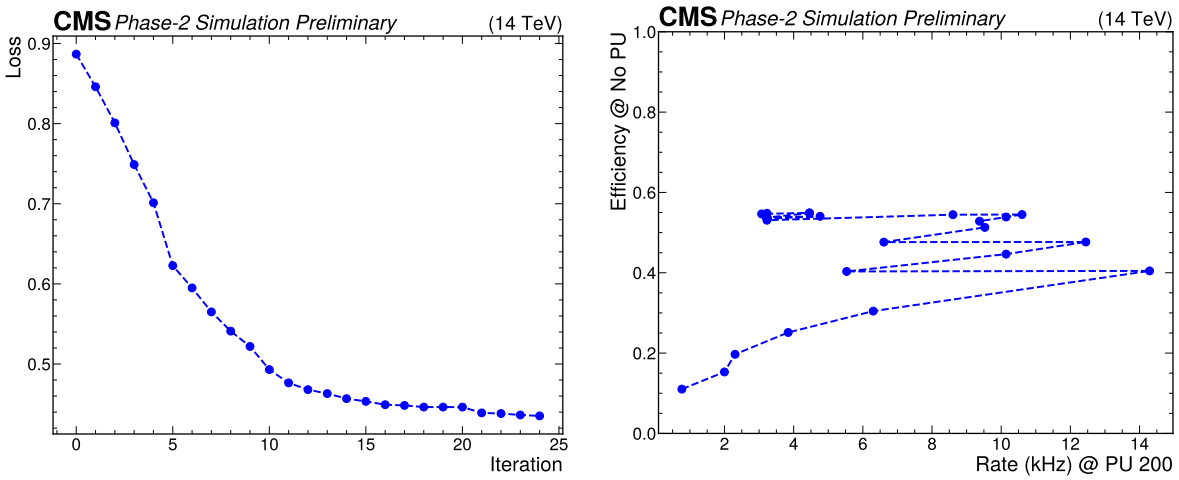


Figure 72: Left: Progression of the scalar loss throughout the optimisation process, displaying the best solution identified at each iteration. The dashed line serves as a visual aid. **Right:** Trajectory of the optimisation in the efficiency-rate space, beginning at iteration 0 (bottom left). The dashed line is included for visual guidance. [110]

The reference point was chosen as the unattainable point of 100% efficiency at a 15 kHz trigger rate. After 25 iterations, the optimised solution reached 54.8% efficiency with a corresponding trigger rate of 3.2 kHz. Validation using the signal sample overlaid with 200 pileup interactions yielded a slightly reduced efficiency of 51.1%, demonstrating robust performance under realistic high-luminosity conditions.

Table 1: Cut values obtained by optimising an algorithm targeting two track-matched muons, using a reference point defined by 100% efficiency and a 15 kHz trigger rate. Note: The pseudorapidity (η) cuts were pruned during the optimisation process, as they were not required to achieve an optimal solution. [110]

Muon 1	$p_T > 3.55$ GeV	$\not\eta >$ pruned $\not\eta <$ pruned	quality = loose
Muon 2	$p_T > 3.48$ GeV	$\not\eta >$ pruned $\not\eta <$ pruned	quality = very loose
Correlations	$m > 4.86$ GeV $m < 6.52$ GeV	$\Delta R > 0$ $\Delta R < 1.64$	$q_1 \neq q_2$ $\Delta z_0 < 1.32$ cm

One key advantage of this cut-based optimisation method over more complex deep neural network approaches is the interpretability and physical plausibility of the resulting cut values.

An analysis of the cut parameters listed in Tab. 1 confirms the physical plausibility of the optimised solution. The selected invariant mass window encompasses the known mass of the B_s^0 meson (5.37 GeV [116]). Furthermore, the requirement that the two muons have opposite charges ($q_1 \neq q_2$) aligns with the neutral nature of the B_s^0. The constraint on the longitudinal impact parameter separation Δz_0 reflects the expectation that both muons originate from a common decay vertex.

Integrating this optimised algorithm into a prototype Level-1 Trigger menu for the High-Luminosity LHC (see Section A.2) leads to an 8% absolute gain in efficiency for $B_s^0 \to \mu^+\mu^-$ decays, with a modest additional (pure) trigger rate of 1.5 kHz. To further enhance signal acceptance, we extended the optimisation strategy to include muon candidates not matched to silicon tracker tracks (see Sections 4.3.1 to 4.3.3). All six pairwise combinations of the three muon types — track-matched, standalone, and standalone displaced — were independently optimised using the same reference point of 100% efficiency and 15 kHz rate with identical cuts. The six optimised algorithms were then combined using a logical "OR" operation to form a unified algorithm.

This composite algorithm achieves a trigger efficiency of 74.1%, a 25.9% absolute improvement over the 48.2% efficiency of the baseline High-Luminosity menu (Section A.2), at the cost of an increased trigger rate of 58.7 kHz. Intermediate solutions can be obtained by adjusting the optimisation reference point.

Tab. 2 provides a comparison with other cut-based algorithms of the prototype menu (Section A.2). The optimised algorithm relying solely on track-matched muons outperforms all existing algorithms in both efficiency and rate for this decay channel. For use cases requiring maximal efficiency, the composite algorithm presents a compelling option, albeit with a correspondingly higher trigger rate.

Table 2: Comparison between the optimised combined and double track-matched muon algorithms and the highest-efficiency $B_s^0 \to \mu^+\mu^-$ algorithms from the High-Luminosity prototype trigger menu (Section A.2).

Algorithm	Efficiency	Rate (kHz)
Optimised (combined)	71.8%	70.4
Optimised (track-matched muons only)	51.1%	3.2
DoubleTkMuon_4_4_OS_Dr1p2	39.9%	20.3
DoubleTkMuon_OS_Er1p5_Dr1p4	25.2%	77.1
TripleTkMuon_5_3_0_DoubleTkMuon_5_3_OS_MassTo9	11.7%	16.7
TripleTkMuon_5_3p5_2p5_OS_Mass5to17	6%	10.4

5.5.4 VBF $\to$ H $\to \tau^+\tau^-$

The decay of the Higgs boson into a tau-lepton pair is of particular interest for multiple reasons. It offers a direct probe of the Higgs-fermion Yukawa coupling, enables tests of coupling universality across various production channels such as gluon fusion (ggF), vector boson fusion (VBF), and associated production (VH), and provides sensitivity to potential sources of CP violation and new physics beyond the Standard Model.

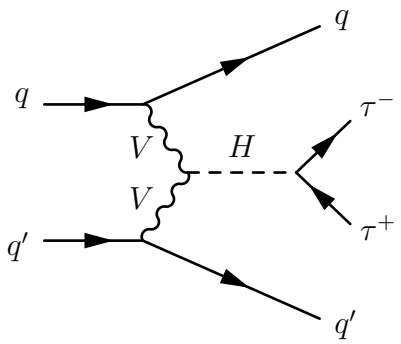


Figure 73: Feynman diagram of the VBF $\to$ H $\to \tau^+\tau^-$ process.

In this study, vector boson fusion (VBF) Higgs production was simulated using the POWHEG framework [117] and interfaced with Pythia 8 [115], handling the Higgs decay into tau lepton pairs and modelling pileup interactions. All simulations were conducted within CMSSW. Three samples were produced: a pure VBF $\to$ H $\to \tau^+\tau^-$ signal sample, the same signal overlaid with an average of 200 pileup collisions, and a "Minimum Bias" background sample, identical to that used in the $B_s^0 \to \mu^+\mu^-$ study.

The algorithm targets events containing two PUPPI tau candidates identified by a neural network (see Section 4.5.2). Selection criteria were applied to both individual candidate variables — such as transverse momentum (p_T), pseudorapidity (η), and neural network score — and to pairwise correlations, including angular separation (ΔR), invariant mass (M), and the combined transverse momentum of the pair. Efficiency during optimisation was estimated using the signal-only sample, while validation was once again carried out with the sample that includes the signal superimposed with 200 pileup collisions.

To map the Pareto front, the optimisation was performed using four distinct reference points. This illustrates a key strength of the method: the ability to steer solutions by adjusting the reference point. While the $B_s^0 \to \mu^+\mu^-$ optimisation could, in principle, demonstrate a similar feature, its solutions tend to saturate near the obtained optimum, making shifts from changing the reference point less apparent.

As shown in Fig. 74, both feasible (below the dashed line) and infeasible reference points (above the dashed line) were explored.

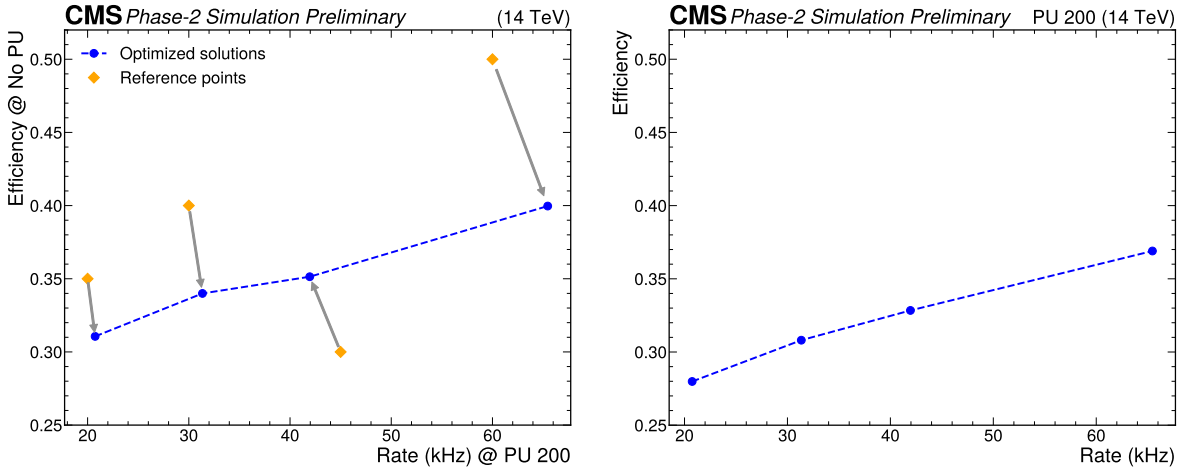


Figure 74: Left: Visualisation of the optimisation process showing the target reference point and the resulting solution in the efficiency-rate space, evaluated without pileup. The dashed line indicates an interpolation between points on the estimated Pareto front. **Right:** Performance of the optimised solutions validated using the signal sample with an average of 200 pileup interactions. The dashed line again shows an interpolation between points on an approximation of the Pareto front. [110]

5.6 Triggering on $\tau \to 3\mu$ using track-matched muons

5.6.1 Introduction

One particularly intriguing decay to investigate using the increased luminosity at the High-Luminosity LHC is the flavour-violating process $\tau \to 3\mu$. This decay is not strictly forbidden, as lepton flavour is not associated with any fundamental symmetry that would enforce conservation via Noether's theorem. While flavour-changing processes have been observed in the neutrino sector through the phenomenon of neutrino oscillations, no such transitions have been observed for charged leptons. Within the Standard Model, the decay $\tau \to 3\mu$ is therefore expected to only occur via neutrino-mediated loop diagrams, which are extremely suppressed by the Glashow-Iliopoulos-Maiani (GIM) mechanism [118]. This suppression originates from the unitarity of the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix U, which governs the mixing between neutrino flavour eigenstates and their corresponding mass eigenstates. When summing over the three neutrino mass eigenstates in the loop, unitarity enforces cancellations between contributions of different generations. As a result, only a tiny remainder proportional to

the neutrino mass splittings survives. The corresponding matrix element can be written schematically as

$$\mathcal{M}_{\tau\to 3\mu} \sim \sum_{i=1}^{3} U_{\mu i} U_{\tau i}^{*} \, F\left(\frac{m_{\nu_i}^2}{M_W^2}\right) \quad , \tag{5.19}$$

where F denotes the loop-function obtained from the one-loop Feynman diagrams contributing to the decay, m_{ν_i} are the neutrino masses of the three mass eigenstates, and M_W is the mass of the W boson.

