Journal of Instrumentation

Development of the CMS detector for the CERN LHC Run 3

To cite this article: A. Hayrapetyan *et al* 2024 *JINST* **19** P05064

Published by IOP Publishing for Sissa Medialab

Received: *September 11, 2023*
Accepted: *November 6, 2023*
Published: *May 23, 2024*

The Large Hadron Collider and the experiments for Run 3 —
Accelerator and experiments for LHC Run3

Development of the CMS detector for the CERN LHC Run 3

The CMS collaboration

E-mail: cms-publication-committee-chair@cern.ch

Abstract: Since the initial data taking of the CERN LHC, the CMS experiment has undergone substantial upgrades and improvements. This paper discusses the CMS detector as it is configured for the third data-taking period of the CERN LHC, Run 3, which started in 2022. The entire silicon pixel tracking detector was replaced. A new powering system for the superconducting solenoid was installed. The electronics of the hadron calorimeter was upgraded. All the muon electronic systems were upgraded, and new muon detector stations were added, including a gas electron multiplier detector. The precision proton spectrometer was upgraded. The dedicated luminosity detectors and the beam loss monitor were refurbished. Substantial improvements to the trigger, data acquisition, software, and computing systems were also implemented, including a new hybrid CPU/GPU farm for the high-level trigger.

Keywords: Calorimeters; Large detector systems for particle and astroparticle physics; Particle tracking detectors

ArXiv ePrint: 2309.05466

https://doi.org/10.1088/1748-0221/19/05/P05064

2024 JINST 19 P05064

Contents

2024 JINST 19 P05064

---

1 Introduction

The CMS detector [1] is a large, multipurpose apparatus located at the CERN LHC. The detector was designed for the study of a variety of physics phenomena, including the search for the Higgs boson, which was discovered in 2012 [2–4], and the measurement of its properties, the exploration of the electroweak sector and vector boson scattering, precision measurements of standard model (SM) particles and interactions, flavor physics, heavy-ion physics, and searches for new physics beyond the SM.

The LHC Run 1 started in 2009 and, until the end of 2012, proton-proton (pp) collision data corresponding to a total integrated luminosity of about 30 fb^{-1} were delivered at center-of-mass energies of 7 and 8 TeV . In addition, CMS successfully recorded data from high-energy lead-lead collisions. After a first long shutdown, referred to as long shutdown 1 (LS1), the second data-taking period, Run 2, followed in 2015–2018 at an energy of 13 TeV, during which an integrated luminosity of about 165 fb^{-1} was delivered with peak instantaneous luminosities up to $2\times10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, twice the original LHC design value. During LS1 and Run 2, the first generation of detector upgrades, referred to as the Phase 1 upgrade program, was implemented.

After the second long shutdown (LS2, 2019–2021), the LHC Run 3 was started in 2022 and is expected to deliver about 250 fb^{-1} of integrated luminosity. In Run 3, the center-of-mass energy for pp collisions is 13.6 TeV . During the third long shutdown (LS3), scheduled to start in 2026, CMS will undergo a major upgrade program, referred to as the Phase 2 upgrade, in preparation for the data taking at the High-Luminosity LHC (HL-LHC), designed to deliver instantaneous luminosities up to $7.5 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ at a pp center-of-mass energy of 14 TeV . At the end of the HL-LHC, the total integrated luminosity is expected to be 3000 fb^{-1}. The Phase 2 upgrade includes a new inner tracking system and a new endcap calorimeter, as well as substantial improvements for most other subsystems of CMS. Upgrades are also in preparation in almost all other areas of CMS. In this paper, we present the various upgrades of the CMS detector since Run 1 that are designed to optimize the detector for sustained or improved performance at increased luminosity and energy.

The CMS detector has an overall length of 22 m, a diameter of 15 m, and weighs 14 000 tons. A schematic view is shown in figure 1. The detector is nearly hermetic, designed to trigger on [5, 6] and identify electrons, muons, photons, and (charged and neutral) hadrons [7–10]. The central feature of the CMS experiment is a superconducting solenoid of 6 m internal diameter and 12.5 m length that provides a magnetic field of 3.8 T with a stored energy of 2.2 GJ. Within the magnetic volume are a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL), each composed of a barrel and two endcap sections. Forward calorimeters extend the pseudorapidity (η) coverage provided by the barrel and endcap detectors. Muons are measured in gas-ionization detectors embedded in the steel flux-return yoke outside the solenoid.

Events of interest are selected using a two-tier trigger system. The first-level (L1) trigger is composed of custom hardware processors and uses information from the calorimeters and muon detectors to select events at a rate of about 100 kHz within a latency of about 4 μs [5]. The second level, known as the high-level trigger (HLT), consists of a cluster of commercial processors running a version of the full event reconstruction software optimized for fast processing. It was originally designed to reduce the event rate to around 1 kHz before data storage [6]. During Run 3, the L1 trigger and HLT operate at typical output rates of 110 kHz and 5 kHz, respectively.

A full description of the CMS detector, together with a definition of the coordinate system used and the relevant kinematic variables, is reported in ref. [1]. In the remaining part of this section, the CMS detector components are briefly introduced.

In section 2, the CMS solenoid magnet is described.

The inner tracking system (section 3) is used to measure the trajectories of charged particles produced in the collisions at the LHC. It is located in the innermost part of the CMS detector, closest to the interaction point. Prior to the Phase 1 upgrade, the pixel detector had three barrel layers and two disks in each endcap. In its current form, the pixel detector is composed of four barrel layers and three disks of silicon sensors on each side of the interaction point, with a total of 124 million readout channels. During LS2, the innermost barrel layer was replaced to ensure optimal performance until the end of Run 3. The strip tracker comprises ten layers of silicon strip sensors in the barrel, arranged in a cylindrical shape, and nine disks in the endcaps on each side of the detector. The strip sensors are segmented into long, thin strips, which are used to measure the trajectories of the particles and

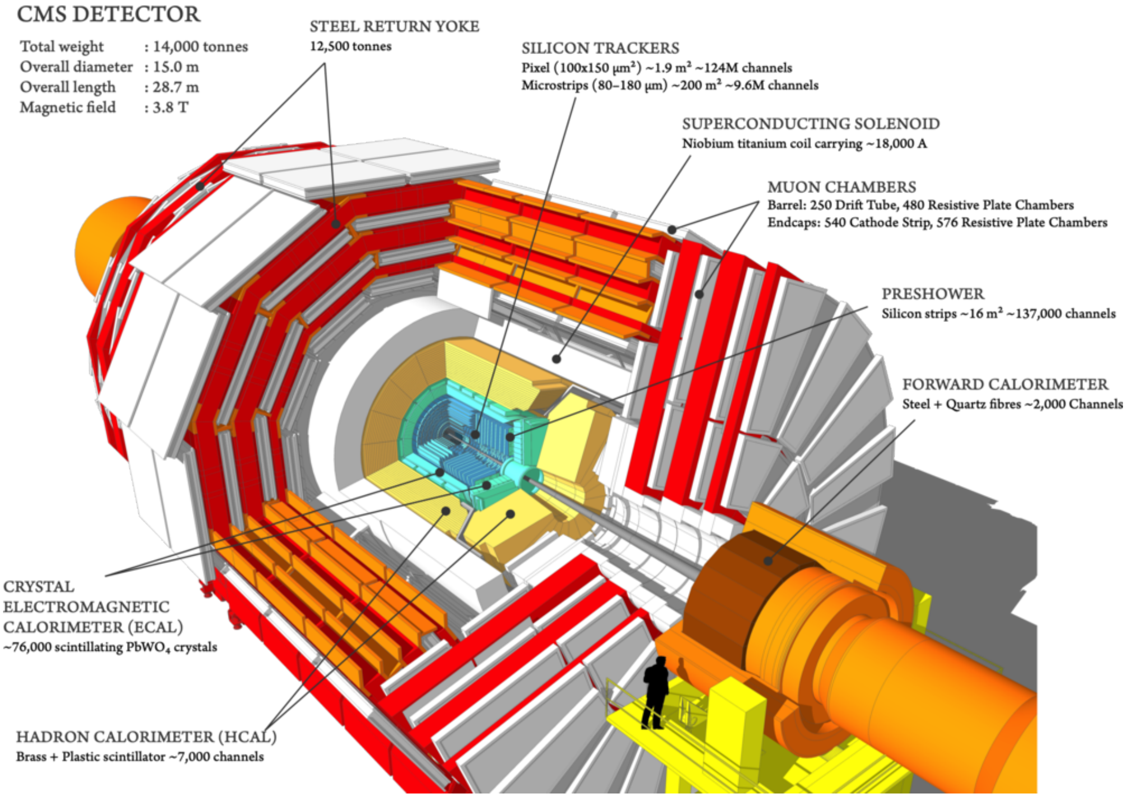


Figure 1. Schematic drawing of the CMS detector. Reproduced from [11]. CC BY 4.0.

provide a hit resolution of 20 μm for charged particles that cross the sensor perpendicularly. The tracker is designed to have excellent momentum resolution and tracking efficiency. It can detect and track particles with transverse momentum p_T as low as 50 MeV within the range $|\eta| < 2.5$. Tracks with a momentum around 100 GeV in the central region of the detector have an impact parameter resolution of about 10 μm, and a transverse momentum resolution near 1%.

The electromagnetic calorimeter (section 4) is made of 75 848 lead tungstate ($PbWO_4$) crystals: 61 200 crystals are located in the barrel (EB) and 7324 in each of the endcaps (EE) that provide a pseudorapidity coverage of $|\eta| < 3$. The lead tungstate crystals have a depth of about 23 cm, corresponding to about 25 radiation lengths X_0. Preshower detectors, consisting of two planes of silicon sensors interleaved with a total of $3X_0$ of lead, are located in front of each EE detector. The ECAL is designed to identify electrons and photons, and measure their positions and energies, which are reconstructed from energy deposits using algorithms that constrain the clusters to the size and shape expected for electrons or photons. The ECAL also contributes significantly to the reconstruction of jets and missing transverse momentum. The electron momentum is estimated by combining the energy measurement in the ECAL with the momentum measurement in the tracker. The momentum resolution for electrons with $p_T \approx 45$ GeV from $Z \to e^+e^-$ decays ranges from 1.6% to 5%. It is generally better in the barrel region than in the endcaps, and also depends on the bremsstrahlung energy emitted by the electron as it traverses the material in front of the ECAL. The ECAL also provides information on the arrival time of the electrons and photons which can be used in physics analyses, such as searches for long-lived particles.

The hadron calorimeter (section 5) is designed to measure the energy of charged and neutral hadrons. It contributes to the identification of hadrons and the measurement of their properties. It also aids in the reconstruction of jets and missing transverse momentum, and the identification of electrons and photons. The HCAL comprises four subdetectors: the hadron barrel (HB), hadron endcap (HE), hadron outer (HO), and hadron forward (HF) calorimeters. The HB and HE are located inside the solenoid magnet of the CMS detector and surround the ECAL. They cover the pseudorapidity ranges $|\eta| < 1.392$ and $1.305 < |\eta| < 3.0$, respectively, and are made of layers of brass plates interleaved with layers of scintillating tiles. The HF, constructed from steel and quartz fibers, is located outside the solenoid at ± 11.5 m from the collision point, and covers the $3.0 < |\eta| < 5.2$ range. Finally, the HO, made of plastic scintillator and the first layer of the barrel flux return covers the $|\eta| < 1.26$ range. The HCAL is designed to have a good hermeticity, with the ability to detect hadrons in nearly the full 4π solid angle. The Phase 1 upgrade of HCAL was installed in stages from 2016–2019. In HB and HE, the hybrid photodiode detectors were replaced with silicon photomultipliers, which reduced anomalous signals, improved radiation tolerance, and allowed for finer longitudinal readout segmentation. The HF photomultiplier tubes were also upgraded to reduce anomalous signals. The readout electronics were upgraded to support the increased channel count, improve the precision, and add signal timing information. When combining information from the entire CMS detector, the jet energy resolution typically amounts to 15–20% at 30 GeV, 10% at 100 GeV, and 5% at 1 TeV.

The muon detectors (section 6) are used in the identification of muons and the measurement of their momenta. The muon system comprises four subsystems: the drift tubes (DTs), the cathode strip chambers (CSCs), the resistive-plate chambers (RPCs), and the recently added gas electron multiplier (GEM) detector. Altogether, the CMS muon detectors comprise almost one million electronic channels.

The DTs consist of chambers formed by multiple layers of long rectangular tubes that are filled with an Ar and CO_2 gas mixture. An anode wire is located at the center of each tube, whereas cathode and field-shaping strips are positioned on its borders. They create an electric field that induces an almost uniform drift of ionization electrons produced by charged particles traversing the gas. The charged-particle trajectory is determined from the arrival time of the currents generated on the anode wires of the readout.

The CSCs are made of layers of proportional wire chambers with orthogonal cathode strips and are operated with a gas mixture of Ar, CO_2, and CF_4. Signals are generated on both anode wires and cathode strips. The finely segmented cathode strips and fast readout electronics provide good timing and spatial resolution to trigger on and identify muons.

The RPCs comprise two detecting layers of high-pressure laminate plates that are separated by a thin gap filled with a gas mixture of $C_2H_2F_4$, i-C_4H_{10}, and SF_6. The electronic readout strips are located between the two layers, and the high voltage is applied to high-conductivity electrodes coated on each plate. The detectors are operated in avalanche mode to cope with the high background rates. Due to their excellent time resolution, they ensure a precise bunch-crossing assignment for muons at the trigger level.

The key feature of the GEM is a foil consisting of a perforated insulating polymer surrounded on the top and bottom by conductors. A voltage difference is applied on the foils producing a strong electric field in the holes. The GEM is operated with a gas mixture of Ar and CO_2. When the gas volume is ionized, electrons are accelerated through the holes and read out on thinly separated strips.

2024 JINST 19 P05064

This structure allows for high amplification factors with modest voltages that provide good timing and spatial resolution, and can be operated at high rates.

The precision proton spectrometer (PPS) (section 7) is designed to detect protons scattered at very small angles in interactions where the protons remain intact and only a small fraction of their initial energy goes into the production of particles at small rapidity. In such events, the reconstruction of the kinematic properties of the protons uses their energy loss to determine the invariant mass of the system produced in the quasi-elastic collision. The PPS detector includes tracking and timing stations, which are located inside the LHC tunnel on both ends of the CMS detector about 200 m from the CMS interaction point. Precision tracking and timing is provided by silicon pixel and diamond detectors, respectively. The detectors are enclosed in movable stations, referred to as "roman pots", within which the detectors can be positioned as close as a few millimeters from the proton beam. The PPS first started in Run 2 as a joint project (CT-PPS) with the TOTEM Collaboration [12]. The initial PPS system consisted of two tracking and one timing station on each side. For Run 3, the PPS was upgraded for improved efficiency and precision with an additional timing station on each side.

The beam radiation instrumentation and luminosity (BRIL) system (section 8) comprises various detectors that measure the instantaneous luminosity and monitor in real time the beam-induced background, beam losses, and timing. Three luminosity detector systems provide robust bunch-by-bunch luminosity measurements in real time. They are: (i) the fast beam condition monitor (BCM1F), which counts hits in silicon pad diodes; (ii) the pixel luminosity telescope (PLT), which counts triple coincidences; and (iii) the HF calorimeter. The HF is instrumented with a dedicated readout for the real-time luminosity measurement and provides hit-tower counting (HFOC) and transverse-energy sums (HFET). The beam condition monitor for losses (BCML), using diamond and sapphire sensors, provides protection against catastrophic beam loss and is part of the LHC beam-abort system. The beam pickup timing device (BPTX) provides logical beam signals to the L1 trigger system. The BRIL system includes the measurement of radiation in the experimental cavern. The measurements are complemented by detailed simulations using the CMS radiation simulation applications.

The data acquisition (DAQ) system (section 9) is responsible for: the readout of all detector data for events accepted by the L1 trigger; the building of complete events from subdetector event fragments; the operation of the filter farm cluster running the HLT; and the transport of event data selected by it to the permanent storage in the Tier 0 computing center. The DAQ consists of: custom-built electronics reading out event fragments; a data-concentrator network transporting the fragments to the surface; a cluster of readout and event-building servers interconnected via the event-building network; the filter-farm cluster of multicore servers connected by the data network running the HLT software; a distributed storage system where event data selected by HLT filtering are buffered; and a transfer system connected to the Tier 0 center via a high-speed network. The DAQ also includes the trigger control and distribution system (TCDS), which distributes timing to the trigger and subdetector electronics, and implements trigger control logic as well as the trigger throttling system (TTS).

The L1 trigger (section 10) consists of electronics responsible for making a fast selection of events based on the presence of high-energy particles in the detector. The L1 trigger receives energy and position information, so-called trigger primitives (TPs), from the calorimeters and the muon detectors. The TPs are evaluated by a trigger processor, which is composed of custom-built electronics and field programmable gate array (FPGA) devices that perform the trigger decision based on a set of predefined trigger algorithms. The L1 trigger operates at trigger rates of about

2024 JINST 19 P05064

110 kHz. During LS2, the L1 trigger was upgraded to also process TPs that are designed to select long-lived particles.

The HLT (section 11) is a software-based system in which the full event information is used to select events of interest based on their physics content. The HLT is implemented as a parallel computing system that processes the event data in real time. Since the start of Run 3, the HLT makes use of graphical processing units (GPUs) in the trigger filter farm. The GPUs facilitate the offloading of specific parts of the reconstruction algorithms, e.g., tracking based on the pixel detector, as well as parts of the calorimeter reconstruction. The use of GPUs has led to a substantial reduction of the overall event processing time. With the performance improvements for Run 3, HLT-reconstructed analysis data, referred to as HLT scouting data, are recorded at a rate of about 30 kHz. In parallel, the storage rate of normal triggers was increased to about 2 kHz. Furthermore, the system also stores extra data samples, the so-called parking data sets, at a rate of around 3 kHz. The parking events will only be reconstructed by the offline computing infrastructure at a later time, when the resources will not be needed for the core activities. Other HLT reconstruction improvements in the areas of muon tracking, b jet tagging, and tau lepton reconstruction were also implemented for Run 3.

The offline computing system (section 12) has three key roles: to process the recorded data; to generate sufficiently large Monte Carlo simulation samples based on theoretical models and detector response modeling; and to facilitate physics data analysis performed at the CMS institutes around the world. The CMS data and simulation samples are stored and processed in a globally distributed network of centers, using an ever-growing array of heterogeneous resources. Continuous improvements are made in software and computing performance. Most notably, multithreaded processing and offloading to GPU resources have been introduced.

2 Solenoid magnet

The superconducting solenoid magnet provides a magnetic field of 3.8 T, and forms the center piece of the CMS experiment. A picture of the open CMS detector with visible magnet cryostat is shown in figure 2.

The original plan for the magnet during LS2 had been to turn it off, but to keep it cooled at 4 K. The plan was changed substantially because of a water leak in the experimental cavern inside a diffusion pump of the magnet cryostat discovered during a routine check. To intervene without putting the magnet at risk, it was decided to warm the magnet up to room temperature.

The procedure took place during the Covid-19 lockdown at CERN in April and May 2020. After an outgassing period, the vacuum volume was brought back to atmospheric pressure, and the diffusion pumps were removed and replaced. The vacuum circuit was also fully cleaned, and modifications were implemented for easier access and improved backup capabilities, with new valves and flanges allowing the connection of backup vacuum pumps if needed, as illustrated in the picture in figure 3 (left).

In parallel to this repair, the new free wheel thyristor (FWT) system [14] was installed on the powering circuit of the magnet, visible in figure 3 (right). The FWT bypasses the power converter in a closed loop in case the converter is in a faulty state, e.g., in the event of a power failure or a lack of cooling, thus avoiding a slow discharge to zero current. The FWT contributes to increasing the magnet's lifetime and the operational time at nominal field.

2024 JINST 19 P05064

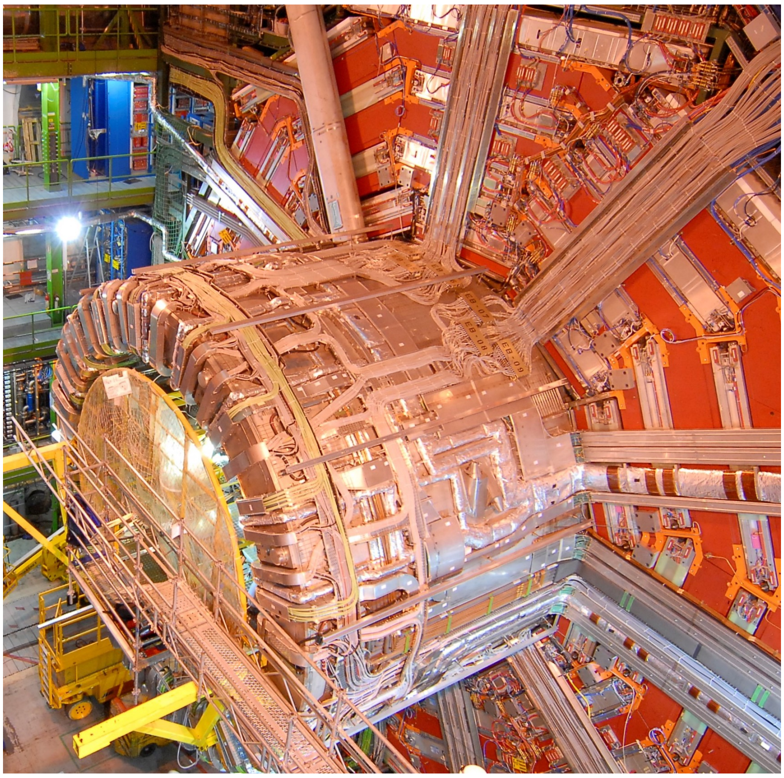


Figure 2. The solenoid magnet cryostat within the open CMS detector. Reproduced with permission from [13].

Figure 3. Photographs of a part of the new vacuum pumping circuit (left) and of the new CMS FWT system installed in the CMS service cavern (right).

As shown in figure 4, a full discharge followed by a ramp up takes about eight hours. A partial ramp down to 9.5 kA, corresponding to 2 T at the pp interaction point, is implemented in case of a cryogenics stop. The idle state of the power converter at constant current, indicated in figure 4 as a dashed line, is devised to provide the time needed for the reconnection of the cold box and refill of the liquid helium dewar, this way avoiding the risk of a fast discharge caused by helium flow fluctuations which may trigger a quench on the power leads and superconducting busbars. Also represented in figure 4 is the "free wheel" mode, when the FWT is triggered, typically after a power glitch, followed by an idle state period after the power converter restart and before the ramp up to nominal field can be resumed.

2024 JINST 19 P05064

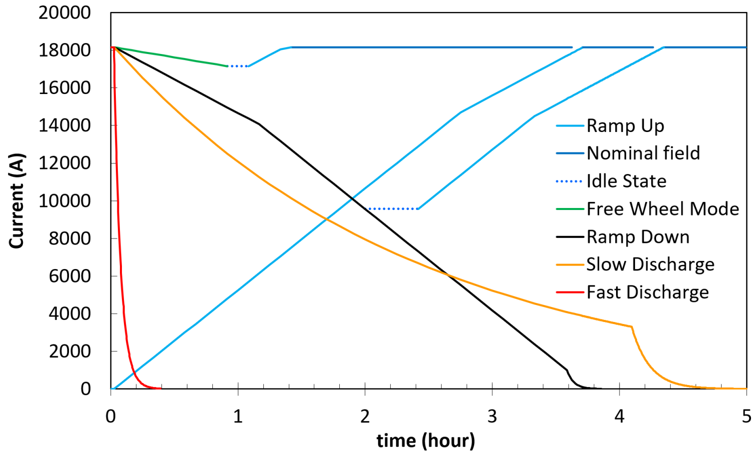


Figure 4. The CMS magnet current ramp and discharge modes representing the various magnet operation procedures and their duration. A current of 18164 A corresponds to a magnetic field of 3.8 T.

The control systems of the magnet and insulation vacuum were also fully upgraded during LS2. New programmable logic controllers were installed, new control electronics for both the FWT and the renewed vacuum pumping circuit were integrated, and the magnetic measurement system, using Hall probes, flux-loops, and nuclear magnetic resonance devices, was consolidated. The cryogenics system inside the cold box was improved by installation of a large filter with a reduced mesh to limit the recurrent clogging of the turbine filters as much as possible.

The magnet was successfully commissioned in September 2021, just ahead of the LHC pilot beam run, when it was operated at full magnetic field for two weeks with its upgraded powering system and repaired vacuum system. In March 2022, the magnet was ramped up to its nominal field of 3.8 T and declared ready for Run 3.

3 Inner tracking system

3.1 Pixel detector

The silicon pixel detector is the innermost part of the CMS inner tracking system. It provides three-dimensional space points close to the LHC collision point, which allow for high precision tracking and vertex reconstruction.

3.1.1 Detector design

The first CMS pixel detector [1], installed in 2008, consisted of three barrel layers at radii of 44, 73, and 102 mm and two endcap disks on each end at distances of 345 and 465 mm from the detector center. It provided three-point tracking for charged particles and performed very well during Run 1. However, already in Run 1 the instantaneous luminosity delivered by the LHC exceeded the design value of $1 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, which resulted in a pixel detector readout inefficiency. In order to maintain good tracking performance, this pixel detector was replaced with a more efficient and robust four-point tracking system. In addition, the radius of the beam pipe was reduced in 2014 from 30 to 23 mm, which allowed the innermost pixel layer to be placed closer to the interaction point. The improved pixel detector was installed at the beginning of 2017.

2024 JINST 19 P05064

The new detector, referred to as the Phase 1 pixel detector [15], consists of four barrel layers (L1–L4) at radii of 29, 68, 109, and 160 mm, and three disks (D1–D3) on each end at distances of 291, 396, and 516 mm from the center of the detector. The layouts of the two detectors, the original and the upgraded one, are compared in figure 5. The new layout provides four-hit coverage, instead of three, for tracks up to an absolute pseudorapidity of 3.0. The details of the Phase 1 pixel detector design and construction have been published in ref. [16].

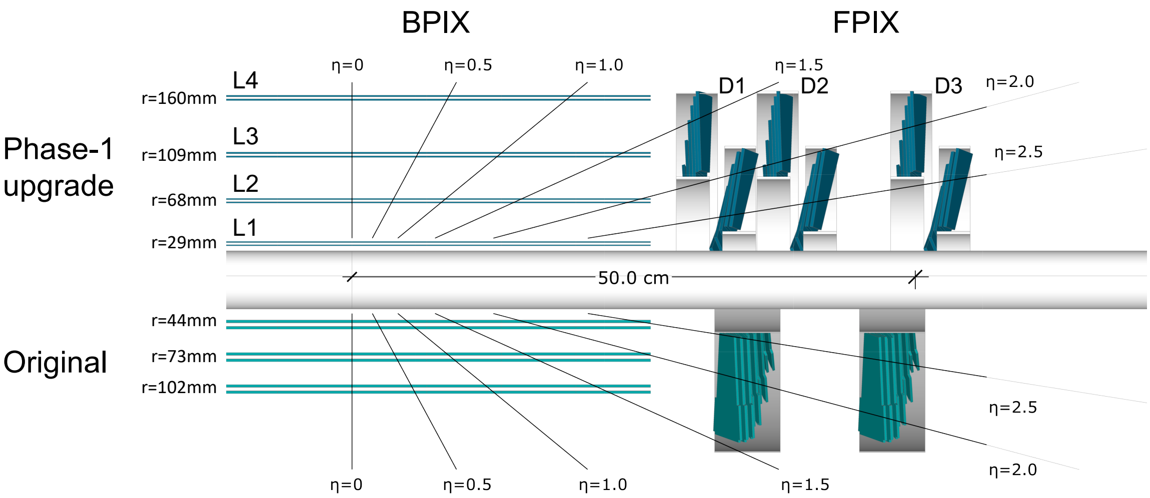


Figure 5. Longitudinal view of the Phase 1-upgraded pixel detector compared to the original detector layout. Reproduced from [16]. The Author(s). CC BY 4.0.

In addition to providing more tracking planes, the new detector contains several other improvements: DC-DC power converters are used to supply the necessary current while reusing the existing cables; a two-phase CO_2 cooling system is used; a 400 Mb/s digital readout system is implemented instead of the 40 MHz analog one; and the readout chips are modified to handle higher data rates.

Table 1 shows the main parameters of the Phase 1 pixel detector. The detector consists of two parts, the pixel barrel (BPIX) and the pixel forward disks (FPIX), which are mechanically and electrically independent. The BPIX detector consists of two 540 mm-long half-barrels divided in the y-z plane. The FPIX detector consists of twelve half-disks with radii ranging from 45 to 161 mm. Each half-disk is further divided into two rings of modules. The basic building block of both the BPIX and FPIX is a silicon sensor module comprised of a sensor with 160×416 pixels and a size of $100\times150\ \mu m^2$, connected to 16 readout chips (ROCs). In total, 1184 and 672 modules are used in the BPIX and FPIX, respectively. The entire Phase 1 pixel detector comprises a total of 124 million readout channels.

The sensor modules are mounted on light-weight mechanical structures, with thin-walled stainless steel tubes used for the CO_2 evaporative cooling (section 3.1.4). Carbon fiber and graphite materials with high thermal conductivity are incorporated into the detector mechanical structures (section 3.1.3). Both the BPIX and FPIX are connected to four service half-cylinders. They host the auxiliary electronics for readout and powering.

The installation of the Phase 1 pixel detector took place during the extended year-end technical stop of the LHC in 2016–2017. The new detector performed successfully during the Run 2 data-taking period in 2017–2018. However, since the detector installation in 2017 two major interventions

Table 1. Summary of the average radius and z position, as well as the number of modules for the four BPIX layers and six FPIX rings for the Phase 1 pixel detector.

BPIX			
Layer	Radius [mm]	z position [mm]	Number of modules
L1	29	−270 to +270	96
L2	68	−270 to +270	224
L3	109	−270 to +270	352
L4	160	−270 to +270	512

FPIX			
Disk	Radius [mm]	z position [mm]	Number of modules
D1 inner ring	45–110	±338	88
D1 outer ring	96–161	±309	136
D2 inner ring	45–110	±413	88
D2 outer ring	96–161	±384	136
D3 inner ring	45–110	±508	88
D3 outer ring	96–161	±479	136

took place. Due to an unexpected failure (section 3.1.4), all 1216 DC-DC converters had to be replaced in the LHC year-end technical stop 2017–2018. The second intervention was done during LS2 and involved the replacement of the layer-1 modules and, again, all the DC-DC converters. The new layer-1 modules, in addition to the planned replacement of the radiation-damaged silicon sensors, also included new versions of the readout ASICs, which fixed some of the shortcomings observed in the earlier versions, described in section 3.1.2. The LS2 was also used to repair several faulty optical fibers and bad power connections, upgrade the electronic boards in the FPIX system, and fix broken FPIX cooling inlets.

With the innermost layer placed at a radius of 29 mm from the beam line, the modules in this region are exposed to very high radiation doses and hit rates, as shown in table 2: the radiation fluence for L1 with 300 fb^{-1} is 2.2×10^{15} $\mathrm{n_{eq}/cm^2}$, corresponding to the operational limit of the installed system [16]. To ensure that the inner layer remains fully operational throughout all of Run 3, as planned from the beginning, the innermost BPIX layer was replaced during LS2 in 2019–2021.

3.1.2 Silicon modules

Schematic drawings of the Phase 1 pixel detector modules are shown in figure 6. A module consists of a $18.6 \times 66.6\,\mathrm{mm}^2$ silicon sensor that is bump-bonded to 2×8 ROCs. Each ROC has 80×52 rectangular pixels with a size of $100 \times 150\,\mu\mathrm{m}^2$, the same as in the original pixel detector. A high-density interconnect (HDI) flex printed circuit is glued to the sensor and wire-bonded to the 16 ROCs. A token bit manager chip (TBM) is mounted on top of the HDI (two TBMs in the case of L1 modules). The TBM controls the readout of a group of ROCs.

2024 JINST 19 P05064

Table 2. Expected hit rate, fluence, and radiation dose for the BPIX layers and FPIX rings. The hit rate corresponds to an instantaneous luminosity of $2.0 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ [16]. The fluence and radiation dose are shown for integrated luminosities of $300\,\mathrm{fb}^{-1}$ for the BPIX L1 and $500\,\mathrm{fb}^{-1}$ for the other BPIX layers and FPIX disks, well beyond the expected integrated luminosities for the detectors at the end of Run 3, of 250 and $370\,\mathrm{fb}^{-1}$, respectively.

	Pixel hit rate [$\mathrm{MHz/cm^2}$]	Fluence [$10^{15}\,\mathrm{n_{eq}/cm^2}$]	Dose [Mrad]
BPIX L1	580	2.2	100
BPIX L2	120	0.9	47
BPIX L3	58	0.4	22
BPIX L4	32	0.3	13
FPIX inner rings	56–260	0.4–2.0	21–106
FPIX outer rings	30– 75	0.3–0.5	13– 28

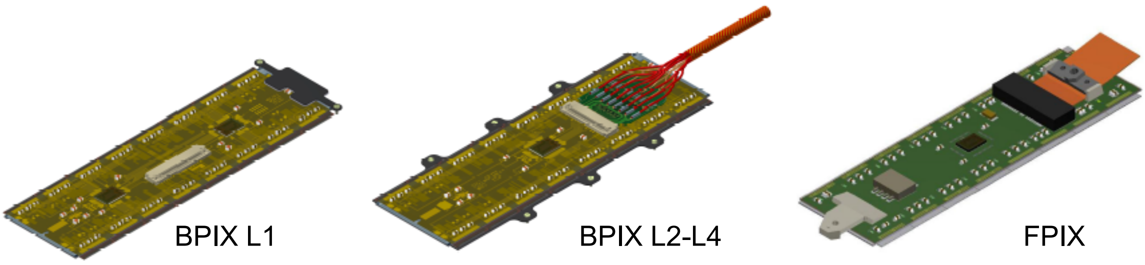


Figure 6. Drawings of the pixel detector modules for BPIX L1, BPIX L2–4, and the FPIX detector. Reproduced from [16]. The Author(s). CC BY 4.0.

Sensors. The silicon sensors are of the n-in-n type [16], with strongly n-doped (n^+) pixelated implants on an n-doped silicon bulk and a p-doped back side. The n^+ implants collect electrons, which has the advantage of being less affected by charge trapping caused by high irradiation [17, 18]. It is also advantageous as it allows sensors to be operated under-depleted after the detector is irradiated (when the so-called type inversion [16] is reached).

For the BPIX sensors, the n-side inter-pixel isolation was implemented through the moderated p-spray technique with a punch-through biasing grid. More details can be found in the references of ref. [16]. The sensors are made from approximately 285 μm-thick phosphorus-doped 4-inch wafers produced in a float-zone (FZ) process (fabricated at CiS). The FPIX sensors use open p-stops for n-side isolation, where each pixel is surrounded by an individual p-stop structure with an opening on one side. They were produced on 300 μm-thick 6-inch FZ-wafers (produced at Sintef).

The radiation resistance of a module is, to a large extent, defined by the possibility of increasing the sensor bias voltage to obtain a sufficiently high signal charge. As already mentioned in section 3.1.1, BPIX L1 modules have the highest radiation exposure and therefore need to be exchanged after approximately $250\,\mathrm{fb}^{-1}$.

During operation in 2017–2018, L1 modules were run with high efficiency at a bias voltage of 450 V, up to an integrated luminosity of almost $120\,\mathrm{fb}^{-1}$. After the replacement of the BPIX L1 during LS2, the new L1 modules must withstand a fluence expected to be about twice as high until the end of Run 3. In order to maintain a high enough signal charge, the pixel detector power supplies were upgraded to deliver a maximum voltage of 800 V during LS2, as described in section 3.1.4.

Readout chip. The upgraded ROCs, PSI46dig and PROC600 [16], are manufactured in the same 250 nm CMOS technology as the ROCs used in the original pixel detector (PSI46 [19]). Their design requirements are summarized in table 3.

Table 3. Parameters and design requirements for the PSI46dig and PROC600.

	PSI46dig	PROC600
Detector layer	BPIX L2–L4 and FPIX	BPIX L1
ROC size	$10.2 \times 7.9\,\mathrm{mm}^2$	$10.6 \times 7.9\,\mathrm{mm}^2$
Pixel size	$100 \times 150\,\mu\mathrm{m}^2$	$100 \times 150\,\mu\mathrm{m}^2$
Number of pixels	80×52	80×52
In-time threshold	$<2000\,e^-$	$<2000\,e^-$
Pixel hit loss	<2% at 150 MHz/cm^2	<3% at 580 MHz/cm^2
Readout speed	160 Mb/s	160 Mb/s
Maximum trigger latency	6.4 μs	6.4 μs
Radiation tolerance	120 Mrad	120 Mrad

The FPIX detector and layers 2–4 of the BPIX use the PSI46dig. Its design follows very closely the original ROC with the readout architecture based on the column-drain mechanism [20]. The pixel cell remains essentially unchanged except for the implementation of an improved charge discriminator. The improved discriminator reduces cross talk between pixels and the time walk of the signal [19] and thus leads to lower threshold operation (below $2000\,e^-$). The main modifications were made in the chip periphery in order to overcome the limitations of the PSI46 at high rate. They included a size increase of the data buffers (from 32 to 80 cells) and time-stamp buffers (from 12 to 24 cells) to store the hit information during the trigger latency, the implementation of an additional readout buffer stage to reduce dead time during the column readout, and the adoption of 160 Mb/s digital readout. In contrast to the previous ROC, the data readout is digital, using an 8-bit analog-to-digital converter (ADC) running at 80 MHz. Digitized data are stored in a 64×23 buffer, which is read out serially at 160 MHz.

The PROC600 is used for BPIX layer 1 and has to cope with hit rates of up to 600 MHz/cm^2. Therefore, the data transfer of pixel hits to the periphery must be much faster than the PSI46dig. This was achieved by a complete redesign of the double column architecture. The pixels within a double column are dynamically grouped into clusters of four and read out simultaneously, which significantly speeds up the readout process.

During operation in 2017 and 2018, the PROC600 delivered high-quality data. However, two shortcomings of the PROC600 were identified, and are discussed in more detail in ref. [16]. The first was cross talk between pixels, which was higher than expected and generated noise at high hit rates. The second was a lower efficiency caused by a rare loss of data synchronization in double columns. Therefore, it was decided to develop for Run 3 a new, revised version of the PROC600. The main change in the new version addresses the rare cases of data synchronization loss. The issue was tracked to a timing error in the time-stamp buffer of the double column, which leads to

2024 JINST 19 P05064

inefficiencies at low and high hit rates. It was corrected in the buffer logic of the revised PROC600. The higher-than-expected noise was traced to inappropriate shielding of the circuitry for calibration pulse injection and has been fixed in the revised version of the PROC600. In addition, the routing and shielding of power and address lines were improved. Both changes led to lower noise and lower cross talk between pixels. The revised PROC600 was used to construct the modules for the new L1, which was installed during LS2 for use in Run 3.

Token bit manager chip. The TBM is a custom, mixed-mode, radiation-hard integrated circuit that controls and reads out a group of 8 (L1) or 16 (L2–L4, FPIX) ROCs. To increase the data output bandwidth from a module, two 160 Mb/s ROC signal paths, with one path inverted, are multiplexed into a 320 Mb/s signal, and then encoded into a 400 Mb/s data stream that is optically transmitted to the downstream DAQ system. The TBM has a single output (TBM08) version for L3, L4, and FPIX, and dual output versions, which are used in L2 (TBM09) and L1 (TBM10). The TBM08 version has two independent 160 Mb/s ROC readout paths, and the TBM09 and TBM10 versions have four separate, semi-independent, 160 Mb/s ROC readout paths.