Detailed calculations incorporating all one-loop diagrams within the Standard Model yielded a predicted branching ratio of the order of 10^{-54} [119] to 10^{-55} [120]. However, some sources have controversially suggested that the branching ratio could reach as high as 10^{-14} within the Standard Model [121].

Even the largest predicted branching ratio among those found in literature remains far beyond the experimental reach of the High-Luminosity LHC, with both CMS and ATLAS expected to set upper limits in the order of 10^{-9} at a 90% confidence level [122]. So, any observed excess in this channel would constitute a clear indication of flavour-changing processes beyond the Standard Model.

5.6.2 Comparing cut-optimised Pareto fronts

As described in Section 5.5, our objective is to retain as many events containing the $\tau \to 3\mu$ decay as possible, while simultaneously minimising the overall trigger rate, which is estimated using "Minimum Bias" events (see Section 3.2). To achieve this, we apply cuts on the three-body invariant mass of track-matched muon triplets. If the three muons originate from the same τ lepton, their combined invariant mass is expected to be close to the τ mass of 1.78 GeV, and the longitudinal impact parameter distance (Δz_0) between all muon pairs should be small. Additional constraints are imposed on the two-body invariant mass combinations and angular separations (ΔR) between muons for completeness. The exact bounds for these variables are determined through the optimisation procedure described in Section 5.5.

For comparison, a second optimisation run was performed using only two-object correlations. This included cuts on the invariant masses and angular separations (ΔR) between all muon pairs within the three-muon system, along with cuts on the aforementioned longitudinal impact parameter distance (Δz_0).

Three different τ lepton production channels were investigated: two heavy-flavour decay modes, in which a $B^{\pm}$ or $D_s^{\pm}$ meson decays into a $\tau^{\pm}$ lepton, and a third channel where a $W^{\pm}$ boson decays into a $\tau^{\pm}$. The production of $B^{\pm}$, $D_s^{\pm}$, and $W^{\pm}$ particles, along with their subsequent decay chains, was simulated using Pythia 8.3 [115] within the CMSSW framework.

For simplicity and because the presence of pileup has only a minor impact on signal efficiency, as discussed in Section 5.5, all signal samples were generated without super-

imposed pileup collisions. In contrast, the "Minimum Bias" sample — used to estimate the trigger rate — includes an average of 200 superimposed pileup collisions.

The $W^{\pm} \to \tau^{\pm}$ sample comprised 6,003 events, pre-filtered in two steps: first, from 20,449 to 8,771 events by selecting only events containing three generated muons visible in the outer silicon tracker (Section 3.4); and second, by requiring at least three reconstructed track-matched muon candidates. The 2,768 events that failed to meet the second criterion likely do so due to the high mass of the W boson (80.37 GeV [7]). Its decay into a τ lepton and an anti-neutrino $\overline{\nu_\tau}$ results in both particles receiving large, oppositely directed momenta. The subsequent decay of the highly boosted τ can produce muons that are strongly collimated. In cases of extreme collimation, two same-sign muons may become indistinguishable, preventing accurate reconstruction and thereby failing the selection criteria.

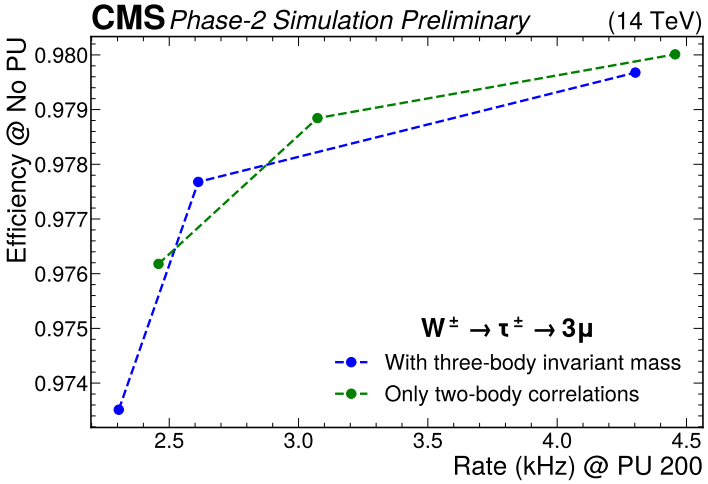


Figure 75: Comparison of the coarsely estimated Pareto fronts obtained by optimising with three different reference points, once including the three-body invariant mass, and once using only single quantity cuts and two-body correlations. The dashed line serves as a visual guide, interpolating between the solution points.

From Fig. 75, we observe that the solutions in both cases nearly overlap, with only minor deviations likely due to convergence to different minima. Comparing the upper-right solutions from both curves, as detailed in Tab. 3 and 4, we find that the resulting cut values are also quite similar. In both cases, the tight ΔR constraints — selecting highly boosted parent particles — combined with a looser requirement on the two-body invariant mass (which lies further from the true τ lepton mass), appear sufficient for accurately selecting this decay channel. This configuration achieves a signal efficiency of 98% with a total rate of only 4.5 kHz. The inclusion of the three-body invariant mass introduces only a very loose constraint ($m_{123} < 4.56$ GeV), accompanied by a relaxation of one of the two-body invariant mass cuts (m_{12}).

Table 3: Cut values obtained by optimising an algorithm targeting the $W^{\pm} \to \tau^{\pm} \to 3\mu$ decay, without using the three-body invariant mass as a cut variable. The chosen reference point corresponds to 100% signal efficiency and a rate of 40 kHz, yielding an optimised solution with 98% signal efficiency at a rate of 4.5 kHz. Note: Some cuts were pruned during the optimisation process, as they were not required for an optimal solution.

Muon 1	$p_T > 4.39$ GeV		
Muon 2	~~$p_T >$~~ pruned		
Muon 3	~~$p_T >$~~ pruned		
Two-body Correlations	~~$\Delta R_{12} >$~~ pruned	$\Delta R_{13} > 0.01$	$\Delta R_{23} > 0.02$
	$\Delta R_{12} < 0.28$	$\Delta R_{13} < 0.36$	$\Delta R_{23} < 0.58$
	~~$\Delta z_{0_{12}} <$~~ pruned	$\Delta z_{0_{13}} < 1.54$ cm	$\Delta z_{0_{23}} < 1.78$ cm
	~~$m_{12} >$~~ pruned	~~$m_{13} >$~~ pruned	~~$m_{23} >$~~ pruned
	$m_{12} < 1.62$ GeV	$m_{13} < 3.15$ GeV	$m_{23} < 3.15$ GeV

Table 4: Cut values were derived by optimising an algorithm for the $W^{\pm} \to \tau^{\pm} \to 3\mu$ decay, incorporating the three-body invariant mass as one of the cut variables. Using a reference point of 100% signal efficiency at a rate of 40 kHz, the optimisation yielded 98% signal efficiency and a rate of 4.3 kHz. Note: Some cuts were again pruned during the optimisation process, as they were not required to obtain an optimal solution.

Muon 1	$p_T > 2.06$ GeV		
Muon 2	~~$p_T >$~~ pruned		
Muon 3	$p_T > 5.08$ GeV		
Two-body Correlations	$\Delta R_{12} > 0$	$\Delta R_{13} > 0.01$	~~$\Delta R_{23} >$~~ pruned
	$\Delta R_{12} < 0.48$	$\Delta R_{13} < 0.33$	$\Delta R_{23} < 0.38$
	~~$\Delta z_{0_{12}} <$~~ pruned	$\Delta z_{0_{13}} < 1.98$ cm	$\Delta z_{0_{23}} < 1.34$ cm
	~~$m_{12} >$~~ pruned	~~$m_{13} >$~~ pruned	~~$m_{23} >$~~ pruned
	$m_{12} < 3.15$ GeV	$m_{13} < 3.15$ GeV	$m_{23} < 3.15$ GeV
Three-body Correlations	$m_{123} < 4.56$ GeV		

The $B^{\pm} \to \tau^{\pm} \to 3\mu$ and $D_s^{\pm} \to \tau^{\pm} \to 3\mu$ samples were analogously pre-filtered to retain only events in which three track-matched muons could be reconstructed. Ultimately, these samples contained relatively few events, 800 for the $B^{\pm}$ decay and 601 for the $D_s^{\pm}$ decay. In both cases, the pre-filtering step resulted in a large number of rejected events: 8.16 million for the $B^{\pm}$ sample and 17.23 million for the $D_s^{\pm}$ sample.

This high rejection rate is primarily due to the low masses of the $B^{\pm}$ (5.28 GeV [116]) and $D_s^{\pm}$ (1.97 GeV [116]) mesons, which result in low transverse momentum (p_T) muons in the final state. These low-p_T muons often lose their energy before reaching the muon stations (see Fig. 17). Only when the parent mesons possess sufficiently high p_T do their decay products appear as reconstructed, track-matched muons.

This limitation highlights the potential for future approaches that rely solely on tracking

information, rather than requiring muon stubs (cf. Section 4.3.4). Such strategies could significantly enhance the overall efficiency for triggering on heavy-flavour decay modes.

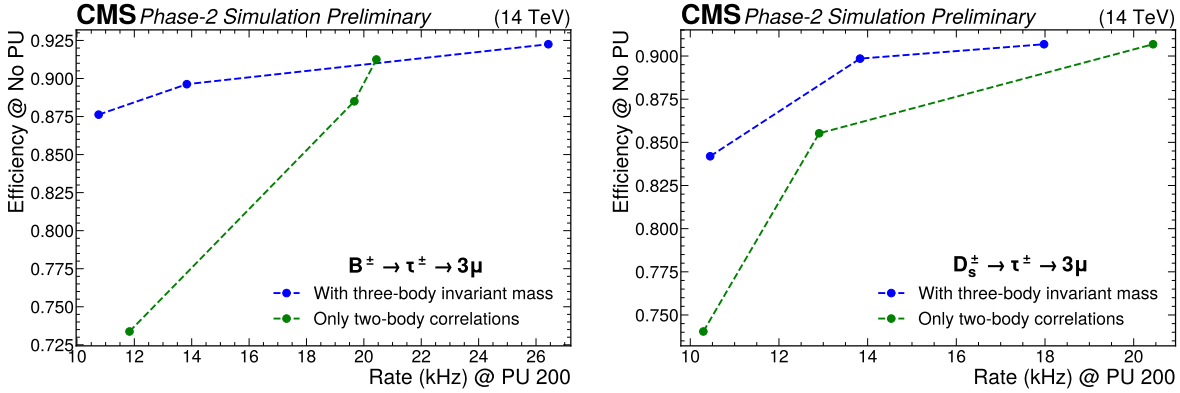


Figure 76: Comparison of approximate Pareto fronts resulting from optimisations performed with three different reference points, once incorporating the three-body invariant mass, and once using only individual variable thresholds and two-body correlations. The dashed line provides a visual interpolation between the solution points. **Left:** Results for the $B^{\pm} \rightarrow \tau^{\pm} \rightarrow 3\mu$ decay. **Right:** Results for the $D_s^{\pm} \rightarrow \tau^{\pm} \rightarrow 3\mu$ decay.

Given the baseline requirement of three reconstructed track-matched muons, a slight improvement can be observed in Fig. 76 when triggering on heavy-flavour decay modes with the inclusion of the three-body invariant mass, particularly at lower trigger rates. Comparing the highest-efficiency cut values for the $D_s^{\pm} \rightarrow \tau^{\pm} \rightarrow 3\mu$ channel in Tab. 5 and 6, we note that the ΔR requirements are less stringent than in the $W^{\pm} \rightarrow \tau^{\pm} \rightarrow 3\mu$ case, which is expected due to the significantly lower mass of the initial-state particle. While the three-body invariant mass cut values in Tab. 6 remain relatively loose, especially considering the true τ lepton mass of 1.78 GeV, their inclusion appears to noticeably reduce the trigger rate.

Table 5: Cut values were derived by optimising an algorithm targeting the $D_s^{\pm} \rightarrow \tau^{\pm} \rightarrow 3\mu$ decay, excluding the three-body invariant mass from the set of cut variables. Using a reference point corresponding to 100% signal efficiency at a rate of 40 kHz, the optimisation yielded a solution with 90.7% signal efficiency and a reduced rate of 20.4 kHz. Note: Pruning was again applied to remove cuts not essential for the optimal solution, which particularly affected the transverse momentum (p_T) cuts due to the low mass of the $D_s^{\pm}$ meson.

Muon 1	~~$p_T >$~~ pruned		
Muon 2	~~$p_T >$~~ pruned		
Muon 3	~~$p_T >$~~ pruned		
Two-body Correlations	$\Delta R_{12} > 0.01$	$\Delta R_{13} > 0.02$	~~$\Delta R_{23} >$~~ pruned
	$\Delta R_{12} < 1.08$	$\Delta R_{13} < 1.08$	$\Delta R_{23} < 0.56$
	$\Delta z_{0_{12}} < 1.46$ cm	$\Delta z_{0_{13}} < 1.93$ cm	$\Delta z_{0_{23}} < 0.92$ cm
	~~$m_{12} >$~~ pruned	~~$m_{13} >$~~ pruned	$m_{23} > 0.72$ GeV
	$m_{12} < 1.62$ GeV	$m_{13} < 1.44$ GeV	$m_{23} < 5.1$ GeV

Table 6: Cut values were determined by optimising an algorithm for the $D_s^{\pm} \to \tau^{\pm} \to 3\mu$ decay, with the three-body invariant mass included among the cut variables. Starting from a reference point of 100% signal efficiency at a rate of 40 kHz, the optimisation resulted in 90.7% signal efficiency and a rate of 18 kHz. Note: Pruning again removed non-essential cuts, most notably those on transverse momentum (p_T), reflecting the low mass of the $D_s^{\pm}$ meson.

Muon 1	~~$p_T >$~~ pruned		
Muon 2	~~$p_T >$~~ pruned		
Muon 3	~~$p_T >$~~ pruned		
Two-body Correlations	$\Delta R_{12} > 0.01$	~~$\Delta R_{13} >$~~	$\Delta R_{23} > 0$
	$\Delta R_{12} < 0.38$	$\Delta R_{13} < 0.67$	$\Delta R_{23} < 0.56$
	$\Delta z_{0_{12}} < 1.49$ cm	$\Delta z_{0_{13}} < 2.12$ cm	$\Delta z_{0_{23}} < 0.89$ cm
	~~$m_{12} >$~~ pruned	~~$m_{13} >$~~ pruned	~~$m_{23} >$~~ pruned
	$m_{12} < 2.25$ GeV	$m_{13} < 1.59$ GeV	$m_{23} < 5.1$ GeV
Three-body Correlations	$m_{123} > 1.25$ GeV	$m_{123} < 8.83$ GeV	

6 Conclusions and outlook

The Global Trigger (GT) for High-Luminosity operation is already in a highly advanced state, incorporating a wide range of cuts on both single-object quantities and object correlations. These cuts have been implemented in firmware as configurable modules, and within the CMSSW emulator. The instantiations of the condition firmware modules were extensively tested with a wide variety of cuts, within a menu of greater complexity than the current prototype menu for High-Luminosity operation (see Section A.2), and with substantially more algorithms (up to 336) than were originally anticipated for a single board. Excellent timing closure was achieved during testing, giving strong confidence in the overall readiness of the system to cope with potentially rapidly evolving run conditions at the High-Luminosity LHC.

The correctness of the implementation was first validated for the CMSSW emulator by comparing it against a standalone floating-point tool developed by the Detector Performance Group. Discrepancies between the two emulators were negligible and could be fully attributed to expected effects, such as bitwidth constraints and the use of LUTs for the computation of certain correlations (M, M_T, $M/\Delta R$, and two-particle P_T).

Additionally, the errors introduced by the use of the final-width and final-depth LUTs were extensively studied, and found to be negligible compared to the intrinsic energy resolution of the detector.

Following the emulator validation, the firmware implementation was cross-checked against the emulator. No discrepancies were observed across a large sample of events, confirming the correctness of the firmware.

A new method was proposed to automate the selection of cut values when targeting specific physics signatures. This method simplifies the management of an extensive menu, expected to include up to 1,000 algorithms for High-Luminosity operation while enabling rapid adaptation to changing run conditions. The problem constitutes a multi-objective optimisation problem, requiring to strike a balance between minimising the trigger rate and maximising the signal efficiency. An achievement scalarizing function was introduced to convert the two competing objectives into a scalar loss function, followed by an optimisation procedure minimising this function. Exemplary results were presented for the $B_s^0 \to \mu\mu$ and VBF $\to H \to \tau\tau$ decay channels, showcasing the method's fast convergence and its strong adaptation to the targeted signal characteristics. It was further highlighted that the obtained solutions can be checked for plausibility and that by systematically varying the reference point, different optimal solutions can be explored.

The implementation of the three-body invariant mass correlation, an especially demanding quantity due to the large number of combinatorial possibilities, was also highlighted. Its application to triggering on the never observed, flavour-violating $\tau \to 3\mu$ decay was discussed as an example of the system's capabilities.

Looking ahead, the set of inputs is expected to grow, particularly with the introduction of new and updated trigger objects leveraging machine learning classifiers. These

developments are planned to bring a range of new tagging scores requiring novel ways to correlate them. In the longer term, similar evolutions are anticipated. Importantly, the current GT structure is flexible enough that no fundamental changes are expected to be necessary; rather, all foreseeable additions can be embedded naturally within the existing CMSSW emulator and firmware framework.

Beyond the planned future additions to the GT firmware and emulator, extensive integration tests with other subsystems are scheduled, building upon the limited tests already performed. These tests aim to incrementally validate each trigger path within the Level-1 Trigger chain, thereby ensuring the overall readiness of the CMS detector for High-Luminosity operation at the LHC.

A particularly notable integration test is planned during the ongoing Run 3 of the LHC. In this test, reconstructed muons from an already-installed sector of the upgraded barrel muon system for High-Luminosity operation will be received and processed by a smaller variant of the new GT. This variant features a limited set of 24 algorithms along with Final-OR functionality, all implemented on a Serenity board equipped with a KU15P FPGA. This test will mark the first time the new system processes real CMS data, approximately five years ahead of the scheduled commencement of High-Luminosity operation at the LHC.

7 Acknowledgement

I would like to express my sincere gratitude to Hannes Sakulin for his continued guidance and coordination, both within our group and with other groups at CERN, throughout my PhD. I am especially thankful for his thorough review of all my documents and presentations.

I would also like to thank my university supervisor Manfred Jeitler for his invaluable insights into the experimental aspects of the CMS detector, as well as for his encouragement and guidance in exploring new ideas for triggering.

My sincere thanks go to my colleagues in the Global Trigger group — Gabriele Bortolato, Elias Leutgeb, and Dinyar Rabady — for the fruitful collaboration, constructive feedback, and excellent ideas that have significantly shaped and improved my work.

I would further like to acknowledge the Level-1 scouting team — Thomas James, Rocco Ardino, and Mateo Migliorini — for their valuable insights into various components of the CMS Trigger, shared through many helpful discussions. My thanks also go to Phillip Brummer for his input regarding the operational aspects of the CMS data acquisition.

Moreover, I am deeply grateful to my parents for their support in helping me pursue this final step in my university education. I would especially like to thank my father and his partner for their logistical assistance during my moves to and from the CERN area. Finally, I extend my heartfelt thanks to Leni for being a reliable source of support and encouragement throughout my time at CERN, even during long periods of absence.

My stay at CERN was primarily funded by the CERN Austrian Doctoral Student Program, with additional support from CERN's CMS DAQ group, for which I am thankful.

Bibliography

[1] G. Bertone and D. Hooper, *History of dark matter*, *Rev. Mod. Phys.* **90** (2018) 045002.