In addition to the increased output bandwidth, several critical features were added to the TBM for Phase 1. As a result of adopting a faster digital readout, finer control over the timing of internal TBM operations and external TBM inputs was needed: delay adjustments were added for the ROC readouts, the token outputs, the data headers and the data trailers, and relative phase adjustments were added between the 40 MHz incoming clock, the 160 MHz clock, and the 400 MHz clock. To prevent very long readouts from blocking the DAQ system, an adjustable token timeout was added that can reset the ROCs and drain buffered data.

Operation of the TBM during collision data taking revealed a vulnerability to a particular single event upset (SEU) that halts the TBM and requires a power cycle of the TBM to recover. An additional iteration of the TBM chips was designed in the spring of 2018 to address this TBM SEU issue and to add an adjustable delay of up to 32 ns to the 40 MHz clock. The delay was added to allow finer adjustment of the relative timing between modules. The new chips are used in the new modules in BPIX L1, incorporated during the consolidation work in LS2.

BPIX module construction. The BPIX detector contains 1184 modules, all having a similar design but coming in three different flavors depending on the requirements of the different layers. The three module types are summarized in table 4.

Table 4. Overview of module types used in the Phase 1 pixel detector.

	ROC	Number of ROCs	TBM	Number of TBMs	Number of 400 Mb/s readout links/module
BPIX L1	PROC600	16	TBM10	2	4
BPIX L2	PSI46dig	16	TBM09	1	2
BPIX L3, L4, & FPIX	PSI46dig	16	TBM08	1	1

A drawing of the detector module cross section for BPIX L2–4 is shown in figure 7. During the module production process, bare modules are made by bump bonding of the 16 ROCs to the sensor. In the next step, the thin four-layer HDI is glued onto the sensor side of the bare module and

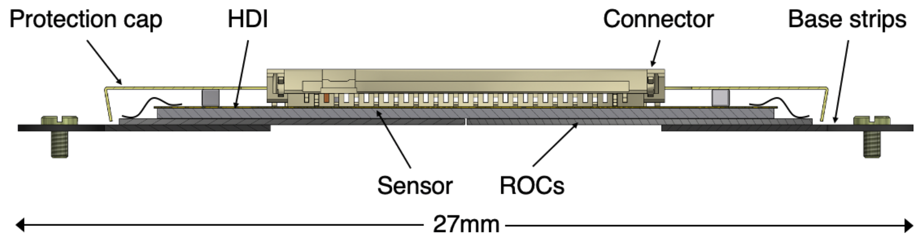


Figure 7. Cross sectional view of a pixel detector module for BPIX L2–4, cut along the short side of the module. Reproduced from [16]. The Author(s). CC BY 4.0.

wire-bonded to the ROCs. The TBMs are already mounted, wired-bonded, and tested on the HDI before joining the HDI to the bare module. The BPIX L2–4 modules are mounted using silicon nitride base strips that are glued under the ROCs on the two long sides of the module (figure 6, center). The tight space requirements in the innermost BPIX layer required a different scheme with clamps between two modules (figure 6, left). A cap (not shown in the figure) made from a 75 μm-thick polyimide foil protects the wire bonds of the ROCs and TBMs against mechanical damage from the cables of other modules when mounted on the BPIX mechanics. A single L2 module, excluding the cable, has a mass of about 2.4 g and represents a thickness corresponding to 0.8% of a radiation length at normal incident angle.

The production of the BPIX modules was shared by five different module-assembly centers operated by groups from Germany, Switzerland, Italy, Finland, and Taiwan. The assembly tools and procedures were mostly standardized, however, each center used different bump-bonding techniques, as described in more detail in ref. [16]. The overall yield of the module assembly ranged from 65 to 85% in the different production centers, a bit lower than expected. The main reason was low-yield phases during the production start-up.

FPIX module construction. The 672 FPIX modules installed in the detector all have the same design and are instrumented with PSI46dig ROCs and a TBM08 with one readout link per module. The FPIX modules (2.9 g/module without cable) use a different sensor, HDI, and aluminum/polyimide flat flex cable and are not interchangeable with BPIX modules. The bump-bonding procedure was done at RTI and used an automated Datacon APM2200 bump bonder. Bare modules were sent from RTI to two FPIX module assembly and testing sites in the U.S.A. Modules were assembled using an automated gantry and pick and placement equipment. The wire bonds were encapsulated to protect against humidity and to mechanically support the wire bonds. The average yield for FPIX production modules was 68%, largely limited by bare module quality, especially in the early production.

3.1.3 Mechanics

BPIX mechanics. The detector mechanics have been described in detail in ref. [16], and a brief outline will be given here. The BPIX detector modules are mounted on ladders, with each ladder supporting eight modules. The ladders are mounted on four concentric layers split into half-cylinders. These are staggered by an alternating arrangement of ladders at smaller and larger radii. Such a ladder placement provides between 0.5 and 1.0 mm of sensor overlap in the r-ϕ plane. Ladders are made from carbon-fiber reinforced polymer (CFRP) and have a length of 540 mm and a thickness of 500 μm. The end rings, on which the ladders are suspended, consist of a CFRP/Airex/CFRP sandwich structure.

The 1.7 m-long BPIX service half-cylinders contain the DC-DC converters, the opto-electronic components, and the module connections. The modules are connected to the service cylinders with micro-twisted-pair copper cables with a length of about 1 m. These are contained in an additional structure placed in between the BPIX detector mechanics and the service half-cylinder.

The BPIX cooling is provided by complex looping tubes that cool the detector modules and the components on the service half-cylinders. The tubes are made out of stainless steel with an inner diameter of 1.7 mm and a wall thickness of 50 μm. The loops are between 957 and 1225 cm long and dissipate a power of up to 240 W each. On the service half-cylinders the tubes have a wall thickness of 200 μm and an inner diameter of 1.8 and 2.6 mm for supply and return lines, respectively.

FPIX mechanics. The three half-disks forming one FPIX quadrant are supported by a carbon-fiber composite service half-cylinder. The FPIX half-disks consist of two turbine-like mechanical support structures with an inner assembly providing a sensor coverage from radii of 45 to 110 mm with 11 blades, while the outer assembly covers radii from 96 to 161 mm with 17 blades (figure 8). The half-disks serve as the cooling isotherms for the sensor modules. One module is mounted on each side of a blade, such that there is a small overlap in coverage at the outer edge of the blade. The overlap is larger for adjacent modules closer to the beam. The flat panel blades are made of 0.6 mm-thick sheets of thermal pyrolytic graphite (TPG) encapsulated between two 70 μm-thick single-ply carbon fiber face sheets. The blades are suspended between an inner and an outer 2.4 mm thick graphite ring. Each graphite ring houses the stainless steel cooling lines and is reinforced on the side facing away from the blades with a carbon fiber skin.

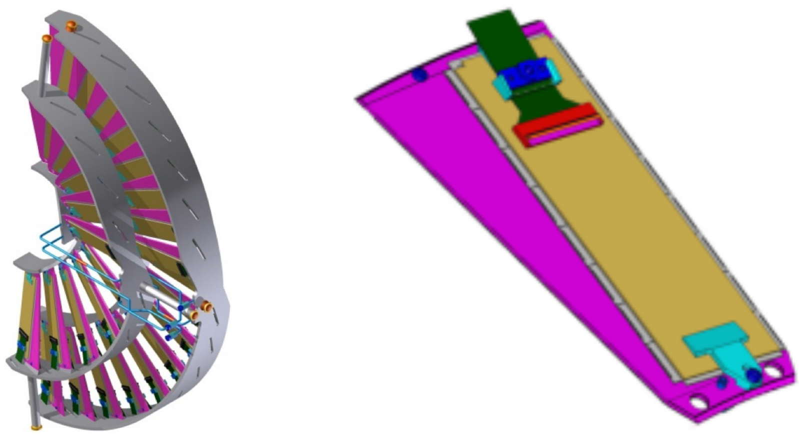


Figure 8. Drawing of an FPIX half-disk made from two half-rings of modules mounted on blades that are suspended between graphite rings (left), and close up view of a module mounted on a blade (right). Reproduced from [16]. The Author(s). CC BY 4.0.

The FPIX service half-cylinders are 2220 mm long, with an outer radius of 175 mm and an inner radius of 165 mm. The service half-cylinders consist of a double-walled carbon fiber rear section that is 1500 mm long for support elements, a corrugated single-wall carbon fiber front section that houses the half-disks, and an aluminum end flange for the cable and cooling tube pass through. A ruby ball sliding in aluminum bushings on each of the three half-disk support stalks provides a kinematic mount. A detailed description of the FPIX detector mechanics can be found in ref. [16].

2024 JINST 19 P05064

3.1.4 Services

Readout architecture and data acquisition system. In order to be efficient, the CMS Phase 1 pixel detector readout architecture was copied from the original detector, with modifications made to handle the 400 Mb/s readout and the increased number of channels. The entirety of the CMS Phase 1 system readout is detailed in ref. [21], and a summary is provided below.

The auxiliary on-detector electronics were modified to include a crystal-driven phase-locked-loop chip (QPLL) [22] that reduced the jitter on the clock signal, a higher bandwidth chip (DLT) [23] to translate the module readout into a level suitable for the laser driver chip, and a new optical transmitter (TOSA from Mitsubishi Electric Corp.) for the higher bandwidth digital signal.

The off-detector VME-based DAQ system used for the original detector was replaced by a μTCA-based system using a common carrier board (FC7) [24] with conventional 10 Gb/s optical network links for the transfer of the data to the CMS central DAQ system. The DAQ system, used to control and read out the full pixel detector, consists of 108 frontend driver modules (FEDs), which receive and decode the pixel hit information; three (2017–2018) to six (≥2021) frontend controller modules (TkFECs), which serve the detector slow control; 16 pixel frontend controller modules (PxFECs) used for module programming and clock and trigger distribution; and 12 AMC13 [25] cards providing the clock and trigger signals. Application-specific mezzanine cards and firmware make an FC7 carrier board into a FED or FEC.

Power system. The Phase 1 pixel detector requires low voltages to supply the readout chips on the pixel modules, a bias voltage to deplete the pixel sensors, and voltages to supply the control electronics components on the supply tubes and service cylinders.

The nominal control voltage of 2.5 V is supplied by A4602 CAEN power supply modules. Power supplies by CAEN of type A4603DH provide the low and high voltages to the pixel modules. The maximum bias voltage that can be delivered was raised from 600 to 800 V during LS2. Each bias voltage channel can provide up to 16 mA, and serves several pixel modules. While the auxiliary and bias voltage supply systems have conceptually not been changed for Phase 1, the low voltage supply system was changed from a direct parallel powering system to a DC-DC conversion powering system. The number of ROCs, and thus the current consumption of the detector, has roughly doubled as compared to the original pixel detector. Thanks to DC-DC conversion factors of 3–4, the currents on the typically 50 m-long supply lines between the detector and the power supplies are reduced with respect to a direct powering scheme.

The DC-DC converter modules are built around the CERN radiation-tolerant FEAST buck converter ASIC [26], and were optimized in terms of dimensions, mass, and performance for the application in the pixel detector [27, 28]. The DC-DC converter modules are installed on the BPIX supply tubes and in the FPIX service cylinders, at a distance of about 1 m from the pixel modules and outside of the sensitive pixel volume. A total of 1216 DC-DC converter modules are used in the pixel detector.

Two types of DC-DC converter modules are needed to supply the pixel modules: one delivering 2.4 V to the analog circuitry of the ROC, and one delivering 3.3 or 3.5 V (depending on the position of the served pixel module in the detector) to the digital circuitry of the ROC and to the TBMs. Each such pair of DC-DC converter modules serves between 1 and 4 pixel modules, depending on the layer and ring of the pixel modules. The DC-DC converter output currents range from 0.4 to 1.7 A

2024 JINST 19 P05064

for the analog supply, and from 1.3 to 2.4 A for the digital supply. The power efficiency is 80–84%, depending on the output voltage and load. The DC-DC converters can be disabled and enabled from an already existing chip (CCU) [29] used both in the original and in the Phase 1 pixel detector; when disabled the DC-DC converters do not provide an output voltage. This feature was used during 2017 to power-cycle the TBM chips suffering from a SEU, which could not be recovered otherwise.

Up to seven DC-DC converters (of one variant) are connected to one low voltage power supply channel. The low voltage part of the original A4603 power supplies has been adapted to the DC-DC conversion powering scheme. The maximum output voltage was raised to 12.5 V and the fast remote sensing was abandoned, while a slow-control loop to compensate for voltage drops along the supply cables is still available.

In general, the DC-DC conversion powering system worked very well. However, starting in October 2017, DC-DC converters started to fail, and at the end of the 2017 data-taking period, about 5% of the DC-DC converters were defective. This was traced back to a problem in the FEAST ASIC, namely a radiation-induced leakage current in a transistor [30]. When the chip is disabled, a voltage above the chip specification can build up on a certain node, damaging the chip. For the 2018 data taking, all DC-DC converters were replaced with (almost) identical ones. The input voltage was reduced from about 11 to 9 V, and disabling of the chip was replaced by power-cycling of the power supplies. Due to these operational changes, no DC-DC converter failed in 2018. During LS2, new DC-DC converters were again produced, featuring a new version of the DC-DC ASIC (FEAST2.3). In this new version, the problem is fixed and operational changes are no longer needed.

The use of DC-DC converters meant that the LV and HV modularity of the detector no longer matched. The LV could be switched off by disabling a DC-DC converter, however, the HV stayed on because of the limited number of HV wires. The failure of individual DC-DC converters affected a small number of pixel modules that were kept under bias voltage in 2017 while their LV was off. Having the sensor biased but the pixel readout amplifiers off damaged the amplifiers in the affected pixel modules. Several of those damaged modules were replaced during LS2. In addition, the LV/HV modularity was harmonized in FPIX; for BPIX this was not possible due to the limited number of HV wires.

Cooling. To keep the silicon sensors below 0°C and remove heat from the other detector elements, CO_2 evaporative cooling utilizing the two-phase accumulator controlled loop (2-PACL) approach [31] is used in the Phase 1 pixel detector. Evaporative CO_2 cooling provides low density, low viscosity, and high heat transfer capacity while allowing the use of an all passive, small-diameter, thin-walled stainless steel pipe network inside the detector volume. This results in a lower contribution of cooling to the overall detector material budget. During normal operation (−22°C) the expected power [15] from the BPIX (6 kW) and FPIX detectors (3 kW) is removed using two dedicated 15 kW CO_2 plants, one for each detector, located in the CMS detector cavern. There is enough flexibility and capacity in the system so that the eight cooling loops in each detector can be connected to either cooling plant with no loss of cooling capacity. The typical temperature at the pixel module surface is about 12°C higher than the coolant. A detailed description of the CMS Phase 1 pixel detector cooling system can be found in ref. [16].

During the testing of the removed FPIX in 2019, one of the inlet cooling pipe connections at the detector end flange was broken. It was therefore decided to replace all the FPIX inlet connectors with a more robust solution that required only a single wrench to make a connection. This sped up the detector installation in 2021 and prevented another occurrence of a broken inlet connector.

3.1.5 Detector operation

Detector live fraction. The detector live fraction, defined as the fraction of working ROCs, was 95.0 and 96.1% for the BPIX and FPIX detectors, respectively, during the first collisions in 2017. The main causes of failures in the BPIX detector were the loss of power due to faulty connectors and modules masked because of readout problems. For the FPIX detector the main problem was an issue with the clock distribution in one sector. Towards the end of 2017, the fraction of nonworking modules was dominated by the failure of the DC-DC converters described in section 3.1.4. The DC-DC problems lowered the working fraction to 90.9 and 85.0% for the BPIX and FPIX detectors, respectively.

During the 2017/2018 LHC year-end technical stop, faulty components were repaired and all DC-DC converters were replaced. A few broken BPIX modules, which were accessible without disassembling the detector layers, were also replaced. These repairs improved the working detector fraction for the 2018 data-taking period. For the BPIX it varied from initially 98 to 93.5% at the year end, with the main drop due to faulty power connectors. The FPIX detector working fraction was stable throughout the entire year at 96.7%.

As already mentioned above, during LS2 a new L1 was installed in the BPIX. Other repairs, in the BPIX and FPIX, involved faulty infrastructure components like bad connectors and broken optical fibers. In addition, eight modules were replaced in BPIX L2. With these improvements, the working fractions for BPIX and FPIX at the start of Run 3 were 99.1 and 98.5%, respectively.

Threshold adjustment. Pixel charge thresholds are an important performance parameter since they directly influence the position resolution. Lower thresholds increase the charge sharing between pixels, resulting in a better resolution. However, too low thresholds result in noise saturating the readout. Therefore, thresholds in all ROCs are adjusted to the lowest possible value, but well above the noise level itself. More details about the threshold adjustment are given in ref. [16].

During the 2017-18 data-taking period, the BPIX L2–4 thresholds were about 1400 e^- and similarly for the FPIX detector at about 1500 e^-. Because of the higher noise in BPIX L1, the thresholds had to be higher, about 2200 e^-. With these pixel thresholds the number of noise hits was very low, below 10 pixels per bunch crossing per layer, resulting in a per pixel noise hit probability of less than 10^{-6}. Individual pixels that showed a hit probability exceeding 0.1% were masked during operation. The total fraction of masked pixels was less than 0.01%.

The revised PROC600 used in the new L1 installed in 2021 has significantly lower cross-talk noise (as discussed in section 3.1.2). The expected threshold, noise, and time-walk behavior of the revised PROC600 version is similar to that of the PSI46dig. These improvements allow the BPIX L1 to be operated in Run 3 with significantly lower thresholds, similar to the ones for the other layers.

3.1.6 Performance of the pixel tracker

The two most important performance parameters for a pixel detector are hit efficiency and position resolution. Both strongly affect the ability to perform pattern recognition and b tagging, two main roles for a well-functioning pixel detector.

2024 JINST 19 P05064

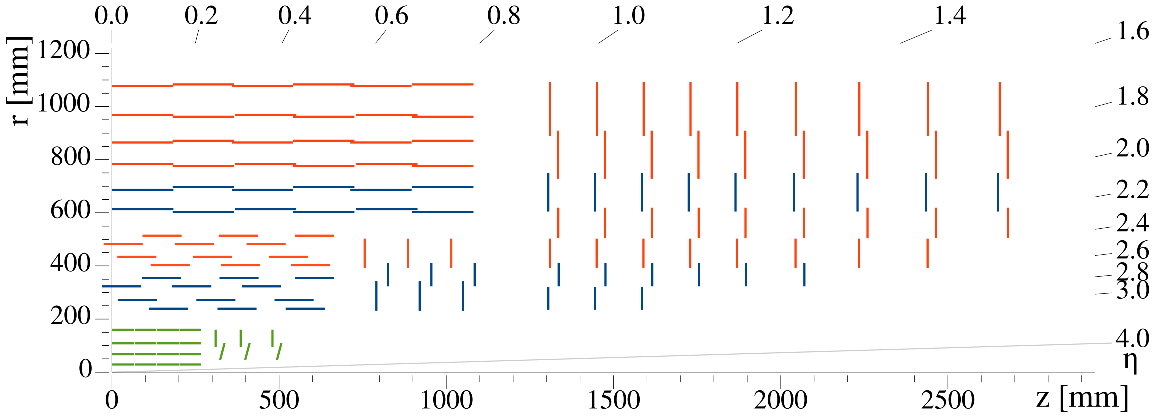


Figure 9. Schematic view of one quadrant in the r-z view of the CMS tracker: single-sided and double-sided strip modules are depicted as red and blue segments, respectively. The pixel detector is shown in green. Reproduced from [32]. CC BY 4.0.

The hit efficiency has greatly improved with the Phase 1 upgrade, mostly due to the ability of the upgraded ROCs to read data at higher speed and operate at lower thresholds, as described in section 3.1.2 and ref. [16]. The hit efficiencies measured at an instantaneous luminosity of $2 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ are 97, 98, 99, and 99.5% for BPIX L1–L4, respectively. For the forward disks, the average efficiency is 99%. These numbers present a big improvement with respect to the Phase 0 detector, where the efficiencies would have been much smaller for the same data rates [16].

The position resolution is measured using the "triplet" method [16], where the expected hit position in a detector layer is interpolated from two other layers. For the BPIX L3, this method yields a position resolution of 11.0 μm in the r-ϕ direction and 24.3 μm in the z direction. For BPIX L1 and L2, the resolutions are somewhat worse due to the higher sensor radiation damage at lower radii. For the forward disks, the resolution is 11.9 μm in the r direction and 21.0 μm in the z direction. These numbers agree with the design expectations and are consistent with simulations.

Pixel sensors suffer from radiation damage induced by the high density of charge particle tracks. In order to maintain good efficiency and position resolution, it is necessary to keep increasing the sensor's bias voltage. The nominal voltage after installation was 150 V for the BPIX modules and 300 V for FPIX. Especially for the inner BPIX layers, these voltages had to be increased in a few steps during data taking. Presently, end of 2023, the voltages are 450, 350, 250, and 250 V for the BPIX L1–L4, respectively, and 350 and 300 V for the inner and outer rings of the FPIX disks. The values are expected to increase further throughout Run 3, eventually reaching the maximum of 800 V for BPIX L1 and 600 V for all other layers.

3.2 Strip detector

3.2.1 Detector description

The silicon strip tracker (SST), together with the pixel detector, provides measurements of charged particle trajectories up to a pseudorapidity of $|\eta| < 2.5$. The layout of the SST is shown in figure 9.

The SST has 9.3 million silicon micro-strips and 198 m^2 of active silicon area distributed over 15 148 modules. Single-sided p-on-n micro-strip sensors are used. The detector is 5 m long and has a diameter of 2.5 m. It has ten layers in the barrel region with four layers in the tracker inner barrel (TIB) and six layers in the tracker outer barrel (TOB). The TIB is supplemented with three tracker inner disks (TID) at each end. In the forward regions, the detector consists of tracker endcaps (TEC). Each TID is composed of three rings of modules and each TEC is composed of up to seven rings. In TIB, TID, and in rings 1–4 of the TECs, sensors with a thickness of 320 μm are used, while in TOB and in rings 5–7 of the TECs, 500 μm thick sensors are used. The modules in the barrel layers measure r and ϕ coordinates, while the modules in the TECs and TIDs are oriented to measure the coordinates in ϕ and z. In four layers in the barrel and three rings in the endcaps, stereo modules are used (figure 9). These modules have a second module mounted back-to-back with a stereo angle of 100 mrad. The stereo modules provide coarse measurements of an additional coordinate (z in the barrel and r in the endcaps).

The analog signals from 128 strips are processed by one APV25 chip. The chip has 128 readout channels, each consisting of a low-noise and charge-sensitive preamplifier, a 50 ns CR-RC type shaper, and a 192-element deep analog pipeline which samples the shaped signals at the LHC frequency of 40 MHz [33]. Signals from two APV25 chips are multiplexed, converted to optical signals by analog opto-hybrids (AOH) [1], and transmitted via optical fibers to front-end drivers (FED), located in the service cavern outside the radiation zone. Pedestal and common mode subtraction, as well as cluster finding, are performed in the FEDs. Clock, trigger information, and control signal are trasmitted to the detector by the frontend controllers (FEC), also located in the service cavern. Configuration data for the modules is distributed via the I2C protocol to communication-and-control units (CCU) [29], grouped in token ring networks (control rings). The modules in the SST are grouped in power groups each of which shares one power supply channel. There are 1944 power groups in total. Each power group has two low-voltage channels with 2.5 and 1.25 V regulators and two high-voltage channels that can be regulated up to 600 V [1]. The detector is cooled with C_6F_{14} monophase coolant by two cooling plants.

The SST has been operated stably and successfully since 2009, and operation is scheduled to continue until the end of Run 3. During Run 1, the SST was operated with its primary cooling at +4°C, significantly above the designed operating temperature, due to insufficient humidity control in the service channels and in the bulkhead region, i.e., the interface region between the detector volume and the outside seal. In 2009, the detector suffered from an over-pressure incident. Both inlet and outlet lines of 90 cooling loops of circulating C_6F_{14} were closed on of the two cooling plants. After that the detector warmed up. As a result of the over-pressure, some of the cooling lines developed leaks or were detached from modules. Due to this incident there are several regions in the detector that have closed cooling loops or degraded cooling contacts.

During LS1 in 2013–2014, a number of engineering changes were carried out on the detector infrastructure that allowed to lower the operating temperature of the SST below 0°C. Most prominently, a dedicated plant was installed to produce dry air or oxygen-depleted air. The plant has a flow of about 250 m^3/h, and is able to meet the dew point requirement of the SST (around −60°C). The plant is the primary source of dry gas injection to the detector and its services. In LS1, the insulation of all service channels and the bulkhead of the detector was significantly improved in order to control the humidity conditions in the closed volume of the detector.

All these modifications to the detector infrastructure facilitated the SST operation at −15°C since the beginning of Run 2. However, with increasing irradiation, the leakage currents began to approach the power supply limits (12 mA) in the regions with no cooling or degraded cooling contacts. Thermal runaway was observed in several power groups of the TIB. As a consequence, since 2018, the SST was operated at -20°C, which sufficiently reduced the leakage currents. By the end of Run 3 it will be necessary to lower the operating temperature to -25°C. During LS2 a test at a temperature of -25°C confirmed that the detector can be operated at this temperature and that there is no degradation of the humidity conditions inside the detector and the service channels.

3.2.2 Performance of the strip tracker

The performance of the SST will be discussed in the following. More details about the results in this section can be found in ref. [34].

Throughout all the years of operation, no SST on-detector components were exchanged because the detector has been inaccessible. The fraction of bad detector components has been largely stable during Run 1 and Run 2. This includes the readout channels that are excluded from the data taking: failing control rings, problems in LV or HV distribution, and individually switched-off modules, single APV25 chips or groups of strips. As can be seen in figure 10, the fraction of bad components was stable throughout Run 2 and amounts to about 4%.

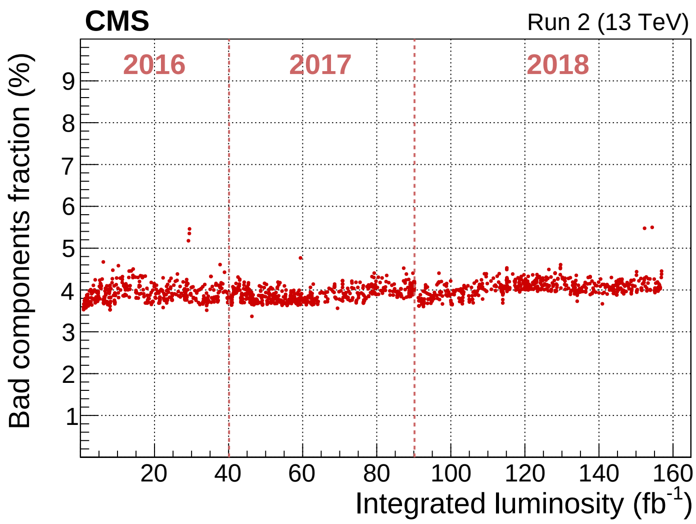


Figure 10. Fraction of bad components for the CMS silicon strip detector as a function of the delivered integrated luminosity. Reproduced with permission from [34].

One of the most important performance characteristics is the signal-to-noise ratio (S/N). The evolution of S/N with accumulated integrated luminosity is shown in figure 11 (left). As expected from irradiation studies [1], the S/N degrades approximately linearly with the integrated luminosity [34]. The decrease observed during Run 2 indicates that the SST will continue to provide high-quality data until its end of life, estimated to be at 500 fb^{-1}, well beyond the expected end of Run 3.

Another important aspect of the SST is the hit efficiency, which is the detection efficiency for a particle traversing a sensor. The measurement of the hit efficiency is performed using tracks that pass the quality criteria as defined in ref. [35]. In order to avoid inactive regions, trajectories that are close to sensor edges or their readout electronics in the studied layer are not considered. The efficiency is determined from the fraction of traversing tracks with a hit in a module anywhere within a range of 15 strips from the expected position. The measured hit efficiency under typical conditions

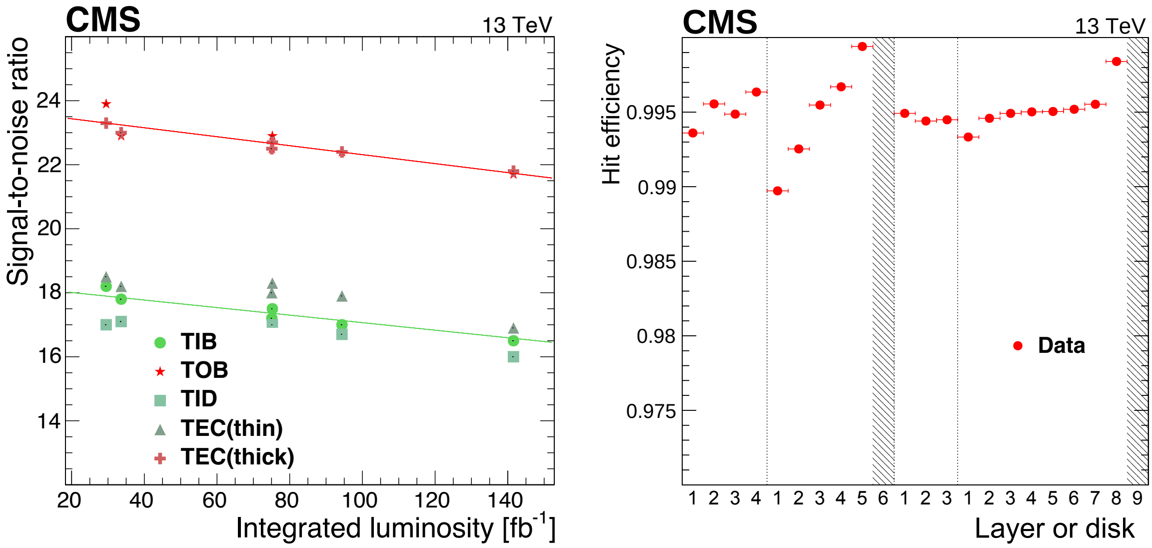


Figure 11. Left: signal-to-noise ratio as a function of integrated luminosity as accumulated in pp collisions during Run 2, separately for the different detector partitions. Triangles and crosses indicate the results for sensors in the TEC of thickness 320 and 500 μm, respectively. Right: hit efficiency of the silicon strip detector taken from a representative run recorded in 2018 [34] with an average hit efficiency under typical conditions at an instantaneous luminosity of $1.11 \times 10^{34}\,\mathrm{cm^{-2}\,s^{-1}}$. The two gray bands represent regions where the hit efficiency is not measured due to the selection criteria of the analysis [34].

during Run 2, at an average instantaneous luminosity of $1.11 \times 10^{34}\,\mathrm{cm^{-2}\,s^{-1}}$, corresponding to about 31 pp interactions per bunch crossing, is shown in figure 11 (right). The average hit efficiency is about 99.5%, depending on the layer. Since the inefficiency mainly depends on the particle flux, the inner layers have a somewhat lower efficiency than the outer ones. Moreover, the inefficiency depends on the sensor thickness and on the pitch.

Radiation effects are also monitored during the operation of the detector, including the increase of the leakage currents in the sensors, the evolution of the full depletion voltage due to the change of the effective sensor doping concentration, and the evolution of the laser-driver performance in the optical readout chain.

During Run 3, due to increasing luminosity, the leakage currents will continue to rise. It can therefore be expected that some modules in regions with closed loops or degraded cooling contact will experience thermal runaway, or that the corresponding HV power-supply channels will reach their limit of 12 mA. Most of the modules are double-sided, and one way to reduce the self-heating effect is to switch off one side of the module. A voltage reduction can also reduce the leakage currents significantly. However, this is possible only if the applied voltage remains above the full depletion voltage. Towards the end of Run 3, it is expected that a lowering of the detector temperature to -25°C will become necessary. It is estimated that this measure will reduce the number of modules experiencing thermal runaway after 500 fb^{-1} of integrated luminosity by roughly a factor of 2.

The sensors of the SST are operated at an applied voltage of 300 V in over-depletion mode, because the sensors are p-on-n type and some of them have undergone type inversion of the bulk material. The full depletion voltage is measured by performing bias voltage scans during pp collisions. A scan of the full detector is done usually twice per year during data taking and once per month on a selected set of modules. The evolution of the full depletion voltage with integrated luminosity of one module in TIB layer 1 is shown in figure 12. The measurements of the full depletion voltage are also compared with simulations, which describe the change with integrated luminosity well.

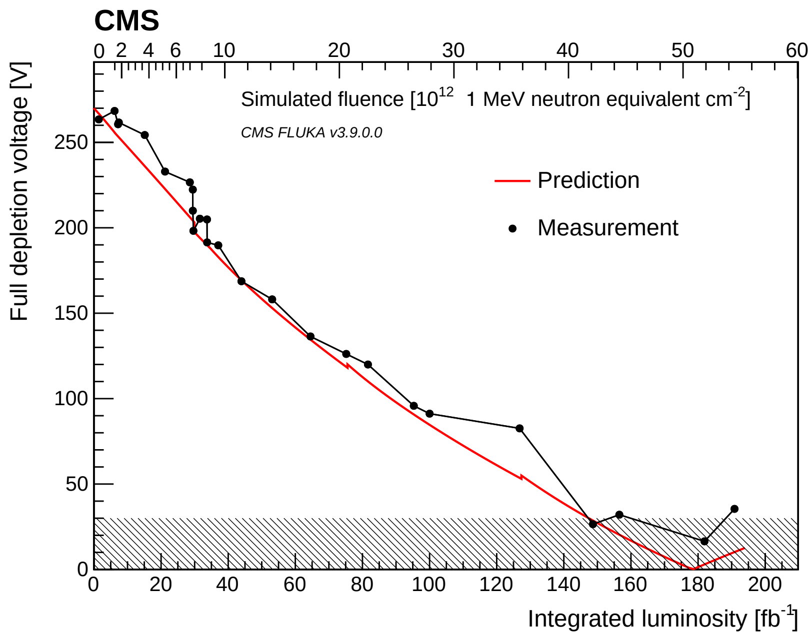


Figure 12. Evolution of the full depletion voltage for one TIB layer-1 sensor as a function of the integrated luminosity and fluence until the end of Run 2. The full depletion voltage is measured from the cluster-width variable, shown as black dots, and the predicted curve is based on a model that uses fluence and temperature history as inputs [36]. The hashed area highlights the region at low values of the full depletion voltage where the analysis loses sensitivity [34].

The installation of the pixel detector and the cooling plant maintenance work caused extended periods of time when the silicon detector was not cooled as well as it would have been desirable from the point of view of radiation damage. In figure 12, small increases due to annealing are visible in the simulation around integrated luminosities of 75 and 130 fb^{-1}, corresponding to the winter shutdown periods. As can be seen from measurements and simulation, at around 200 fb^{-1}, the TIB layer-1 sensors are close to the inversion point. The overall situation with the reduction of the full depletion voltage in the SST is shown in figure 13. For each subdetector a decrease of the full depletion voltage is observed that depends on the distance from the interaction point. It is observed that the regions of the detector that are closest to the interaction point, namely TIB layer 1, TID ring 1, and TEC ring 1, are affected the most, as expected.

In summary, the SST has been delivering high quality data for the reconstruction of charged particle tracks since the start of the LHC operation. The performance of the system continues to be excellent also after more than 200 fb^{-1} of integrated luminosity. Since the beginning of Run 3, the detector has been operated at -20°C. It is expected that the operation temperature will be lowered further to -25°C in order to reduce the leakage current. While radiation effects are visible in all parts of the detector, the margins are large enough for the detector to be operated safely and efficiently, and to provide high-quality data until the end of Run 3.

2024 JINST 19 P05064

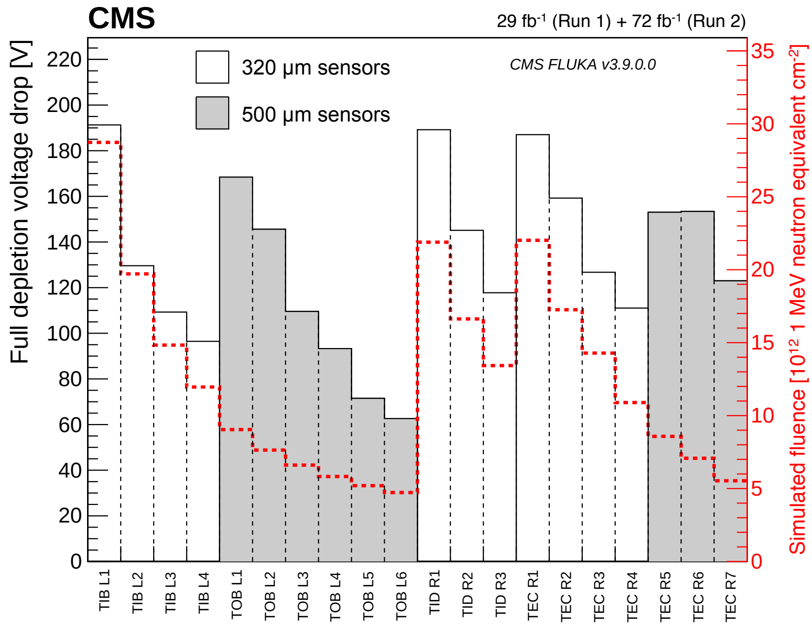


Figure 13. Decrease of the full depletion voltage for each layer, computed as the difference between the values measured at the time of the tracker construction and the values obtained by the analysis of a bias-voltage scan performed in September 2017 on all the tracker modules. The white (gray) histograms represent modules with 320 (500) μm thick sensors. The average fluence for each layer is shown by the red line.

4 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) is placed outside the inner tracking system of CMS. It provides a measurement of the energy of electrons and photons, as well as their impact position and arrival time at the crystals.

4.1 Experimental challenges

The increase in the instantaneous and integrated luminosity, experienced during Run 1 and Run 2 of the LHC and expected to continue in the future, poses operational challenges for the ECAL. The radiation dose deposited in the detector reduces the average light transmission of the $PbWO_4$ crystals, lowering the signal-to-noise ratio of the electronics readout. The radiation also induces an increase in leakage currents in the barrel photodetectors, which are avalanche photodiodes (APDs), with a corresponding increase in the electronic noise [37, 38]. The instantaneous luminosity reached $2.1 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ during Run 2, compared to $0.75 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ achieved in 2012. The increase in luminosity also poses challenges to the level-1 (L1) trigger system. Specifically, signals from direct energy deposition by particles in the APDs, termed “spikes”, must be rejected. Such spikes occur at a rate that is proportional to the luminosity. The higher radiation has also caused the silicon sensors of the preshower detector to have increased bulk currents, which require regular updates of the HV bias and correspondingly of the calibration of their response.

Additionally, the number of multiple pp interactions in a single bunch crossing (BX), termed pileup, has increased on average from 21 (up to 40) during Run 1 to 34 (up to 80) during Run 2. The bunch spacing in the machine reached its nominal value of 25 ns at the beginning of Run 2, half of what it was in Run 1. Since the typical signals from the calorimeter, after shaping by the electronics, fall to 10% of their peak value in about 250 ns, the changes in the LHC operation have resulted in an increased number of overlapping signals from neighboring BXs, referred to as out-of-time (OOT) pileup.

These effects will be discussed in more detail in the following sections, along with the improvements in the calibration of the calorimeter and the final performance achieved during Run 2.

4.2 Response monitoring

The $PbWO_4$ crystals of the ECAL, when subjected to irradiation, undergo transparency changes. This is discussed in greater detail in ref. [39] and can be ascribed to the formation of color centers, which cause absorption bands in the crystal that reduce the light attenuation length. The creation of color centers is a dynamic process depending on the dose rate absorbed by the crystals. Its annealing process spontaneously takes place at room temperature and results in partial recovery of the transmittance. Since the scintillation process remains unaltered, a reference light signal can be used to measure and monitor the transparency and response changes, and corrections can be applied to equalize the crystal-to-crystal response.

To monitor and correct the response of the ECAL, a dedicated laser monitoring system is used that operates primarily at a wavelength of 447 nm, near the peak of the scintillating light spectrum. Additional monitoring wavelengths have been used, in particular a near-infrared one at 796 nm and a green one at 527 nm. These probe the transparency in regions that are much less sensitive to radiation damage (infrared) and more sensitive to the permanent component of the radiation damage (green).

The laser is operated at 100 Hz. To avoid interference with signals from beam collisions, the light is injected into the crystals during the LHC abort gap where there are no bunches in either beam, in intervals of at least 3 μs. The abort gap is necessary to accommodate the beam abort kicker rise time and is available in all LHC filling schemes. The power of commercial lasers operating at a suitable repetition rate allows the injection of light into a few hundred crystals simultaneously. This is achieved using a system of optical fibers and diffusing spheres acting as homogeneous splitters. Light from a group of 200 fibers is measured by two p-n diodes. The variation in response is obtained by comparing the signal acquired by the APDs with the reference p-n diode. The time-dependent correction factor $LC_i(t)$, derived from the monitoring system for each crystal i, is defined as: $LC_i(t) = [R_i(0)/R_i(t)]^{\alpha}$, where $R_i(t)$ is the measured response to laser light at time t, and α is a parameter that takes into account the difference in path between the laser and scintillation light. Figure 14 summarizes the long-term evolution of the ECAL response to laser light during Run 1 and Run 2.

4.3 Noise evolution

The electronic noise in the endcaps (EE) is approximately constant. In the barrel (EB), radiation induces damage to the structure of the APD silicon lattice, causing an increase in the leakage current. The evolution of the leakage current is shown in figure 15 (left) as a function of the integrated luminosity for the central rapidity region and the most forward region of the barrel. It is well in line with the expectation from irradiation studies shown in figure 15 (right). The studies were performed using a pair of APDs, equivalent to ones on each of the ECAL barrel crystals. Measurements were done in CMS in situ for the points below 10 μA, while the points at higher currents are based on laboratory measurements of irradiation with neutrons at different fluences, as indicated in the figure.

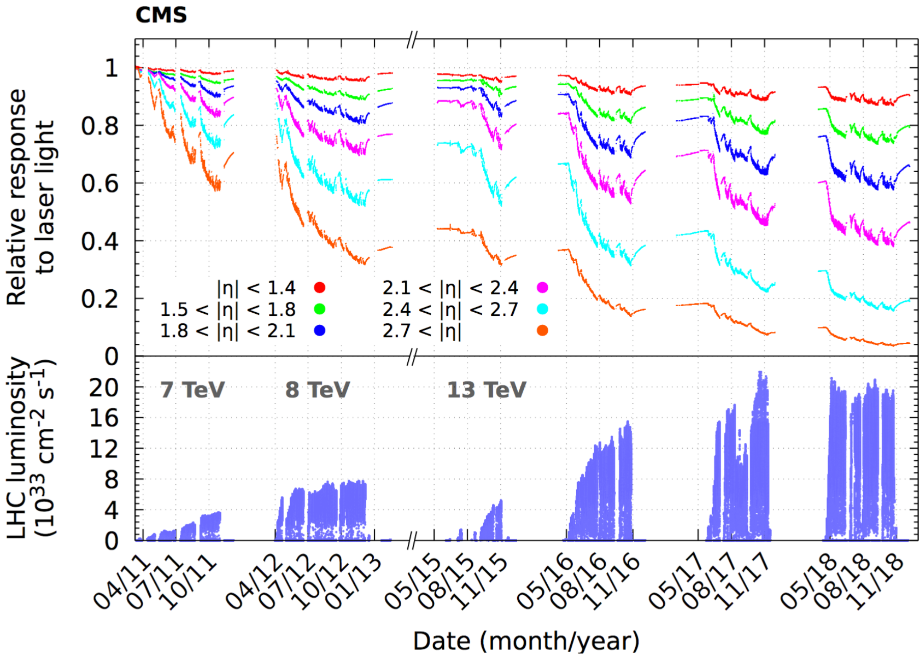


Figure 14. Relative response to laser light injected into the ECAL crystals, measured by the laser monitoring system, averaged over all crystals in bins of $|\eta|$. The response change observed in the ECAL channels is up to 13% in the barrel, $|\eta| < 1.5$, and reaches up to 62% at $|\eta| \approx 2.5$, the limit of the CMS inner tracker acceptance. The response change is up to 96% in the region closest to the beam pipe. The recovery of the crystal response during the periods without collisions is visible. These measurements, performed every 40 minutes, are used to correct the physics data. The lower panel shows the LHC instantaneous luminosity as a function of time.

When the signal is corrected for the reduction in average light yield, the electronics noise is effectively amplified. The effective noise, expressed in terms of equivalent energy and equivalent transverse energy, is shown in figure 16. The measurements are extracted from the fluctuation of the signal baseline, and are shown as a function of the pseudorapidity, covering the barrel and endcap regions. The three main data-taking periods of Run 2 are shown, along with the cumulative integrated luminosity since the beginning of Run 1. The shape as a function of $|\eta|$ is the result of the noise increase as a function of rapidity, a consequence of the larger radiation dose received by the forward regions, and of the conversion to equivalent transverse energy, the relevant quantity for physics measurements.

4.4 Signal reconstruction

The signals, after analog processing, are digitized every 25 ns. Upon a trigger, ten consecutive samples are transmitted to the backend electronics [40]. In order to cope with the increased OOT pileup, a novel amplitude reconstruction algorithm was developed for Run 2, based on a template fit, called "multifit" [41]. The multifit algorithm replaced the Run 1 method based on a digital-filtering technique [42], which estimated the energy by weighting five consecutive samples around the pulse maximum and subtracting the pedestals computed on the first three samples before the signal.

The multifit algorithm uses a template fit to extract the amplitude of the in-time pulse and the pulses coming from interactions occurring up to five BXs before and four BXs after, all within

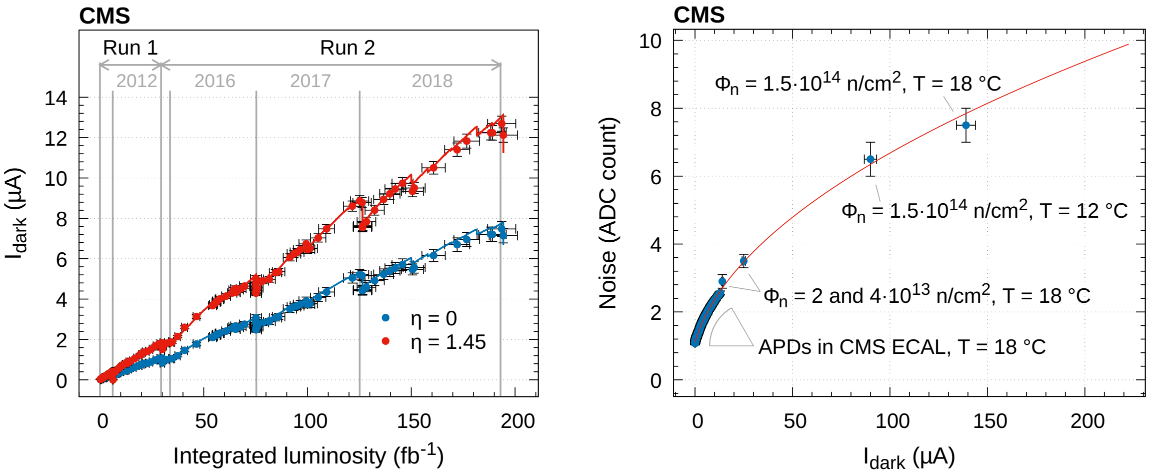


Figure 15. Left: evolution of the APD dark current as a function of the integrated luminosity since the beginning of Run 1. The gray vertical lines represent the ends of the Run 1 and Run 2 data-taking periods. Spontaneous annealing of the radiation-induced defects is visible as vertical steps in the measurements and corresponds to long stops in the data taking, e.g., year-end technical stops. Right: measurement of the electronic noise in the most sensitive amplification range as a function of the measured leakage current of an APD pair. The measurements are explained in the text. The red line is a fit to the data with a square root function. The maximum expected fluence for Run 3 is 4×10^{13} n_{eq}/cm^2 at $|\eta| = 1.45$.

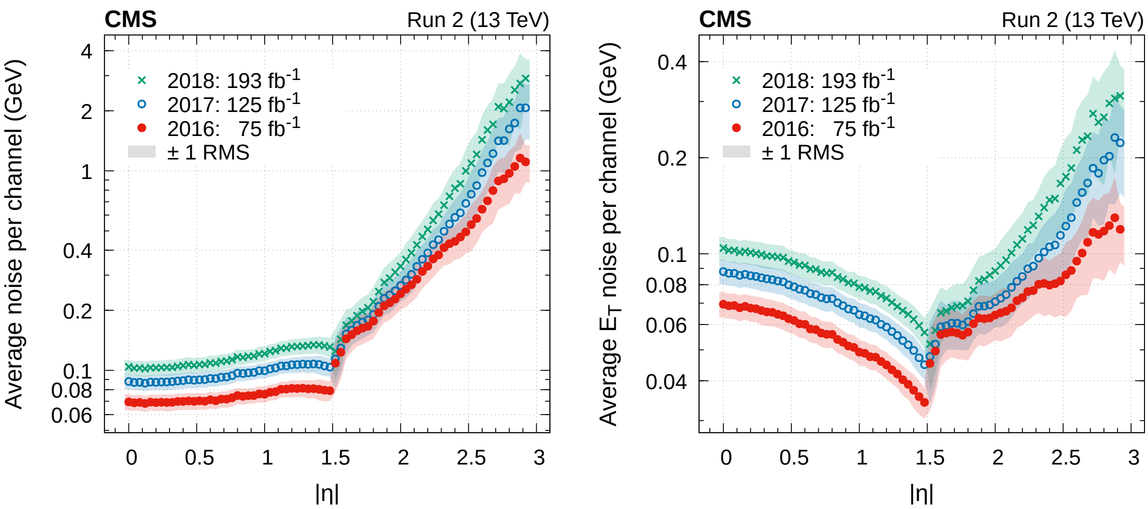


Figure 16. Average noise per channel as extracted from data using variations (RMS) of the signal baseline at the end of each year of Run 2 data taking. The amplitude RMS is converted into an energy equivalent (left) and a transverse-energy equivalent (right). The integrated luminosity accumulated since the start of Run 1 is indicated. The spread of the equivalent noise at a given pseudorapidity is indicated by the light-colored bands.

the 10-sample digitization window and potentially contributing to the total signal of one channel. Because of the number of samples and free parameters in the fit, the baseline value of the signal is not computed dynamically but is instead obtained from regular measurements performed during data taking at least once per run. Additional inputs to the fit are the template signal shapes, one per channel, and the noise covariance matrix. Both inputs are regularly measured from data and updated when found to differ significantly from those in use. This happens typically a few times per year, depending on the luminosity profile of the LHC.

The performance of the multifit algorithm has been measured using events from $\pi^0 \to \gamma\gamma$ and $Z \to ee$ decays. The energy resolution is excellent, as is the stability as a function of the OOT pileup. The improvement with respect to a nonoptimized digital-filtering technique is more significant for low-amplitude pulses, where the relative contribution of OOT pileup pulses is larger. The algorithm is sufficiently fast to be used in the HLT, and was adapted for execution on GPUs in the new processor farm used in Run 3.

The arrival time of the signal relative to the digitization window is measured by a digital-filtering technique based on the ratio of consecutive samples [43]. The timing information is subsequently corrected for its dependency on the pulse amplitude, as derived from simulations.

4.5 Trigger

The ECAL provides crystal energy sums, termed trigger primitives (TPs) [44], to the CMS L1 trigger for every BX. The trigger primitives are computed from energy sums of groups of 5×1 crystals, referred to as "strips" [45]. Each strip is served by an individual FENIX chip that performs energy intercalibration, E-to-E_T conversion, amplitude estimation, and BX assignment functions. In the EB, a sixth FENIX chip sums five strip-sums to compute the 5×5 "trigger tower" transverse energy, calculates the "fine-grain" electromagnetic bit based on the compatibility of the deposits with those from an electromagnetic shower, and computes the strip fine-grain bit for the rejection of signals from direct energy deposition in the APDs ("spike killing") [46]. The strip fine-grain electromagnetic bit is configured to return a 0 for a spike-like energy deposit (a single channel above a configurable transverse energy threshold) or a 1 for a shower-like energy deposit (multiple channels above threshold). In the EE, the five strip sums are transmitted to the off-detector trigger concentrator card (TCC) to complete the formation of the trigger towers.

The TCC is responsible for the transmission of the barrel and endcap TPs to the L1 calorimeter trigger every BX via the optical synchronization and link board (oSLB) mezzanine cards. The TCC also performs the classification of each trigger tower, its transmission to the selective readout processor at each L1 trigger accept signal, and the storage of the trigger primitives for subsequent reading by the data concentrator card.

The ECAL L1 TPs are corrected for the effects of crystal and photodetector response changes due to LHC irradiation. Correction factors are derived using measurements from the laser calibration system, and the same corrections are also applied in the HLT. These corrections were first applied in 2012 only in the endcaps, for 22 individual rings of crystals of the same pseudorapidity, and were updated once per week. During Run 2, because of the higher beam intensities and correspondingly larger response losses, the TP corrections were applied per crystal and extended to the EB. From 2017 onwards, an automated validation procedure was developed to check the impact of the updated conditions on the L1 and HLT trigger rates, and the frequency of the updates was increased to twice per week to better track the response losses versus time.

Radiation-induced changes in the ECAL signal pulse shapes, in particular in the most forward regions of the EE, caused a continually growing probability for the BX to be misassigned in the ECAL TPs. This effect, termed trigger primitive pre-firing, resulted in an inefficiency for recording potentially interesting events of about 0.1% in any given primary data set, and about 1% for events with two high-energy forward jets of invariant mass around 200 GeV [5]. Following the discovery of this issue in early 2018, η-dependent timing offsets were applied to the ECAL frontend electronics, and throughout 2018, periodic updates to these offsets were made during every LHC technical stop, in order to minimize the level of pre-firing in both the EB and EE.

The ECAL spike-killer algorithm has been retuned for the more challenging beam and detector conditions of Run 2. Spike-like energy deposits are rejected in the formation of the ECAL TPs by exploiting the additional functionality of the FENIX ASICs, the strip fine-grain electromagnetic bit. If the deposit is considered spike-like, and the tower energy is above a second configurable threshold, the tower energy is set to zero and does not contribute to the triggering of the corresponding event.

The spike-killer parameters were updated in 2016 to account for the higher LHC luminosity and the larger single-channel noise observed in the EB during Run 2. These new thresholds reduced the contamination of spikes in the ECAL TPs, corresponding to a transverse energy E_T of more than 30 GeV, by a factor of 2, with negligible impact on the triggering efficiency of electromagnetic signals with $E_T > 20$ GeV.

The spike-killing efficiency is sensitive to drifts in the ECAL signal baseline. Periodic updates in 2018 of up to twice per year of the baseline measurements used in the TP formation were therefore required in order to maintain a stable spike-killing efficiency. By periodically updating the baseline values, the spike contamination for TPs with $E_T > 30$ GeV was maintained below 20% during the 2018 run. These improvements in TP calibration and spike rejection, together with improvements in the L1 trigger system itself, allowed the L1 electron/photon trigger to operate with high efficiency and at the lowest possible E_T thresholds throughout Run 2 [5].

4.6 Channel calibration and synchronization

While the principles and methods of the ECAL calibration have not changed and are described in ref. [47], a brief summary and update is given here to help discuss the results.

The calibration of the calorimeter proceeds in several steps: (i) channels are corrected as a function of time for response changes as measured by the laser monitoring system; (ii) the response of channels at the same pseudorapidity, i.e., within the same ϕ-ring, is intercalibrated using specific physics channels as reference; (iii) ϕ-rings are intercalibrated with each other; and (iv) the absolute energy scale of the detector is fixed. In a separate and independent procedure, channels are synchronized by using the average arrival time of particles in minimum-bias events. Energy selections are applied to ensure an adequate signal-to-noise ratio, and additional criteria remove outliers in the timing distributions and ensure that the pulses have a good shape.

To complete the aforementioned steps (iii) and (iv), $Z \to ee$ events are used. For step (ii) the combination of a number of independent techniques is employed and is briefly summarized in the following.

The position of the two-photon invariant mass peak from π^0 decays is a good physics standard candle for intercalibration, even if at low energy. At the LHC, π^0 are produced in abundance, and a dedicated trigger and data acquisition stream allow for an efficient collection of a large data set. Events from this stream are saved in a reduced data format that contains only ECAL information in the proximity of the selected photon pair, to optimize the bandwidth at the HLT. Starting from L1 electromagnetic candidates, the reconstruction applies a simplified clustering algorithm that collects energy in a 3×3 crystal matrix centered around an energy deposit, called a "seed", greater than 0.5

(1.0) GeV in the barrel (endcaps). An offline analysis applies a correction derived from simulation to take into account effects of the readout, e.g., channel zero-suppression, energy lost in the vicinity of the detector boundaries, and dead channels. An iterative fit to the invariant mass of the diphoton pair is performed, varying in each iteration the intercalibration coefficients and recomputing the clustered energy and candidate selection, until the variation from one step to the following is negligible.

The E/p method exploits the distribution of the ratio between the reconstructed calorimeter energy E and the momentum p measured in the tracker of high-energy electrons from W and Z boson decays. In order to obtain a pure electron sample, electron candidates are selected using kinematic, identification, and isolation requirements. The algorithm evaluates the intercalibration in an iterative way. In each iteration, the intercalibration coefficients are updated to constrain the peak of the E/p distribution to equal unity, and the clustered energy is recalculated. A correction is applied to take into account ϕ-dependent biases in the momentum measurement due to the presence of inhomogeneous tracker support structures. The correction is calculated from $\mathrm{Z} \to \mathrm{ee}$ events using the tracker momentum measurement in a specific ϕ region for one of the electrons and the ECAL energy measurement in any ϕ region for the other. The nonuniformity in ϕ is on the order of 1%.

Electrons in $\mathrm{Z} \to \mathrm{ee}$ events can also be used for the calibration of the detector. Low-bremsstrahlung electrons are selected to minimize the influence of detector material upstream of the ECAL. The definition of low-bremsstrahlung electrons is based on the narrowness of the electromagnetic shower detected in the calorimeter. The well known invariant mass peak and distribution of the dielectron decay provide an almost background-free reference channel. An unbinned likelihood is built for the distribution observed in data, assuming the invariant mass is well described by a classical Breit-Wigner function convolved with a Gaussian function that accounts for detector effects. The resolution and scale of the Gaussian function peak value are the free parameters of the likelihood. The granularity at which the parameters are allowed to vary permits the determination of the crystal intercalibration within ϕ rings, the relative calibration between rings, and the absolute energy scale of the detector. Compared to the other two, this method of intercalibration is particularly effective at large pseudorapidities. The calibration between rings (η-scale) is derived approximately every $5\,\mathrm{fb}^{-1}$, in order to correct for drifts in the detector response.

The azimuthal symmetry of the energy flow in minimum-bias events, which was successfully used during Run 1, became more challenging during Run 2 because of the increased effective noise. While not competitive in precision for intercalibration purposes, it has been used to monitor the single-crystal response over time and provide useful insights into identifying and correcting residual imperfections in the light corrections. An example of such imperfections are the slow regional drifts of approximately 1% over one year of data taking, depending on the integrated luminosity and currently attributed to response changes in the p-n reference diode.

4.6.1 Intercalibration precision

Each of the intercalibration methods described above produces a set of constants with a corresponding statistical and systematic uncertainty. The statistical precision can be evaluated by comparing sets of constants derived from disjoint event samples. In the case of the $\mathrm{Z} \to \mathrm{ee}$ method, the precision can be obtained from the fitting procedure.

The π^0 method can be used to provide intercalibration constants for the barrel with a statistical precision that ranges from 0.1 to about 0.3%, slowly increasing with pseudorapidity, for $10\,\mathrm{fb}^{-1}$

of integrated luminosity. In the endcaps the precision is around 1%, due to the larger particle multiplicity, which requires a tighter event selection, and larger detector noise.

The E/p method requires around $50\,\mathrm{fb}^{-1}$ (about a calendar year in Run 2) to derive a set of intercalibration constants. The corresponding statistical precision varies from 0.2% for electrons in the inner barrel region to 0.4% for electrons in the outer barrel. The nonuniformity in ϕ of the material in front of the ECAL introduces a systematic uncertainty.

The $\mathrm{Z} \to \mathrm{ee}$ method also requires about $50\,\mathrm{fb}^{-1}$ to provide a set of constants of precision comparable to the E/p ones in the barrel, while in the endcap it provides a precision far better than the other methods. In Run 2, thanks to a sufficient integrated luminosity, it was possible to intercalibrate crystals in regions of $|\eta| > 2.5$ in the endcaps, where the other methods cannot be used either because of very high pileup contaminations or because the region is outside the tracker coverage.

The calibration constants determined using the three methods are combined using a weight that is proportional to the inverse square of the estimated precision. For the π^0 and E/p methods, the systematic uncertainties are evaluated by studying the impact of the intercalibration constants on the Z boson lineshape, and found to be dominant for π^0 and comparable to the statistical precision for E/p. The intercalibration precision achieved in 2018 with the three methods combined is better than 0.5% for the entire barrel, and between 0.5 and 1% for the endcaps.

4.7 Run 2 operations summary

The ECAL DAQ operated during Run 2 with a luminosity-weighted efficiency larger than 99.6% for the EB and EE, and larger than 99.2% for the preshower detector. The ECAL trigger system also operated with high efficiency and availability during Run 2. The luminosity-weighted efficiency of the trigger system, accounting for trigger downtime and deadtime, i.e., automatic throttling of the readout decisions due to too high input rates, was larger than 99.9%. The fraction of ECAL channels that contributed to the DAQ was larger than 98.6% at the end of Run 2, with a loss of less than 0.2% over the course of the four years of Run 2 operations. The fraction of channels that contributed to the trigger was larger than 99%, and only a few problematic towers, strips, and individual channels were permanently masked.

A number of improvements to the firmware and software of the TCCs [48] were implemented to achieve and maintain these high efficiencies in the more challenging beam conditions of Run 2. These involved the automatic detection and masking of noisy or problematic signals from the frontend readout via configurable thresholds, without the need for manual intervention. The algorithms allowed the setting of individual thresholds per strip in the EE, such that they could be adapted to changing LHC conditions, as well as to increased radiation-induced noise in the forward regions of the EE. As a result of these improvements, which were fully implemented in both the EB and EE before the 2018 run, the number of incidents requiring manual intervention, as well as the deadtime and downtime associated with the ECAL trigger system, were significantly reduced in 2018 compared to 2017 [48].

Additional improvements were made to the data acquisition boards, firmware, and software to be more resilient against and make automatic the recovery from single-event upsets. Despite the luminosity increase, these remained a negligible source of downtime throughout Run 2.

To increase the reliability and ease maintenance, the crates of the high voltage system for the barrel and endcaps have been upgraded from the CAEN SY1525 to the CAEN SY4527, and the low voltage system has been made more redundant.

2024 JINST 19 P05064

4.8 Run 2 performance

The performance of the ECAL in terms of energy resolution and stability of the energy scale is evaluated using $Z \to ee$ events reconstructed using the ECAL information alone. The energy resolution is affected by pileup, noise, and accuracy of the calibration, in relative order of decreasing importance. The resolution as a function of pseudorapidity is shown in figure 17 for electrons with low bremsstrahlung emissions and for an inclusive electron sample.

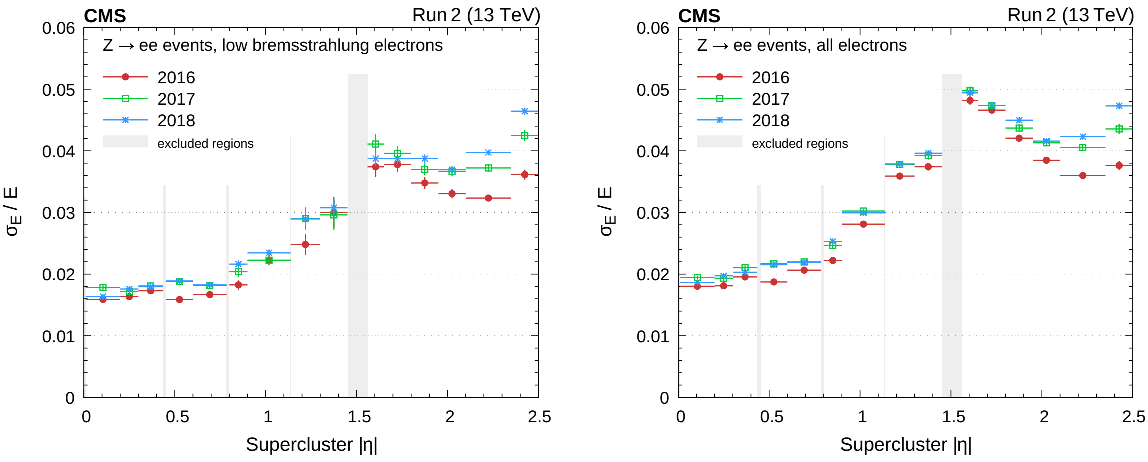


Figure 17. ECAL resolution for electrons having low bremsstrahlung emissions (left) and for an inclusive selection of electrons (right). The horizontal bars show the bin width.

The stability of the energy scale is monitored as a function of time by measuring the position of the peak in the dielectron invariant mass for $Z \to ee$ events, and is found to be within 0.1% in the barrel and a few times 0.1% in the endcaps. This figure is valid for electromagnetic deposits that are used in the reconstruction of jets and missing transverse momentum. For precision physics studies involving electrons and photons, the energy scale is further corrected using $Z \to ee$ events; after corrections, the stability is improved by a factor of 2 to 4, depending mostly on the pseudorapidity.

The ECAL time resolution is measured using $Z \to ee$ events by comparing the arrival times of the two electrons, defined as the time of the seed crystal in the supercluster. This time is corrected for time-of-flight, determined from the two electron tracks from the primary vertex of the event. Additional reconstruction quality and kinematic criteria ensure a pure sample of electrons. The time resolution resulting from the analysis of 2018 data is displayed in figure 18. It is shown as a function of the effective amplitude normalized to the noise, defined as:

$$\frac{A_{\text{eff}}}{\sigma_n} = \sqrt{\frac{2}{(\sigma_1/A_1)^2 + (\sigma_2/A_2)^2}}, \tag{4.1}$$

where A_1 and A_2 are the amplitudes of the two electron signals, and σ_1 and σ_2 the electronics noise of the corresponding channels. Notable analyses profiting from the ECAL time resolution are those looking for long-lived particles, such as that described in ref. [49].

2024 JINST 19 P05064

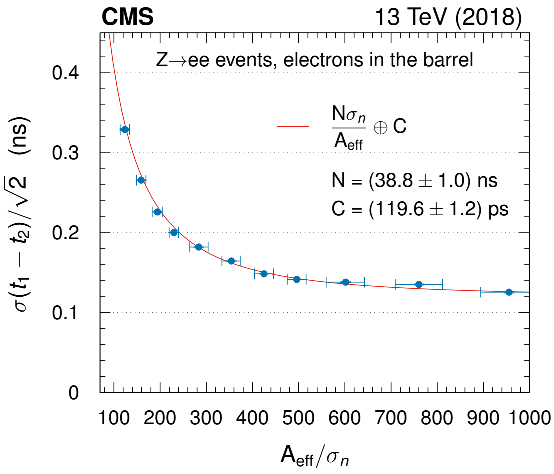


Figure 18. ECAL timing resolution as measured from Z → ee events by comparing the arrival time of the two electrons. The performance has constantly improved over the years due to the frequency at which the synchronization constants have been updated, and is shown for 2018, the last year of Run 2. Updates in the constants are necessary to compensate for pulse shape changes induced by radiation. Vertical bars on the points showing the statistical uncertainties are too small to be seen in the plot. The red line correspond to a fit to the points with the parametrization of the resolution shown in the legend.

4.9 Preparation for Run 3

During LS2, the ECAL activities focused on improving the detector safety and control systems, on the algorithm to determine the trigger primitives, and on the development of a system to automatically compute, validate, and deliver updated calibrations.

4.9.1 Safety and control system

A significant upgrade of the ECAL safety and detector control systems, DSS and DCS, was performed during LS2. The sensor readout system for temperature, humidity, voltage, and current levels, composed of custom readout units, was replaced by industrial analog input (AI) modules. The 12 readout units based on RS-485 interfaces were exchanged by 45 AI modules, which are standard Siemens-certified peripherals (Simatic S7-300 analog input SM 331) connected through Profibus communication buses. This type of connection provides access to extensive diagnostic information and enables the readout of a full sensor at sampling intervals as fast as 0.1 s. This is about one order of magnitude faster than the previous system. Following the update of the readout method, the programmable logic controller (PLC) of the safety system was reprogrammed with completely new software, and is now part of the pool of PLC framework applications in CMS. The action matrix that defines the behavior of the PLC in terms of input and output signals and raises interlocks to protect the detector in case of alarm conditions, did not change, but benefits from the improved hardware. The new system was extensively validated during the regular cosmic ray data-taking campaigns in 2020 and 2021.

4.9.2 Trigger

Several improvements to the ECAL trigger-primitive formation and calibration were implemented for Run 3. These include a further optimization of the spike-killer thresholds (section 4.5) for the expected Run 3 pileup and noise levels, optimization of the digital-filter weights used to compute the trigger-primitive energies accounting for radiation-induced changes in pulse shapes, and the possible use of a second set of amplitude weights to further improve spike rejection and for the potential tagging of out-of-time signals.

More frequent corrections to account for crystal and photodetector response changes were also implemented for Run 3, and the frequency of the updates was increased from twice a week to once per LHC fill. These corrections are important to maintain stable trigger rates and efficiencies, and improve the energy resolution of the related L1 and HLT objects, particularly electron/photon candidates.

4.10 Calibration

With the aim of reducing, as much as possible, the need for multiple reconstruction of CMS data sets with updated calibrations, a framework has been setup to provide automatic execution and bookkeeping of the workflows necessary to compute and validate updated and refined detector conditions as soon as enough data is available. Not only does this allow to follow closely the prompt reconstruction of CMS data with the best foreseeable conditions, but it also permits to have the best conditions available for the data (re)reconstruction as soon as the data taking is finished. While the techniques and physics standard candles used for calibrating the ECAL were well consolidated during Run 2, novelty and optimization have been introduced at the technical level of the data analysis needed to provide detector conditions. Additionally, with a layer to validate conditions before deployment, key figures of detector performance can be controlled, such as stability, resolution, and projected rates in the HLT. The workflows is synchronized with the online data taking and orchestrated by a Jenkins instance, deployed through Red Hat OpenShift technologies, based on an InfluxDB backend and a Python server.

5 Hadron calorimeter

5.1 The hadron calorimeter in Run 1 and Run 2

The CMS hadron calorimeter [50] (HCAL), shown schematically in figure 19, is composed of four major subdetectors: the hadron barrel (HB) [51], hadron endcap (HE) [50], hadron forward (HF) [52], and hadron outer (HO) calorimeters [53]. The HB and HE cover the pseudorapidity regions $|\eta| < 1.392$ and $1.305 < |\eta| < 3.0$, respectively. The HO provides a measurement of the shower tails in the region $|\eta| < 1.26$, and the HF covers $3.0 < |\eta| < 5.2$.

The HB and HE primarily use brass as the absorber, except for the inner and outer layers of HB, which are constructed from steel. The HB absorber is shown in figure 20 (left). The signals are produced in plastic scintillating tiles (figure 20, right), and the resulting blue light is shifted to green via embedded wavelength-shifting fibers. The towers in HB (HE) have up to 17 (18) scintillator layers, as shown in figure 19. Sequential layers are grouped into "depth" segments: the light from the layers in a given depth segment is optically summed and read out by a single photodetector. Clear plastic fibers send the signal to the hybrid photodetectors (HPDs) in the original design or silicon photomultipliers (SiPMs) after the upgrades. The segmentation is a tower structure in η-ϕ space. The η segmentation is indicated by the black solid lines in figure 19. The towers are referenced using integer indices ieta and iphi, where the ieta assignments are given in the figure and iphi runs from 0 to 71, corresponding to the 72 divisions

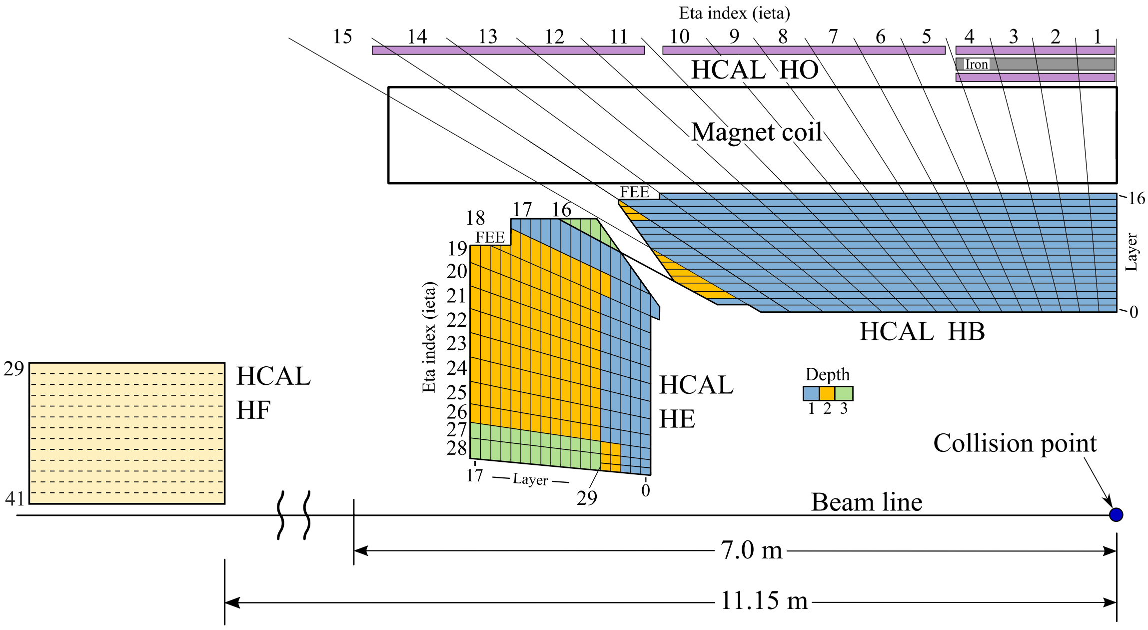


Figure 19. Schematic view of the HCAL as of 2016, showing the positions of its four major components: HB, HE, HO, and HF. The layers marked in blue are grouped together as "depth 1," i.e., the signals from these layers of a given tower are optically summed and read out by a single photodetector. Similarly, the layers shown in yellow and green are combined as depths 2 and 3, respectively, and the layers shown in purple are combined for HO. The notation "FEE" denotes the locations of the HB and HE frontend electronics readout boxes. The solid black lines, roughly projective with the interaction point, denote the η divisions in the tower η-ϕ segmentation, and the numbers at the edge of the tower denote the ieta index. Reproduced from [54]. © 2020 CERN for the benefit of the CMS collaboration. CC BY 3.0.

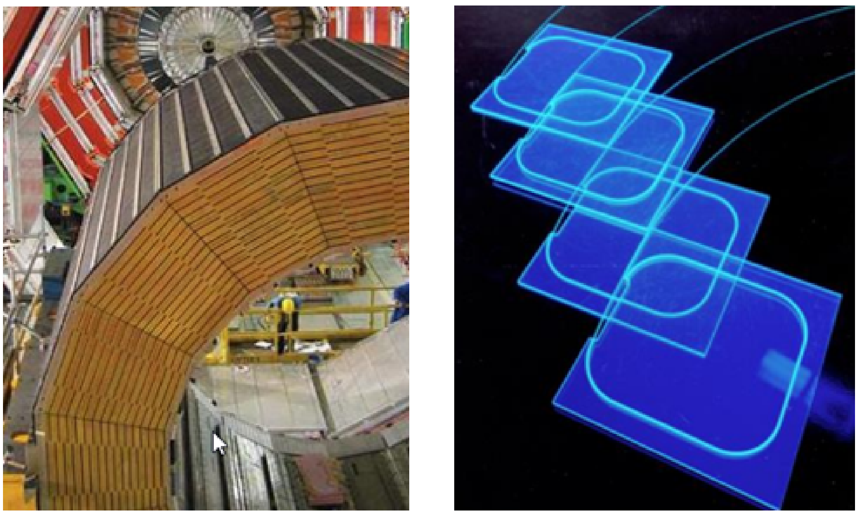


Figure 20. Left: brass absorber for the hadron barrel calorimeter HB. Right: scintillating tiles with wavelength shifting fibers used as the active media in the barrel, endcap, and outer hadron calorimeters.

in ϕ. Physically, the scintillators are arranged in "megatiles", which are trays that support an array of scintillator tiles, along with the fibers that route the light to the photodetectors. An HE megatile is shown in figure 21. All channels in a subdetector with the same iphi are defined as residing in the same wedge.

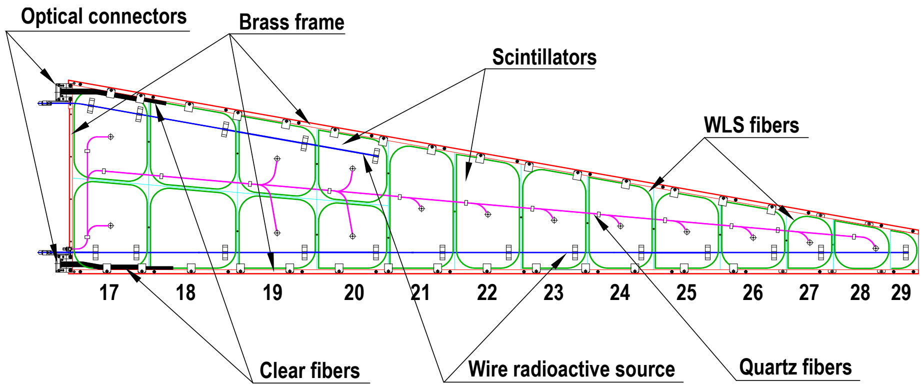


Figure 21. Physical arrangement of the scintillator and wavelength-shifting fibers into tiles for an HE megatile. Reproduced from [54]. © 2020 CERN for the benefit of the CMS collaboration. CC BY 3.0.

The HF, shown schematically in figure 22, is a 1.65 m-long sampling calorimeter with steel absorber. Plastic-clad quartz fibers with a diameter of 0.6 mm are the active elements, producing Cherenkov light. The fibers are inserted into 1 mm grooves drilled in the steel absorber, 5 mm apart. The fibers are parallel to the beam line. To enable discrimination of electrons and photons from hadrons, the fibers have two lengths: 1.65 m "long fibers" and 1.40 m "short fibers". The short fibers run from the outside of the HF to 25 cm from the front face of the HF. The HF photodetectors are photomultiplier tubes (PMTs).