[2] N. Jarosik, C. L. Bennett, J. Dunkley, B. Gold, M. R. Greason, M. Halpern et al., *SEVEN-YEAR WILKINSON MICROWAVE ANISOTROPY PROBE (WMAP) OBSERVATIONS: SKY MAPS, SYSTEMATIC ERRORS, AND BASIC RESULTS*, *The Astrophysical Journal Supplement Series* **192** (2011) 14.

[3] B. D. Fields, K. A. Olive, T.-H. Yeh and C. Young, *Big-Bang Nucleosynthesis after Planck*, *Journal of Cosmology and Astroparticle Physics* **2020** (2020) 010.

[4] A. D. Sakharov, *Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe*, *Pisma Zh. Eksp. Teor. Fiz.* **5** (1967) 32.

[5] A. D. Sakharov, *Violation of CP invariance, C asymmetry, and baryon asymmetry of the universe*, *Soviet Physics Uspekhi* **34** (1991) 392.

[6] N. Arkani-Hamed, S. Dimopoulos and G. Dvali, *The hierarchy problem and new dimensions at a millimeter*, *Physics Letters B* **429** (1998) 263.

[7] P. D. Group, R. L. Workman, V. D. Burkert, V. Crede, E. Klempt, U. Thoma et al., *Review of Particle Physics*, *Progress of Theoretical and Experimental Physics* **2022** (2022) 083C01.

[8] G. Degrassi, S. Di Vita, J. Elias-Miró, J. R. Espinosa, G. F. Giudice, G. Isidori et al., *Higgs mass and vacuum stability in the Standard Model at NNLO*, *Journal of High Energy Physics* **2012** (2012) .

[9] S. Rugh and H. Zinkernagel, *The quantum vacuum and the cosmological constant problem*, *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* **33** (2002) 663.

[10] R. J. Adler, B. Casey and O. C. Jacob, *Vacuum catastrophe: An elementary exposition of the cosmological constant problem*, *American Journal of Physics* **63** (1995) 620.

[11] K. Uzawa, Y. Morisawa and S. Mukohyama, *Excitation of the Kaluza-Klein gravitational mode*, *Phys. Rev. D* **62** (2000) 064011.

[12] O. S. Brüning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole et al., *LHC Design Report*, CERN Yellow Reports: Monographs. CERN, Geneva, 2004, 10.5170/CERN-2004-003-V-1.

[13] Particle Data Group collaboration, J. Beringer, J. F. Arguin, R. M. Barnett, K. Copic, O. Dahl, D. E. Groom et al., *Review of Particle Physics*, *Phys. Rev. D* **86** (2012) 010001.

[14] J. Erler, *Mass of the Higgs boson in the standard electroweak model*, *Phys. Rev. D* **81** (2010) 051301.

[15] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A. Abdelalim et al., *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Physics Letters B* **716** (2012) 1.

[16] S. Chatrchyan, V. Khachatryan, A. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo et al., *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Physics Letters B* **716** (2012) 30.

[17] E. Lopienska, "The CERN accelerator complex." https://cds.cern.ch/record/2800984, 2022.

[18] CERN, "The accelerator complex." `https://home.cern/science/accelerator-complex`.

[19] D. Jacquet, *Injection*, .

[20] A. Schaeffer, "The waltz of the LHC magnets has begun." https://home.cern/news/news/accelerators/waltz-lhc-magnets-has-begun, 2019.

[21] L. Evans and P. Bryant, *LHC Machine*, *Journal of Instrumentation* **3** (2008) S08001.

[22] D. Boussard, E. Chiaveri, E. Häbel, H. P. Kindermann, R. Losito, S. Marque et al., *The LHC superconducting cavities*, .

[23] W. Herr and B. Muratori, *Concept of luminosity*, .

[24] R. Calaga, *Crab Cavities for the High-luminosity LHC*, *18th International Conference on RF Superconductivity* (2018) THXA03.

[25] B. Muratori and T. Pieloni, *Luminosity levelling techniques for the LHC*, in *ICFA Mini-Workshop on Beam-Beam Effects in Hadron Colliders*, pp. 177–181, 2014, `1410.5646`, DOI.

[26] S. Fartoukh, S. Kostoglou, M. Solfaroli Camillocci, G. Arduini, H. Bartosik, C. Bracco et al., *LHC Configuration and Operational Scenario for Run 3*, tech. rep., CERN, Geneva, 2021.

[27] O. Aberle, I. Béjar Alonso, O. Brüning, P. Fessia, L. Rossi, L. Tavian et al., *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*, CERN Yellow Reports: Monographs. CERN, Geneva, 2020, 10.23731/CYRM-2020-0010.

[28] CMS collaboration, "CMS Luminosity - Public Results." https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults, 2024.

[29] T. Sakuma and T. McCauley, *Detector and Event Visualization with SketchUp at the CMS Experiment*, *Journal of Physics: Conference Series* **513** (2014) 022032.

[30] CMS collaboration, S. Chatrchyan, G. Hmayakyan, V. Khachatryan, A. M. Sirunyan, W. Adam, T. Bauer et al., *The CMS experiment at the CERN LHC*, *Journal of Instrumentation* **3** (2008) S08004.

[31] ATLAS collaboration, *Expected pileup values at the HL-LHC*, tech. rep., CERN, Geneva, 2013.

[32] D. Contardo, M. Klute, J. Mans, L. Silvestris and J. Butler, *Technical Proposal for the Phase-II Upgrade of the CMS Detector*, tech. rep., Geneva, 2015. 10.17181/CERN.VU8I.D59J.

[33] I. Neutelings. https://tikz.net/author/izaak.

[34] A. Rossi, *The CMS Tracker for the High Luminosity LHC*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1048** (2023) 167950.

[35] CMS collaboration, *The Phase-2 Upgrade of the CMS Tracker*, tech. rep., CERN, Geneva, 2017. 10.17181/CERN.QZ28.FLHW.

[36] H. Kolanoski and N. Wermes, *Teilchendetektoren*. Springer Berlin Heidelberg, 2016.

[37] E. Longo and I. Sestili, *Monte Carlo calculation of photon-initiated electromagnetic showers in lead glass*, *Nuclear Instruments and Methods* **128** (1975) 283.

[38] S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce et al., *Geant4—a simulation toolkit*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** (2003) 250.

[39] J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce Dubois, M. Asai et al., *Geant4 developments and applications*, *IEEE Transactions on Nuclear Science* **53** (2006) 270.

[40] J. Allison, K. Amako, J. Apostolakis, P. Arce, M. Asai, T. Aso et al., *Recent developments in geant4*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **835** (2016) 186.

[41] P. A. al., *Energy resolution of the barrel of the CMS Electromagnetic Calorimeter*, *Journal of Instrumentation* **2** (2007) P04004.

[42] CMS collaboration, A. Hayrapetyan, A. Tumasyan, W. Adam, J. W. Andrejkovic, B. Arnold, H. Bergauer et al., *Development of the CMS detector for the CERN LHC Run 3*, *JINST* **19** (2024) P05064 [2309.05466].

[43] A. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, T. Bergauer, J. Brandstetter et al., *Calibration of the CMS hadron calorimeters using proton-proton collision*

data at $\sqrt{s}$ *= 13 TeV*, *Journal of Instrumentation* **15** (2020) P05002.

[44] S. Abdullin, V. Abramov, B. Acharya, N. Adam, M. Adams, P. Adzic et al., *The CMS barrel calorimeter response to particle beams from 2 to 350 GeV/c*, *The European Physical Journal C* **60** (2009) 359.

[45] A. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, T. Bergauer, J. Brandstetter et al., *Calibration of the cms hadron calorimeters using proton-proton collision data at* $\sqrt{s}$ *= 13 tev*, *Journal of Instrumentation* **15** (2020) P05002.

[46] S. Abdullin, V. Abramov, B. Acharya, M. Adams, N. Akchurin, U. Akgun et al., *Design, performance, and calibration of cms forward calorimeter wedges*, *The European Physical Journal C* **53** (2007) 139–166.

[47] CMS collaboration, *The Phase-2 Upgrade of the CMS Endcap Calorimeter*, tech. rep., CERN, Geneva, 2017. 10.17181/CERN.IV8M.1JY2.

[48] N. Akchurin, *Detailed results of the response of a CMS HGCAL silicon-pad electromagnetic calorimeter prototype to 20-300 GeV positrons.*, tech. rep., CERN, Geneva, 2021.

[49] B. Acar, G. Adamov, C. Adloff, S. Afanasiev, N. Akchurin, B. Akgün et al., *Performance of the CMS High Granularity Calorimeter prototype to charged pion beams of 20-300 GeV/c*, *Journal of Instrumentation* **18** (2023) P08014.

[50] CMS collaboration, "The Phase-2 Upgrade of the CMS Level-1 Trigger." CERN-LHCC-2020-004, CMS-TDR-021, 2020.

[51] CMS collaboration, *The Phase-2 Upgrade of the CMS Muon Detectors*, tech. rep., CERN, Geneva, 2017.

[52] CMS collaboration, J. G. Layter, *The CMS muon project: Technical Design Report*, Technical design report. CMS. CERN, Geneva, 1997.

[53] CMS collaboration, *Performance of the cms drift tube chambers with cosmic rays*, *Journal of Instrumentation* **5** (2010) T03015.

[54] F. Gasparini, R. Giantin, R. Martinelli, A. Meneguzzo, G. Pitacco, P. Sartori et al., *Bunch crossing identification at lhc using a mean-timer technique*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **336** (1993) 91.

[55] P. Arce, M. Bellato, M. Benettoni, A. Benvenuti, D. Bonacorsi, M. Bontenackels et al., *Bunched beam test of the cms drift tubes local muon trigger*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **534** (2004) 441.

[56] J. Smith, W. L. van Neerven and J. A. M. Vermaseren, *Transverse Mass and Width of the W Boson*, *Phys. Rev. Lett.* **50** (1983) 1738.

[57] CMS collaboration, V. Khachatryan et al., *The CMS trigger system*, *JINST* **12** (2017) P01020 [1609.02366].

[58] A. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter et al., *Particle-flow reconstruction and global event description with the CMS detector*, *Journal of Instrumentation* **12** (2017) P10003.

[59] CMS collaboration, *The Phase-2 Upgrade of the CMS Barrel Calorimeters*, tech. rep., CERN, Geneva, 2017.

[60] P. Kumar and B. Gomber, *The cms level-1 calorimeter trigger for the hl-lhc*, *Instruments* **6** (2022) .

[61] I. Ehle and on behalf of the CMS collaboration, *Design of the CMS High Granularity Calorimeter trigger primitive generator system*, *Journal of Instrumentation* **19** (2024) C02009.