The HCAL readout control chain is shown schematically in figure 23. In the frontend readout boxes (RBXs), mounted on the detector, the electronics chain begins with analog signals from the photodetectors. The optical decoder unit (ODU) maps the incoming fibers from the scintillators to the photodetectors. These signals are digitized by the QIE chips (QIE8, QIE10, and QIE11) [56–59] in the readout modules (RMs), which provide both energy information via an analog-to-digital converter (ADC) and, for the QIE10 and QIE11 chips, rising-edge timing information via a time-to-digital converter (TDC). The clock, control, and monitoring (CCM) module provides the LHC clock, synchronous fast commands, slow control, and monitoring for the detector frontend electronics. Each RBX also contains a calibration module (CM), which provides calibration signals from two sources: an in situ LED, which sends light directly to the photosensors, and an off-detector laser, which can send light either to the photosensors or into the scintillator tiles for some of the detector layers. The CM also provides a reference measurement of the LED and laser calibration signal intensities.

5.2 Upgrades

5.2.1 Motivation and overview of the upgrade

The motivation for and the design of the CMS calorimeter system Phase 1 upgrades are given in refs. [60, 61]. The main goals of the HB and HE upgrades were: to replace the HPD photodetectors, which produced anomalous signals [62] and showed signal degradation [63]; to increase the segmentation to that shown in figure 24, to allow for both layer-dependent corrections for the

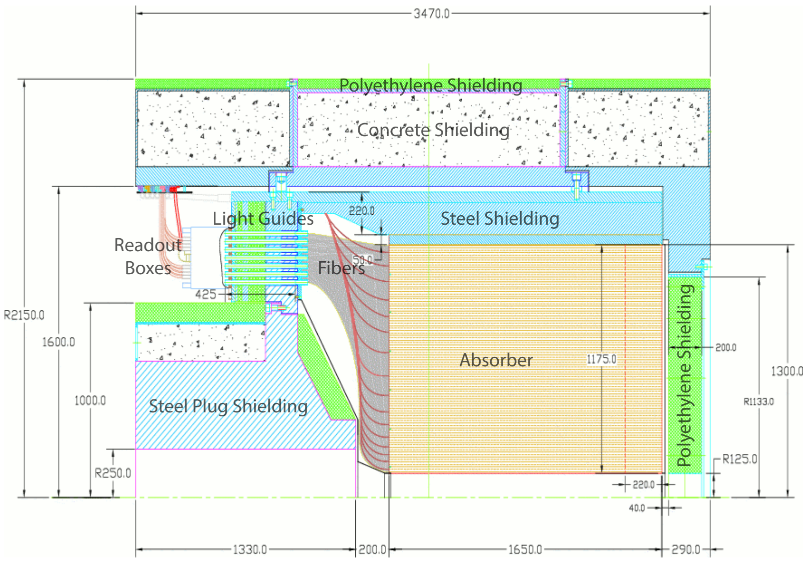


Figure 22. Schematic view of the CMS hadron forward calorimeter, HF. The yellow region represents the steel absorber with embedded quartz fibers; the grey shaded area to the left represents fibers which deliver the signals to light guides that penetrate the steel plug shielding; the white rectangle to the left of the light guides represents the frontend readout boxes, which house the photomultiplier tubes. The dimensions shown in the diagram are in millimeters.

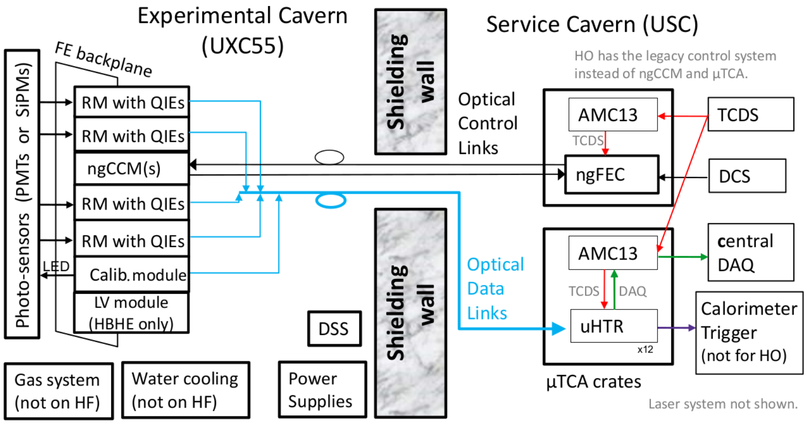


Figure 23. Schematic diagram of the HCAL readout. The frontend electronics chain begins with analog signals from the HPDs, SiPMs, or PMTs. The HB and HE used HPDs prior to the Phase 1 upgrade and SiPMs after, while the HF uses PMTs. These signals are digitized by the QIE chips in the readout modules (RMs). The "next-generation clock and control module" (ngCCM) is part of the frontend control system. The digitized data are sent to the backend μHTRs (μTCA HCAL trigger and readout cards). Data from multiple μHTRs are concentrated into the AMC13 cards and forwarded to the CMS central data acquisition (DAQ) system. The μTCA cards also send data to the trigger system. The AMC13 cards also distribute the fast commands arriving from the CMS timing and control distribution system (TCDS) within each μTCA crate. The detector control system (DCS) software communicates with the "next-generation frontend controller" (ngFEC).

observed radiation damage to the scintillating tiles [64] and better rejection of energy deposits from pileup interactions; to increase the readout bandwidth to allow for the larger number of channels; to add signal arrival time measurements in the HB, HE, and HF; and to standardize the readout electronics across the different calorimeter systems. For the HF, the PMTs were also replaced because they were a source of anomalous signals [62].

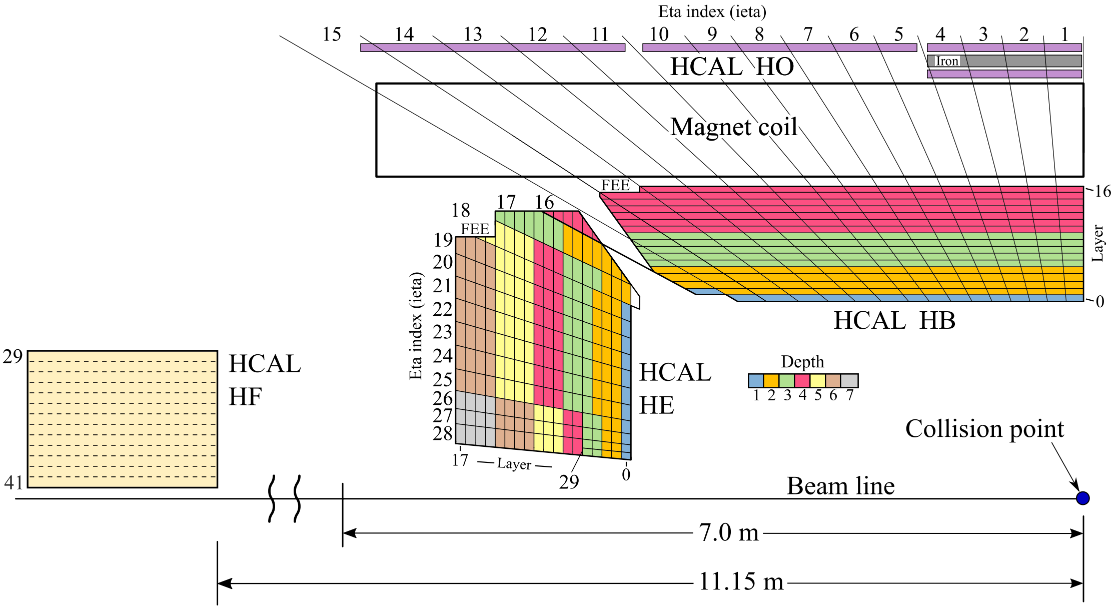


Figure 24. The longitudinal and transverse HCAL segmentation for Run 3. Within a tower, layers with the same color are routed to the same SiPM. The location of the frontend electronics is indicated by the letters FEE. Reproduced from [54]. © 2020 CERN for the benefit of the CMS collaboration. CC BY 3.0.

The HCAL plastic scintillators are subject to radiation damage, which results in the reduction of the signal output. The relevant primary characteristics of the LHC operation are the total delivered integrated luminosity, which determines the radiation dose received by the scintillator tiles, and the average instantaneous luminosity, which controls the dose rates. Radiation effects were evaluated by studying the tile performance during the 2017 LHC operation, corresponding to a delivered integrated luminosity of about $50\,\mathrm{fb}^{-1}$. Signal reduction was studied as a function of dose rate. These measurements provide unique information on the radiation damage at dose rates significantly lower than previously studied. The HE tile results were obtained using several complementary methods: a movable radioactive source that can access all the tiles to compare their signal output before and after the 2017 data-taking period; inclusive and isolated-muon energy deposits produced during pp collisions; and a laser calibration system.

The laser system consists of a triggerable excimer laser and light distribution system that delivers UV light (351 nm) to the scintillator tiles in layers 1 and 7 via quartz fibers, as well as directly to the photodetectors. During the 2017 data-taking period, pulses of laser light were injected between LHC fills when there were no collisions. Laser data were collected throughout the 2017 data-taking period. Figure 25 (left) presents the relative signals in layer 1 versus dose for tiles in the ieta range 21–27. The signals show an approximately exponential decrease during periods of stable luminosity, with slopes that depend on the dose rate. Tiles at smaller ieta show more damage per dose than those at larger ieta, implying that at a fixed dose the damage to the scintillators increases with decreasing dose rate, within the range of our measurements.

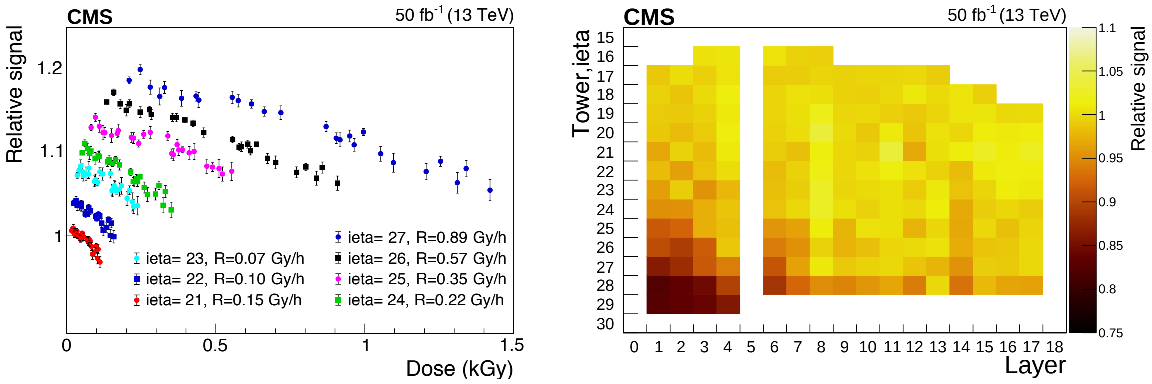


Figure 25. Left: relative laser light signal versus the accumulated dose for scintillator tiles in layer 1 and ieta range 21–27. The average dose rate R for each set of points is given in the legend. The vertical scale is logarithmic and subsequent sets are shifted up by a factor of 1.03 relative to the previous set for better visibility. Each set starts at a dose corresponding to an integrated luminosity of $7\,\mathrm{fb}^{-1}$. The vertical bars give the scaled statistical uncertainties. Right: ratio of the signals from the ^{60}Co source observed before and after the 2017 data-taking period for scintillator tiles in the HE as a function of ieta and layer number. Tubes in layers 0 and 5 have obstructions and cannot be accessed.

Each individual tile in the HCAL is designed to be serviced by a movable ^{60}Co radioactive source using small tubes that are integrated into the calorimeter. The ^{60}Co source provides photons with energies of 1.17 and 1.33 MeV. The source is attached to a wire that guides it through the tubes. All tiles except those in layers 0 and 5, whose tubes have obstructions, can be accessed. The data were collected during the periods when the LHC did not operate, before both the 2017 and 2018 data-taking periods. Values of the ratio, averaged over iphi as a function of the scintillating tile layer number and tower index ieta, are shown in figure 25 (right). The signal loss is smaller for tiles at larger radial distance from the beam and for layers that are deeper in the calorimeter.

Because the damage depends on the layer number, increased segmentation allows for more accurate damage corrections to be applied as a function of integrated luminosity, reducing the degradation of the resolution.

The on-detector upgrade components must operate in a radiation field and withstand cumulative effects from both the total ionizing dose (TID) and nonionizing energy loss (NIEL), as well as single-event effects (SEE). The required tolerance for NIELs is quoted as that corresponding to an equivalent fluence (cm^{-2}) of 1 MeV neutrons. Expected SEE rates are characterized by the fluence (cm^{-2}) of hadrons with energy exceeding 20 MeV. Requirements for the HB, HE, and HF are shown in table 5. Since the HE will be replaced after Run 3, the numbers given for the HE are the requirements for Run 3, while the HB and HF values correspond to the full HL-LHC duration.

5.2.2 HB/HE/HO photodetector upgrade

A significant part of the HCAL Phase 1 upgrade was the replacement of the HPD photodetectors with SiPMs for HB, HE, and HO. All three subdetectors used HPDs during Run 1 [65], and were converted to SiPMs at different times. The HO SiPMs were installed prior to the start of Run 2. For

2024 JINST 19 P05064

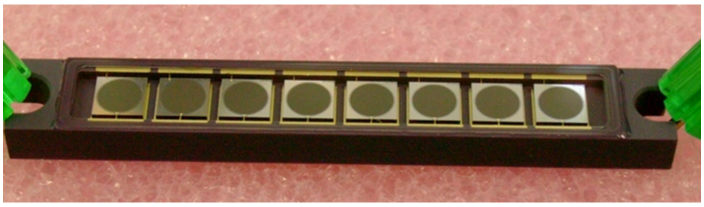
Table 5. Radiation requirements for the Phase 1 upgrade. The HE numbers are for Run 3, while the HB and HF values correspond to the full HL-LHC duration.

Detector	TID [krad]	NIEL [cm^{-2}]	SEE [cm^{-2}]
HB	3.1	1.1×10^{12}	2.0×10^{11}
HE	0.9	9.0×10^{10}	1.6×10^{10}
HF	4.1	7.0×10^{11}	1.8×10^{11}

HE, a ϕ sector of 20° was instrumented with SiPMs during the 2016/2017 year-end technical stop, and the rest of the detector was converted during the 2017/2018 technical stop. The HB used HPDs for all of Run 2, with SiPMs installed prior to the start of Run 3. The grouping of layers in a tower to individual photosensors, i.e., the longitudinal readout segmentation, is indicated by color in figure 19 prior to the upgrade and following the upgrade in figure 24. Prior to Run 3, most of the HB towers had a single readout segment, except for ieta = 15 and ieta = 16. The HB readout segmentation increased from one to four channels starting in Run 3. For HE, the readout segmentation increased from two to three channels per tower in Run 1 to six to seven channels during Run 2 as the SiPMs were installed.

The SiPM has many advantages over the HPD, including high photon detection efficiency (PDE), high gain, a large linear dynamic range, rapid recovery time, somewhat better radiation tolerance, and insensitivity to magnetic fields. The active area of each SiPM is circular, with diameters of 2.8 mm (sensing up to four fibers) and 3.3 mm (sensing up to seven fibers). In each HE readout unit, 19 HPD channels were replaced by 48 channels of SiPMs. For the HB, each readout unit has 64 channels of SiPMs.

Arrays of eight individual SiPMs were placed into a ceramic carrier known as the "SiPM package". The packages for the HE and HB are identical. They were designed by CMS and fabricated at Kyocera. Figure 26 (left) shows a top-and-side view of an eight-channel SiPM array in its package. Each SiPM is connected electrically to the package via two wire bonds, one for the signal and one for the bias voltage. The 16 pins at the bottom of the package connect the SiPMs to the downstream electronics. The package precisely locates each SiPM on the ODU, so its mechanical dimensions are important. Before sending them to the SiPM vendor, the mechanical tolerances of each package were measured at the CERN Metrology Laboratory, and more than 30 parameters were compared with the design specifications. Figure 26 (right) shows 48 ceramic SiPM packages prepared for measurement. The yield of good packages was very high with less than one percent rejected for being out of tolerance.



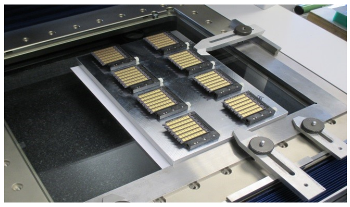


Figure 26. Left: top and side views of an eight-channel SiPM array in its ceramic package. Right: ceramic packages at the CERN Metrology Laboratory for characterization measurements.

Before choosing a vendor and entering final SiPM production for the HE, and then later the HB, careful studies were carried out on preproduction samples from several vendors. The preproduction samples were of order 1500 channels. Key SiPM parameters included PDE, gain, recovery time, uniformity of breakdown voltage, and dark current [66]. A small sample of SiPMs was also irradiated for radiation damage studies [67]. Based on the results, the Hamamatsu SiPMs were chosen for final production.

The quality control program was virtually identical for the HE and HB. We measured the signal response (PDE times gain), breakdown voltage, dark current, forward resistance, and capacitance for each channel. We also looked for spurious noise pulses in every channel. A total of 1400 production HE arrays were tested, and 104 arrays, or 7.43%, were rejected. Since there are eight SiPM channels per array, this corresponded to a yield of good SiPMs better than 99%. For the HB production, a total of 1680 production arrays went through the same quality control testing, with only 82 rejected, again giving a yield of good SiPMs better than 99%. The uniformity was also excellent throughout the two production runs. As an example, figure 27 shows the signal response for 3600 HE SiPMs, with an RMS of about 1.5%. After quality control, a total of 864 (1152) SiPM arrays were installed into the HE (HB).

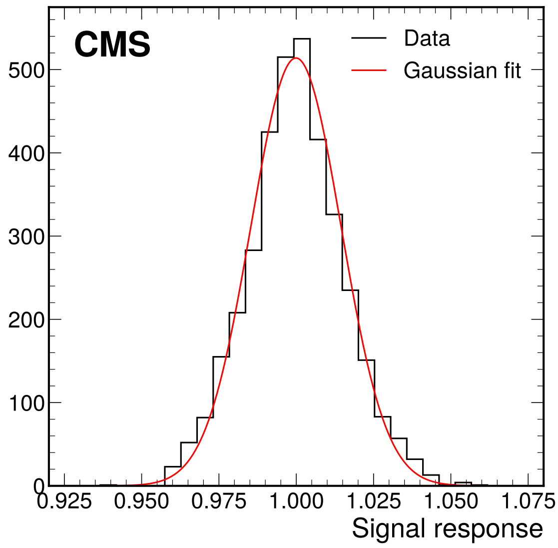


Figure 27. Distribution of the signal response (photon detection efficiency times gain) for 3600 HE SiPMs.

5.2.3 Readout box

The RBX is an aluminum shell with snake-shaped grooves machined on the surface for insertion of copper cooling pipes. The RBX houses the RMs, control electronics, and the QIE-based frontend electronics, which are connected to the backend in the control room via control and data fibers. The HB and HE frontend electronics, shown schematically in figure 28 are very similar, and primarily differ to reflect the mechanical, structural, and channel multiplicity differences between the two subsystems.

Each HE, HB, and HO RM has 48, 64, and 18 SiPMs, respectively. Each RM houses a single SiPM control card and four frontend readout ADC cards (three cards for the HO). Each RM also houses an ODU. As shown in figure 23, the RBX consists of four RMs installed together with a ngCCM and a CU. In HB, two ngCCMs are installed per RBX, each independently controlling two

2024 JINST 19 P05064

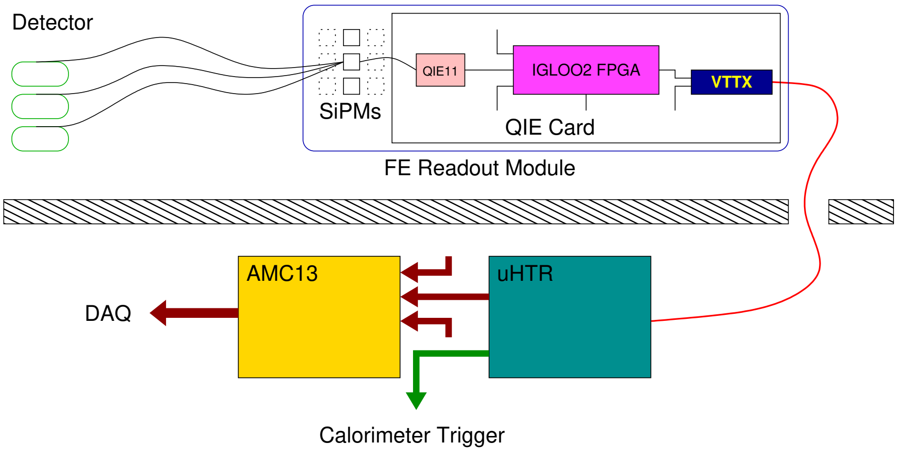


Figure 28. Upgraded HCAL endcap (HE) and barrel (HB) DAQ readout chain, including the SiPMs, frontend readout electronics, and backend system. Reproduced from [60]. CC BY 3.0.

RMs. All RBX components receive the power and communicate via a backplane PCB. The RBX receives its fast and slow control signals through the ngCCM module. This module receives the clock and fast reset signals via optical links from the backend, and distributes these signals to each card in the RBX. Configuration information for the frontend is received over the optical link as well. Finally, the ngCCM is responsible for aggregating and returning all status information about the RBX, which is transmitted to the backend via an optical link.

Each RBX is installed on top of a scintillator megatile in the locations indicated in figure 24. One RBX covers 20° in ϕ. The total number of RBXs is 108, excluding those used for the HF. The RBX is the part of the common CMS infrastructure; therefore, each RBX has individual power, water cooling, and dry gas lines connected to common water and gas manifolds in order to provide proper and stable cooling for the HCAL electronics and decrease the relative humidity around the SiPMs. The maximum temperature of RBX components does not exceed 35°C, and the relative humidity in the SiPM region stays constant within 5%, except on rare occasions when there are cooling or dry gas issues.

The first upgraded RBX for the initial sector of the HE instrumented with SiPMs was installed in 2017. One HE RBX was upgraded at the end of 2016, the rest of the HE in 2017/2018, and the HB RBXs in 2019 after Run 2.

5.2.4 Photodetector control

The use of SiPMs required the design of completely new high-density photodetector control electronics. The new design was done initially for the HO photodetector upgrade, and consisted of an additional board placed next to the QIE cards in the RM, shown schematically in figure 29 (left) and as constructed in figure 29 (right). The HE and HB subdetectors reuse the same design topology to minimize design, testing, and validation time, although small changes were needed to accommodate the different SiPM types. The control board uses a single +8 V power voltage from the RBX backplane and is controlled via an I2C-compatible slow-control connection. The board provides individual bias voltage regulation and leakage current measurements, limiting of bias voltage current for SiPM protection, a Peltier cooler control, and remote temperature/humidity readout.

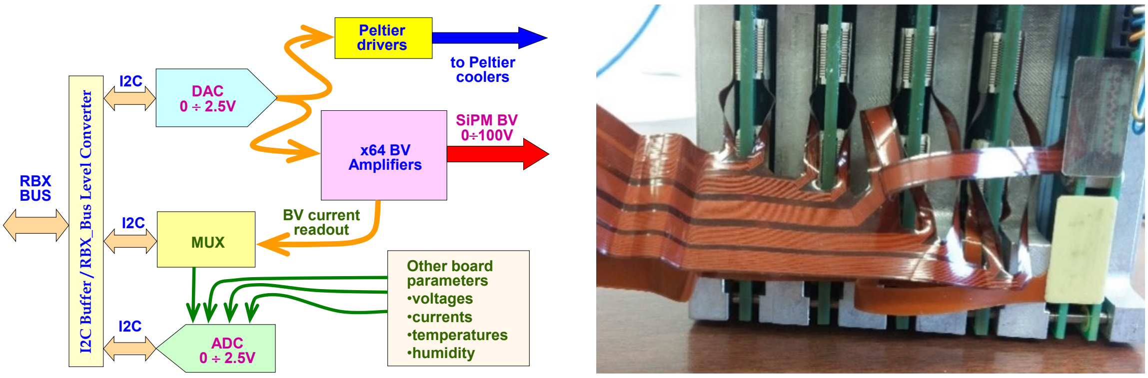


Figure 29. Left: control board block diagram. Right: HB card pack with one control and four frontend boards. The flex cables provide signal and bias voltage connections to 64 SiPMs.

Bias voltage and temperature regulation strongly affect the SiPM noise and stability. In order to limit the variation in the gain of all the detector channels to 2%, the bias voltage and temperature are strictly controlled, with several key parameters shown in table 6. The measured in situ temperature stability was 0.3°C.

Table 6. HCAL SiPM requirements for the Phase 1 upgrade. The design (measured) parameter values are shown. For the design (measured) value of the BV stability, the RMS (peak-to-peak) is quoted.

Parameter	HB	HE
Channel number per RM	48 (48)	64 (64)
Bias voltage [V]	65–75 (0–80)	65–70 (0–80)
BV resolution [mV]	40 (20)	40 (20)
BV stability [mV]	40 (26)	40 (26)
BV ripple/noise (RMS above 10 kHz) [mV]	7 (3–7)	7 (3–7)
Temperature monitoring [°C]	0.25 (0.1)	0.25 (0.1)
Max operating current [μA]	500 (500)	1000 (1000)
BV current resolution [nA]	122 (122)	244 (244)
Total ionizing dose [krad]	3.1 (17)	0.9 (17)
1 MeV-equivalent neutron fluence [cm^{-2}]	9×10^{11} (2×10^{12})	1.1×10^{12} (2×10^{12})
Power for Peltier cooler	(6.5 V × 1 A)	(6.5 V × 2 A)

5.2.5 Optical decoder unit

The optical decoder unit (ODU) is the interface between the clear optical fibers bringing light signals from the HCAL scintillator and the SiPMs that convert the light into electrical signals. Each ODU contains a network of short (about 10 cm) clear plastic fibers that remap the planar geometry of the incoming light signals from the megatiles into a tower geometry that is more appropriate and useful for physics analysis. While the size, shape, and some details are different for the HE and HB ODUs, the basic features are the same. Each RM in HB and HE contains one ODU, for a total of 144 ODUs each in HB and HE.

2024 JINST 19 P05064

The ODU is essentially a box containing the network of fibers described above. The fibers are fabricated into cables with polished connectors at each end, which mate with the fiber connectors from the detector. The cables are cut in half to create “pigtails”, which form the basic units for assembly into the ODU. Each HB (HE) pigtail has up to 18 (12) fibers. The routing and connections of the fibers are shown in figure 30 (left). At the top of the picture, the pigtail connectors are attached through open slots in a machined aluminum patch panel, where they connect one-to-one to the fibers coming from the HCAL detector layers. At the left of the picture, the fibers are mapped to their designated holes in the so-called “cookie,” a precisely machined piece of plastic containing 64 (48) holes for the HB (HE) which directs the light to individual SiPMs. The HE cookie is a solid piece of polyether ether ketone plastic, while the HB cookie has two plastic layers with an insulating foam layer in the middle to prevent heat from leaking to the SiPMs. The number of fibers in each hole varies from one to seven, depending on the depth segment being read out. The fibers are glued in place in the cookie, which is optically finished with a diamond flycutter. All the precision-machined parts were fabricated in local industry, while the ODU assembly, inspection, and testing were done by CMS. A completely assembled production HE ODU is shown in figure 30 (right).

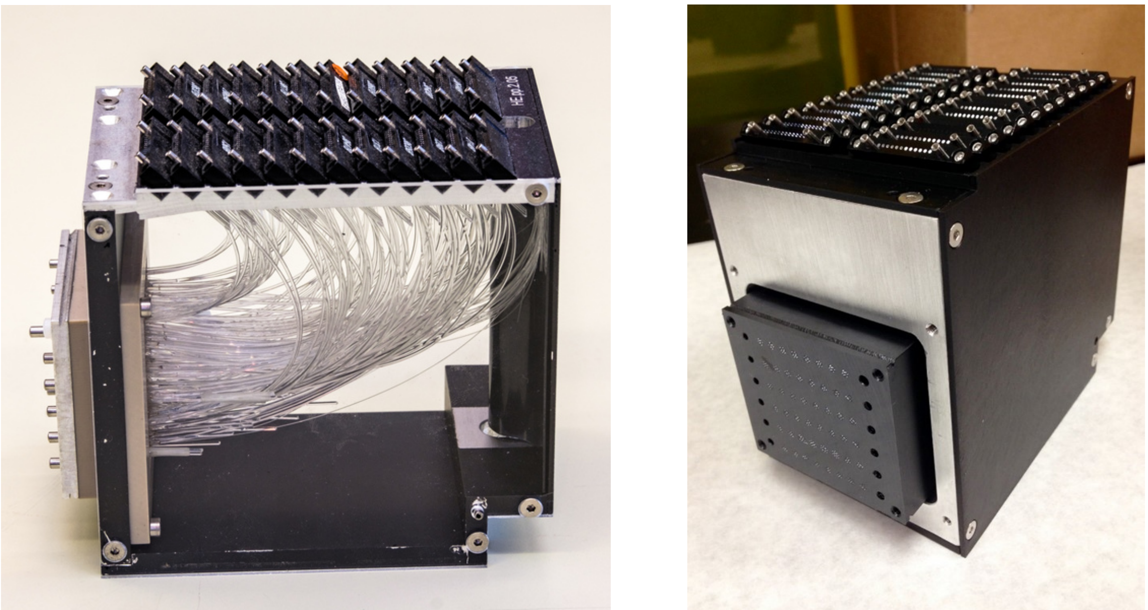


Figure 30. Left: view of a spare HE optical decoder unit (ODU), showing its light box and fiber mapping. The fibers route from the patch panel at the top to the “cookie” at the left. The side panels are clear rather than opaque for display purposes. Right: a production HE ODU. The clear fiber pigtail connectors attached to the patch panel are visible at the top. The plastic “cookie” is seen at the front of the ODU.

Extensive quality control testing of both the individual fibers and the fully assembled ODUs was performed to ensure the integrity of the light transmission. Each cable was tested by connecting it to a similar cable of wavelength-shifter fibers that were illuminated individually by LEDs. The light throughput of each fiber in the cable was measured with a photodiode. To make sure there were no problems with the connectors at either end, the cable was reversed, reconnected, and measured again. To test the fully assembled ODUs, light was injected into the ODU through the connectors in the patch panel, fiber by fiber, and measured by a large photodiode at the face of the cookie. A total of 144 fully qualified ODUs were installed into both the HB and HE.

5.2.6 Frontend readout card

The signals from the SiPMs are integrated and digitized by the QIE11 ASICs [56–59]. The QIE11 integrates negative input charge pulses in 25 ns buckets and digitizes the result with approximately constant resolution over a large dynamic range. The QIE11 achieves an effective 17-bit dynamic range with approximately 1% resolution with only 8 bits through the use of a custom floating point analog-to-digital converter (ADC) with pseudo-logarithmic response. The ADC has a programmable gain via user-configurable input current shunts and a low input impedance (< 15 W) that is suitable for use with SiPMs. A 6-bit time-to-digital converter (TDC) is also included on the QIE11 chip, consisting of a programmable-threshold discriminator which detects arrival time of the input pulse in 0.5 ns bins. The phasing of the charge integration window relative to the input clock can be adjusted by the user in 0.5 ns steps over the whole 25 ns window. The nominal sensitivity of QIE11 is 3.1 fC per count at the low end (with no programmable input shunt selected). The gain of any particular chip can vary from this nominal value due to process variations, so all chips are calibrated with better than 1% precision. The nominal maximum charge that can be digitized is approximately 350 pC, yielding the nearly 17-bit dynamic range.

Each readout card in the HE (HB) supports 12 (16) SiPM channels and contains a corresponding number of QIE11 ASICs. In addition, each card contains one Microsemi ProASIC3L FPGA that acts as an I2C bridge between the slow-control unit (ngCCM) and the frontend chips, one or two Microsemi IGLOO2 FPGAs that serialize and format the data, and one VTTx [68] module providing two 4.8 Gb/s optical links. The necessary power for operating the frontend readout card is supplied by the CERN FEASTMP DC-DC converters, which are radiation and magnetic field tolerant.

Digital data from the QIE chips are serialized and formatted by two (one) Microsemi IGLOO2 FPGAs in the HB (HE). The flash-based IGLOO2 FPGA achieves sufficient radiation tolerance for the HCAL frontend, better than typical SRAM-based FPGA technologies. In the HE, there is sufficient bandwidth to transmit all 8 bits of ADC and 6 bits of TDC data per channel per bunch crossing; however, in the HB, bandwidth constraints require the reduction of TDC information to two bits, which are used to encode four arrival time scenarios: prompt, slightly delayed, and significantly delayed times of arrival, plus the case where no valid pulse is present. After formatting by the IGLOO2, data from each QIE card are transmitted to off-detector backend electronics via 5 Gb/s VTTx optical links.

5.2.7 Slow and fast-control systems

The control systems for the HB, HE, and HF were updated to support the upgraded frontend electronics. For the Phase 1 upgrade of the frontend, the fast-control system, synchronized with the LHC clock, and the slow-control system (also referred to as the DCS), which runs at a much slower frequency than the LHC clock, are handled together within the same hardware. The fast-control system delivers the LHC clock with a maximum jitter of 1 ns, sends the orbit synchronization signal, delivers the so-called "warning test enable" (WTE) signal for the recording of calibration data, provides reset capabilities, and provides fast monitoring.

These requirements are similar for the HB, HE, and HF with a few exceptions: the HF has PMTs instead of SiPMs, and these do not need to receive configuration commands; the HB and HE have a

2024 JINST 19 P05064

secondary, redundant control link that can replace the primary control link; and each subsystem has a different number of channels and cards. As a result, a similar architecture is used for the HB, HE, and HF control systems, although the physical realization of the hardware is different. The redundant-link scheme has changed from what was described in the CMS HCAL Phase 1 Upgrade TDR [60]. The redundant link was removed from the HF because it is relatively accessible. In HB and HE, the hardware configuration has been simplified thanks to the installation of more optical fibers. The redundant HE control link was successfully used during the last year of Run 2, when a primary control link failed due to a malfunction of its optical transmitter [69]. For Run 3, the control link with the highest optical power between primary and secondary is used.

The hardware includes the ngCCM modules in the frontend, the ngFEC modules in the backend, and pairs of optical fibers linking the modules (each pair supports bidirectional serial communication), as shown in figure 31. The ngFEC module is a μTCA baseboard with mezzanine cards and pluggable optical transceivers (SFP+), and is based on the FC7 Kintex 7 FPGA AMC board [24]. Its main functions are to receive TCDS signals from the μTCA backplane, provide an interface to the DCS computers, merge fast control and slow control over the same bidirectional link used to communicate with the ngCCM, maintain a fixed latency for the fast-control signals across power cycles and ngFEC optical ports, make use of the ngCCM redundant scheme, and support up to twelve bidirectional links.

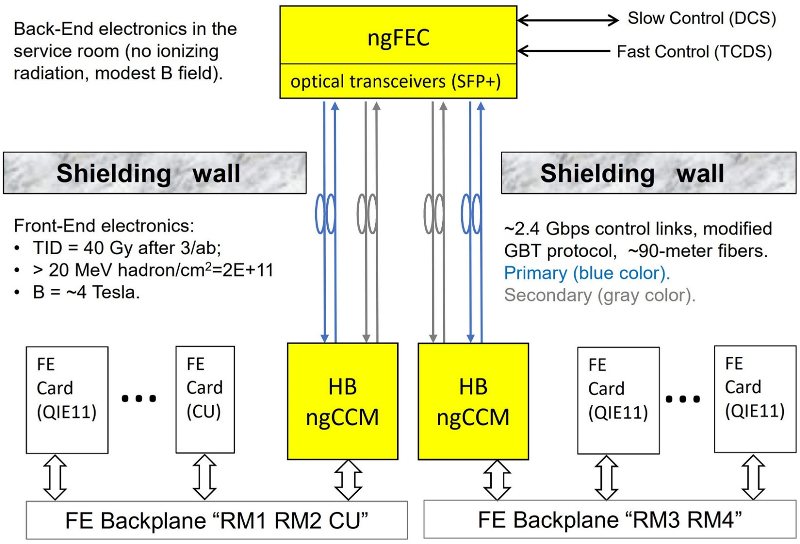


Figure 31. Block diagram of the HB controls. The ngFEC and ngCCM modules are needed to run and monitor the frontend electronics. All control links connecting the service rooms to the experimental cavern are optical. A secondary link connecting the ngFEC and the ngCCM is available in case of a primary link failure.

The ngCCM is connected electrically to the rest of the frontend over backplanes. The optical links run a modified scheme of the CERN-developed 4.8 Gb/s GBT communication protocol [68, 70]. The GBT protocol was tailored for the GBTx frontend ASIC; however, the GBTx was not available in time for the original HCAL Phase 1 schedule. In the ngCCM, the Microsemi IGLOO2 FPGA was used as a replacement for the GBTx. The performance of the dynamic characteristics of the IGLOO2 FPGA are marginal at 4.8 Gb/s; for CMS purposes, to allow an ample safety factor, it was decided to modify the protocol and run it at half speed. The reduced bandwidth is still sufficient for

2024 JINST 19 P05064

all control requirements of the HCAL frontend. The Phase 1 upgrade of the HCAL frontend and backend was successfully completed in October 2019.

5.2.8 Backend electronics: readout

During Run 2, the HCAL backend electronics were upgraded from VME-based to components based on the μTCA architecture. This replacement was the first step in the Phase 1 upgrade, and the μTCA-based electronics have operated with a mixture of the original and upgraded frontend throughout Run 2. The design of the backend is described in detail in ref. [60].

In Run 3, the backend and frontend host a much-increased channel count and data volume. With the completion of the Phase 1 upgrade, the number of HB (HE) channels is increased by roughly a factor of 3.5 (2.6), and each channel has a larger number of bits from the upgraded QIE10/11 ADC and TDC. This increase required efficient algorithmic development to calculate the trigger primitives within latency constraints and with the available resources of the μHTR FPGA. This firmware was achieved, along with the enhancements to the trigger primitive calculations discussed in section 5.3.

The HCAL data acquisition system is managed by an online software suite that controls, configures, and monitors the HCAL electronics. Development of the HCAL online software began in 2005, and it continues to evolve significantly; for example, new components were recently developed to support the upgraded frontend and backend electronics. The software has two main layers that provide complementary functionality. The first layer is composed of custom software written in C++ using the XDAQ [71] framework for distributed data acquisition. For communication with the μTCA electronics, the IPBus architecture [72] is used within XDAQ. Configuration parameters are provided in human-readable format by a version-controlled system, and larger hardware parameter sets (e.g., SiPM bias voltages, phase delays, etc.) are read from databases. The second software layer allows for coordination between different parts of the HCAL, and with other CMS subsystems. It is written in Java, and based on the RCMS framework [73]. The XDAQ and RCMS frameworks are described in detail in section 9.10.

The frontend electronics are controlled and configured by the ngCCM server, a separate program based on the C++ actor framework [74]. It detects and corrects for single-event upsets automatically, and stabilizes the SiPM temperatures by controlling the Peltier voltages. The ngCCM server communicates via IPBus with special command processors implemented in the ngFEC firmware for each I2C and JTAG channel in the frontend backplanes. The XDAQ-based software communicates with the ngCCM server via JSONRPC over WebSocket [75, 76] to control, configure, and monitor the frontend electronics. The HCAL DCS WinCC-OA communicates with the ngCCM server via a raw socket protocol to obtain voltage, current, temperature, and humidity readings and set the Peltier target temperature. A command-line interface to the ngCCM server additionally allows updates to be uploaded to the frontend FPGAs.