[62] CMS collaboration, C. Fernandez Bedoya, *Upgrade of the CMS Drift Tube electronics for the High Luminosity LHC*, tech. rep., CERN, Geneva, 2024.

[63] C. Foudas, P. Katsoulis, T. Lama, S. Mallios, G. Karathanasis, I. Papavergou et al., *Upgrade of the CMS Barrel Muon Track Finder for HL-LHC featuring a Kalman Filter algorithm and an ATCA Host Processor with Ultrascale+ FPGAs*, p. 139, 06, 2019, DOI.

[64] A. Irshad, *The CMS GEM Detector Front-end Electronics – Characterization and Implementation*, Ph.D. thesis, Universite Libre de Bruxelles, 2021.

[65] J. F. Low, "EMTF++: pT assignment through neural networks." https://indico.cern.ch/event/932396/contributions/3918150/attachments/2062535/3460481/2020-06-23_ml_cms_l1_muon_trigger_v2.pdf, 2020.

[66] K. Bunkowski, *The algorithm of the CMS Level-1 Overlap Muon Track Finder trigger*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **936** (2019) 368.

[67] P. Leguina, *Firmware implementation of a displaced muon reconstruction algorithm for the Phase-2 Upgrade of the CMS muon system*, *Journal of Instrumentation* **18** (2023) C12005.

[68] P. A. Fokow and P. L. Lopez, "Firmware implementation of Phase-2 Overlap Muon Track Finder algorithm for CMS Level-1 trigger ." https://indico.ific.uv.es/event/6923/contributions/20246/attachments/10872/14878/Trigger-CMS_DisplacedMuons.pdf, 2024.

[69] CMS collaboration, A. Hayrapetyan et al., *Performance of the CMS high-level trigger during LHC Run 2*, 2410.17038.

[70] CMS Collaboration, "CMSSW GitHub repository: Level-1 Trigger emulators." https://github.com/cms-sw/cmssw/tree/

b21360aa34fa649a85f724c56243e229e50c0c41/L1Trigger.

[71] CMS collaboration, C. Brown, *CMS Level-1 Track Finder for the Phase-2 Upgrade*, *PoS* **VERTEX2023** (2024) 022.

[72] R. Aggleton, L. Ardila-Perez, F. Ball, M. Balzer, G. Boudoul, J. Brooke et al., *An fpga based track finder for the l1 trigger of the cms experiment at the high luminosity lhc*, *Journal of Instrumentation* **12** (2017) P12019.

[73] T. James, *Level-1 Track Finding with an all-FPGA System at CMS for the HL-LHC*, 2019.

[74] C. Brown, A. Bundock, M. Komm, V. Loncar, M. Pierini, B. Radburn-Smith et al., *Neural Network-Based Primary Vertex Reconstruction with FPGAs for the Upgrade of the CMS Level-1 Trigger System*, *Journal of Physics: Conference Series* **2438** (2023) 012106.

[75] CMS collaboration, R. E. Mccarthy, *Displaced Vertex Track Trigger for the CMS Phase-2 Upgrade*, tech. rep., CERN, Geneva, 2024.

[76] G. Petrucciani, *Particle Flow reconstruction in the CMS Level-1 Trigger for the HL-LHC*, in *EPJ Web of Conferences*, vol. 214, p. 01019, EDP Sciences, 2019.

[77] C. Herwig and on behalf of the CMS collaboration, *Particle flow reconstruction for the CMS Phase-II Level-1 Trigger*, *Journal of Instrumentation* **18** (2023) C01037.

[78] D. Bertolini, P. Harris, M. Low and N. Tran, *Pileup per particle identification*, *Journal of High Energy Physics* **2014** (2014) .

[79] CMS collaboration, *Jet Reconstruction with the Seeded Cone algorithm in the CMS Phase-2 Level-1 Trigger*, .

[80] S. Summers, I. Bestintzanos and G. Petrucciani, *Reconstructing jets in the Phase-2 upgrade of the CMS Level-1 Trigger with a seeded cone algorithm*, *EPJ Web of Conferences* **295** (2024) 02024.

[81] A. Chambers, D. Rankin, C. C. Team et al., *A neural network-based tagger for the identification of bottom quarks in the CMS Level-1 trigger*, in *APS April Meeting Abstracts*, vol. 2022, pp. H09–007, 2022.

[82] C. Collaboration, *The Phase-2 Upgrade of the CMS Beam Radiation Instrumentation and Luminosity Detectors*, tech. rep., CERN, Geneva, 2021.

[83] C. Collaboration, *The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger*, tech. rep., CERN, Geneva, 2021.

[84] A. Rose, D. Parker, G. Iles, O. Sahin, P.-A. Bausson, A. Tsirou et al., *Serenity: An ATCA prototyping platform for CMS Phase-2*, in *Proceedings of Topical Workshop on Electronics for Particle Physics — PoS(TWEPP2018)*,

TWEPP2018, p. 115, Sissa Medialab, May, 2019, DOI.

[85] Serenity Collaboration, "EMP framework." https://serenity.web.cern.ch/serenity/emp-fwk.

[86] G. Fedi, S. Fiorendi, M. Holmberg, A. Howard, G. Iles, D. Monk et al., *Lessons learnt from the first vertical slice of the CMS Outer Tracker*, *Journal of Instrumentation* **18** (2023) C01041.

[87] T. Mehner, L. Ardila-Perez, M. Balzer, G. Fedi, M. Fuchs, A. Howard et al., *Lessons learned while developing the Serenity-S1 ATCA card*, *Journal of Instrumentation* **19** (2024) C02018.

[88] Samtec Inc., "Firefly™ micro flyover system™." https://www.samtec.com/optics/optical-cable/mid-board/Firefly.

[89] CMS collaboration, P. S. and W. T., "Technical coordination & online software documentation for the phase-2 level-1 trigger." https://cms-l1t-phase2.docs.cern.ch.

[90] C. G. Larrea, K. Harder, D. Newbold, D. Sankey, A. Rose, A. Thea et al., *IPbus: a flexible Ethernet-based control system for xTCA hardware*, *Journal of Instrumentation* **10** (2015) C02019.

[91] G. Bortolato, C. Deldicque, D. Gigi, B. Huber, E. Leutgeb, A. Lobanov et al., *Architecture and prototype of the CMS Global Level-1 Trigger for Phase-2*, *Journal of Instrumentation* **18** (2023) C01034.

[92] G. Bortolato, M. Cepeda, J. Heikkilä, B. Huber, E. Leutgeb, D. Rabady et al., *Adaptability and efficiency of the CMS Level-1 Global Trigger firmware implementation for Phase-2*, *Journal of Instrumentation* **19** (2024) C03007.

[93] G. Bortolato, M. Cepeda, J. Heikkilä, B. Huber, E. Leutgeb, D. Rabady et al., *Design and implementation of neural network based conditions for the CMS Level-1 Global Trigger upgrade for the HL-LHC*, *Journal of Instrumentation* **19** (2024) C03019.

[94] G. Bortolato, *A new Global Trigger for the CMS Experiment at the HL-LHC: Neural Network algorithms and Hardware Implementation*, Master's thesis, Università degli studi di Padova, 2022.

[95] CMS collaboration, "CMSSW: Framework and Event Data Model Offline Guide." https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideFrameWork.

[96] C. Jones, M. Paterno, J. Kowalkowski, L. Sexton-Kennedy and W. Tanenbaum, *The new CMS event data model and framework*, *Proceedings for Computing in High-Energy Physics (CHEP'06), Mumbai, India* **13** (2006) .

[97] D. Spiga, S. Lacaprara, W. Bacchi, M. Cinquilli, G. Codispoti, M. Corvo et al., *CRAB: the CMS distributed analysis tool development and design*, *Nuclear*

Physics B - Proceedings Supplements **177-178** (2008) 267.

[98] CMS collaboration, G. Bortolato, M. L. Cepeda, J. K. Heikkilae, B. Huber, E. Leutgeb, D. S. Rabady et al., *The Level-1 Global Trigger for Phase-2 Algorithms, configuration and integration in the CMS offline framework*, .

[99] E. Leutgeb, "The CMS Phase-2 Global Trigger: Hardware, infrastructure and physics." unpublished Ph.D. thesis, Technische Universität Wien, 2025.

[100] W. Jakob, J. Rhinelander and D. Moldovan, "pybind11 – Seamless operability between C++11 and Python." https://github.com/pybind/pybind11, 2017.

[101] Advanced Micro Devices, Inc., "UltraScale Architecture Configurable Logic Block User Guide (UG574)." https://docs.amd.com/r/en-US/ug574-ultrascale-clb.

[102] Advanced Micro Devices, Inc., "UltraScale Architecture Memory Resources User Guide (UG573)." https://docs.amd.com/v/u/en-US/ug573-ultrascale-memory-resources.

[103] Advanced Micro Devices, Inc., "UltraScale Architecture DSP Slice (UG579)." https://docs.amd.com/v/u/en-US/ug579-ultrascale-dsp.

[104] J. Alimena, J. Beacham, M. Borsato, Y. Cheng, X. C. Vidal, G. Cottin et al., *Searching for long-lived particles beyond the Standard Model at the Large Hadron Collider*, *Journal of Physics G: Nuclear and Particle Physics* **47** (2020) 090501.

[105] L. Lee, C. Ohm, A. Soffer and T.-T. Yu, *Collider searches for long-lived particles beyond the Standard Model*, *Progress in Particle and Nuclear Physics* **106** (2019) 210.

[106] M. Jeitler, A. Taurok, H. Bergauer, C. Deldicque, J. Erö, M. Ghete et al., *The level-1 global trigger for the CMS experiment at LHC*, *Journal of Instrumentation* **2** (2007) P01006.

[107] J. Wittmann, G. Aradi, B. Arnold, H. Bergauer, M. Jeitler, T. Matsushita et al., *Design and performance of the phase I upgrade of the CMS Global Trigger*, *Journal of Instrumentation* **12** (2017) C01046.

[108] CMS collaboration, A. M. Sirunyan et al., *Performance of the CMS Level-1 trigger in proton-proton collisions at* $\sqrt{s} = 13$ *TeV*, *JINST* **15** (2020) P10017 [2006.10165].

[109] C. Lundy, "Primary Vertex dZ Cut Firmware for CMS p2gt." https://indico.cern.ch/event/1313804/contributions/5526659/attachments/2696642/4680373/CERN%20Presentation%203%20(1).pdf, 2023.