5.2.9 Voltage source upgrades

The Run 1 HCAL frontend electronics power system was based on the CAEN Easy infrastructure. The RBX electronics were fed with two power lines, nominally at 6.5 and 5 V, requiring a total power of around 90 W per RBX. The HPD high voltage (HV) and bias voltage (BV) and the HF PMT HV were provided using power supplies custom made in Bulgaria. The HB, HE, and HO RBXs were powered using CAEN A3016 power supplies, each module having six power channels

2024 JINST 19 P05064

capable of powering three RBXs. The HF readout crates were powered using four CAEN A3100 modules for each HF module (positive and negative η). Successive photodetector, frontend, and backend upgrades required some changes to the powering system.

The HO photodetector upgrade, with the replacement of HPD the sensors with SiPMs, required new power supplies, since the new frontend generates the BV internally from the low voltages (LV). No change was required for the LV system, since the A3016 modules were able to provide the extra power demanded by the SiPMs.

The HF LV system for the new frontend electronics also required some changes. Unlike the old electronics, the new electronics only use one supply voltage, in the range 8–10 V, but with more power. During Run 2, the power was provided by eight A3100 units operating at 8 V, four each of the two HF detectors. Each of the sets was in an CAEN EASY crate fed with an individual A3486 power converter. However, a problem was observed during operation, correlated with instantaneous luminosity. The power supplies, located in the HF racks near the detector, were subject to high levels of radiation, causing occasional single-event upsets (SEUs), i.e., incidents in which the control of the modules was lost, leading to the loss of power in the frontend electronics. This resulted in the loss of about $150\,\mathrm{pb}^{-1}$ of data in 2017. For 2018, a mitigation protocol was developed, consisting of a fast software detection, reset, and recovery of the power supplies, followed by a reconfiguration of the frontend electronics. The protocol reduced the time of data loss to typically 1–2 minutes, and reduced the total data loss to about $50\,\mathrm{pb}^{-1}$. The rate of SEUs scaled linearly with instantaneous luminosity, with about 20 SEU events in 2018; further, the power cycling was expected to shorten the lifetime of the power supplies, and the high level of irradiation made maintenance prohibitively difficult. Hence a more robust solution was implemented in LS2: the power supplies were moved to an area under CMS where radiation is minimal. Long cables from this location lead the power to the frontends. To compensate for the voltage drop on the cables, the A3100 power supplies have been replaced by A3100HBP units, described below, which can provide a high enough voltage to ensure the frontends receive the nominal operational voltage of 8 V. Following these changes, no SEU incidents were observed during the first year of Run 3.

The HE and HB upgrades, with the increase in the number of SiPM readout channels, led to more changes to the powering system. The HB and HE custom made power supplies were replaced by Keysight N6700C mainframes with N6736B modules. Twenty N6700C units currently provide the BV for the HB and HE SiPMs. The HE A3016 modules were replaced by A3100HBP units. The CAEN A3100HBP model provides one channel with 8–14 V, 50 A, and 600 W output and can power up to two HE RBXs in parallel. These units are also used in the HF, with ten units operating at 10 V used in each half of the HF. A transient voltage spike in one of the HE power supplies, caused by a power cut in June 2018, damaged the inputs of two RBXs. This affected HEM15 and HEM16, which cover 2% of the total HCAL acceptance and 3% of the HB+HE. After identifying the cause of the problem, CAEN introduced a modification to the power supplies to prevent such transients from occurring again. Independently, the HCAL Collaboration also developed an external circuit to suppress such transients. Currently, all the HE RBXs are protected with both the CAEN modification and the external overvoltage protection circuit.

The extra readout channels of the HB SiPMs required more power than what was practical with the A3100HBP units, and Wiener Marathons (PL 508 with five channels 5–15 V/40 A) were chosen as replacements. Eight Marathons are used to power the entire HB frontend electronics. To prevent

2024 JINST 19 P05064

damage due to transients, overvoltage protection circuits similar to the ones for the HE are installed on each of the Marathon output channels. New LV cables between the supplies and the RBXs were also installed to keep the voltage drop coming from the additional power within the safe operating range of the frontend electronics.

5.2.10 Photodetector and system calibration instrumentation

Changes in the light yield are expected due to aging of the scintillators, radiation damage, and variations in the photodetector response. The HCAL calibration and monitoring systems are designed to determine the absolute energy scale, to monitor the calorimeter system for changes during the lifetime of the detector, and to derive the energy scale correction factors. In particular, two complementary methods are used for monitoring. The first method consists of a movable radioactive wire source, with source tubes installed in every megatile in such a way that the tubes cross all of its tiles. Measurements for all tiles can be made during long LHC shutdowns to validate the readout-channel mapping. This system was used during the Phase 1 upgrade to provide a relative calibration when the SiPMs replaced the HPDs. The second method is based on light injection from two sources, a UV laser and an LED system. The UV laser light is injected into two layers of each wedge to monitor radiation damage to the scintillator, and both the UV laser and LED light is injected into the optical decoder box that has the photodetectors, to monitor damage to the SiPMs. A calibration module (CM) inside the frontend RBX is responsible for delivering the UV laser and LED light to the photodetectors. The CM was redesigned as part of the Phase 1 upgrade.

Each RBX contains one CM, which includes an LED pulser for the SiPMs, as well as a laser light distribution system. Fibers from the CM are connected to each of the interstitial microfibers in the light mixers, allowing for the LED or laser signals to be delivered to the SiPMs. The CMs also contain pin diodes which directly measure the amplitude of the LED or laser signals, providing a reference value for the calibration. The design of the LED pulser board is the same as that of the original HB/HE calibration module. The controls for the pulser, however, are significantly enhanced. In particular, the CMS global WTE signal is used by the CM to generate LED pulses during the LHC orbit gap. The LED pulse can be positioned between 2 and 65 535 triggering clock synchronization counts (MCLK) from a resetting WTE signal. This pulse can be delayed relative to the MCLK signal from 0 to 25 ns in 0.5 ns increments. The pulse amplitude and width are programmable from about 85 mV to 4.75 V and 0.5 ns to 25 ns, respectively. The pulser board contains six TE Connectivity AMP 147323-1 connectors providing a total current of 0.75 A at 5.5 V DC and a ground for powering the PIN diodes. The pulser board receives communication and MCLK from the ngCCM through the backplane header.

5.2.11 HF upgrade

When shower particles pass through the HF quartz fibers, Cherenkov radiation is produced. However, anomalous signals can also be produced by muons from pp collisions or beam halo interactions whose trajectories pass in the vicinity of the RBXs. Relativistic muons generate Cherenkov radiation when they pass through the glass window of the PMTs. Since the sampling correction for the HF calorimeter is large, these "window events" result in very large signals. Signals produced by muons can be distinguished from shower light through timing: the shower signal is delayed relative to the muon signal due to the longer path length and the lower speed of light in quartz. The PMTs used in the original HF detector were Hamamatsu Corporation R7525s, with a glass window that is 1.2 mm thick at the center, increasing to 6 mm at the edge. The HF PMT upgrade replaced these with Hamamatsu R7600-M4s. The new PMTs have thinner windows (1 mm of UV glass) and four anodes arranged in a 2×2 grid. The light from the fibers of a given ieta-iphi tower is spread out over the face of the PMT and generates signals on all four anodes; the diagonal anodes in the 2×2 grid are grouped together and read out by a single QIE10 chip. The resulting dual-anode readout allows for the discrimination of signals from muons interacting directly with the PMT, which will typically produce a signal concentrated on a single anode closest to the impact point. The amplitude of signals from muon window events between the old and new PMTs is compared in figure 32.

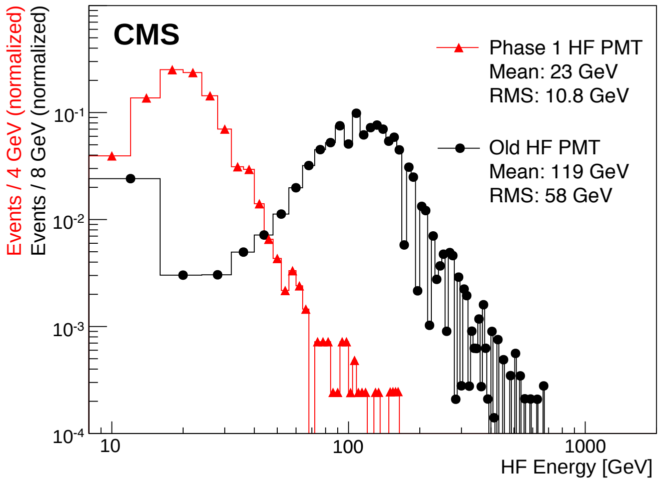


Figure 32. Cherenkov signals generated in the PMT windows from a muon test beam. Thin windows in the Phase 1 upgrade four-anode PMTs produce smaller signals (red) than those produced in the original PMTs with thick windows (black).

The new PMTs were extensively tested in several beam tests [77, 78] and in CMS during data taking. Algorithms utilizing the signal imbalance between the anodes were proposed and tested during the beam tests. However, a new readout system was needed to take advantage of the additional channels. To be cost-effective, frontend readout cards were redesigned for two-channel readout of the four-anode PMTs (where each readout channel sums two anodes). The backend was also changed from the old VME-based to the new μTCA-based system.

The new frontend cards included the upgraded QIE10 ASIC, which is similar to the QIE11, except for having a constant 20 Ω input impedance and no programmable current shunting. As described earlier, the QIE10 chips also have a TDC to measure the arrival time of the signals. Each frontend readout card for HF contains 24 QIE10 ASICs, each digitizing a single dual-anode PMT channel, as well as two Microsemi IGLOO2 FPGAs and three VTTx modules providing six 4.8 Gb/s optical links. Similar to the upgraded electronics in HB and HE, the power is supplied by CERN FEASTMP DC-DC converters. In addition to the replacement of the PMTs and the readout electronics, improvement in data transfer was planned to handle the increased load due to the TDC and two-channel readout.

The HF calorimeter is in a high-radiation area, leading to doses of 1000 Gy in fibers near the beam pipe. To monitor radiation damage, some of the quartz fibers are equipped with a special laser system that allows both the incident laser light intensity and the intensity of the laser light after traversing the quartz fiber to be measured with the same PMT. In Run 1, the light was provided by a laser system located in the underground service cavern, which is several meters away from the HF calorimeters. The light was then split into several fibers before it reached the HF quartz fibers.

A new laser device was developed that utilized solid state laser diode technology with a wavelength of 450 nm (allowing for the removal of the wavelength shifters used in the old system), and is based on the HF Phase 1 LED calibration unit concept. The laser diode driver and the control circuit reside on a mezzanine card mounted on the QIE board. Optics mix the light and distribute it to four output fibers. The module is installed in the HF frontend crates, which are attached to the HF calorimeters. Figure 33 (left) shows a schematic view of the laser daughter board as part of the QIE frontend card. The upgraded HF light distribution system for monitoring radiation damage is shown in figure 33 (right). The settings of the new system are more flexible, since it is an integral part of the upgraded frontend, allowing for, e.g., a 32 ns delay range adjustable in 0.5 ns steps, and a pulse width adjustable from less than 3 to 30 ns full-width at half-maximum. There is also an on-board PIN diode to independently monitor the performance of the light source.

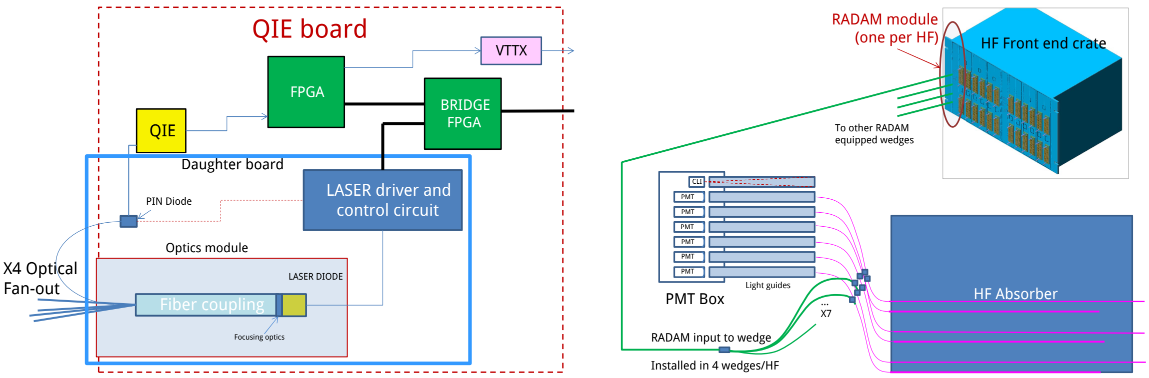


Figure 33. Left: schematic of the new HF QIE laser board, used in the HF radiation monitoring system. Right: sketch of the upgraded HF radiation damage monitoring system light distribution and electronics chain.

5.3 Trigger

The HCAL trigger is based on the HCAL calorimeter towers, which, in the HB and HE, are composed of projective calorimeter readout channels summed together in depth. While most readout channels have $\Delta\phi = 5^\circ$, the HE readout channels with $|\eta| > 1.74$ have $\Delta\phi = 10^\circ$. To form trigger towers with $\Delta\phi = 5^\circ$, thereby matching the rest of the HCAL, these HE readout channels are split into two in ϕ, with each unit assigned one-half of the measured energy. In the HF, trigger towers are formed from six physical HF towers, with energies of the long and short fibers summed.

The HCAL trigger primitives (TPs) are formed for each trigger tower by combining the information from individual calorimeter towers. Since the HCAL QIE10/11 chips use 8-bit nonlinear ADCs with several ranges, the energy reported by the QIE is first linearized as a 10-bit word via a dedicated look-up table. The linearization depends on several parameters, including the gain, the conversion from the collected charge in fC to ADC counts, and the pedestal. After linearization, the various depth samples are summed, and the amplitude of the pulse is estimated using a pulse filtering scheme that

2024 JINST 19 P05064

subtracts out-of-time pileup. These sums define the various trigger channels, and a corresponding TP is produced by the μHTR. Finally, the trigger tower transverse energies are compressed to the range 0 to 128 GeV with a least significant bit (LSB) of 0.5 GeV, and transmitted to the L1 trigger.

The TP amplitude is reconstructed by taking the sample-of-interest (SOI) bunch crossing as the main contribution to the total signal amplitude. The measurement from the SOI alone works well for HF, where the pulses are shorter than the 25 ns bunch spacing, but for HB and HE, a significant fraction of the pulse leaks into the subsequent bunch crossing. In Run 2, the sum of energies in the SOI and SOI+1 was used to reconstruct the TP energy; this method accounts for the signal leakage into SOI+1, but also incorporates the energy from out-of-time pileup interactions in SOI+1. For Run 3, a new algorithm was developed that instead uses the SOI and SOI-1, i.e., the preceding bunch crossing. The scheme subtracts a weighted amount of the measurement from SOI-1, mitigating the inherent leakage of the out-of-time pileup from SOI-1 in the SOI. To account for the signal leakage into SOI+1, a correction factor is derived from the known pulse shapes rather than using the measurement from SOI+1, thereby avoiding incorporating contributions from out-of-time pileup.

In addition to the TP generation in the HB and HE, six feature bits of information are also generated that can be transmitted to the L1 trigger. These bits facilitate encoding information about: (i) the longitudinal shower profile data for use in calibration, lepton isolation, and identification of minimum ionizing particles, and (ii) the shower time data constructed from the TDC information available in each constituent channel of the trigger tower. Configurable look-up tables determine which time windows within the bunch crossing of interest are represented by the available TDC codes. These time window boundaries have a granularity of 0.5 ns. Starting in Run 3, a subset of the feature bits is used to flag signals characteristic of exotic long-lived particle decays, using either the TDC timing to mark hits with late arrival times or the shower profile data to mark distinctive energy deposits in the various layers of the HCAL. These bits are used by the L1 trigger to select hadronic signatures from long-lived particles with decay lengths of 1–2 m which decay prior to or within the HCAL. The timing precision that can be achieved is estimated to be within 1–2 ns. The resolution is dominated by interchannel synchronization uncertainties and shower-by-shower fluctuations, as determined from studies done using highly energetic hadronic showers in 2018 data.

5.4 System and beam tests

Beam tests were performed with the upgraded Phase 1 frontend at the H2 beam line of the SPS at CERN. The upgrade electronics were installed in the 20°-prototype HE wedge at H2, and were read out by four RMs split between two RBXs. One RBX and two RMs were replaced with the upgraded electronics, while the remaining ones were left unchanged as a reference. The system was tested with 150 GeV muons and with pions having energies ranging from 30 to 300 GeV. Six time samples were recorded for each event, with the beam timed to arrive in the fourth time sample. The total charge for an event was taken as the sum of the last three time samples centered around the beam arrival. The pedestal for each event was estimated from the sum of the first three time samples before the beam trigger and was subtracted from the total charge.

The 2015 beam tests served as the first full test of the entire upgraded readout system and proved the system was functional. The 2017 test beam with production frontend electronics served to quantify the final performance of the upgraded detector. Representative results from this data are shown in figure 34. Using the muon data, the response of the detector instrumented with SiPMs was measured to be 4.3 photoelectrons per layer. The pion data were used to derive the energy response and resolution of the upgraded detector. For pions, the number of photoelectrons per GeV was measured to be 32.1. The shower energy was measured by taking the sum of all charge collected by each channel in a 3×3 grid of towers around the beam direction. The relative contribution of each detector channel to the energy sum was adjusted using its response to muons to account for differences in the detector response between channels. The beam test was also used to make a unique measurement of the shower profile as a function of depth with a special ODU that allowed the individual readout of each layer in one tower. These studies were particularly important because they measured the maximum energy that can be deposited in any particular layer.

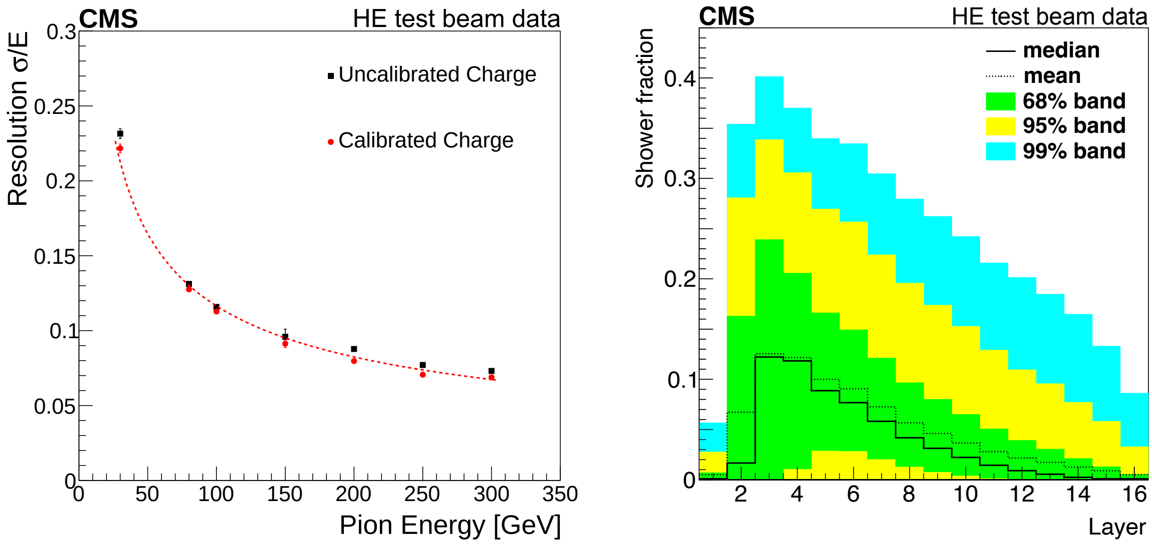


Figure 34. Left: energy resolution of the upgraded prototype detector as a function of the pion beam energy, shown with and without the channel calibrations derived from the response to muons. Right: longitudinal shower profile measured using the special ODU. Bands containing 68, 95, and 99% of all events for each layer are shown.

Irradiation tests of the full suite of frontend electronics were performed at the CERN High Energy Accelerator Mixed-Field (CHARM) facility with the goal of determining the expected error rates from SEEs, as well as the expected lifetime performance related to the cumulative edamage from ionizing and nonionizing radiation. The electronics were subjected to an ionizing dose of 202 Gy, a 1 MeV neutron equivalent fluence of 1.96×10^{12} cm^{-2}, and a > 20 MeV hadron equivalent fluence of 5.88×10^{11} cm^{-2}. At the end of the irradiation, the frontend electronics showed no indication of permanent damage, aside from the expected inability to program the FPGAs. The SEE rates were determined based on bit error counting in the encoded data stream and the pseudo-random bit sequence in the control link. The combined SEE rate was estimated to be less than 30 per fb^{-1} of data. Other upsets requiring a reset or other interventions are estimated to occur at a rate of about one per day.

5.5 Performance

5.5.1 Endcap photodetector performance in Run 2

The photoelectron peaks in the pedestal and low-intensity LED runs, shown in figure 35 (left), can be used to measure the SiPM gains using a single fit function to the observed multipeak charge spectrum. The gain was found to be stable during the 2017 data taking at the 1% level, corresponding to 42 ± 1 fC per photoelectron, where the uncertainty is the RMS across the 184 channels equipped with SiPMs. The SiPM dark current was also monitored during data taking, as shown in figure 35 (right). The slope of the fitted line is proportional to the SiPM area. The deviations from linearity are due to SiPM annealing when the beam is off and to variations in instantaneous luminosity. With this SiPM dark current growth rate, we expect 110 MeV of noise at the end of Run 3 (500 fb^{-1}).

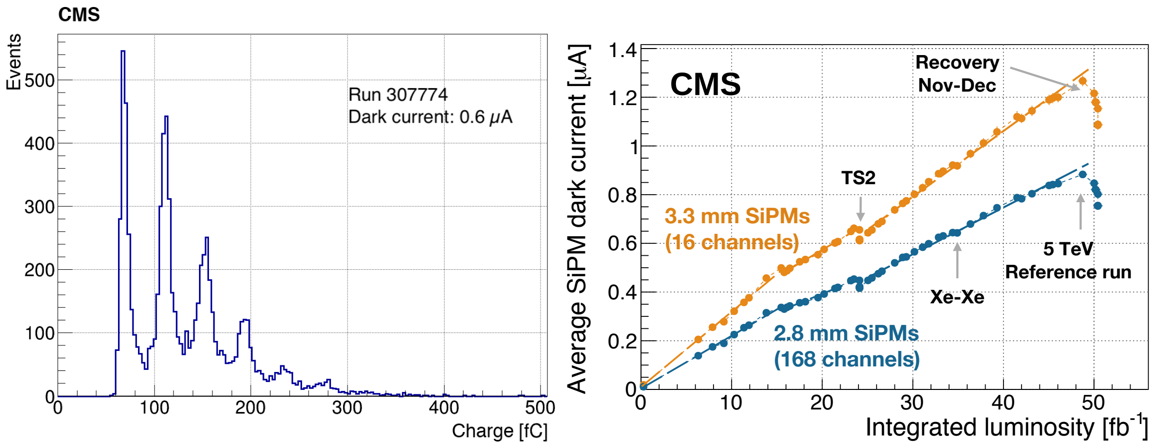


Figure 35. Left: pedestal distribution for a channel with QIE11 + SiPM readout. The charge is integrated in a time window of 100 ns. The QIE pedestal and photoelectron peaks are visible. Right: dark current increase with the integrated luminosity in 2017, where the slope of the fitted line is proportional to the SiPM area. The deviation from linear behavior is due to SiPM annealing in the absence of beam and variation in the instantaneous luminosity.

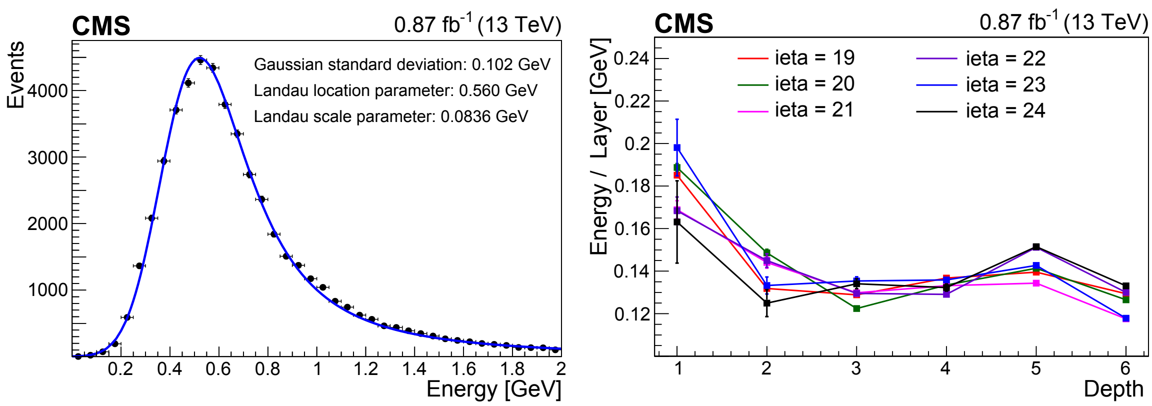


Figure 36. Left: energy deposit from muons in pp collision events in the HE tower corresponding to ieta = 20 and depth = 5. The energy spectrum is fitted using the convolution of a Gaussian function with a mean of zero and a Landau distribution. The fitting function has three free parameters: the Landau location parameter, the Landau scale parameter, and the width of the Gaussian. Right: the most probable value of the muon energy deposit per layer in HE towers as a function of depth for different η regions. The vertical bars represent the statistical uncertainty. Muons from collision events are considered when their trajectory is contained within a single HCAL tower. The muon signal peak is fitted with the convolution of a Gaussian and Landau functions. The Landau location parameter is divided by the number of scintillator layers in the considered depth.

2024 JINST 19 P05064

In Run 1 and Run 2, the HCAL calibration used E/p from isolated tracks for the energy scale measurement versus η and ϕ-symmetry for equalizing the response along ϕ [54]. After the Phase 1 upgrade, new calibration techniques are needed because of the increased depth segmentation. The scintillator and SiPM aging depends on η and depth, which necessitates depth intercalibration. The signals from isolated, minimum-ionizing muons traversing the HCAL provide an excellent probe for interdepth calibration, corresponding to approximately five photoelectrons per layer. As shown in figure 36 (left), a clear minimum ionizing peak from muons can be observed in collision events when the muon traverses one and only one HCAL tower. The energy deposit per layer is shown in figure 36 (right). The HE detector is homogeneous in depths 2–6, while depth 1 has a thicker and brighter scintillator layer. Because of this, only depths 2–6 are equalized.

5.5.2 HCAL performance in Run 2

For the combined ECAL and HCAL systems, including barrel and endcap, the relative charged pion energy resolution obtained from the test beam can be described as

$$\frac{\sigma}{E} = \frac{84.7\%}{\sqrt{E}} \oplus 7.6\%, \tag{5.1}$$

where E is in GeV. Corrections for the nonlinearity of the calorimetry system due to its noncompensating response to hadronic and electromagnetic energy depositions have been made [79]. The time resolution for energy deposits in the HB and HE, calculated by weighting the QIE digitization times by the associated energies, is 1.2 ns [80]. The HF energy resolution from the test beam [52] is

$$\frac{\sigma}{E} = \frac{280\%}{\sqrt{E}} \oplus 11\%. \tag{5.2}$$

The absolute calibration constants derived from the test beam modules were transferred to the full calorimeter system using an intercalibration from a radioactive ^{60}Co source [64]. The calibration was improved using 13 TeV collision data from 2016, as described in ref. [54]. From the ϕ-symmetry of energy flow in minimum-bias events, the HCAL was intercalibrated to within 3%. An absolute calibration uncertainty of 2% was determined using isolated pions with track momenta between 40 and 60 GeV showering in the HCAL. The energy resolution was found to be 19.4, 18.8, and 23.6% in the HB, HE, and transition region, respectively. Signal loss due to radiation damage is relevant in the HE at high $|\eta|$. It is monitored using a laser calibration system and the response to a ^{60}Co source, and is cross-checked with hadrons and muons from pp collisions [64].

During Run 2, new algorithms for reducing the effect of out-of-time pileup on reconstructed HCAL energies were developed. The HCAL response to incoming particles rises to its maximum within 10 ns, followed by an exponential decay, with 90% of the pulse contained within two 25 ns time samples (TSs). In Run 1, the bunch spacing was 50 ns, and the energy of hits was reconstructed by a simple sum of charges in the SOI and SOI+1 after the contribution from leakage currents was subtracted, with a correction factor applied to account for the tail extending beyond two TSs. In Run 2, the bunch spacing changed to 25 ns, and hence this method was expected to yield a poor energy resolution, since it incorporates contributions from pulses from preceding bunch crossings that overlap with the pulse from the SOI. The new algorithms developed for Run 2 subtract the energy of out-of-time pileup using pulse template fits.

In 2016–2017, two separate pulse-template fitting algorithms, referred to as Method 2 and Method 3, were deployed for offline reconstruction and in the HLT, respectively. The most recent algorithm, called “minimization at HCAL, iteratively” (MAHI), was developed and deployed for data taking in 2018 [81]. Notably, MAHI performs well enough to be executed within the HLT latency requirements, and, for Run 3, has also been ported to run on GPUs for further reduction in processing time. The MAHI algorithm was used for the legacy reprocessing of the CMS Run 2 data. Detailed results are presented in ref. [82].

5.5.3 HF performance in Run 2

The performance of the HF upgrade was evaluated during Run 2. Figure 37 (left) shows the distribution of the signal arrival time as a function of the collected charge [83]. Signals with a time around 7 ns are from showers in the calorimeter, while those at earlier times are due to Cherenkov radiation in the PMT window. The arrival time can thus be used to identify window events. While most signals due to PMT Cherenkov radiation come early, there is a tail to later times, due to imperfect synchronization. Elimination of such background events can be improved by comparing the signals obtained in both channels of the PMT, as shown in figure 37 (right). The charge asymmetry between the PMT channels is calculated as the difference divided by the sum of the signals in the two channels. For genuine events, this value should be close to zero. On the other hand, background events are observed well away from the central peak at zero since they are produced by stray muons hitting one quadrant or a side of the four-anode PMTs. The combination of the arrival time and asymmetry methods improves the elimination of these background events.

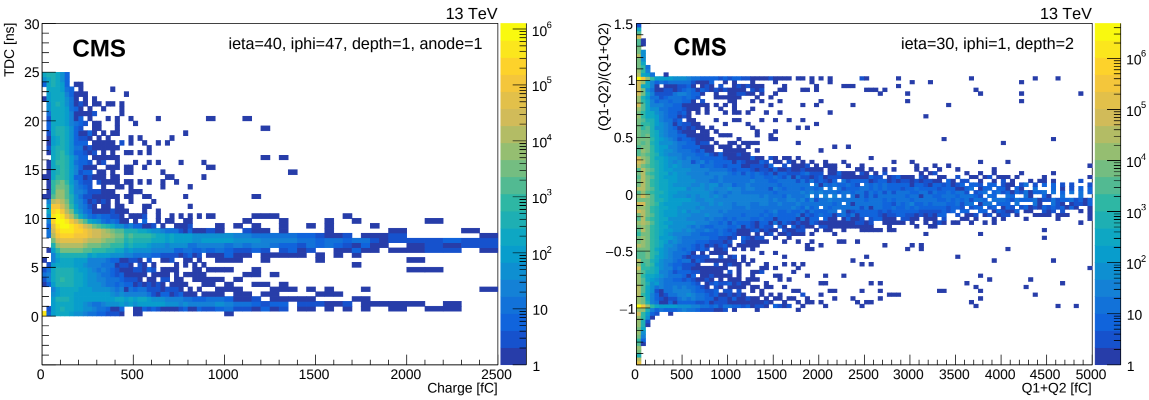


Figure 37. Left: HF signal arrival time, as measured in the TDC, versus the collected signal charge. All signals arriving within less than 5 ns are “window events”. The color indicates the number of events using the scale to the right of each plot. The data were taken in early 2017. Right: charge asymmetry between the two channels of a PMT versus the total charge in both channels. The light from genuine collision events is well mixed in the light guides before falling on the PMT, hence similar signals are expected in all four anodes, which are grouped into two channels. The so-called “window events” due to Cherenkov radiation in the PMT window most likely fall on one or two anodes, producing asymmetric signals.

Figure 38 shows the evolution of the missing transverse momentum with improvements to the HF anomalous signal identification based on the arrival time criteria (TDC filters), topological filters, and combined criteria (TDC, charge asymmetry, and topological filters). Topological filters have

2024 JINST 19 P05064

been used since the beginning of Run 1 and are based on ratios of energies in the long and short fibers. The new filters, based on the timing and ratios of the PMT channel energies, are as effective as this topological selection. The combination of all the anomalous signal reduction techniques gives the best performance. Additional topological filters based on the shape of jets versus η and ϕ were developed to reject additional noise that escapes these filters [84].

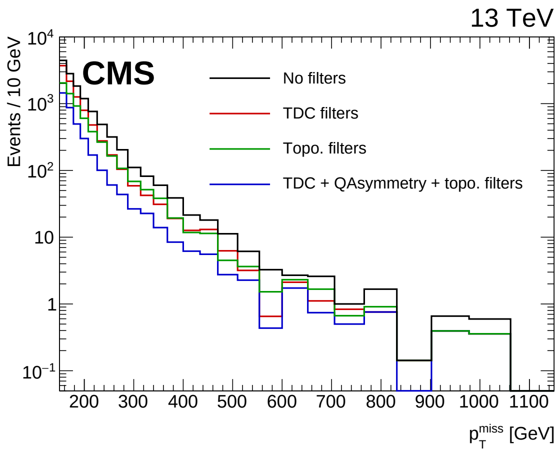


Figure 38. Effect of filters on the HF anomalous energy contributions to the missing transverse momentum measurement. The methods developed based on hardware improvements installed as part of the Phase 1 upgrade are as effective as the topological selections used previously. Including both the new and old filters further reduces the anomalous missing transverse momentum.

6 Muon system

A central feature of the CMS experiment is a powerful system for triggering on and detecting muons. In the previous runs of the LHC, muons have been crucial to many of the physics results of CMS and have contributed to hundreds of published results, including the discovery of the Higgs boson. The importance of muons in the CMS physics program continues to remain high in Run 3 and beyond.

The objectives of the CMS muon system are to identify muons, measure their momenta, and provide signals for triggering on them. These goals are achieved with four complementary detector systems arranged in the steel flux-return yoke of the CMS solenoid. These systems provide efficient detection of muons over a large range of pseudorapidity. The location in the magnetized steel behind the calorimeters and solenoid ensures a low probability of penetration to the muon detectors by particles other than muons and neutrinos.

The physical arrangement of the muon detectors is shown in figure 39. The central section is configured in a barrel geometry with four roughly cylindrical stations at different radii from the beam axis. The endcap section is arranged in four planar stations in z in each endcap.

The drift tube (DT) system in the barrel covers $|\eta| < 1.2$ and is composed of drift chambers with rectangular cells. The DTs provide precise spatial measurements, as well as trigger information. This system is described in more detail in section 6.1.

The cathode strip chamber (CSC) system in the endcap comprises multiwire proportional chambers having cathode strips with an R-ϕ geometry and covering the region $0.9 < |\eta| < 2.4$. The

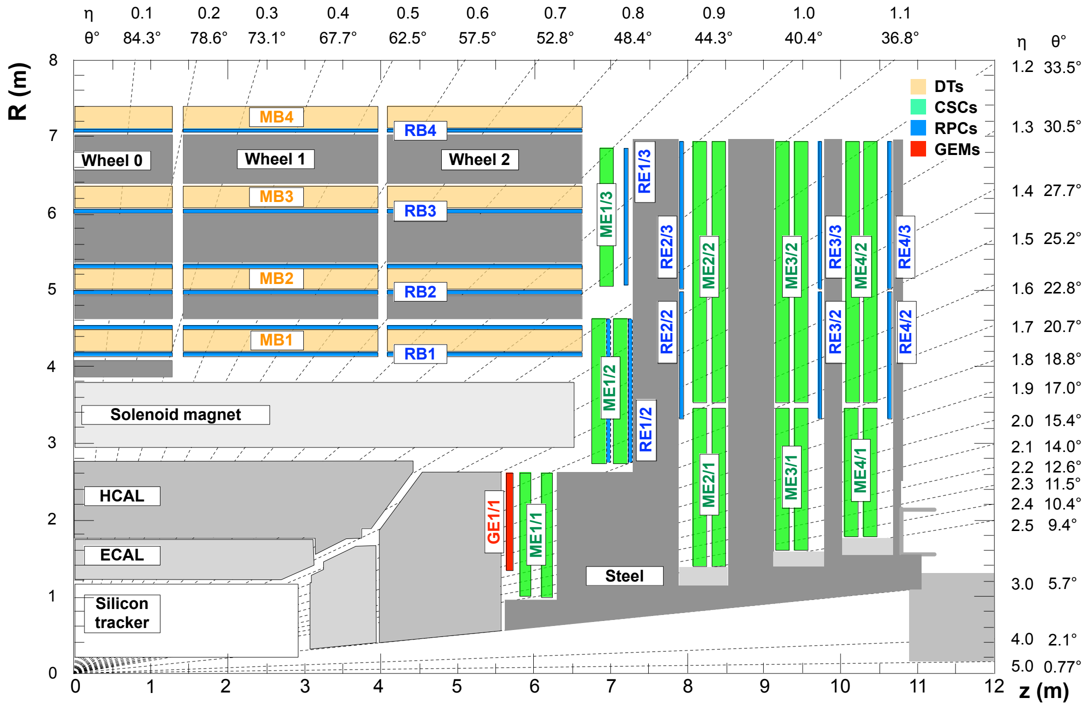


Figure 39. Schematic view in the r-z plane of a CMS detector quadrant at the start of Run 3. The interaction point is in the lower left corner. The locations of the various muon stations are shown in color: drift tubes (DTs) with labels MB, cathode strip chambers (CSCs) with labels ME, resistive plate chambers (RPCs) with labels RB and RE, and gas electron multipliers (GEMs) with labels GE. The M denotes muon, B stands for barrel, and E for endcap. The magnet yoke is represented by the dark gray areas. Reproduced from [8]. © 2018 CERN for the benefit of the CMS collaboration. CC BY 3.0.

CSC system provides both trigger and precision position information. Because of the higher flux of particles in the endcap region, the CSCs are designed to have a faster response time than the DTs. More details on the CSC system are found in section 6.2.

The resistive plate chambers (RPCs) are double-gap chambers operated in avalanche mode. The RPCs are located in both the barrel and endcap regions, and they complement the DTs and CSCs with a very fast response time that can be used to unambiguously identify the bunch crossing corresponding to a muon trigger candidate. The RPCs are further described in section 6.3.

Finally, GE1/1, a station of gas electron multiplier (GEM) chambers, is located in front of the inner ring of CSC chambers in the first endcap station. The GEMs have both a fast response and good spatial resolution, and they augment the CSCs in a region of high particle flux. More information about the GE1/1 station can be found in section 6.4.

The different groups of muon chambers are labeled with two letters: MB and RB for DTs and RPCs in the barrel region; ME, RE, and GE for CSCs, RPCs, and GEMs in the endcap regions. For MB and RB, these letters are followed by a single number (1–4) indicating the barrel index. Two indices follow ME, RE, and GE, where the first indicates the station in $|z|$ and the second indicates the ring in R. The value of each index increases with increasing distance from the center of the detector.

For each bunch crossing, fast trigger data ("trigger primitives") are sent from each muon detector system (DT, CSC, RPC, and GEM) to the dedicated level-1 (L1) trigger muon track-finder hardware in the barrel (BMTF), endcap (EMTF), and overlap (OMTF) regions. The L1 muon trigger is described in more detail in section 10.2.

The muon system is summarized in table 7, which gives the number of chambers for each subsystem, the number of readout channels, and the spatial and time resolution.

Table 7. Properties of the CMS muon system at the beginning of Run 3. The resolutions are quoted for full chambers, and the range indicates the variation over specific chamber types and sizes. The spatial resolution corresponds the precision of the coordinate measurement in bending plane. The time resolution of the RPC of 1.5 ns is currently not fully exploited since the DAQ system records the hit time in steps of 25 ns.

Muon subsystem	Drift tube (DT)	Cathode strip chamber (CSC)	Resistive plate chamber (RPC)	Gas electron multiplier (GEM)
$\lvert\eta\rvert$ range	0.0–1.2	0.9–2.4	0.0–1.9	1.55–2.18
Number of chambers	250	540	480 (barrel) 576 (endcap)	72
Number of layers/chamber	8 (R-ϕ) 4 (z, MB1–3)	6	1 2 (RB1, RB2)	2
Surface area of all layers	18 000 m^2	7000 m^2	2300 m^2 (barrel) 900 m^2 (endcap)	60 m^2
Number of channels	172 000	266 112 (strips) 210 816 (wire groups)	68 136 (barrel) 55 296 (endcap)	442 368
Spatial resolution	100 μm	50–140 μm	0.8–1.3 cm	100 μm
Time resolution	2 ns	3 ns	1.5 ns	<10 ns

The performance of the muon system in Run 1 and the first part of Run 2 is documented in refs. [8, 85]. Much of the muon system is unchanged from that used in Run 1 and described in 2008 [1], but there have been additions and improvements. Notably, there were three major additions to the endcap detector suite: the outer ring of CSC chambers in station four ("ME4/2"), the outer rings of RPC chambers in station four ("RE4/2" and "RE4/3", collectively "RE4"), and the GEM system in station one ("GE1/1"). In addition, there were important upgrades to the electronics and trigger in many subsystems, which are described in the corresponding sections.

6.1 Drift tubes

6.1.1 General description

Drift tubes (DTs) equip the barrel part of the CMS muon detector, serving as offline tracking devices and providing standalone trigger capabilities. The basic DT detector unit is a rectangular drift cell with a transverse size of $4.2 \times 1.3\,\text{cm}^2$, whose layout is shown in figure 40 (left). A gold-plated stainless steel anode wire, with a diameter of 50 μm, is located at the center of the cell, and cathode strips are placed on its side walls. Additionally, electrode strips are located at the top and the bottom of each cell to shape the drift field. The cathodes and electrode strips are set at a voltage of −1200 and 1800 V, respectively, whereas the anode wires operate at applied voltages that vary between 3500 and 3600 V, depending on the individual chambers (more details are discussed in section 6.1.3). Cells are filled

2024 JINST 19 P05064

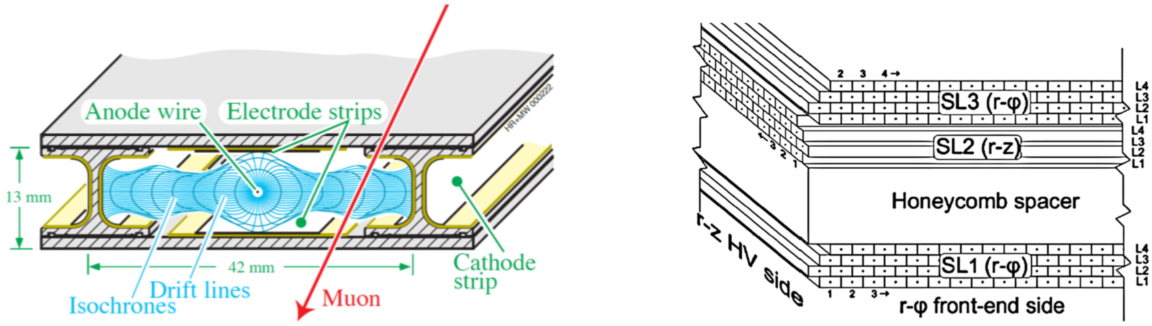


Figure 40. Left: layout of a CMS DT cell showing the drift lines and isochrones. Reproduced from [1]. © 2008 IOP Publishing Ltd and SISSA. All rights reserved. Right: schematic view of a DT chamber. Reproduced from [86]. © 2010 IOP Publishing Ltd and SISSA. All rights reserved.

with a gas mixture of 85% Ar and 15% CO_2, which offers good quenching properties and, under the operational conditions described above, is characterized by a saturated drift velocity around 55 µm/ns.

A schematic layout of a DT chamber is shown in figure 40 (right). Within a chamber, cells are arranged parallel to each other to form layers (L), and groups of four layers, staggered by a half-cell, form superlayers (SL). Each DT chamber is equipped with two SLs that measure the muon trajectory along the bending plane (R-ϕ). Additionally, chambers from the three innermost detector stations host an SL that measures the position along the longitudinal (r-z) plane as well. A total of 250 DT chambers, covering a pseudorapidity range up to $|\eta| < 1.2$, are arranged in five wheels with an identical layout. The wheels are placed parallel to each other along the CMS global z axis, and are labelled W−2, −1, 0, +1, and +2. Within each wheel, chambers are organized in four concentric station rings, labelled from inside-out as MB1 to MB4, and segmented into 12 sectors (S) along the CMS global ϕ coordinate.

The performance of the DT system, measured over Run 1 and Run 2, was found to be remarkably stable and in line with the design expectations. This is documented in refs. [8, 85]. Track segments are typically reconstructed offline with an efficiency above 99%, and they are characterized by spatial and time resolutions around 100 μm and 2 ns, respectively. The efficiency to reconstruct a standalone DT segment in the trigger (also called a trigger primitive) and to correctly identify its bunch crossing (BX) of origin is above 95%. The position (direction) resolution of the DT trigger segments is approximately 1 mm (5 mrad).

Several upgrades, mostly concerning the off-detector electronics, occurred after the end of Run 1 to cope with the twofold increase in LHC instantaneous luminosity during Run 2 with respect to its original design. They are part of the CMS Phase 1 upgrade [87] and are described in section 6.1.2.

Even if a replacement of the DT on-board electronics will be required as part of the Phase 2 CMS muon system upgrade [88], the DT chambers themselves will operate unchanged throughout the HL-LHC period. For this reason, strategies to extend the detector longevity, and maximize the DT performance over a period longer than the one originally expected, were put in place in Run 2 and LS2. They are documented in section 6.1.3.

6.1.2 Phase 1 upgrades of the DT electronics

The increase in instantaneous luminosity during Run 2 required some changes to both the trigger and readout chains, particularly to the electronics hosted in the balconies surrounding the DT wheels in the experimental cavern, the so-called sector collector. To accommodate the schedule of access opportunities in LS1 and the end-of-year technical stops, and to benefit also from the latest digital technologies, these upgrades took place in multiple stages: first a relocation of the sector collector and, subsequently, the trigger and readout upgrades.

On the other hand, with a single exception, no upgrade of the on-board DT detector electronics [1] was performed or is foreseen until LS3. This holds true for both the frontend (FE) electronics and the HV distribution, which are physically embedded in the chamber gas volume, and for the first level of the DT readout and local trigger electronics, which are hosted in aluminum structures attached to the DT chambers, called minicrates.

The only change that occurred in the minicrates was the replacement of 48 theta trigger boards (TRB), corresponding to the ones located in MB1 of W±2, which was performed over LS1. The new TRBs host a Microsemi FPGA that offers the same functionality as the ASIC-based boards originally installed on the detector. This replacement insures the availability of spares for the FE trigger components of the ASIC TRBs until the end of their operation.

A schematic view of the Phase 1 DT detector electronics architecture, as it was in Run 1 and is now in Run 3, is presented in figure 41. Details about this schema are described throughout the following sections.

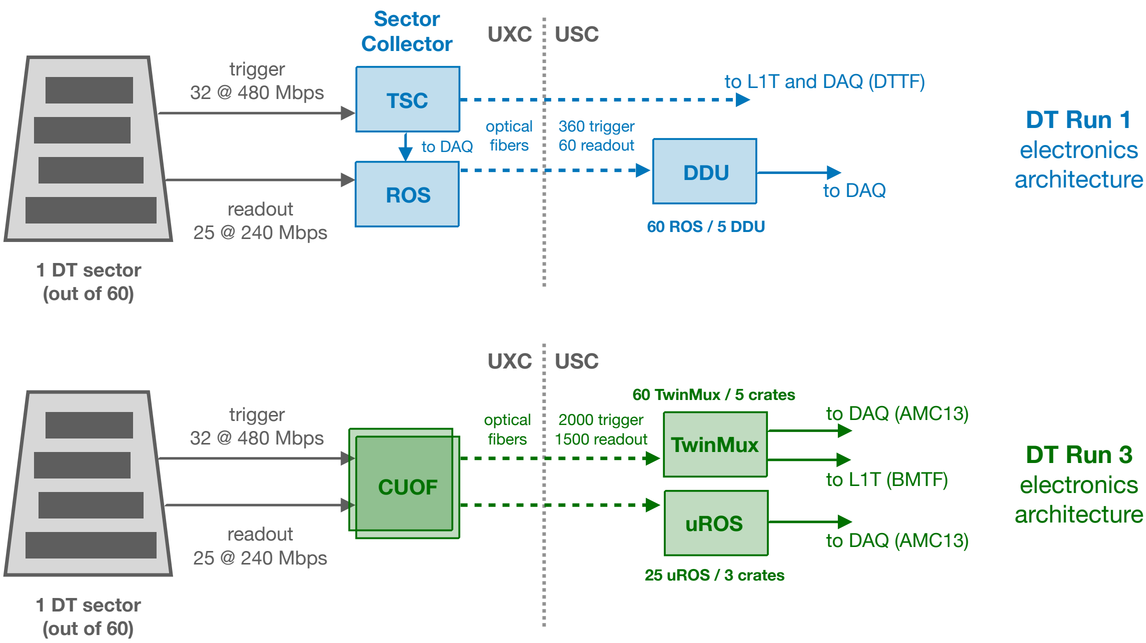


Figure 41. Schematic views of the Run 1 and Run 3 DT off-detector electronics architectures.

Relocation of the DT sector collector. The first stage of the DT electronics upgrade consisted of a relocation of the so-called sector collector (SC). In its original layout, the SC was composed of ten 9U VME crates, hosting the second level of the DT trigger, the trigger sector collector (TSC) and the readout, the readout server (ROS) electronics, as well as different boards in charge of slow control, monitoring, timing signal distribution, and power distribution. The SC was located on the balconies surrounding the detector in the experimental cavern (UXC).

The relocation project [89] consisted of moving the TSC and ROS boards to the service cavern (USC). As a consequence, the length of the readout and trigger links from the minicrates was increased by approximately 60 m. To cope with the increased length, a dedicated system that converts electrical signals from the chambers into optical ones (CUOF) was installed in the racks previously occupied by the SC. Data were then transmitted by means of optical fibers to the USC, where they were converted back to electrical signals by a second set of dedicated boards (OFCU) before being injected into the ROS and TSC.

The SC relocation brought an increase of operational reliability to the DT system. Moving the TSC and ROS boards to the USC made them always accessible during LHC running periods, allowing the prompt solution of possible problems without the need to wait for technical stops when access to the UXC was possible. It also paved the way for further stages of the Phase 1 DT upgrade, which led to an overall performance improvement. Firstly, the region in which the SC was originally installed was characterized by high radiation levels (up to 0.2 Gy per year in nominal LHC conditions). Without a relocation, these constraints would have led to a limitation in the choice of the electronics components that could be used for the upgrade to radiation-hard units. Secondly, a nonnegligible residual magnetic field (up to 40 mT) is also present on the detector balconies, imposing further restrictions on the use of magnetic components such as inductors and ferrites. The choice of cooling turbines used in the CMS balconies is also constrained by the need to operate in the presence of a magnetic field. If the SC had remained in the UXC, this would have put stringent limits on the power consumption of the upgraded system, since the power dissipation of the original SC had already challenged the cooling capacity of the turbines operating on the balconies. Thanks to the relocation, these constraints were relaxed, leading to more freedom in the design of the upgraded trigger and readout electronics.

The readout and trigger links from the minicrates are based on DS92LV1021 serializers from National Semiconductors, which have an embedded clock. Serial words of 12 bits (10 bits for payload and overhead) are clocked at 20 (40) MHz for the readout (trigger), resulting in a bit rate of 240 (480) Mb/s. Unless an L1 accept trigger is issued, the readout transmits an idle payload that was designed to maximize the DC balancing, resulting in a 40% duty cycle. Instead, in the trigger payload, streams of zeros are preferentially transmitted unless trigger segments are built. This results in a significant DC imbalance in the trigger links, which can be tolerated by operating the CUOF optical transmitters in a low-bias mode, among other improvements upstream, e.g., those to the TwinMux concentrator described below.

A CUOF board consists of a 9U motherboard where four mezzanine cards are typically plugged in. Each mezzanine card carries out the conversion of information received from up to eight links, organized in two FTP cables from either the readout or trigger of one DT chamber. The electrical signals enter the CUOF from RJ45 connectors located at the front of the crate. They are then routed into a line equalizer, restoring levels and compensating for the distortions of the electrical transmission line, and finally are injected into a laser driver that controls the laser diodes. Vertical cavity surface emitting laser (VCSEL) diodes are used. There are eight VCSEL diodes for each mezzanine card, which are connected to a fiber fan-out using LC-type connectors. Each CUOF motherboard also hosts two A3P600L ProASIC3L FPGAs from Microsemi, which control the configuration and the monitoring of the laser drivers. Fine tuning of the drivers' bias and modulation settings is of prime importance to ensure correct transmission of the DC-unbalanced information from the trigger link.

2024 JINST 19 P05064

A total of ten CUOF crates (corresponding to two crates per wheel) is installed in the UXC balconies. Each of them hosts thirteen CUOF boards, six (seven) of them dedicated to the transmission of readout (trigger) information, and covering a total of six DT sectors. Two A3050 CAEN modules provide power to the system in each crate. Each of them delivers two independent power supply channels, thus, for a given wheel, four power partitions exist. The power consumption of the CUOF system corresponds roughly to half of that of the previous SCs. A picture of the two CUOF crates instrumented in the balconies surrounding W−1, is shown in figure 42 (left).

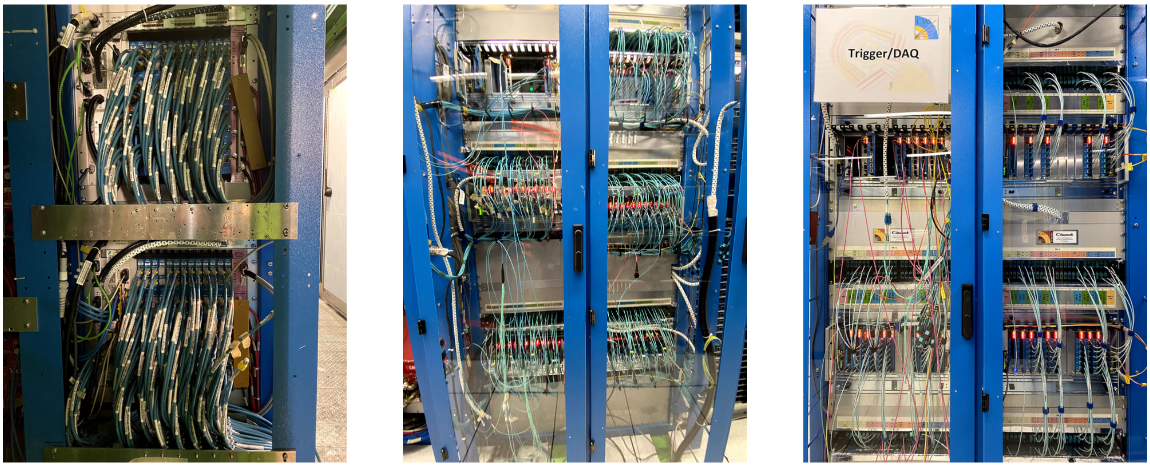


Figure 42. Left: front view of two out of the ten DT CUOF crates located in the UXC balconies surrounding the muon barrel (W−1). Center: front view of the five μTCA crates of the TwinMux in the USC. Right: front view of the three μTCA crates of the μROS in the USC.

Optical communication occurs as 850 nm transmission over OM3 multimode fibers with a 50 (125) μm core (cladding) diameter. Individual fibers are organized in MTP cords, each containing twelve fibers, which get further assembled in groups of eight to form trunk cables of 96 fibers. Considering the arrangement of fibers into MTP cords, as well as the need for spares, a total of 60 trunk cables is used. This corresponds to a total of 5760 fibers and thus easily covers the minimal need of the system of 3500 fibers. The installation of the UXC-to-USC optical fibers through the trigger tunnels and cable chains to the wheels was labor- and access-intensive, but also leaves a legacy infrastructure that will be exploited by the DT Phase 2 upgrade in LS3. A very important requirement, made at the time of purchasing from the vendor, was that the relative length of different fibers had to be carefully equalized. This was needed to preserve the phase relationship between the signals coming from the different trigger links of a single chamber, which could not be compensated for at the input of the TSC. The propagation delay along different fibers was measured, finding an excellent uniformity, with variations between fibers from the same MTP cord of under 1 ns [90].

The OFCU conversion was performed by dedicated boards that output LVDS signals into RJ45 connectors. These, in turn, were used to transmit information for input to the ROS and TSC boards. Due to different requirements for the readout and trigger electronics, different OFCU boards were developed, but they were both based on commercial parallel optics receivers from AVAGO (HFBR-782BEPZ). At later stages of the Phase 1 DT upgrade, such receivers were re-used as components of the upgraded readout and trigger electronics. The upgrade also included the

2024 JINST 19 P05064

design of various slow-control electronics for both the CUOF and OFCUs, and a link to maintain the injection of trigger data into the readout chain.

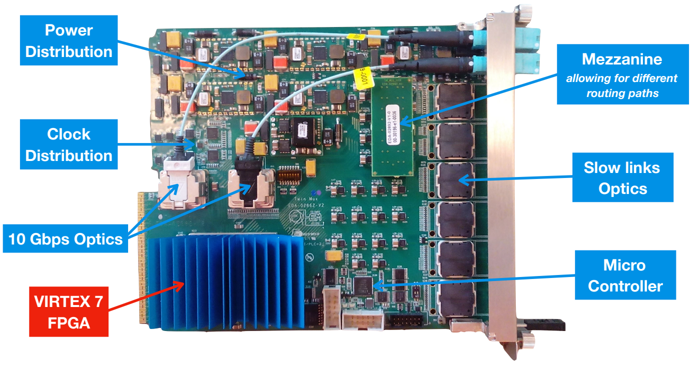
The entire relocation of the DT SCs took place during LS1, between February 2013 and August 2014. Commissioning with cosmic rays and calibration runs were performed shortly after its completion. It was confirmed that the performance of the upgraded system was consistent with the original one, and the DT system operated very successfully throughout Run 2.

Upgrade of the muon barrel local trigger: the TwinMux concentrator. In the original CMS level-1 (L1) muon trigger architecture, tracks were reconstructed using three different track finders, each one mostly exploiting information from a single muon detector: DT, CSC, or RPC. Candidates from the different track finders were only merged at the last stage of the muon trigger logic. For the Phase 1 L1 trigger upgrade [91], described in detail in section 10, a different layout was chosen. The overall muon trigger chain was designed to exploit information from all the detectors covering the area crossed by a given muon as early as possible in the online reconstruction. This approach was adopted to maximize overall performance and to better control the data acquisition rates. In the case of the muon barrel, the DTs provide excellent position resolution, whereas the RPCs are characterized by excellent time resolution. Hence, in the Phase 1 L1 trigger, information from both detectors is combined by dedicated electronics that provide primitives of superior performance (called super-primitives) already at the input of the barrel muon track finder (BMTF).

To accomplish this, the DT TSC was replaced by a new component, called TwinMux [92], which acts as a concentrator for the data coming from both the DT and barrel RPC chambers. The TwinMux combines information from the two into superprimitives and transmits them to the BMTF and the overlap muon track finder (OMTF) using the 10 Gb/s link protocol exploited by the Phase 1 L1 trigger system.

For the TwinMux, a single slot double-width full-height μTCA board, called TM7, is used. A TM7 can reach a maximum of 96 optical connections thanks to six front panel Avago optical receivers (72 links limited to 2.7 Gb/s) and two Minipods for high-speed data transmission and reception (up to 13 Gb/s). Figure 43 (left) shows a picture of a TM7 board, where its main components are highlighted. The TM7 board is based on a Xilinx Virtex-7 FPGA that, in the case of the trigger, achieves the merging of several 480 Mb/s links to higher speed serial links and compensates delays to provide BX alignment of the trigger data coming from different inputs. Twelve of the 72 inputs are optionally routable to GTH Gigabit Transceiver inputs [93] in order to handle the GOL-based [94] 1.6 Gb/s links that receive the RPC links. A small mezzanine PCB allows the desired path to be chosen for lines routed to one of the front optical transceiver. Four additional GTHs are used to transfer data on the μTCA backplane.

From each minicrate, DT trigger information is transmitted as described above. For the RPC detector, five link board masters (LBMs) compress the trigger hit data relative to one muon barrel sector and serialize it through the GOL transmitters. The TwinMux is in charge of forwarding this data to the BMTF and OMTF, by applying a scale-up in the transmission rate (and hence a reduction in the number of links). It is also responsible for duplicating the data to be sent to different track finder processors (up to four times for the sectors of the outer wheels where DT data is shared between the barrel and overlap track finders). Such redundancy is included to reduce the connections between the track finder processors and hence to increase the reliability of the system (which proved



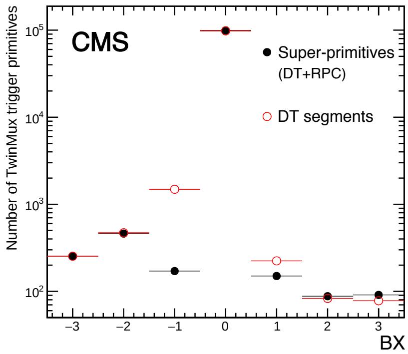


Figure 43. Left: picture of a TM7 board with the main modules highlighted. Right: BX distribution of L1 trigger primitives reconstructed in the muon barrel [95]. Open (red) circles show the performance of trigger primitives reconstructed using information from the DT detector only. Filled (black) circles show the same performance figure for super-primitives built combining information from the DT and RPC detectors.

to be a weak point in the design of the legacy DT track finder). The minimum bandwidth required for forwarding trigger data of one sector is 16 (8) Gb/s for the DT (RPC), implying the need for a total of three 10 Gb/s links. The clock distribution is based on two very-low-jitter PLLs that can broadcast two different clocks to all the FPGA transceivers for performing synchronous or asynchronous data transmission. Finally, a microcontroller is responsible for managing the IPMI interface on the backplane. It handles low-level operations like the hot swap of single boards without the need of switching off a full crate, or monitoring the temperature sensors.

To cover the full barrel, 60 TM7 TwinMux boards are hosted in five μTCA crates. Each of them is equipped with an AMC13 for clock and slow-control command distribution and for providing a connection to the CMS DAQ. Each crate is also equipped with a commercial μTCA carrier hub (MCH), a redundant power module, and a JTAG switch used for remotely programming the boards. A picture of the five μTCA crates hosting the TwinMux is presented in figure 42 (center).

The TwinMux firmware consists of six major blocks. The DT and RPC receiver blocks handle input channels from the two detectors, which correspond to a given muon barrel sector, performing in both cases a link alignment to the same BX. In the case of the DT receiver, a factor 3 oversampling, leading to a frequency of 1.44 Gb/s, is applied to cope with potential data integrity errors caused by the aforementioned DC imbalance of trigger signals. For the RPC, the GOL protocol is decoded. Moreover, clusters are formed of nearby RPC hits and their coordinates are converted to the ones used by the local DT trigger. A so-called super-primitive generation block combines information from the DT and RPC detectors according to algorithms described in more detail later in this section. The output transmitter block formats and sends the trigger output data to the track finders using a 10 Gb/s link protocol. A readout implementation is also included that allows trigger-primitive readout through the CMS DAQ system. Finally, an IPbus-based slow-control block also exists. The TwinMux needs 34.5 BXs of latency for data to go through the FPGA from input to output. Of these, only four BXs are actually used for super-primitive generation, whereas 26.5 are taken for handling of RPC data by the RPC receiver block. Finally, four additional BXs are taken by the serialization of the data sent to the track finders. The DT receiver block works in parallel with, and in the shadows of, the RPC one.

The TwinMux was deployed in production, as part of the Phase 1 L1 trigger upgrade, starting from 2016, after having been tested in a slice of the detector that was sending signals both to the TSC OFCU and TwinMux boards during the 2015 run. The development of the algorithms to build super-primitives, which are used as input to the BMTF, occurred in stages. In 2016, no combination was performed, hence only trigger primitives built by the DT on-board minicrate electronics were fed into the BMTF.

In 2017, an algorithm that combines information from the DT and RPC detectors was deployed to improve the super-primitive BX identification. Within such algorithm, the compatibility of the DT trigger primitives and RPC clusters that are built in nearby chambers is checked by comparing the difference in azimuthal angle ($\Delta\phi$) between them. If a match within a programmable window is found between a DT trigger primitive and (at least) one RPC cluster, and if the difference in terms of BXs between the DT trigger primitive and the RPC cluster is within ± 1 BX, a super-primitive is built using the DT trigger segment position and direction, but the RPC BX. We note that no time correction is attempted if a DT trigger primitive is built exploiting hits from all layers of the ϕ-SLs of a given DT chamber. In any case, a dedicated quality flag, documenting the successful matching with the RPC, is set. This combination better exploits the complementarity of the DT and RPC detectors, relying on the spatial resolution of the former and the time resolution of the latter. The impact, in terms of BX identification performance for trigger primitives caused by muons from pp collisions, is shown in figure 43 (right). The asymmetry in the BX distribution of DT primitives (red open dots), mostly due to the occasional presence of δ rays that can spoil the reconstruction of standalone DT trigger segments, is mitigated when the combination with the RPC is effective (black filled dots). Because of the BX correction, the muon barrel trigger primitive BX assignment efficiency is also increased, on average, by 1.4%.

In 2018, a further improvement was put into production. In the MB1 and MB2 stations, where two RPC layers cover both surfaces of each DT chamber, as described in section 6.3.1, the generation of RPC-only super-primitives is attempted if no DT trigger primitives are present in a given BX. In that case, pseudosegments are built out of RPC cluster pairs that are reconstructed at the same BX in the two different RPC layers. If more pseudosegments are generated at a given BX, only the one whose direction is closest to that of a straight track coming from the CMS interaction point (ϕ_b) is retained. Finally, only pseudosegments for which ϕ_b is below a programmable threshold are accepted. Additionally, also in this case, the RPC-only primitive is marked with a dedicated quality flag. Upon deployment of this algorithm in data taking, a further increase of the super-primitive efficiency around 3% was observed in the MB1 and MB2 stations. Due to the redundancy of the muon system, such local trigger efficiency improvement resulted in a marginal improvement of the BMTF efficiency. Nevertheless, a rate reduction of a few percent was observed for the lowest unprescaled L1 muon trigger. This is because, in a few cases, BMTF tracks now get built using a larger number of points along the muon trajectory, improving the measurement of the track transverse momentum.

Given the fact that it uses a different track-building algorithm, no super-primitive generation is attempted on primitives that are sent to the OMTF, since the OMTF receives a full list of DT trigger segments and combines them with RPC information directly at the track-finding step, as also described in section 6.3.2. The super-primitive algorithm that operated throughout 2018 was also deployed at the start of Run 3.

Upgrade of the DT readout: the μROS system. Simulation studies have shown that the ROS was the most severe bottleneck in the DT readout chain. The ROS combined information from a total of

25 readout boards (ROB) providing the minicrates with data for a full DT sector. The time needed by a ROS to perform event building from all the input links depended on how hits were distributed within the ROBs. A high noise rate from sporadic small groups of channels could be easily sustained. However, high overall background hit rates, which come with a rather uniform distribution within each of the chambers of a sector, take much more time to be processed. In addition, muons crossing the DT can produce up to 44 hits within a single ROS. When both the background and muon hit rates are considered, taking as a proxy an LHC instantaneous luminosity of $2 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, the maximum acquisition rate that a ROS can sustain becomes close to the 100 kHz limit imposed by the DAQ and L1 trigger.

For this reason, throughout Run 2, the ROS, as well as the downstream components of the DT readout chain, called device-dependent units (DDU), were replaced with a new system, based on the μTCA architecture, named μROS. For the μROS [96], the same TM7 boards designed for the TwinMux are used, but a different firmware is deployed to implement the functionalities needed by the DT readout.

Each TM7 features six 12-fiber MTP receivers, for a total of 72 input links. A total of 25 ROBs provide the inputs from a given sector. The data from each wheel are thus processed by five μROS boards, four of them receiving three sectors each (24 channels per sector, 72 links), and the fifth receiving the 25$^{\text{th}}$ channel for each of the twelve sectors. The production system comprises three μTCA crates (central, positive, and negative wheels) and 25 μROS boards. Each crate is equipped with an AMC13 that provides clock and slow-control distribution, as well as a connection to the CMS DAQ. A picture of the μROS system is presented in figure 42 (right). With this architecture, no further components of the DT readout system (DDU) are needed. Under the conditions reached over Run 2, the maximum payload bandwidth varied, depending on the wheel, between 0.3 Gb/s in W0 and 0.6 Gb/s in W+2 and −2, remaining well within the AMC13 limits.

In terms of firmware, the components needed to handle the slow control and general board functionality are inherited from the TwinMux. Special care was instead put into the design of the block in charge of data deserialization. This firmware can recover input data with high quality and minimal data losses with respect to the input stream. Carrying data at 240 Mb/s, the receiver samples data at 1.2 Gb/s (a factor of 5 oversampling). Majority filtering is performed on the three central samples of each bit before reassembling the original input word. Bits where a weak majority is found (two instead of three) are marked as transmission error candidates. If the data frame parity error check fails and only one bit is marked as a transmission error candidate, the latter gets corrected. The firmware implements a full verification of the ROB protocol and provides statistics for the different ROB failure cases, which are used for monitoring. Finally, while the legacy ROS system masked channels in case of transmission errors until a resync was issued, the event builder from the μROS is capable of recovering from all types of errors as soon as the condition disappears.

The transition to the μROS occurred during the 2017/2018 year-end technical stop. Prior to that, in 2017, signals from the ROBs for a slice of the detector were split and a μROS slice-crate was instrumented and integrated into the CMS DAQ as a separate unit. It was used to develop the μROS prior to the deployment of the full system, which allowed a smooth transition. In LS2, the FE signals of a sector in the external wheels (W+2 S12) were split to allow the continuation of the same strategy for the Phase 2 upgrade, and another slice-test system will be operated in Run 3. The data of such a Phase 2 slice-test are already read by the CMS DAQ.

The impact, in terms of DT performance, of the transition to the μROS is presented in figure 44. The two plots show a chamber-by-chamber map of the DT segment reconstruction efficiency, as measured with a tag-and-probe method. In the measurement, no masking of chambers with hardware or readout issues is applied. The left (right) side of the figure refers to results computed using 2017 (2018) data, collected before (after) the transition to the μROS. Bins where the efficiency is significantly lower than 99% are due to possibly sporadic problems that affected chambers or their readout. A better performance was observed after the upgrade to the μROS, mostly thanks to a reduction of the number of chambers affected by readout problems.

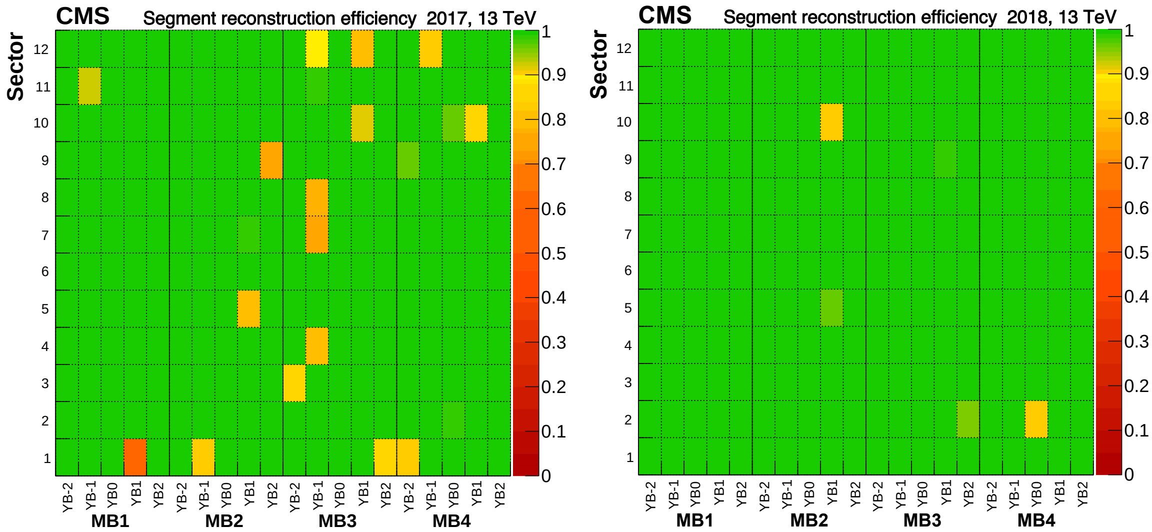


Figure 44. DT segment reconstruction efficiency measured with the tag-and-probe method using data collected by CMS in 2017 (left) [97] and 2018 (right) [98], before and after the transition to the μROS. The efficiency is usually above 99% except for chambers affected by hardware problems, mostly coming from the DT readout. After the deployment of the μROS, the overall efficiency improves.

6.1.3 Detector longevity for Run 3 and beyond

The DT system was designed and validated to sustain ten years of LHC operation in nominal conditions, corresponding to approximately 500 fb^{-1} of integrated luminosity [99]. Accelerated aging studies, carried out prior to the installation of the chambers in CMS, indicated that no degradation in the performance of the present detector (and its electronics) is to be expected under those assumptions. Furthermore, at the boundaries of long data-taking periods, data are regularly collected by varying the HV applied to the anode wires (HV scans) and measuring the detector efficiency at each HV point to assess its stability and exclude potential effects due to early aging. Up to the end of Run 2, no efficiency degradation was observed during the HV scans. Additionally, the original longevity studies were complemented with more stringent ones, performed using the CERN gamma irradiation facility (GIF++) [88], which are summarized in the following paragraph. Given the results of all the above studies, and the current projections in terms of integrated luminosity expected for Run 3, the DT performance is foreseen to remain almost constant over the coming run period.

Nonetheless, though several electronics components will be replaced as part of the CMS Phase 2 upgrade program, existing muon detectors will operate throughout the HL-LHC era. For this reason,

2024 JINST 19 P05064

further accelerated longevity studies are being performed at GIF++ on two spare DT chambers, as mentioned above. Due to the complexity of the physics and chemistry phenomena driving the aging processes, accelerated studies have large uncertainties, typically taken into account by irradiating with a substantial excess with respect to the needed integrated values (safety factors). The reference targets for instantaneous and integrated luminosities used in the aging studies to derive the DT performance are evaluated assuming safety factors of two, which double both the nominal HL-LHC luminosity ($5 \times 10^{34}\,\mathrm{cm^{-2}\,s^{-1}}$) and the total integrated luminosity ($3000\,\mathrm{fb^{-1}}$). Considering these safety factors, by the end of the HL-LHC, the DT hit detection efficiency can possibly drop from its present value of around 96% to approximately 70% in the MB1 stations of W±2 (corresponding roughly to 10% of the entire system). Moreover, in a further 20% of the detector (corresponding to the MB4 stations of the DT sectors covering the top half of CMS and to the MB1 stations of W±1), the efficiency is projected to range between 85 and 90%. Finally, in the rest of the system, efficiencies above 90% are expected. Given the redundancy in terms of the number of DT layers per chamber and of chamber stations in the muon system, the maximum inefficiency due to DT aging, localized in a narrow region around $|\eta| = 1.0$, is expected to be within 2 (5)% for standalone offline reconstruction (trigger). Though these expected losses are not very large, mitigation strategies, which are described in the following sections, have been put in place during Run 2 and LS2, to maximize the longevity of the DT detector and ensure the highest achievable performance in the long term.

Optimization of the operational working points. For gaseous detectors, deterioration due to aging becomes more significant as the integrated charge released in the gas volume increases. In turn, the integrated charge depends on the rate and type of particles crossing the different regions of the detector, the collected charge per particle, and the total integrated running time.

In the case of the DT detector, background dominates the hit rates and is largest in the MB1 stations of W±2 and the MB4 stations for the sectors covering the top half of CMS. The collected charge per particle depends on both the particle type and the gas amplification factor, which is driven by the HV settings of the cell anode wires. Therefore, for a fixed integrated luminosity, reducing the HV working points of the anode wires can result in a significant reduction of the integrated charge. Of course, such an optimization can be performed only within the limits where the impact on detector performance is deemed acceptable.

During the 2017 and 2018 LHC runs, the HV of the DT anode wires was progressively lowered with respect to the default 3600 V used until 2016, in the chambers most exposed to background. In 2017, the wires in MB1 of W±2, and the ones in MB4 of sectors 3, 4, and 5 for all wheels were operated at 3550 V. In 2018, a further reduction was applied, reaching the HV values given in table 8. The discrimination threshold values applied in the DT FE electronics were also lowered from 30 to 20 mV for the entire detector.

For the MB1 of the external wheels, lowering the HV of the anode wires to 3500 (3550) V resulted in a relative reduction of 58 (45)% in the drained current with respect to the original 3600 V setting.

The overall performance of the system under the updated operational conditions was thoroughly studied using pp collision data. Firstly, the reduction of the FE thresholds to 20 mV resulted in a marginal increase of overall detector noise, which was handled by masking a few specific noisy wires. Reducing the gain and the FE threshold resulted in shifts of a few ns of the effective drift

Table 8. HV settings for the anode wires of the different DT stations and wheels used for the 2018 LHC run.

	W−2	W−1	W0	W+1	W+2
MB1	3500 V	3550 V	3550 V	3550 V	3500 V
MB2	3550 V	3550 V	3600 V	3550 V	3550 V
MB3	3550 V	3600 V	3600 V	3600 V	3550 V
MB4	3550 V	3550 V	3550 V	3550 V	3550 V

time measured by the electronics caused by opposite-sign effects. These partially cancel but were nevertheless carefully corrected for, by updating the trigger synchronization and offline calibration. Fine-tuned corrections were needed to maintain the optimal performance of the system in terms of BX identification efficiency and time resolution of the offline reconstruction. These remained well in line with what is reported in section 6.1.1. The DT hit detection efficiency was then measured. Results for the ϕ-SLs of the MB1 stations are given as function of the total integrated luminosity in figure 45. Different colors in the plot refer to different DT wheels. The efficiency is stable throughout the different run years, apart from step variations dominated by changes in the FE and HV, which amount at most to 1%. This was proven to be true for the rest of the detector as well and, given the redundancy of layers forming a DT chamber, a negligible impact on the DT local reconstruction efficiency was observed. Finally, the effect of the updated settings on the DT hit resolution was also assessed. In the case of ϕ-SLs, where hit resolution impacts the muon transverse momentum measurement, only a slight worsening was observed. In the chambers where the HV was lowered to 3550 V, the degradation is around 10%. In the MB1 of the external wheels, where the HV was set to 3500 V, the effect is slightly larger, but the hit resolution remains within 250 μm, in line with the design expectations.