[110] B. Huber, G. Bortolato, E. Leutgeb, D. Rabady, H. Sakulin and on behalf of the CMS collaboration, *Optimizing cut-based algorithms to specific physics acceptance regions*, *EPJ Web of Conferences* (2025) in press.

[111] A. P. Wierzbicki, *A mathematical basis for satisficing decision making*, *Mathematical Modelling* **3** (1982) 391.

[112] K. Miettinen and M. M. Mäkelä, *On scalarizing functions in multiobjective optimization*, *OR Spectrum* **24** (2002) 193.

[113] R. Hooke and T. A. Jeeves, *"Direct Search" Solution of Numerical and Statistical Problems*, *J. ACM* **8** (1961) 212.

[114] A. J. Buras, J. Girrbach, D. Guadagnoli and G. Isidori, *On the Standard Model prediction for* $\mathcal{B}(B_{s,d} \to \mu^+\mu^-)$, *The European Physical Journal C* **72** (2012) .

[115] C. Bierlich, S. Chakraborty, N. Desai, L. Gellersen, I. Helenius, P. Ilten et al., *Codebase release 8.3 for PYTHIA*, *SciPost Physics Codebases* (2022) .

[116] Particle Data Group collaboration, S. Navas et al., *Review of particle physics*, *Phys. Rev. D* **110** (2024) 030001.

[117] P. Nason and C. Oleari, *NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG*, *JHEP* **02** (2010) 037.

[118] S. L. Glashow, J. Iliopoulos and L. Maiani, *Weak interactions with lepton-hadron symmetry*, *Phys. Rev. D* **2** (1970) 1285.

[119] S. Petcov, *The Processes* $\mu \to e\gamma$, $\mu \to ee\overline{e}$, $\nu' \to \nu\gamma$ *in the Weinberg-Salam Model with Neutrino Mixing*, *Sov. J. Nucl. Phys* **25** (1977) 340.

[120] P. Blackstone, M. Fael and E. Passemar, $\tau \to \mu\mu\mu$ *at a rate of one out of* 10^{14} *tau decays?*, *The European Physical Journal C* **80** (2020) .

[121] X.-Y. Pham, *Lepton flavor changing in neutrinoless* τ *decays*, *The European Physical Journal C* **8** (1999) 513.

[122] A. Cerri, V. V. Gligorov, S. Malvezzi, J. M. Camalich, J. Zupan, S. Akar et al., *Opportunities in Flavour Physics at the HL-LHC and HE-LHC*, 2019.

List of acronyms

ADC	Analog-to-digital converter
ALICE	A Large Ion Collider Experiment
ATCA	Advanced Telecommunications Computing Architecture
ATLAS	A Toroidal LHC Apparatus
ASIC	Application-specific integrated circuit
BC	Barrel Calorimeter
BCP	Barrel calorimeter processor board
BCT	Barrel Calorimeter Trigger
BDT	Boosted Decision Tree
BMTF	Barrel Muon Track Finder
BMTL1	Barrel Muon Trigger Layer-1
BPTX	Beam pickup timing for experiments
BRAM	Block random-access memory
BRIL	Beam radiation, instrumentation, and luminosity detectors
BX	Bunch crossing
CERN	Conseil Européen pour la Recherche Nucléaire
CE-E	Calorimeter Endcap – Electromagnetic
CE-H	Calorimeter Endcap – Hadronic
CLB	Configurable logic block
CMS	Compact Muon Solenoid
CMSSW	Compact Muon Solenoid Software
CSC	Cathode strip chamber
CRAB	CMS Remote Analysis Builder
DAQ	Data acquisition
DSP	Digital signal processing
DT	Drift tube
ECAL	Electromagnetic Calorimeter
EDM	Event Data Model
EM	Electromagnetic

EMP	Extensible, Modular (data) Processor
EMTF	Endcap Muon Track Finder
FPGA	Field-programmable gate array
GCT	Global Calorimeter Trigger
GEM	Gas electron multiplier
GMT	Global Muon Trigger
GT	Global Trigger
GTT	Global Track Trigger
HB	Hadron Calorimeter Barrel section
HCAL	Hadron Calorimeter
HF	Hadron Forward Calorimeter
HE	Hadron Calorimeter Endcap section
HGCAL	High-Granularity Calorimeter
HO	Hadron Calorimeter Outer section
HLT	High-Level Trigger
HT	Hadronic transverse energy
iRPC	Improved resistive plate chamber
KU15P	Xilinx Kintex UltraScale+ KU15P
L1T	Level-1 Trigger
LEIR	Low Energy Ion Ring
LEP	Large Electron-Positron Collider
LHCb	LHC-beauty
LHC	Large Hadron Collider
LUT	Look-up table
LSb	Least significant bit
MET	Missing transverse energy
MHT	Missing hadronic transverse energy
MIP	Minimum ionizing particle
OMTF	Overlap Muon Track Finder
PCB	Printed circuit board
PU	Pileup

PUPPI	Pileup per particle identification
PS	Proton Synchrotron
PSB	Proton Synchrotron Booster
PV	Primary vertex
RPC	Resistive plate chamber
RAM	Random-access memory
RF	Radio frequency
SiPM	Silicon photomultiplier
SL	Super layer
SLR	Super Logic Region
SMASH	Serenity MAnagement SHell
SPS	Super Proton Synchrotron
SRL	Shift register look-up table
SUSY	Supersymmetry
SV	Secondary vertex
TCDS	Trigger Control and Distribution System
TMUX	Time multiplexing
VHDL	Very High Speed Integrated Circuit Hardware Description Language
VU9P	Xilinx Virtex UltraScale+ VU9P
VU13P	Xilinx Virtex UltraScale+ VU13P

Appendix

A.1 List of implemented cuts

The tables below outline the implemented cuts in both CMSSW and firmware. The listed names correspond to parameter names within CMSSW, which may differ slightly from their firmware counterparts. Likewise, the specified data types are those used for configuring condition modules and are converted to hardware values using the operations described in the "hardware conversion" columns. The "expression" column details the actual calculation and comparison operations performed in both hardware and software, where X denotes a placeholder for the cut value. Note: In the actual implementation, any constant factors are integrated into X during the hardware conversion process.

Table 7: Implemented cuts on single object quantities.

Name	Expression	Datatype	Hardware conversion
minPt	$p_T > X$ or $\lvert\sum \vec{p}_T\rvert > X$	double	floor (X/LSb_{p_T})
maxPt	$p_T < X$ or $\lvert\sum \vec{p}_T\rvert < X$	double	ceil (X/LSb_{p_T})
minEta	$\eta > X$	double	floor (X/LSb_{η})
maxEta	$\eta < X$	double	ceil (X/LSb_{η})
minPhi	$\phi > X$	double	floor (X/LSb_{ϕ})
maxPhi	$\phi < X$	double	ceil (X/LSb_{ϕ})
minZ0	$z_0 > X$	double	floor (X/LSb_{z_0})
maxZ0	$z_0 < X$	double	ceil (X/LSb_{z_0})
minScalarSumPt[1]	$\sum p_T > X$	double	floor $\left(X/\mathrm{LSb}_{\sum p_T}\right)$
maxScalarSumPt[1]	$\sum p_T < X$	double	ceil $\left(X/\mathrm{LSb}_{\sum p_T}\right)$
minQualityScore	qualityScore $> X$	uint32	X
maxQualityScore	qualityScore $< X$	uint32	X
qualityFlags	qualityFlags $\wedge X = X$	uint32	X
minAbsEta	$\lvert\eta\rvert > X$	double	floor (X/LSb_{η})
maxAbsEta	$\lvert\eta\rvert < X$	double	ceil (X/LSb_{η})
minIsolationPt	$P_T^{\mathrm{isol}} > X$	double	floor $\left(X/\mathrm{LSb}_{P_T^{\mathrm{isol}}}\right)$
maxIsolationPt	$P_T^{\mathrm{isol}} < X$	double	ceil $\left(X/\mathrm{LSb}_{P_T^{\mathrm{isol}}}\right)$
minRelIsolationPt	$P_T^{\mathrm{isol}} > X \cdot p_T$	double	floor $\left(2^{18} \cdot X \cdot \mathrm{LSb}_{p_T}/\mathrm{LSb}_{P_T^{\mathrm{isol}}}\right)$
maxRelIsolationPt	$P_T^{\mathrm{isol}} < X \cdot p_T$	double	ceil $\left(2^{18} \cdot X \cdot \mathrm{LSb}_{p_T}/\mathrm{LSb}_{P_T^{\mathrm{isol}}}\right)$
minPrimVertDz[2]	$\lvert z_0 - Z_0\rvert > X$	double	floor (X/LSb_{z_0})
maxPrimVertDz[2]	$\lvert z_0 - Z_0\rvert < X$	double	ceil (X/LSb_{z_0})
minPtMultiplicityCut[3]	$\sum (p_T > X) \geq N$	double	floor (X/LSb_{p_T})

[1]The sum is computed by an upstream trigger.

[2]Requires an additional uint32 parameter "primVertex" specifying the index within the primary vertex collection, whose longitudinal distance from the interaction point is denoted by Z_0.

[3]Requires an additional uint32 parameter "minPtMultiplicityN", denoting the multiplicity count.

Table 8: Implemented η-region dependent cuts, enabling different thresholds for different regions. To use these cuts, the parameter "regionsAbsEtaLowerBounds" must be defined, specifying the lower bounds of the η-regions. This parameter is a vdouble, with its length determining the number of η-region. Each region extends from its specified lower bound (inclusive) to the next region's lower bound (exclusive). The upper bound of the last region is always the maximum allowed $|\eta| = 2\pi$.

Name	Expression	Datatype	Hardware conversion
regionsMinPt	$p_T > X$ or $\lvert\sum \vec{p}_T\rvert > X$	vdouble	floor (X/LSb_{p_T})
regionsQualityFlags	qualityFlags $\wedge\, X = X$	vuint32	X
regionsMaxRelIsolationPt	$P_T^{\mathrm{isol}} > X \cdot p_T$	vdouble	ceil $\left(2^{18} \cdot X \cdot \mathrm{LSb}_{p_T}/\mathrm{LSb}_{P_T^{\mathrm{isol}}}\right)$

Table 9: Implemented cuts on topological correlations between two objects. Note: $\Delta\eta = |\eta_1 - \eta_2|$ while $\Delta\phi$ is the smallest azimuthal angle between two momentum vectors, accounting for periodicity, i.e. $\phi = \pi = -\pi$. Most hardware conversions are performed during the VHDL translation step. However, the quantities M, M_T, $M/\Delta R$, and the combined two-object P_T are partially converted within VHDL itself, from *real* datatypes to circumvent the 32-bit integer limitation of VHDL-2008 and to simplify module instantiation.