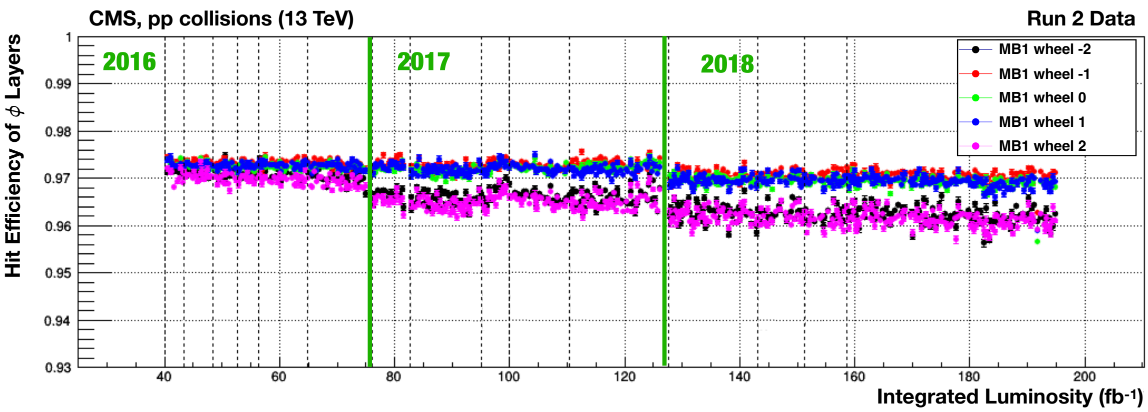


Figure 45. DT hit detection efficiency, computed as a function of the total CMS integrated luminosity, for the ϕ-SLs of MB1 chambers [98]. Different colors refer to different DT wheels. The plot summarizes how the efficiency evolved during Run 2, mostly as a consequence of the different updates of the detector HV and FE threshold settings.

Given the reduction in terms of expected integrated charge, as well as the very mild impact on the overall performance, the operational working points used in 2018 were chosen as the default configuration for Run 3 as well.

2024 JINST 19 P05064

Gas system upgrade and open-loop operation. The DT gas system [1] includes a facility on the surface that stores individual gas components (Ar and CO_2) and combines them to obtain the gas mixture of 85% Ar and 15% CO_2 used by the DT. A second facility, located in the USC, is dedicated to the predistribution of the final gas mixture to the five wheels, as well as to the gas analysis system. Finally, five distribution racks located in the UXC provide the gas flows to the 250 DT chambers, with a flux around 40 l/h to each chamber. The primary goal of the gas analysis system is ensuring the stability of the gas mixture. It consists in the determination of the contamination of O_2 and H_2O, and the drift velocity measurement.

The loss of efficiency caused by aging, described in section 6.1.3, is due to a decrease of gas gain which, in turn, is caused by deposits that form around the anode wires of the DT cells. It is believed that this effect is mostly due to out-gassing of components inside a DT chamber. For this reason, the re-injection and spread from hot detector regions of pollutants contaminating the DT gas should be avoided. Therefore, starting from the 2018 run, the DT gas system operates in the, so-called, open-loop mode, where there is no re-circulation of used gas. Prior to that, the system was operated in a closed-loop mode, where 85% of the used gas was recirculated and only 15% of fresh gas was injected. The closed-loop mode can still safely be used during technical stops or longer periods of inactivity, where no aging due to out-gassing is expected. In order to operate in open-loop mode, an upgrade of the gas system was performed. It allows the intentional introduction of air from a bottle on the surface, to maintain the desired level of O_2, corresponding to 80 ppm. Furthermore, a humidifier with a bypass that ensures the H_2O concentration is kept at 800 ppm was also installed. The presence of small quantities of O_2 and H_2O helps prevent effects that induce chamber aging (such as polymerization) [100], and the target values for such contaminants were derived based on experience from the data taken in Run 1 and Run 2. Because of this upgrade, prompt monitoring of the gas stability has become even more important than in the past.

The O_2 and H_2O measurements are performed using commercial sensors, which have to be recalibrated every year by injecting gas with known components of oxygen and humidity. The drift velocity measurement is done by a more complex system, called VDC [101], which aims to deliver, every ten minutes, a drift velocity measurement. There is one O_2 sensor, one H_2O sensor, and one VDC for each of the five DT wheels, and a separate sampling line for each wheel. The gas sampling permits the selection of the output of any individual chamber, as well as the global output (or the global input) of an entire wheel. The direct measurement of the drift velocity performed by the VDC system is sensitive to possible unknown components or contaminants in the gas. The drift velocity depends on the Ar/CO_2 ratio, as well as on the level of contaminants, such as O_2 and N_2. The drift velocity is a key parameter for the DT local reconstruction, so that any deviation spotted by the VDC system must be investigated and eventually requires immediate intervention on the operation of the DT gas system. An example from the monitoring by the VDC system is presented in figure 46. The first transition from closed-loop to open-loop operation was correctly detected by the drift velocity measurement. In this case, the gas analysis showed very quickly the impact of the injection of air and humidity into the mixture for the case of the open-loop mode, which resulted in a sub-percent variation of the drift velocity.

Other than direct monitoring of the gas mixture and drift velocity, the stability of the gas is also evaluated indirectly with event data, by looking at the stability of the performance of the DT local reconstruction. If a track-segment fit is performed using a sufficient number of hits from a DT

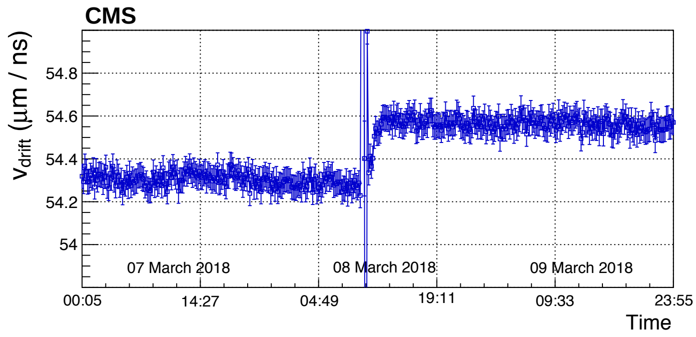


Figure 46. Drift velocity measurement using the fresh gas analyzed by the VDC system. The variation on 8th March 2018 corresponds to the transition between closed-loop and open-loop operation of the DT gas system.

chamber, additional parameters, other than the segment position and direction, can be extracted from the fit itself. In this way, the need for residual corrections on top of the calibration parameters used by the reconstruction, such as the drift velocity, can be evaluated on a segment-by-segment basis. Overall biases in the distribution of such residual corrections are then measured run-by-run for each DT chamber, and their stability across different runs is monitored.

Finally, the DT chambers are operated since Run 1 in a differential pressure mode. The gas control system presently ensures that a positive differential pressure of +3 mbar is applied at the bottom of the wheels. This value, which is well within the mechanical maximum limit (50 mbar), protects against possible contamination. A dedicated system of differential pressure sensors monitors continuously these values, which are transmitted to the DT online monitoring system.

Installation of shields over the outer MB. As mentioned earlier, besides the MB1 stations of the external wheels (W±2), the parts of the MB characterized by the highest level of background are the MB4 stations in the sectors covering the top half of CMS (sectors 1 to 7). Studies based on pp collision data, as well as results from simulations, corroborate the hypothesis that background in this region is mostly generated from interactions of low-energy neutrons permeating the cavern. Therefore, a strategy was put in place to effectively protect the top DT chambers by installing proper absorbing shields.

With the aim of designing such shields, absorbing layers were installed on top of the MB4 stations in sector 4 of W+2 and W−2 during Run 2. Different configurations in terms of material and thickness of absorbers, suggested by simulations of the radiation field in the cavern, were tested each year from 2015 through 2018, and the impact in terms of background reduction was studied. An example of these studies is shown in figure 47 (left), where the magnitude of the linear dependence of the currents in each chamber with respect to the LHC instantaneous luminosity is presented for the 2018 run. The impact of the test shielding configurations installed over the MB4 stations in sector 4 of W±2 is clearly visible. Based on the results of these investigations and considering the outcome of simulations of the mechanical stress that different shield setups would induce on the supporting structures, a layout was proposed that is expected to reduce the background by about a factor 2.

The installation of the final DT MB4 shields was performed during LS2 and completed in October 2020. A schematic view of the shielding layout is presented in figure 47 (right). The figure shows the shield installed over W−2, −1, +1 and +2. In those wheels, shielding cassettes consisting

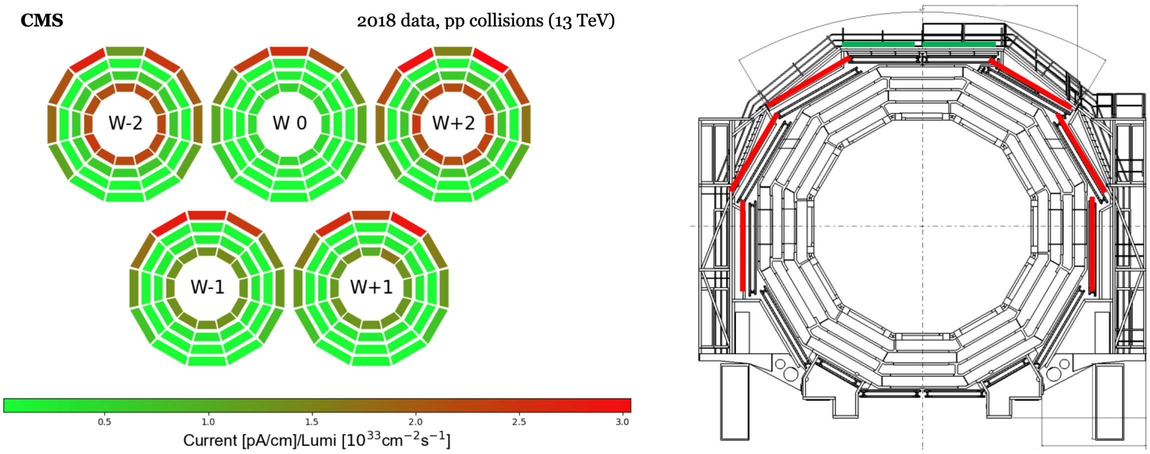


Figure 47. Left: magnitudes of the linear dependence between the currents from each DT chamber of the MB and the LHC instantaneous luminosity [102]. Results are computed after the optimization of the operational working points. Right: transverse view of an MB wheel highlighting the layout of the MB4 shield, as installed in W−2, −1, +1, and +2. Red (green) lines represent shield layers consisting of thin (thick) cassettes.

of 9 mm of lead and 9 cm of borated polyethylene (thick cassettes) are put on top of sector 4 stations (green lines), whereas cassettes consisting of 9 mm of lead and 3 cm of borated polyethylene (thin cassettes), cover sectors 1, 2, 3, 5, 6, and 7. Due to mechanical restrictions imposed by the presence of services from the vacuum tank and inner detectors, the shields deployed to cover the central wheel have a different layout. In this case, no shielding is applied on top of sector 4, whereas, in the other six sectors, shields consisting of thin cassettes only cover the lowest part of half of each chamber. The deformation on the supporting structures induced by the MB4 shields, which have a weight of approximately 40 tons in total, was measured accurately at the weakest points in the structure and found to be always smaller than 1 mm, in good agreement with finite-element calculations.

In summary, during Run 2, the DT off-detector electronics underwent a set of upgrades that improved the performance of both the readout and trigger. Strategies to increase the detector longevity were also put in place to maximize the DT performance through the end of HL-LHC running. Following a thorough recommissioning over LS2, the DT system successfully entered Run 3, showing offline and online tracking performances that are remarkably consistent with the ones achieved at the end of the previous LHC run.

6.2 Cathode strip chambers

6.2.1 General description

The cathode strip chambers (CSCs) are multiwire proportional chambers with a finely segmented cathode strip readout. The strips run radially in order to measure the muon position in the bending plane, the plane perpendicular to the colliding beams axis, while the anode wires are oriented in azimuth and provide a coarse measurement in the radial direction. A precise measurement of the muon coordinate in the azimuthal direction is obtained from charges induced on the cathode strips. Each CSC module consists of six gas layers, each layer having a plane of radial cathode strips and a

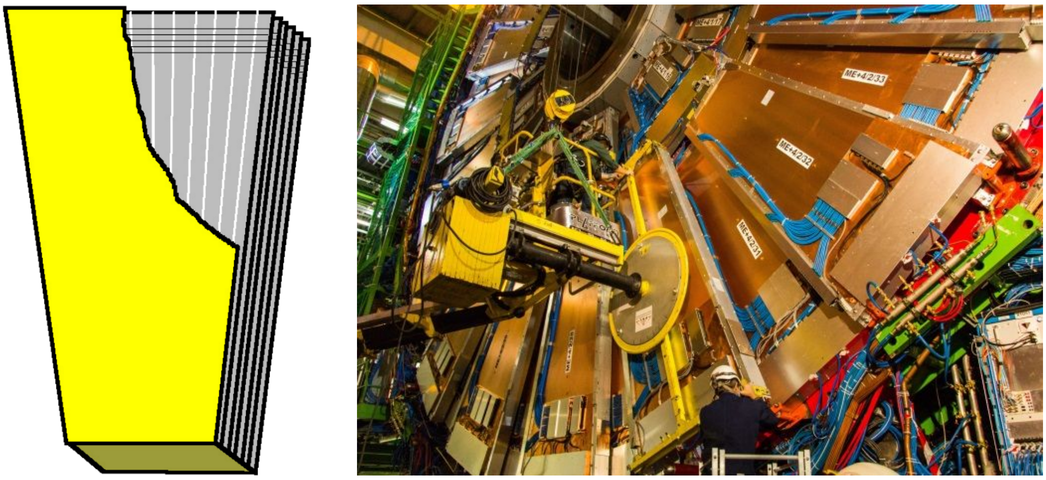


Figure 48. Left: layout of a CSC chamber, with seven trapezoidal panels forming six gas gaps. Only a few wires (lines running from left to right) and strips (gray band running from top to bottom) on the upper right corner are shown for illustration. Reproduced from [103]. CC BY 4.0. Right: installation of the outer CSC chambers (ME4/2) during LS1. Reproduced with permission from [104].

plane of anode wires running perpendicular to the strips. Figure 48 shows the CSC physical layout and a photograph of the installation of some CSCs into the CMS detector.

The gas mixture is 40% Ar, 50% CO_2, and 10% CF_4. The CO_2 component is a nonflammable quencher needed to achieve large gas gains, while the main function of the CF_4 is to prevent anode aging caused by polymerization processes on the wires. Argon is the working gas that is ionized by traversing charged particles. The primary design consideration for the CSC detectors is the ability to provide good spatial and temporal resolution for triggering and identification of a muon. The typical position and time resolutions are 50–140 μm, depending on chamber type, and 3 ns per chamber, respectively.

Figure 39 shows the geometry of the CSC system. At the start of the LHC operation, a total of 468 trapezoidal CSC modules, placed between the steel magnetic flux-return plates, were arranged into four disks (stations). The first station, closest to the interaction point, is further segmented into three rings (ME1/1, ME1/2, and ME1/3), while stations 2 through 4 are separated into just two rings. The rings closest to the beam line (named ME1/1, ME2/1, ME3/1, and ME4/1, or collectively as ME1234/1) are subject to the highest particle rates. Each of the chambers in the ME1/1 ring contains strips split at $|\eta| = 2.1$, which are read out separately and denoted ME1/1a ($2.1 < |\eta| < 2.4$) and ME1/1b ($1.6 < |\eta| < 2.1$). Some key parameters of the CSCs are summarized in table 9.

The ME234/1 chambers each cover 20° in ϕ; all others cover 10°. All chambers, except for the ME1/3 ring, overlap by five strips at each edge and hence provide ϕ coverage without gaps. The original design included 72 ME4/2 chambers that were descoped in the initial construction of the CSC system, and instead were built and installed during LS1 that occurred during 2013–2014. The presence of ME4/2 chambers provides an additional track segment of six hits along the muon trajectory. This allowed for a more robust muon p_T assignment in the L1 trigger described in section 10, so that the rate with a given p_T threshold in the $1.2 < |\eta| < 1.8$ region decreased during Run 2, as shown in figure 49.

As illustrated in figure 39, in the $1.1 < |\eta| < 2.4$ region, almost all possible muon paths cross at least three CSC chambers. In the $0.9 < |\eta| < 1.1$ overlap region, a combination of DT and CSC

Table 9. Key parameters for different types of CSCs.

	ME1/1	ME1/23	ME234/1	ME234/2
Wire diameter [µm]	30	50	50	50
Wire spacing [mm]	2.5	3.2	3.1	3.2
Strip width (narrow) [mm]	3.2	6.6–11.1	6.8–8.6	8.5
Strip width (wide) [mm]	7.6	10.4–14.9	15.6	16.0
Gap between strips [mm]	0.35	0.5	0.5	0.5
Angle subtended by each strip [mrad]	2.96–3.88	2.15–2.32	4.65	2.33
Gas gap [mm]	7	9.5	9.5	9.5
Operating HV [V]	2900	3600	3600	3600

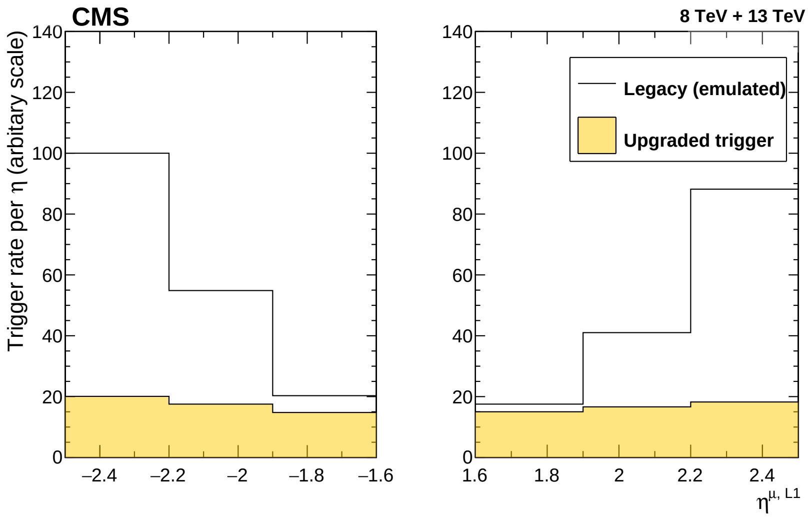


Figure 49. Emulation of the Run 1 algorithms compared to the upgraded Run 2 algorithms, as a function of the L1 muon η. The most common L1 single-muon trigger threshold used in 2017 was $p_T^{\mu,\mathrm{L1}} \geq 25\,\mathrm{GeV}$.

modules also typically provide at least three position measurements along muon paths. Moreover, the four layers of RPCs between the CSCs provide additional muon hits at the trigger level in the $0.9 < |\eta| < 1.6$ region. The extension of the RPC layers and the addition of GEM detectors (GE1/1 in Run 3) further enhance the muon triggering and reconstruction in the $1.6 < |\eta| < 2.2$ region. The high redundancy of the muon system is a central feature of the design and is responsible for the robustness of the muon triggering and reconstruction in CMS.

The present electronics readout system of all CSC chambers, shown schematically in figure 50, share the same architecture as in Run 1. This is briefly summarized below, with a more detailed description available in ref. [1]. On each CSC chamber, an anode local charged track (ALCT) board finds patterns among the six-layer anode hits (sampled by the ALCT board, based on the digitized pulses sent by the on-chamber anode frontend boards) that are consistent with muon tracks, and then assigns a quality rank to each of them. These patterns are called ALCTs. The ALCT board can then send up to two such tracks per each bunch crossing (BX) with the highest quality to the trigger motherboard (TMB). Upon receiving a level-1 accept (L1A) signal from the trigger, coinciding

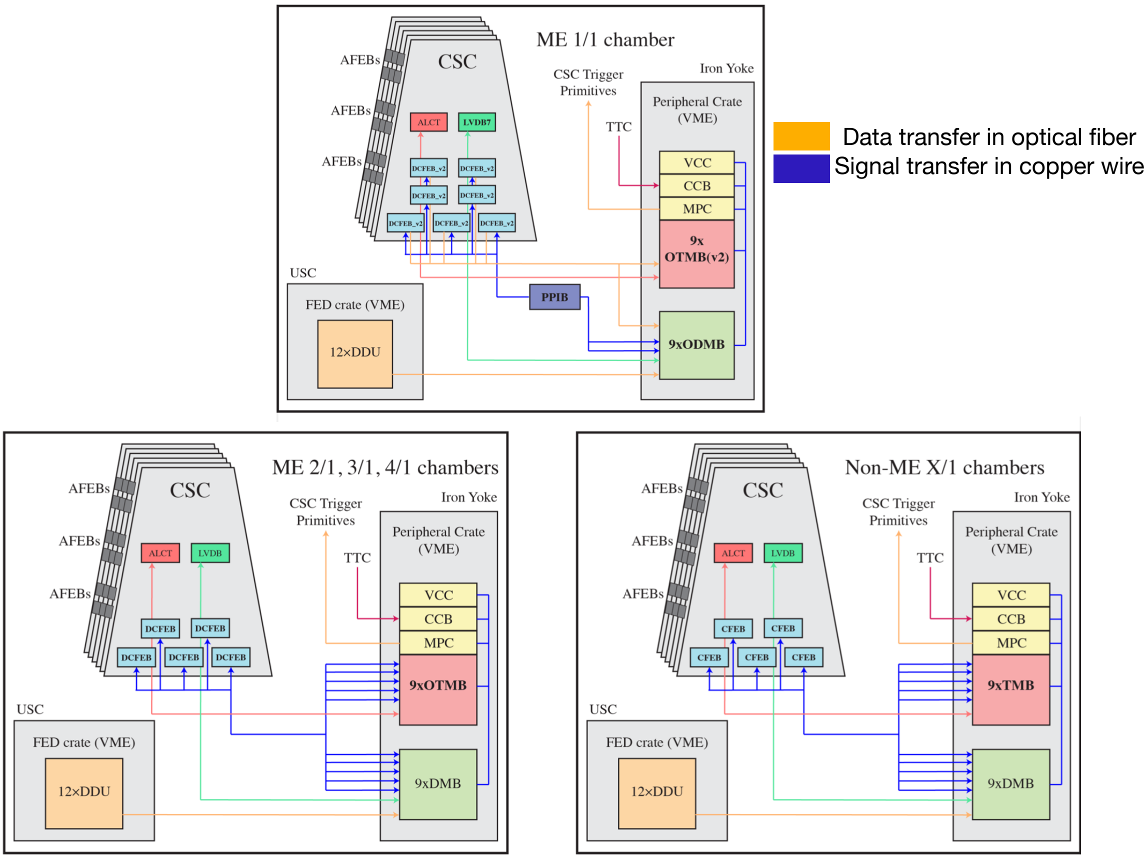


Figure 50. Schematic of the CSC electronics readout systems for Run 3: ME1/1 (upper), ME234/1 (lower left), all other chambers, ME1234/2 and ME1/3 (lower right). Reproduced with permission from [105].

within 3 BXs with an ALCT, the ALCT board sends all wire hits and ALCTs that have been found in a predefined time window to the data acquisition motherboard (DMB). Both the TMB and DMB are located in VME crates mounted on the edge of the endcap steel yoke.

The cathode signals are processed by another set of on-chamber electronics, the cathode frontend boards (CFEB). Each CFEB amplifies the signals from 16 strips on each of six layers of a CSC, and sends fast trigger information and charged particle hits localized within a half-strip width, as assessed by comparator ASICs (comparator data), to the TMB. On receiving an L1A coinciding within 5 BXs with a cathode local charged track (CLCT), the CFEB digitizes the strip signal waveforms over a 400 ns long period, which is sampled every 50 ns by a 12-bit ADC, and sends the digitized samples to the DMB. The transfer of CFEB data to the TMB and DMB comprises the trigger and data paths of the strip signals, respectively. During Run 1, each CSC had five CFEBs, except for ME1/3, which had four per CSC (since the ME1/3 CSCs have only 64 strips).

The low-voltage power for the on-chamber electronics is distributed at the appropriate voltage levels by the on-chamber low-voltage distribution board (LVDB).

The TMB builds cathode hit patterns into CLCTs and finds coincidences with anode hit patterns that form ALCTs to make local charged tracks (LCTs), also called the CSC trigger primitives. It sends the two LCTs with the highest quality per BX to the muon port card (MPC) and, on receiving

2024 JINST 19 P05064

an L1A, to the DMB. The DMB controls the CFEBs on a chamber and collects anode, cathode, and trigger information to send to the detector-dependent unit (DDU), on the arrival of an L1A. The DDU, situated in the frontend driver crates located in the underground service cavern, collects data from 15 DMBs in the CSC system and sends the information through the global DAQ path.

The MPC, situated in the VME crate, collects LCTs from each of up to nine TMBs in a trigger sector, and sends these trigger primitives to the muon track finders described in section 10.2. There is one MPC per peripheral VME crate. The operation of the VME crate is supported by the clock-control board (CCB) and custom VME crate controller (VCC). The CCB serves as the interface between the CSC system and the trigger control and distribution system (TCDS) of CMS. The VCC receives VME commands from the control room and distributes them to the other boards in the peripheral crate via the backplane.

6.2.2 Upgrade of the CSC system since Run 1

While the CSC system operated stably throughout Run 1 and Run 2, some CSC readout electronic boards needed to be upgraded in order to handle the expected longer latency and more stringent trigger requirements at the HL-LHC. Specifically, if the electronics were not altered, the longer latency requirements would fill up and overflow the pipelines of the frontend boards in certain CSC stations, while the higher L1 trigger rates would overwhelm the output bandwidth of various on- and off-chamber electronics boards. To reduce the installation load during LS3, the bulk of the CSC electronics upgrades that required chamber access (dismounting and re-installation) was already performed during LS2.

The CFEBs use switch capacitor arrays to store the charge induced on the cathode strips. These capacitor arrays are capable of storing 96 charge measurements (corresponding to six events worth of data) during the L1 trigger latency. As mentioned above, the digitization of the analog signals and subsequent readout by the DMB only happens when a CFEB receives an L1A that is in coincidence with a CLCT. As a consequence, for CSCs that are closest to the beam (ME1234/1) where the background rate is high, frequent memory overflows and large data losses are expected at the HL-LHC trigger latency. Figure 51 shows the average event loss fraction for ME234/1 rings as a function of the instantaneous luminosity. These curves are based on a statistical model that has been verified by measuring the loss rates in bench tests that emulate the expected background and L1 trigger rates. These results show that data loss would be a severe issue at the HL-LHC with the original CFEBs. The outer CSC chambers will not suffer from these problems because the trigger primitive rates in these rings are lower than those in the inner rings by factors of more than 3.

The CFEBs on ME1/1 have already been vulnerable to data loss at the highest instantaneous luminosities (up to $2 \times 10^{34}\,\mathrm{cm^{-2}\,s^{-1}}$) in Run 2. To mitigate such losses, the four CFEBs on the ME1/1b section of the chambers were replaced by four digital cathode frontend boards (DCFEBs) during LS1. The DCFEBs use fast flash 12-bit ADCs that continuously digitize the cathode signals at 20 MHz, as well as having more powerful Virtex-6 FPGAs with large internal memory resources. The resulting digital pipeline can hold up to 700 events, and thus there should be negligible dead time at the HL-LHC. Two optical links running at 3.2 Gb/s are employed per DCFEB to transmit the raw data for DAQ and the fast comparator data for building the L1 trigger to the new optical data acquisition motherboard (ODMB) and optical trigger motherboard (OTMB), respectively, which replaced the original DMB and TMB.

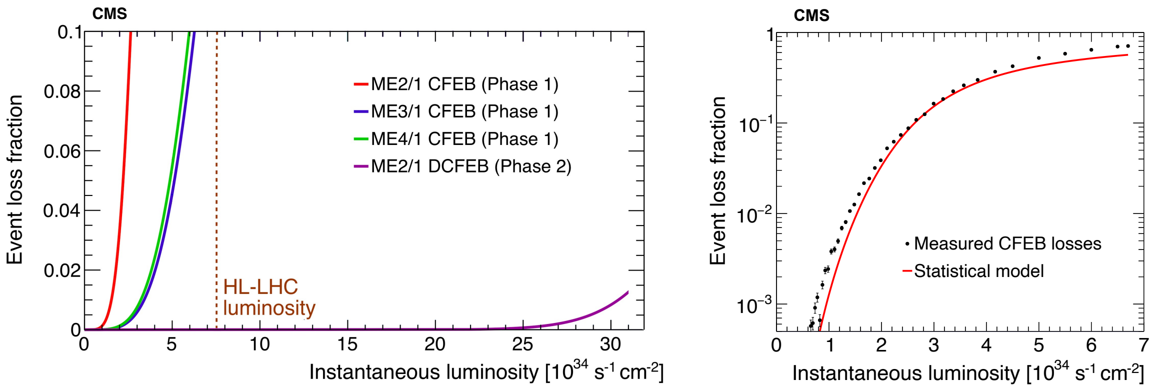


Figure 51. Left: event loss fraction as a function of instantaneous luminosity for different type chambers after different upgrades. The vertical dashed brown line indicates the design HL-LHC luminosity. Right: event loss rate measured in a CFEB under HL-LHC conditions for an ME2/1 chamber, compared to the statistical model [100].

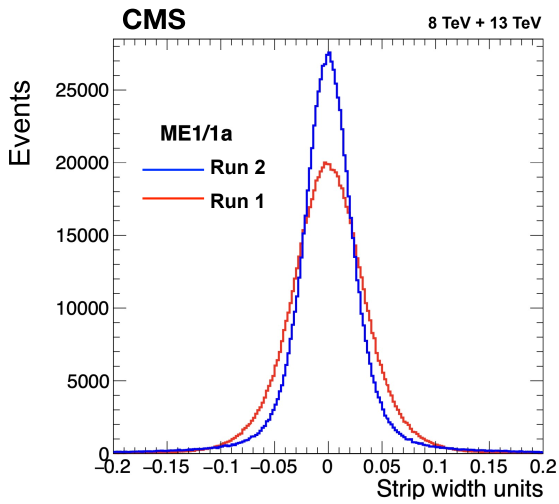


Figure 52. Difference between the position of a reconstructed hit in layer 3 of an ME1/1a chamber and the position obtained by fitting a segment with hits from the other five layers for Run 1 (red) and Run 2 (blue). The spatial resolution is improved by 27% from $\sigma = 64\,\mu$m in Run 1 to 46 μm in Run 2. This is due to the removal of the triple-grouping of strips in ME1/1a, which reduces the capacitance and hence the frontend noise.

In Run 1, due to cost constraints, the 48 strips in the ME1/1a section of the chamber were joined every 16 strips into groups of three, resulting in 16 readout channels served by one CFEB. To remove the ambiguity in triggering and reconstruction, the grouping of the ME1/1a strip readout was removed during LS1, and the single CFEB was replaced by three DCFEBs for strip readout. The ungrouping of the strips in ME1/1a leads to an improvement in spatial resolution of about 20% (figure 52), as well as a decrease in the L1 trigger rate by at least a factor of 3 (figure 49).

Operational experience during Run 2, and additional radiation tests, showed that the programmable read-only memory (PROM) employed in the current DCFEB may not withstand the radiation dose expected at the HL-LHC. Thus, the new DCFEBs (DCFEBv2s) were designed with an option for remote programming of their FPGAs using the CERN-designed GBTx ASIC. The

2024 JINST 19 P05064

performance of the two optical transceivers on the original DCFEBs has been occasionally unreliable, particularly due to single-event upsets (SEUs). While mitigation measures in software and firmware have improved the reliability of these transceivers in the DCFEBv2s, they were replaced by the VTTx, a radiation-hard twin optical transmitter designed by CERN. These new DCFEBv2s were installed in the ME1/1 chambers during LS2, thus mitigating the risks associated with the longevity of their PROMs.

The DCFEBs used in the ME1/1 ring during Run 2 were moved to the chambers in the ME2/1, ME3/1, and ME4/1 rings, where radiation levels are lower. In Run 3, although the ME234/1 chambers are fitted with DCFEBs, the new ODMB is not scheduled to be installed before LS3, so these DCFEBs still send data to the DAQ via the original DMBs.

The ALCTs store raw wire-hit information in a pipeline within the FPGA while waiting for an L1A. The ALCTs installed before Run 1 would also suffer significant data loss during HL-LHC operation. This is because they do not have sufficient FPGA memory resources to hold all raw hit information before sending them to the DMB during the HL-LHC L1 trigger latency. Moreover, the output bandwidth for the boards in the inner rings (ME1234/1) would not be capable of handling the expected HL-LHC data rates. Both of these problems were solved by replacing the mezzanine cards in the affected chambers.

During LS1, the ME1/1 chambers were equipped with new ALCT mezzanine cards having more powerful FPGAs (Spartan 6), with 9–12 times more memory than those used previously (Virtex-E). This, in turn, allows the pipeline to be deep enough to satisfy the HL-LHC latency requirements. Similar ALCT mezzanine cards were also installed on the ME4/2 chambers during the same period.

During LS2, as a second phase of the ALCT upgrade, the ALCT mezzanine cards originally installed in ME1/1 during LS1 were moved to ME1/3, and all other mezzanine cards, except for those used in the ME4/2 rings, were replaced. The new ALCT mezzanine cards are largely based on the ALCT design used for cards installed in ME1/1 during LS1, with a few exceptions discussed below. The ALCT mezzanine cards servicing the ME234/1 rings use two new VTTx optical transmitters, each running at 3.2 Gb/s, to increase the output bandwidth for the expected HL-LHC data rates. The new ME1/1 ALCTs, in addition to a different FPGA from those used for the ME234/1 chambers, have a VTRx transceiver instead of two VTTx transmitters on each mezzanine card. This allows for the same remote FPGA programming option (based on the GBTx ASIC) that the DCFEBv2s have, thus mitigating any risks associated with the aging of the PROMs in that ring. From Run 4, the VTRx transceiver (as well as the VTTx transmitters from the ME234/1 ALCTs) will be connected to an updated ODMB board that transmits data to the CMS DAQ over an optical link at 4.8 GB/s. The new mezzanine cards serving the ME234/2 chambers have the same FPGA as those used for the ME1/1 ALCTs, but they will send data to the DMB through copper links even in the future runs due to the lower expected data rate. Since the new design maintains backwards compatibility, the anode raw data sent to the DAQ from the ME1234/1 ALCTs can be transmitted to the DMBs via the copper path during Run 3, and to the new ODMBs via the optical link(s) from Run 4 onwards.

The DCFEBs that were installed on the ME1/1 chambers in LS1 transmitted data to the CMS DAQ and comparator data to the CMS trigger at 3.2 Gb/s via optical fibers. Since the previously installed TMBs and DMBs were designed to work with the CFEBs and have no optical receivers, the TMBs and DMBs for the ME1/1 chambers were upgraded to OTMBs and ODMBs during LS1 as well. They receive optical data from the seven DCFEBs and use a more powerful FPGA. During

2024 JINST 19 P05064

LS2, all TMBs used for ME234/1 chambers were upgraded to OTMBs to match the upgrade of their CFEBs to DCFEBs. New OTMBv2s were produced for the ME1/1 and ME2/1 chambers. These are almost identical to the OTMBs installed in LS1, except that obsolete components were replaced and the ability to build trigger primitives including GEM data was added, as described in section 6.4.5. The OTMBs serving the ME1/1 chambers during Run 2 were relocated to the ME34/1 rings.

The new electronics consume more power than the old so the low-voltage (LV) systems were upgraded to provide higher maximum currents at the appropriate voltage levels to ensure reliable operation of the CSCs at the HL-LHC.

For Run 1 and Run 2, the CSC high-voltage (HV) system comprised two independent subsystems: a custom one providing HV to the non-ME1/1 chambers (11 016 channels), and a commercial subsystem supplying the ME1/1 chambers (432 channels) [106]. In both systems, the voltage can be regulated on all channels individually, and the currents drawn are continuously monitored.

The commercial system is no longer supported by the manufacturer and does not allow monitoring of the currents with as fine a precision as the HV system used for the non-ME1/1 chambers, as can be seen in figure 53. For these reasons, during LS2, the commercial HV subsystem was replaced by a custom-made HV one, extending the system used for the non-ME1/1 CSCs. Besides ensuring precise current measurements, such a replacement makes the entire CSC HV system homogeneous and simpler to operate and maintain.

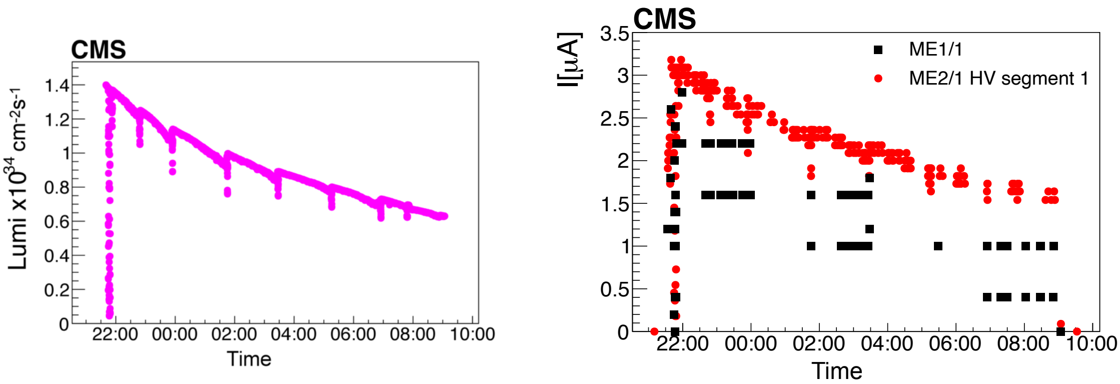


Figure 53. Left: instantaneous luminosity versus time for one of the LHC fills in 2016. Right: current (in nA) in an ME2/1 chamber (the HV segment closest to the beam) for the same fill, as measured with the custom-made HV subsystem used for non-ME1/1 chambers; current (in µA) in an ME1/1 chamber (one plane) for the same fill, as measured with the commercial HV subsystem.

6.2.3 Longevity studies

The CSC chambers will experience a much higher radiation dose than the muon chambers located in the barrel region. The highest neutron fluence and total ionization dose, corresponding to 10 years of HL-LHC running, will reach a level of $1 \times 10^{14}\,\mathrm{n_{eq}/cm^2}$ and 10 Gy, respectively. The total charge released in the gas volume per unit wire length will be about 0.3 C/cm for the CSCs closest to the beam line, corresponding to an integrated luminosity of $4000\,\mathrm{fb}^{-1}$, as maximally achievable by the end of the HL-LHC.

2024 JINST 19 P05064

Longevity tests of CSCs have been underway since 2016 at the CERN GIF++ facility [107]. Typical ME1/1 and ME2/1 chambers have been used. The ME1/1 chambers differ from those of the other rings, of which an ME2/1 chamber is typical, in having thinner wires, smaller gas gaps, and being constructed of different materials. The combination of a photon flux at high rates provided by the GIF++ gamma source and a muon beam periodically provided by the CERN SPS allows the study of not only the longevity of the CSCs but also their performance in an HL-LHC-like environment. The CSCs installed in GIF++ have the same chain of trigger and data acquisition electronics as those installed in CMS during Run 2. The working gas mixture of 40% Ar, 50% CO_2, and 10% CF_4 is supplied with a closed-loop gas system replicating the CSC gas system at CMS. The gas flow and gas refreshing rate are kept at the same level as for nominal CSC operation. While operating at the nominal gas gain under an intense irradiation for more than a year, the chambers accumulated a charge per wire length of 0.33 C/cm; this corresponds to about 1.1 and 1.7 times the amount of charge expected for ten years of HL-LHC running for the ME1/1 and ME2/1 chambers, respectively. Figure 54 shows a modest deterioration of spatial resolution at higher background rates and no deterioration with an accumulated charge up to 0.33 C/cm, the measurements were made with the nominal CSC gas mixture. The average current is proportional to the background intensity, which can be adjusted using a set of filters. The expected typical currents for Run 3 (at $2\times10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$) and HL-LHC (at $7.5\times10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$) background conditions are about 4 and 30 μA for the ME1/1 chamber, and 4.4 and 22 μA for the ME2/1 chamber. No significant deterioration of key chamber parameters such as gas gain, detection efficiency, spurious signal rates, strip-to-strip resistance, or dark currents has been observed. Thus, it is expected that the CSCs themselves will function throughout the planned HL-LHC operation.