Name	Expression	Datatype	Hardware conversion
minDEta	$\lvert\eta_1 - \eta_2\rvert > X$	double	floor (X/LSb_{η})
maxDEta	$\lvert\eta_1 - \eta_2\rvert < X$	double	ceil (X/LSb_{η})
minDPhi	$\Delta\phi > X$	double	floor $\left(X/\mathrm{LSb}_{\phi}\right)$
maxDPhi	$\Delta\phi < X$	double	floor $\left(X/\mathrm{LSb}_{\phi}\right)$
minDz	$\lvert z_{0_1} - z_{0_2}\rvert > X$	double	floor (X/LSb_{z_0})
maxDz	$\lvert z_{0_1} - z_{0_2}\rvert < X$	double	ceil (X/LSb_{z_0})
minDR	$\Delta\phi^2 + \Delta\eta^2 > X^2$	double	floor $\left(X^2/\mathrm{LSb}_{\eta}^2\right)$
maxDR	$\Delta\phi^2 + \Delta\eta^2 < X^2$	double	ceil $\left(X^2/\mathrm{LSb}_{\eta}^2\right)$
minInvMass	$p_{T_1}\, p_{T_2}\, [\cosh(\Delta\eta) - \cos(\Delta\phi)] > X^2/2$	double	floor $\left(\frac{X^2 \cdot C_{\mathrm{LUT}}}{2 \cdot \mathrm{LSb}_{p_T}^2}\right)$
maxInvMass	$p_{T_1}\, p_{T_2}\, [\cosh(\Delta\eta) - \cos(\Delta\phi)] < X^2/2$	double	ceil $\left(\frac{X^2 \cdot C_{\mathrm{LUT}}}{2 \cdot \mathrm{LSb}_{p_T}^2}\right)$
minTransMass	$p_{T_1}\, p_{T_2}\, [1 - \cos(\Delta\phi)] > X^2/2$	double	floor $\left(\frac{X^2 \cdot C_{\mathrm{LUT}}}{2 \cdot \mathrm{LSb}_{p_T}^2}\right)$
maxTransMass	$p_{T_1}\, p_{T_2}\, [1 - \cos(\Delta\phi)] < X^2/2$	double	ceil $\left(\frac{X^2 \cdot C_{\mathrm{LUT}}}{2 \cdot \mathrm{LSb}_{p_T}^2}\right)$
minCombPt	$p_{T_1}^2 + p_{T_2}^2 + 2p_{T_1}\, p_{T_2} \cos(\Delta\phi) > X^2$	double	floor $\left(\frac{X^2 \cdot C_{\mathrm{LUT}}}{\mathrm{LSb}_{p_T}^2}\right)$
maxCombPt	$p_{T_1}^2 + p_{T_2}^2 + 2p_{T_1}\, p_{T_2} \cos(\Delta\phi) < X^2$	double	ceil $\left(\frac{X^2 \cdot C_{\mathrm{LUT}}}{\mathrm{LSb}_{p_T}^2}\right)$
minInvMassOverDR	$M^2/2 > X^2 \cdot \Delta R^2/2$	double	floor $\left(\frac{2^{19} \cdot X^2 \cdot C_{\mathrm{LUT}} \cdot \mathrm{LSb}_{\eta}^2}{2 \cdot \mathrm{LSb}_{p_T}^2}\right)$
maxInvMassOverDR	$M^2/2 < X^2 \cdot \Delta R^2/2$	double	ceil $\left(\frac{2^{19} \cdot X^2 \cdot C_{\mathrm{LUT}} \cdot \mathrm{LSb}_{\eta}^2}{2 \cdot \mathrm{LSb}_{p_T}^2}\right)$
os	$q_1 \neq q_2$	bool	
ss	$q_1 = q_2$	bool	

Table 10: Implemented cuts on three-object correlations.

Name	Expression	Datatype	Hardware conversion
minInvMass	$\frac{M_{12}^2}{2} + \frac{M_{13}^2}{2} + \frac{M_{23}^2}{2} > \frac{X^2}{2}$	double	floor $\left(\frac{X^2 \cdot C_{\text{LUT}}}{2 \cdot \text{LSb}_{p_T}^2}\right)$
maxInvMass	$\frac{M_{12}^2}{2} + \frac{M_{13}^2}{2} + \frac{M_{23}^2}{2} < \frac{X^2}{2}$	double	ceil $\left(\frac{X^2 \cdot C_{\text{LUT}}}{2 \cdot \text{LSb}_{p_T}^2}\right)$
minTransMass	$\frac{M_{T12}^2}{2} + \frac{M_{T13}^2}{2} + \frac{M_{T23}^2}{2} > \frac{X^2}{2}$	double	floor $\left(\frac{X^2 \cdot C_{\text{LUT}}}{2 \cdot \text{LSb}_{p_T}^2}\right)$
maxTransMass	$\frac{M_{T12}^2}{2} + \frac{M_{T13}^2}{2} + \frac{M_{T23}^2}{2} < \frac{X^2}{2}$	double	ceil $\left(\frac{X^2 \cdot C_{\text{LUT}}}{2 \cdot \text{LSb}_{p_T}^2}\right)$

Table 11: Implemented sum cuts that aggregate variables from N objects, where N corresponds to the number of objects targeted by a specific condition ($N = 2$ for double-object conditions, $N = 3$ for triple-object conditions, and $N = 4$ for quad-object conditions).

Name	Expression	Datatype	Hardware conversion
minQualityScoreSum	$\sum_{i=1}^{N} \text{qualityScore} > X$	unit32	X
maxQualityScoreSum	$\sum_{i=1}^{N} \text{qualityScore} < X$	unit32	X

A.2 First prototype menu for High-Luminosity operation

The first prototype menu, maintained by the Detector Performance Group, as it exists within CMSSW at the moment [70]. It is a slightly modified version of the one listed in the technical design report [50] and is commonly referred to as "Step-1 menu".

Note that in the following tables, transverse momentum (p_T) cuts are based on the p_T calibration of the current Run 3 system. This calibration is intentionally not centred and, on average, overestimates the true p_T value to reduce the likelihood of excluding certain "good" candidates due to reconstruction uncertainties. The actual configured p_T cut values, as found in [70], are rescaled versions of their Run 3 counterparts, while the algorithm naming and table listing follow the old Run 3 system.

Table 12: Leptonic algorithms of the current prototype menu for High-Luminosity operation. Note: Electron and Photon ID denote certain working points of the calorimeter reconstruction.

Name	Trigger objects	Cuts
SingleTkMuon22	Track-matched muons	$p_T > 22$ GeV, $\lvert\eta\rvert < 2.4$

Continued on next page

Table 12: Leptonic algorithms of the current prototype menu for High-Luminosity operation. Note: Electron and Photon ID denote certain working points of the calorimeter reconstruction. (Continued)

Name	Trigger objects	Cuts
DoubleTkMuon15_7	1: Track-matched muons 2: Track-matched muons	$p_{T_1} > 15$ GeV, $\lvert\eta_1\rvert < 2.4$, $p_{T_2} > 7$ GeV, $\lvert\eta_2\rvert < 2.4$, quality$_2$ = loose, $\lvert z_{0_1} - z_{0_2}\rvert < 1$ cm, $\Delta R > 0$
TripleTkMuon5_3_3	1: Track-matched muons 2: Track-matched muons 3: Track-matched muons	$p_{T_1} > 5$ GeV, $\lvert\eta_1\rvert < 2.4$, quality$_1$ = loose, $p_{T_2} > 3$ GeV, $\lvert\eta_2\rvert < 2.4$, quality$_2$ = loose, $p_{T_3} > 3$ GeV, $\lvert\eta_3\rvert < 2.4$, quality$_3$ = loose, $\lvert z_{0_1} - z_{0_2}\rvert < 1$ cm, $\lvert z_{0_1} - z_{0_3}\rvert < 1$ cm, $\Delta R_{12} > 0$, $\Delta R_{13} > 0$, $\Delta R_{23} > 0$
SingleEGEle51	Correlator photons	$p_T > 51$ GeV, $\lvert\eta\rvert < 2.4$, Barrel: Electron ID, Endcap: Photon ID
DoubleEGEle37_24	1: Correlator photons 2: Correlator photons	$p_{T_1} > 37$ GeV, $\lvert\eta_1\rvert < 2.4$, Barrel: Electron ID, Endcap: Photon ID, $p_{T_2} > 24$ GeV, $\lvert\eta_2\rvert < 2.4$, Barrel: Electron ID, Endcap: Photon ID, $\Delta R > 0.1$
IsoTkEleEGEle22_12	1: Correlator electrons 2: Correlator photons	$p_{T_1} > 22$ GeV, $\lvert\eta_1\rvert < 2.4$, Barrel: $\frac{P_{T_1}^{\text{isol}}}{p_{T_1}} < 0.13$, Endcap: $\frac{P_{T_1}^{\text{isol}}}{p_{T_1}} < 0.28$, $p_{T_2} > 12$ GeV, $\lvert\eta_2\rvert < 2.4$, Barrel: Electron ID, Endcap: Photon ID, $\Delta R > 0.1$
SingleTkEle36	Correlator electrons	$p_T > 36$ GeV, $\lvert\eta\rvert < 2.4$, Electron ID
SingleIsoTkEle28	Correlator electrons	$p_T > 28$ GeV, $\lvert\eta\rvert < 2.4$, Barrel: $\frac{P_T^{\text{isol}}}{p_T} < 0.13$, Endcap: $\frac{P_T^{\text{isol}}}{p_T} < 0.28$
SingleIsoTkPho36	Correlator electrons	$p_T > 36$ GeV, $\lvert\eta\rvert < 2.4$, Barrel: $\frac{P_T^{\text{isol}}}{p_T} < 0.25$, Electron ID Endcap: $\frac{P_T^{\text{isol}}}{p_T} < 0.205$, Photon ID
DoubleTkEle25_12	1: Correlator electrons 2: Correlator electrons	$p_{T_1} > 25$ GeV, $\lvert\eta_1\rvert < 2.4$, Barrel: Electron ID, $p_{T_2} > 12$ GeV, $\lvert\eta_2\rvert < 2.4$ Barrel: Electron ID, $\lvert z_{0_1} - z_{0_2}\rvert < 1$ cm

Continued on next page

Table 12: Leptonic algorithms of the current prototype menu for High-Luminosity operation. Note: Electron and Photon ID denote certain working points of the calorimeter reconstruction. (Continued)

Name	Trigger objects	Cuts
DoubleIsoTkPho22_12	1: Correlator photons 2: Correlator photons	$p_{T_1} > 22$ GeV, $\lvert\eta_1\rvert < 2.4$, Barrel: Electron ID $\frac{P_{T_1}^{\text{isol}}}{p_{T_1}} < 0.25$, Endcap: Photon ID, $\frac{P_{T_1}^{\text{isol}}}{p_{T_1}} < 0.205$ $p_{T_2} > 12$ GeV, $\lvert\eta_2\rvert < 2.4$ Barrel: Electron ID, $\frac{P_{T_2}^{\text{isol}}}{p_{T_2}} < 0.25$, Endcap: Photon ID, $\frac{P_{T_2}^{\text{isol}}}{p_{T_2}} < 0.205$
DoublePuppiTau52_52	1: Correlator τ_h 2: Correlator τ_h	$p_{T_1} > 52$ GeV, $\lvert\eta_1\rvert < 2.17$, Neural-net score > 0.22, $p_{T_2} > 52$ GeV, $\lvert\eta_2\rvert < 2.17$ Neural-net score > 0.22, $\Delta R > 0.5$

Table 13: Hadronic and missing transverse energy algorithms of the current prototype menu for High-Luminosity operation.

Name	Trigger objects	Cuts
SinglePuppiJet230	Correlator jets	$p_T > 230$ GeV, $\lvert\eta\rvert < 2.4$
DoublePuppiJet112_112	1: Correlator jets 2: Correlator jets	$p_{T_1} > 112$ GeV, $\lvert\eta_1\rvert < 2.4$, $p_{T_2} > 112$ GeV, $\lvert\eta_2\rvert < 2.4$ $\Delta\eta < 1.6$
DoublePuppiJet160_35_Mass620	1: Correlator jets 2: Correlator jets	$p_{T_1} > 160$ GeV, $\lvert\eta_1\rvert < 5$, $p_{T_2} > 35$ GeV, $\lvert\eta_2\rvert < 5$ $M > 620$ GeV
PuppiHT450	Correlator MHT	$\sum p_T > 450$ GeV
PuppiMHT140	Correlator MHT	$\lvert\sum \vec{p}_T\rvert > 140$ GeV
PuppiMET200	Correlator MET	$\lvert\sum \vec{p}_T\rvert > 200$ GeV
PuppiHT400_QuadJet70_55_40_40	1: Correlator MHT 2: Correlator jets 3: Correlator jets 4: Correlator jets 5: Correlator jets	$\sum p_{T_1} > 400$ GeV $p_{T_2} > 70$ GeV, $\lvert\eta_2\rvert < 2.4$, $p_{T_3} > 55$ GeV, $\lvert\eta_3\rvert < 2.4$ $p_{T_4} > 40$ GeV, $\lvert\eta_4\rvert < 2.4$ $p_{T_5} > 40$ GeV, $\lvert\eta_5\rvert < 2.4$

Table 14: Hadronic cross leptonic algorithms of the current prototype menu for High-Luminosity operation.

Name	Trigger objects	Cuts
TkMuonPuppiHT6320	1: Track-matched muons 2: Correlator MHT	$p_{T_1} > 6$ GeV, $\|\eta\| < 2.4$, quality$_1$ = loose, $\|z_{0_1} - Z_0\| < 1$ cm, $\sum p_{T_2} > 320$ GeV
TkMuTriPuppiJet_12_40_ dRMax_DoubleJet_dEtaMax	1: Track-matched muons 2: Correlator jets 3: Correlator jets 4: Correlator jets	$p_{T_1} > 12$ GeV, $\|\eta_1\| < 2.4$, quality$_1$ = loose, $\|z_{0_1} - Z_0\| < 1$ cm, $p_{T_2} > 40$ GeV, $\|\eta_2\| < 2.4$, $p_{T_3} > 40$ GeV, $\|\eta_3\| < 2.4$, $p_{T_4} > 40$ GeV, $\|\eta_4\| < 2.4$, $\Delta R_{12} < 0.4$, $\Delta\eta_{34} < 1.6$
TkMuPuppiJetPuppiMet_ 3_110_120	1: Track-matched muons 2: Correlator jets 3: Correlator MET	$p_{T_1} > 3$ GeV, $\|\eta_1\| < 2.1$, quality$_1$ = loose, $\|z_{0_1} - Z_0\| < 1$ cm, $p_{T_2} > 110$ GeV, $\|\eta_2\| < 2.4$, $\|\sum \vec{p}_{T_3}\| > 120$ GeV
DoubleTkMuPuppiJetPuppi Met_3_3_60_130	1: Track-matched muons 2: Track-matched muons 4: Correlator jets 4: Correlator MET	$p_{T_1} > 3$ GeV, $\|\eta_1\| < 2.4$, quality$_1$ = loose, $\|z_{0_1} - Z_0\| < 1$ cm, $p_{T_2} > 3$ GeV, $\|\eta_2\| < 2.4$, quality$_2$ = loose, $\|z_{0_2} - Z_0\| < 1$ cm, $p_{T_3} > 60$ GeV, $\|\eta_3\| < 2.4$, $\|\sum \vec{p}_{T_4}\| > 130$ GeV, $\Delta R_{12} > 0$
DoubleTkMuPuppiHT_ 3_3_300	1: Track-matched muons 2: Track-matched muons 3: Correlator MHT	$p_{T_1} > 3$ GeV, $\|\eta_1\| < 2.4$, quality$_1$ = loose, $\|z_{0_1} - Z_0\| < 1$ cm, $p_{T_2} > 3$ GeV, $\|\eta_2\| < 2.4$, quality$_2$ = loose, $\|z_{0_2} - Z_0\| < 1$ cm, $\sum p_{T_3} > 300$ GeV, $\Delta R_{12} > 0$
DoubleTkElePuppiHT_ 8_8_390	1: Correlator electrons 2: Correlator electrons 3: Correlator MHT	$p_{T_1} > 8$ GeV, $\|\eta_1\| < 2.4$, Barrel: Electron ID, $\|z_{0_1} - Z_0\| < 1$ cm, $p_{T_2} > 8$ GeV, $\|\eta_2\| < 2.4$, Barrel: Electron ID, $\|z_{0_2} - Z_0\| < 1$ cm, $\sum p_{T_3} > 390$ GeV
TkEleIsoPuppiHT_ 26_190	1: Correlator electrons 2: Correlator MHT	$p_{T_1} > 26$ GeV, $\|\eta_1\| < 2.1$, Barrel: Electron ID, $\frac{P_T^{\text{isol}}}{p_T} < 0.13$, Endcap: $\frac{P_T^{\text{isol}}}{p_T} < 0.28$, $\|z_{0_1} - Z_0\| < 1$ cm, $\sum p_{T_2} > 190$ GeV

Continued on next page

Table 14: Hadronic cross leptonic algorithms of the current prototype menu for High-Luminosity operation. (Continued)

Name	Trigger objects	Cuts
TkElePuppiJet_28_ 40_MinDR	1: Correlator electrons 2: Correlator jets	$p_{T_1} > 28$ GeV, $\lvert\eta_1\rvert < 2.1$, Barrel: Electron ID, $\frac{P_T^{\text{isol}}}{p_T} < 0.13$, Endcap: $\frac{P_T^{\text{isol}}}{p_T} < 0.28$, $\lvert z_{0_1} - Z_0\rvert < 1$ cm, $p_{T_2} > 40$ GeV, $\lvert\eta_2\rvert < 2.4$ $\Delta R > 0.3$
NNPuppiTauPuppiMet_ 55_190	1: Correlator τ_h 2: Correlator MET	$p_{T_1} > 55$ GeV, $\lvert\eta_1\rvert < 2.17$, Neural-net score > 0.22, $p_{T_2} > 190$

Table 15: Bottom-quark-centric algorithms of the current prototype menu for High-Luminosity operation.

Name	Trigger objects	Cuts
DoubleTkMuon_OS_ Er1p5_Dr1p4	1: Track-matched muons 2: Track-matched muons	$\lvert\eta_1\rvert < 1.5$, quality = loose, $\lvert\eta_2\rvert < 1.5$, quality = loose, $0 < \Delta R < 1.4$, $\lvert z_{0_1} - z_{0_2}\rvert < 1$ cm, $q_1 \neq q_2$
DoubleTkMuon_4_4_ OS_Dr1p2	1: Track-matched muons 2: Track-matched muons	$p_{T_1} > 4$, $\lvert\eta_1\rvert < 2.4$, quality = loose, $p_{T_2} > 4$, $\lvert\eta_2\rvert < 2.4$, quality = loose, $0 < \Delta R < 1.2$, $\lvert z_{0_1} - z_{0_2}\rvert < 1$ cm, $q_1 \neq q_2$
DoubleTkMuon_4p5_ 4p5_OS_Er2_Mass7to18	1: Track-matched muons 2: Track-matched muons	$p_{T_1} > 4$, $\lvert\eta_1\rvert < 2$, quality = loose, $p_{T_2} > 4$, $\lvert\eta_2\rvert < 2$, quality = loose, 7 GeV $< M <$ 18 GeV $\Delta R > 0$, $\lvert z_{0_1} - z_{0_2}\rvert < 1$ cm, $q_1 \neq q_2$
TripleTkMuon_5_3_0_ DoubleTkMuon_5_3_ OS_MassTo9	1: Track-matched muons 2: Track-matched muons 3: Track-matched muons	$p_{T_1} > 5$, $\lvert\eta_1\rvert < 2.4$, quality = loose, $p_{T_2} > 4$, $\lvert\eta_2\rvert < 2.4$, quality = loose, $\lvert\eta_3\rvert < 2.4$, quality = loose, $M_{12} < 9$ GeV, $\Delta R_{12} > 0$, $q_1 \neq q_2$ $\lvert z_{0_1} - z_{0_2}\rvert < 1$ cm, $\lvert z_{0_1} - z_{0_3}\rvert < 1$ cm, $\Delta R_{13} > 0$, $\Delta R_{23} > 0$
TripleTkMuon_5_3p5_ 2p5_OS_Mass5to17	1: Track-matched muons 2: Track-matched muons 3: Track-matched muons	$p_{T_1} > 5$, $\lvert\eta_1\rvert < 2.4$, quality = loose, $p_{T_2} > 4$, $\lvert\eta_2\rvert < 2.4$, quality = loose, $p_{T_2} > 2$, $\lvert\eta_3\rvert < 2.4$, quality = loose, $\lvert z_{0_1} - z_{0_2}\rvert < 1$ cm, $\Delta R_{12} > 0$ 5 GeV $< M_{13} <$ 17 GeV, $\Delta R_{13} > 0$, $\lvert z_{0_1} - z_{0_3}\rvert < 1$ cm, $q_1 \neq q_3$, $\Delta R_{23} > 0$