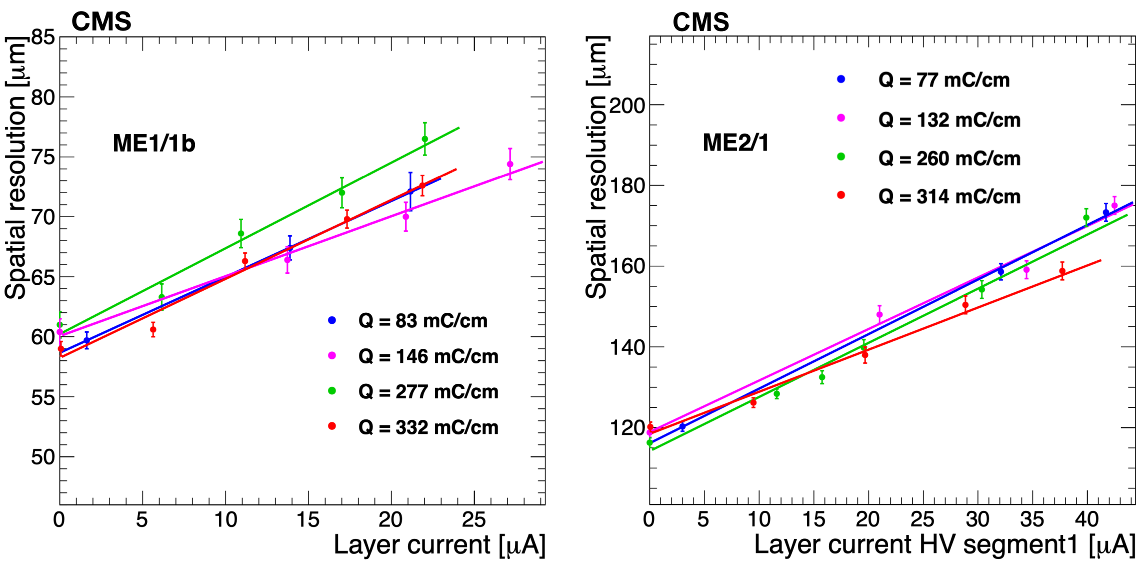


Figure 54. Spatial resolutions for an ME1/1b (left) and ME2/1 (right) chamber using a muon beam while being uniformly illuminated by a ^{137}Cs photon source to simulate the background from high luminosity pp collisions, as a function of the average current per layer in one HV segment. The results for four different accumulated charges per unit wire length are shown, along with linear fits to each set.

The chamber gas gain is set by the HV system, which can be individually adjusted in each of up to 30 segments per CSC. Except for the ME1/1 chambers, each wire plane in other CSCs is divided

2024 JINST 19 P05064

into three or five independent HV segments, which allows the independent regulation or turning off the HV on any of the individual sections. To maximize the lifetime of a CSC it should be operated at the lowest HV compatible with full efficiency. Until mid-2016, all CSCs operated at the same HV, and the average gas gains in different HV segments varied by up to a factor of 2, as shown by the wide blue distribution in figure 55. In 2016, the gas gains of individual CSCs were modified by tuning each HV channel to reduce the spread, as shown by the narrow red distribution in figure 55. This optimized the CSC gas gains for good efficiency without having unnecessarily high HV on any chamber, thus maximizing chamber longevity. Later, in 2018, the overall HV of all rings except ME234/1 was reduced by about 30 V, which decreased the gas gain by about 20%. The chambers now operate just above the knee of the efficiency plateau in gas gain versus HV and hence remain fully efficient, while the CSC system lifetime is expected to be extended by 20%.

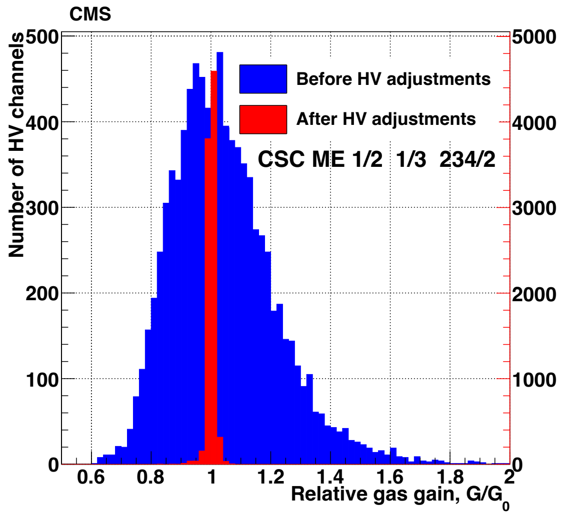


Figure 55. Relative gas gain distribution in CSCs before and after the gas gain equalization campaign in 2016 [100]. Each entry in the histogram presents the mean value of gas gain in each HV channel. The scale of the blue histogram is on the left while the scale of the red histogram is on the right.

The CSC readout electronics can also degrade after exposure to large radiation doses. A systematic program of irradiating the CSC electronic components was performed to identify any of those unable to operate reliably in the HL-LHC radiation environment. All the components used in both the old and upgraded readout boards were found to withstand more than 3 times the expected HL-LHC doses, except for the PROMs used in the frontend boards installed in the ME1/1 chambers, which can only withstand 1–1.5 times the expected dose.

Before the production of the new electronics boards for the Phase 2 upgrade, radiation tests were carried out at the Texas A&M cyclotron and nuclear reactor [108], and at the UC Davis cyclotron. During the tests at the Texas A&M cyclotron, the digital components on the test boards were operated with active data readout while being irradiated with 55 MeV protons. The components tested for SEUs and single-event latch-ups (SELs) included the FPGAs, PROMs, level adapters, and optical transmitters and receivers. In reactor tests, components were exposed to neutrons with energies up to a few MeV, with exposures equivalent to a TID of 30 kRad. This corresponds to a level of neutron radiation equivalent to about 50 years of that expected at the HL-LHC at the location where

CSC electronics are exposed to the highest radiation flux (the inner portion of the ME1/1 chambers). These tests targeted mostly nondigital components such as voltage regulators and power diodes. The results showed that the components selected for the new electronics will operate reliably in the CMS radiation environment at the HL-LHC.

During Run 2, a new campaign of radiation testing was initiated for the upgraded electronics described above. Particular attention was paid to the PROMs, which are known to have some failures after large radiation exposure [109]. The PROMs in the DCFEBs and ALCTs installed during LS1 were tested at the CHARM II mixed radiation facility at CERN, the Texas A&M reactor, and the UC Davis cyclotron. The PROMs performed well up to an exposure of 10 kRad, but both types of PROMs showed some failures at exposures of 15–30 kRad. To mitigate such failures, the option to perform promless programming of the frontend board FPGAs for the ME1/1 chambers, where the radiation is most severe, has been provided. The optical receivers used on the OTMBs upgraded during LS2 were also tested at the UC Davis cyclotron and proved able to sustain the expected HL-LHC radiation dose.

The CSC system has been successfully operating during Run 3 after an enormous upgrade effort in LS1 and LS2. Its longevity has been studied extensively, and it is expected to remain reliable throughout the future running until the end of the HL-LHC era.

6.3 Resistive plate chambers

6.3.1 General description

The CMS resistive plate chambers (RPCs) are gaseous detectors equipped with two gas gaps each having a width of 2 mm and a copper readout plane in between, as shown in figure 56. High voltage is applied to the graphite electrodes, which are coated on the surface of high-pressure (HPL) laminate plates with bulk resistivity in the range of 2–5 $\times 10^{10}$ Ωcm. The chambers are operated in avalanche mode with a gas mixture of 95.2% $C_2H_2F_4$, 4.5% i-C_4H_{10}, and 0.3% SF_6. This allows the chambers to cope with high background rates and ensures an excellent time resolution, as summarized in table 7, facilitating a precise bunch-crossing assignment.

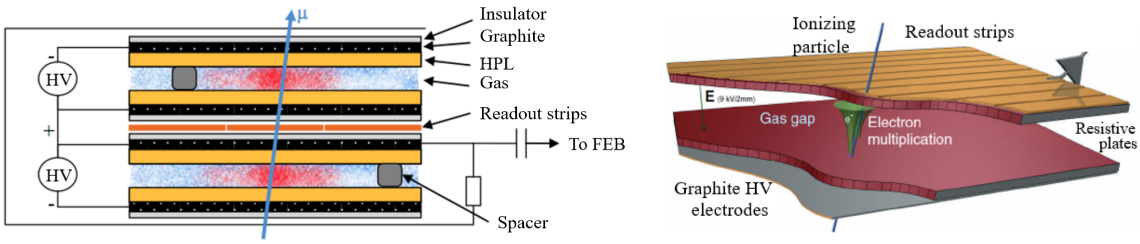


Figure 56. Left: schematic of the double-layer layout of the RPC chambers. Reproduced with permission from [110]. Right: illustration of the RPC technology. Reproduced with permission from [111].

The RPC barrel detector is divided in the direction along the beam axis into five separate wheels labelled W±2, W±1, and W0. The RPCs in each wheel consist of six layers. The first four layers, called RB1in, RB1out, RB2in, and RB2out, are located on the inner and outer sides of the inner two stations of the DT chambers. The other two layers, labelled RB3 and RB4, are located on the inner side of the third and fourth stations of the DTs. There are four disks in each endcap, called RE±4, RE±3, RE±2, and RE±1. Spanning the ϕ direction, each wheel is divided into twelve sectors and each disk into 36 sectors. Due to requirements in the trigger logic, the chambers are divided into two or three pseudorapidity (η) partitions, called rolls. In most of the barrel, there are two rolls: forward and backward. Only RB2in in W±1 and W0 and RB2out in W±2 are divided into three rolls, called forward, middle, and backward. The endcaps are divided into three rolls: A, B, and C. The RPC system consists of 1056 chambers, covering an area of about 3950 m^2, equipped with 123 688 readout strips. In the barrel, the strips are rectangular in shape with a pitch in the range between 2.28 and 4.10 cm, while in the endcaps they are trapezoidal with a pitch between 1.74 and 3.63 cm.

6.3.2 RPC system upgrades since Run 1

Endcap fourth station upgrade. One of the major upgrades of the CMS experiment, performed during the first long shutdown (LS1) of the LHC in 2013–2014, was to add 144 new RPCs to the fourth stations of the detector endcaps. Adding these stations increased the overall robustness of the muon spectrometer and improved the trigger efficiency in the endcap region over the range $1.2 < |\eta| < 1.6$.

The new RE4 RPCs inherit their design from the previously installed endcap chambers [112]. They are grouped in 72 supermodules, 36 per disk, each consisting of two detectors and covering 10° in azimuthal angle. Prior to their installation in CMS, all RE4 chambers (both RE ±4) passed a series of tests at all production stages in order to ensure their quality and performance.

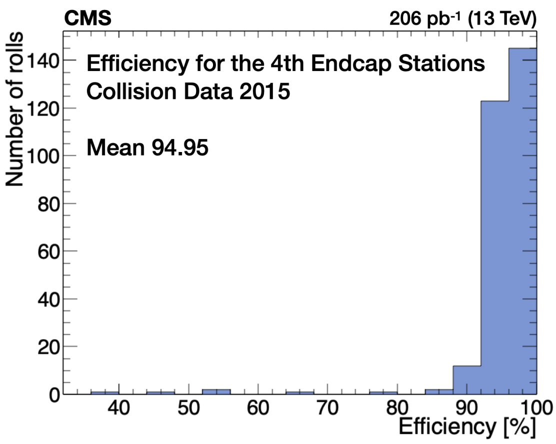


Figure 57. Efficiency distribution of the RE±4 stations in their first year of operation in 2015.

This newest part of the RPC system was successfully commissioned in 2015 at the beginning of the Run 2 data taking. The performance of RE4, shown in figure 57, is in good agreement with expectations [113].

Readout changes and Phase 1 RPC trigger contributions. As mentioned in section 10.2, the Phase 1 level-1 (L1) trigger upgrade [91] moved from a muon detector-based scheme to a geometry-based system. The three muon trigger systems were replaced by three track finders, each covering a specific pseudorapidity region.

The RPC pattern comparator trigger (PACT) has the function of reading out the information from the RPC frontend boards (FEBs). These are installed at chamber level via the link board system, located in the CMS experimental cavern in the balconies next to the detector. In the link boards, the low-voltage differential signaling (LVDS) signals from the FEBs are synchronized with the LHC clock, converted to optical signals, and sent to the RPC trigger boards in the service cavern. Since the beginning of Run 2, the RPC signals coming from the link boards are split and sent to the muon processors TwinMux, as described in section 6.1.2, and to the concentration pre-processing and fan-out (CPPF) [114]. Both TwinMux and CPPF do the hit clustering and cluster selection and forward the produced clusters to the track finders (TwinMux to the BMTF and CPPF to the EMTF). The OMTF, described in section 10.2, reads the RPC data straight from the link boards, then combines them with the full list of DT trigger segments directly at the track-finding step and performs the clustering and selection.

The contribution of the RPC system to the refactored L1 muon trigger architecture is different for the three muon track finders and can be summarized as follows:

- BMTF: RPC timing information is used to improve the DT trigger primitives' bunch crossing assignment. In the first two stations, where two RPC layers are present, a segment is built from the coincidence of hits in the inner and outer chambers. The latter provides redundancy in case of DT inefficiencies.
- OMTF: RPC hit position information is used standalone from the eight available chambers (five in the barrel and three in the endcap), per ϕ division (sector).
- EMTF: RPC hit position information is used in case there is no corresponding CSC trigger primitive.

The combination of information from the barrel muon detectors at an early stage allows the exploitation of the system redundancy already at the step of building the trigger primitives. As shown in the left plot of figure 58 [95], the combination of DT and RPC leads to an average increase of the station-1 barrel trigger-primitive efficiency of about 1.4%, raising the plateau value from 95% to about 96.5%. However, this is not the only gain from the new architecture. The trigger also benefits from detector complementarity since the use of RPC timing information reduces the number of out-of-time DT trigger primitives, as shown in figure 43.

For the OMTF, the use of the RPC complementarity is even more prominent. Figure 58 (right) shows the efficiency as a function of the reconstructed muon p_T in the η region of the OMTF. If RPC data are not present, the efficiency decreases by about 15%.

6.3.3 RPC system longevity

Detector stability studies. Continuous studies are performed throughout the data-taking periods to ensure correct operation and performance stability of the RPC system.

RPC working point calibration. To ensure the most stable performance possible, the operational high voltage V_{app} of the RPCs is controlled such that the effective high voltage V_{eff} is constant even when environmental conditions change during the data taking [116]. The relation between V_{app} and V_{eff} is as follows [117]:

$$V_{app} = V_{eff}\left[1 - \alpha + \alpha\left(\frac{P}{P_0}\right)\left(\frac{T_0}{T}\right)\right], \tag{6.1}$$

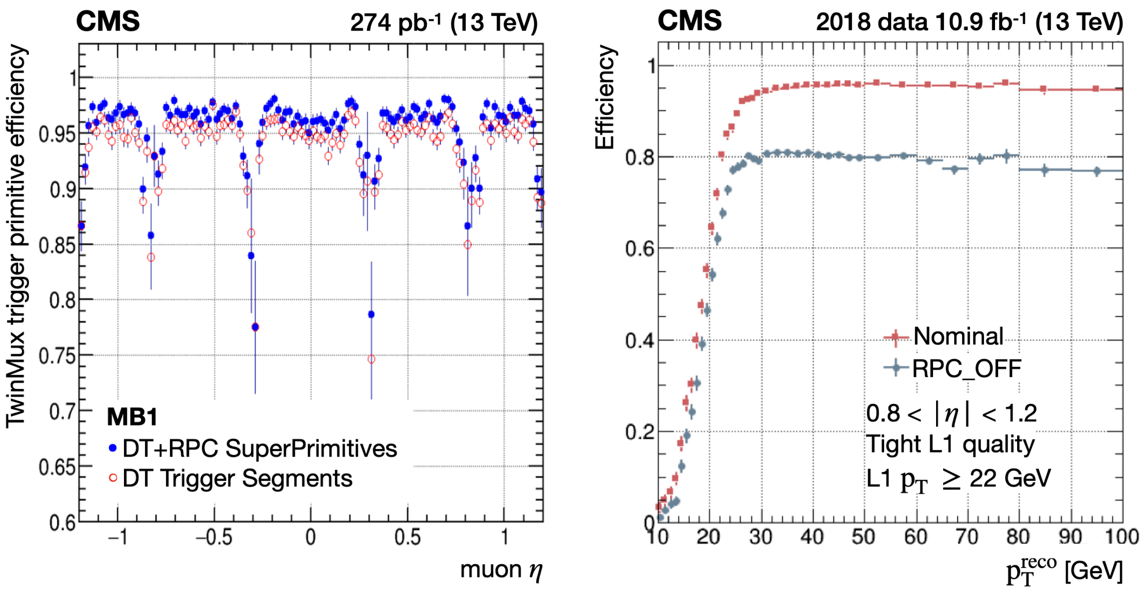


Figure 58. Left: station-1 barrel trigger-primitive efficiency as a function of muon pseudorapidity. Right: trigger efficiency as a function of muon p_T for the OMTF, derived from a trigger emulation applied to real data, using (red) and not using (blue) RPC information. Reproduced from [115].

where P and T are the actual pressure and temperature in the CMS cavern, α is a special pressure correction that equals 0.8, and P_0 and T_0 correspond to values of 965 mbar and 273 K, respectively.

To determine the optimal operating voltage for every chamber, a series of high voltage scans are regularly performed, typically in low and high luminosity conditions. Data are collected for particular detector configurations: the V_{eff} values are equidistantly chosen within a range of 8.8 and 9.8 kV. The data are selected using triggers from the DT and CSC and then reconstructed using the standard CMS muon reconstruction [8]. During HV scan periods, no correction due to pressure or temperature is applied to the voltage. The segment extrapolation method [85] is used to measure the RPC hit efficiency for each HV point. Segments from the nearest DT and CSC chambers are extrapolated to the RPC planes, and the reconstructed RPC clusters (rechits) are matched to the extrapolated points. The efficiency is then calculated as a ratio of the numbers of matched RPC clusters to the extrapolated segments. The resulting efficiency distribution (ε) is fitted for every chamber using a sigmoid function defined as

$$\varepsilon(V_{\text{eff}}) = \frac{\varepsilon_{\text{max}}}{1 + \exp[-\lambda(V_{\text{eff}} - V_{50\%})]}, \tag{6.2}$$

where λ characterizes the slope of the sigmoid at the inflection point, ε_{max} represents the asymptotic efficiency for $V \to \infty$, and $V_{50\%}$ is the inflection point for which ε_{max} reaches an efficiency of 50%.

The optimal working point (WP) for each chamber is determined by interpolating the efficiency distribution ε using the fitting parameters in eq. (6.2), and calculating the voltage corresponding to 95% of ε_{max}. Then, the WP for each chamber is defined as $V_{95\%}$ + 100 V (120 V) for the barrel (endcap) chambers, respectively. The application of individually different offsets leads to similar overall efficiencies in the barrel and endcaps [116, 117].

Since each HV channel supplies two chambers in the endcaps (with a few exceptions), the WP in the endcaps is the average of the WP of the two corresponding chambers for each HV channel. All chambers that operate in single-gap mode or have HV problems are excluded from this method.

Figure 59 shows the evolution of the efficiency as a function of time, as determined from the HV scan data, evaluated at the WP and at $V_{50\%}$ during the LHC Run 1 and Run 2. Despite the changes in the environmental and luminosity conditions, the different calibrations implemented in the detector allowed us to keep the efficiency stable.

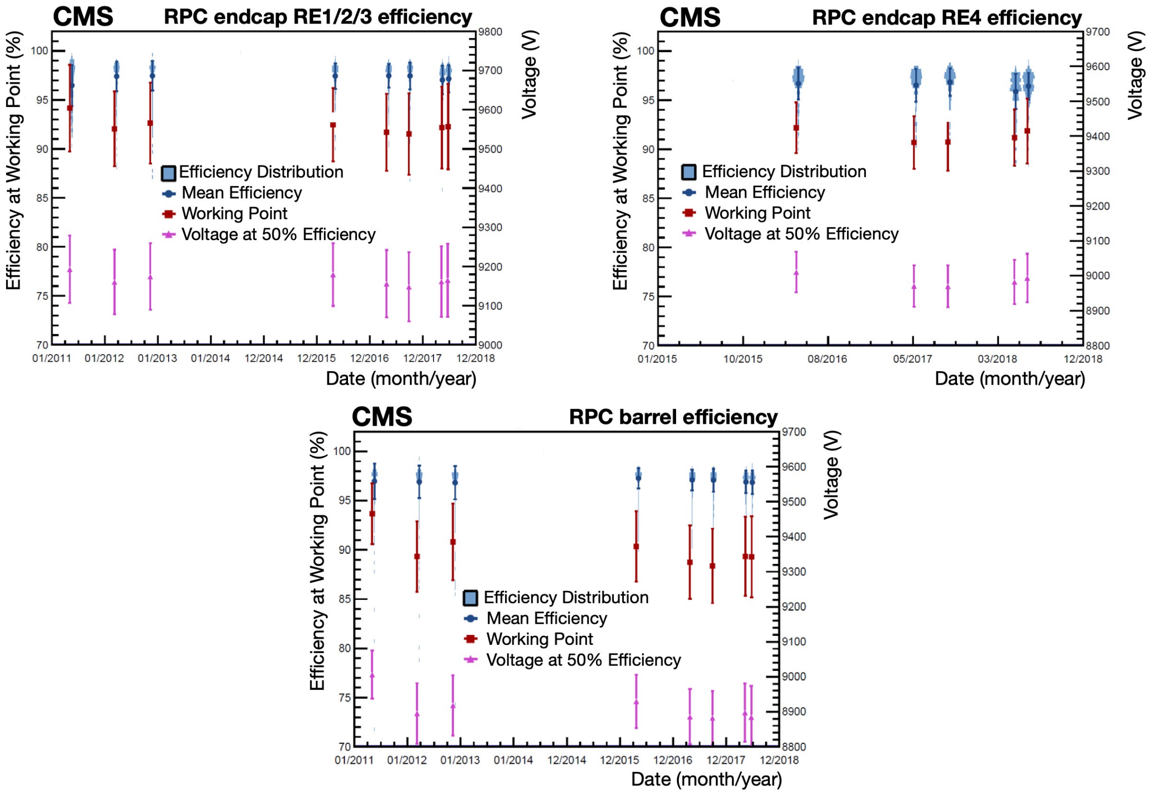


Figure 59. Temporal evolution of efficiencies determined from HV scan data at the WP and at $V_{50\%}$, for RE1/2/3 (upper left), RE4 (upper right), and the barrel (lower). The light blue bands show the histograms of the distributions, where the width of the band represents the population of channels having the corresponding efficiency value.

RPC efficiency and cluster size stability. One of the most important measures of the RPC system performance is the efficiency of the rolls, described in section 6.3.1. The efficiency of a roll can be measured as the ratio of the number of reconstructed hits to the number of muons passing through the roll. To obtain the position of the muon trajectory in the RPC layers, the muon track is extrapolated using either the track parameters or the direction of the track segment from the closest other subdetector in the muon system. During the Run 2 data-taking periods, the analysis used both methods: the track [118] and segment [85] extrapolation.

Figure 60 shows the overall efficiency distribution of the RPC rolls in Run 2. The numbers are for all well-performing RPC rolls. Rolls with an efficiency lower than 70% are excluded if the efficiency drop is caused by a known hardware problems, such as a gas leak.

2024 JINST 19 P05064

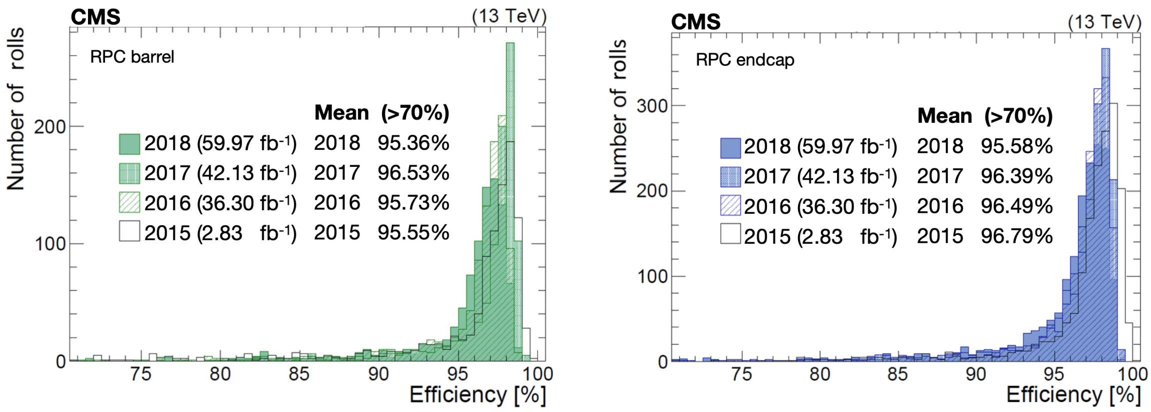


Figure 60. Distributions of the overall RPC efficiencies in the barrel (left) and endcaps (right) during pp data taking in Run 2 [119]. Reproduced from [120].

The cluster size (CLS) is another important quantity since it affects the RPC spatial resolution. It is defined as the number of adjacent strips firing when a discharge is produced in the RPC. The RPC system has an average cluster size of less than three strips, constant over the years, in agreement with expectations [1]. Keeping the cluster size stable over time is one of the most important prerequisites in the correct operation of the RPC system. During Run 2, the chamber cluster size was monitored run-by-run to guarantee the stability of the system.

The efficiency and cluster-size history for the Run 2 data taking are shown in figure 61. The history follows the changes of the applied HV WPs and changes of the isobutane concentration in the gas mixture. The spread in the cluster size distribution in 2015 was caused by a threshold control problem, which was resolved in October 2015, while the small drop of the RE+4 cluster size in 2016 was due to a temporary LV problem, solved at the end of August 2016. The drop of efficiency in the period 01–19 August 2018 came from a configuration problem. Detailed Run 2 performance results can be found in refs. [119, 121].

One of the main objectives of the RPC analysis is to monitor the system performance with respect to the luminosity. Figure 62 displays the barrel and negative endcap efficiency and cluster size as a function of the instantaneous luminosity during pp data taking in 2016 and 2017. For this comparison, data recorded with the same WP are used. The lower efficiency and cluster size for the barrel in 2016 are caused by the higher isobutane concentration during that year.

The comparison between the 2016 and 2017 results shows a stable efficiency and cluster size. The obtained results can be linearly extrapolated to the expected luminosity at the HL-LHC of $7.5 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. A reduction in efficiency of 1.35% is predicted for the barrel and 3.5% for the endcaps. No change is expected for the cluster size under the HL-LHC conditions.

Monitoring of RPC currents. The current drawn by the RPC detectors is one of the base parameters to be considered, both as a measure of the working condition and as an indicator of possible problems. The current should be kept as low as possible to guarantee long-term stability of the RPC gas gaps, and its time evolution is one of the most important input data for longevity studies.

The ohmic current of the RPC system is defined as the current with no beam in the accelerator, at a HV around 7000 V, where the current follows an ohmic law and there is no contribution from

2024 JINST 19 P05064

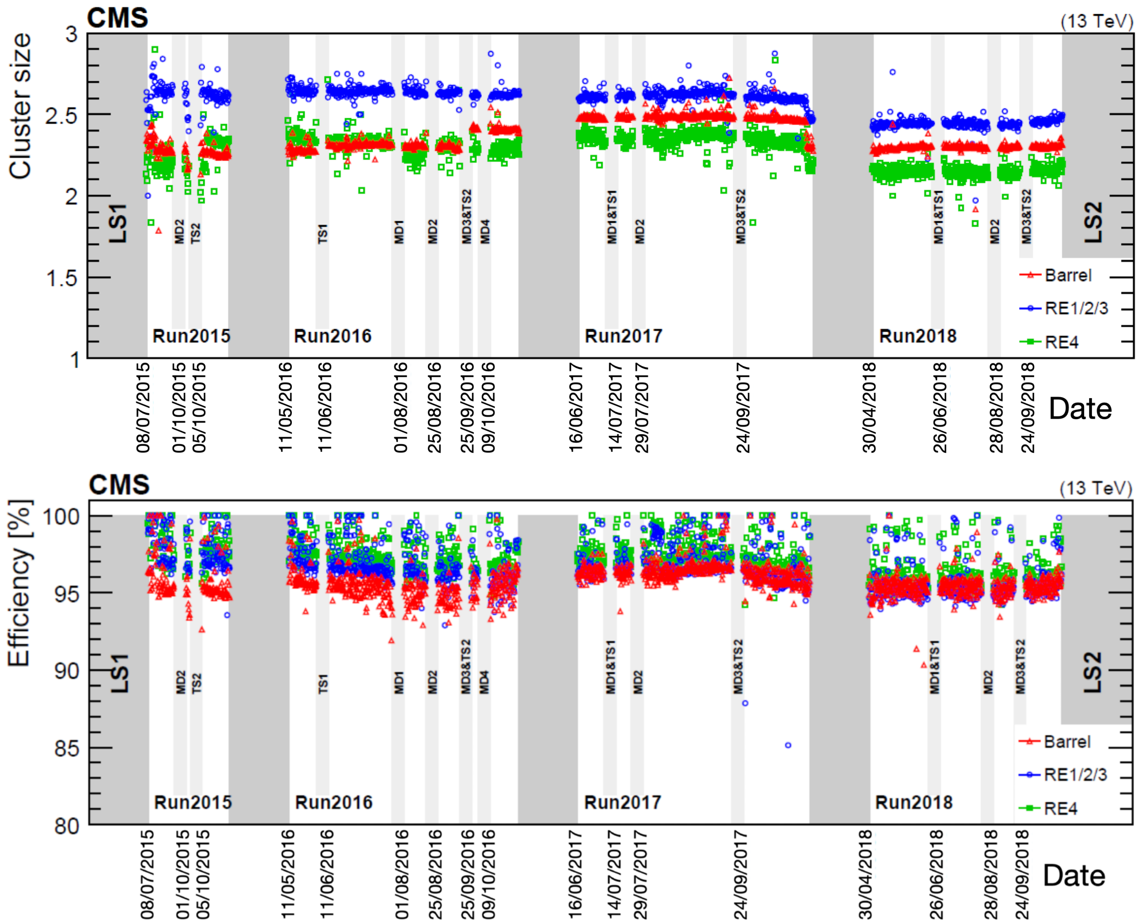


Figure 61. History of the RPC efficiency (upper) and cluster size (lower) during Run 2. Gray areas correspond to the scheduled technical stops.

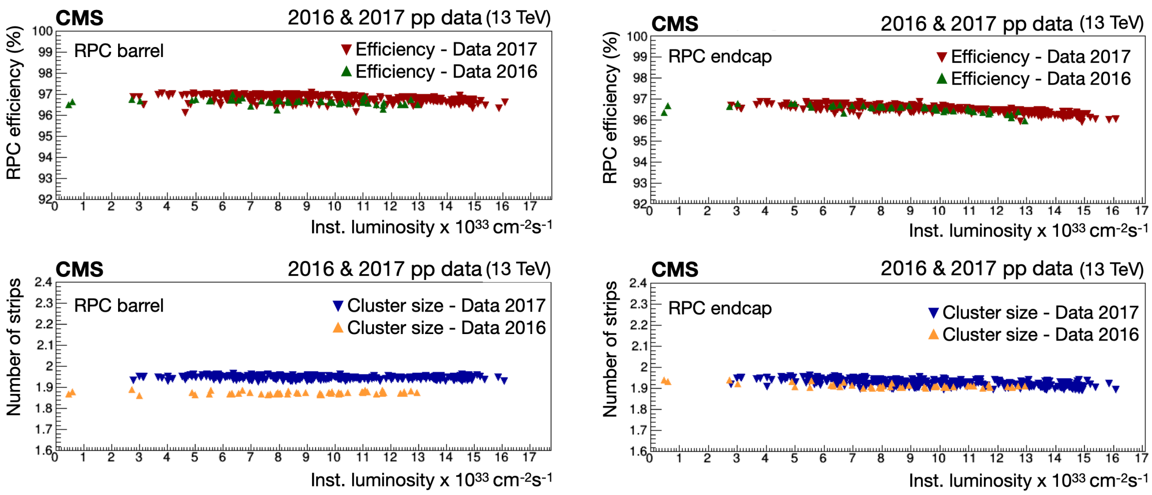


Figure 62. RPC barrel and endcap efficiency (upper) and cluster size (lower) as a function of the LHC instantaneous luminosity for pp collisions in 2016 and 2017. The linear extrapolation to the instantaneous luminosity expected at the HL-LHC of 7.5×10^{34} cm^{-2} s^{-1} shows a 1.35% reduction in efficiency for the barrel and 3.5% for the endcap. Reproduced from [121]. © 2019 IOP Publishing Ltd and Sissa Medialab. All rights reserved.

gas amplification. In practice, the ohmic current values are monitored at 6500 V. The "cosmic current" is defined as the current without beam, at the WP voltage, in the region of gas amplification.

For Run 3, the monitoring capability for the RPC detector currents was improved. An especially designed tool was implemented that automatically stores and classifies the currents in a database as a function of different parameters, such as the CMS magnetic field, the instantaneous luminosity, and other parameters. The tool can identify blocks of data of special importance, such as the periods of HV scans, described in section 6.3.3.

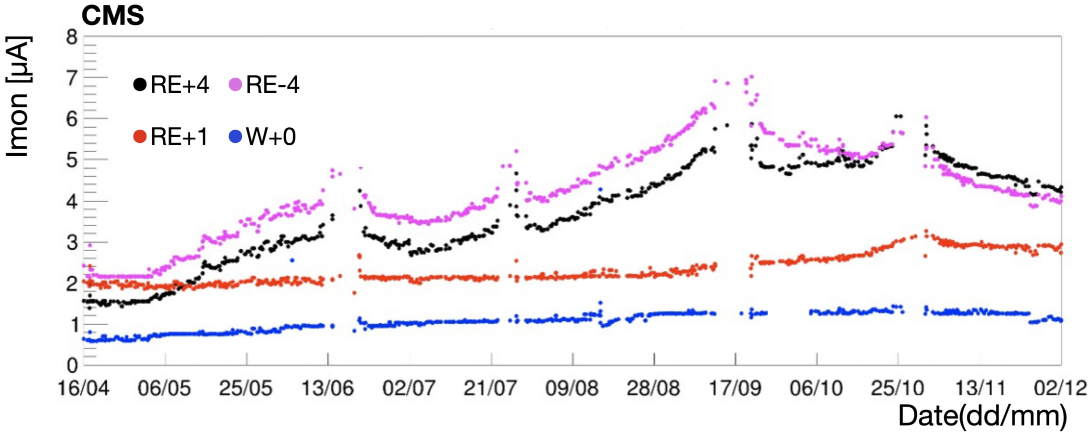


Figure 63. Ohmic current history in W+0, RE+1, RE+4, and RE−4. Reproduced from [120].

Figure 63 shows the currents measured in four RPC stations, W+0 in the barrel, and RE+1, RE+4, and RE−4 in the endcaps. The measured currents are shown as a function of time. An increase of the ohmic currents is observed, in particular for chambers that are more exposed to background, and a linear dependence of the rate on the instantaneous luminosity is seen. In low-background regions such as W+0, the ohmic current increases very slowly. In RE+1 and W+0, the background rate is less than 10 Hz/cm^2. Both have similar gas flows, corresponding to a volume exchange per hour (v/h) of 0.7 and 0.6, respectively. In contrast, in RE4, the background rate is about 40 Hz/cm^2 and the gas flow is 0.35 v/h, which is lower than the rest of the system, due to a wrong calibration of the flow cells. These rates are the maximum values measured in the upper sectors, i.e., the detector sectors above the beam pipe, at instantaneous luminosities of 1.5–2 × 10^{34} cm^{-2} s^{-1}.

The RPC currents depend linearly on the instantaneous luminosity [122]. For each LHC fill, the distributions were fit to a linear function to obtain the slope P_1, also known as the physics current ($i = P_1 L$). Due to the nature of the linear fit, an offset P_0 absorbs the cosmic current (offset + ohmic + gas gain). The slopes P_1 as a function of time for the endcap stations are shown in figure 64. The slope of the RPC current distribution is stable in time. The changes in the middle of August 2018 are due to different HV WPs. Endcap stations that are located at equal distances from the interaction point along the beam pipe have similar slopes. They also have similar rates [122]. No increase due to the integrated luminosity is observed for the slopes over the entire year.

Hydrogen fluoride (HF) is produced in the gas under high electrical discharge. The HF has a high chemical reactivity and electrical conductivity [123], and it is therefore expected to be a source of inner-detector surface damage and relative ohmic current increase that accelerates detector aging.

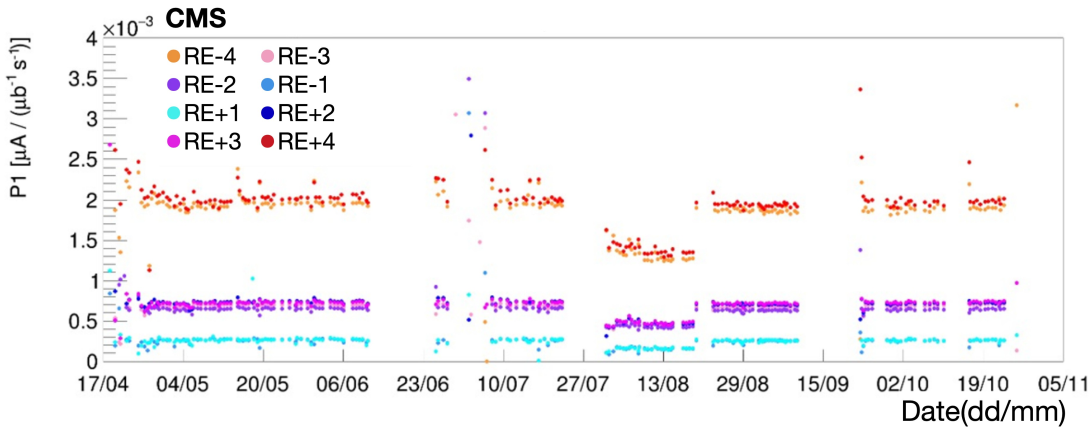


Figure 64. RPC physics current history in RE±1, RE±2, RE±3 and RE±4. Reproduced from [120].

In summer 2018, measurements of the HF concentration were performed at the gas exhausts of three regions: W+0 in the barrel, and RE+1 and RE+4 in the positive endcap.

The ohmic currents as a function of HF concentration are shown in figure 65. The RE+1 and W+0 chambers have a similar HF concentration and gas flow (0.7 and 0.6 v/h), and a background of less than 10 Hz/cm^2. In RE+4, the amount of HF accumulated is higher by a factor 2 at a background of 40 Hz/cm^2, and the gas flow is lower (0.35 v/h) than in W+0 and RE+1. There is a clear linear dependence between the ohmic current and the HF concentration, which implies that HF trapped in the gas gap may form a thin conductive layer. The HF can be efficiently removed by increasing the gas flow in the chambers, depending on the background rate.

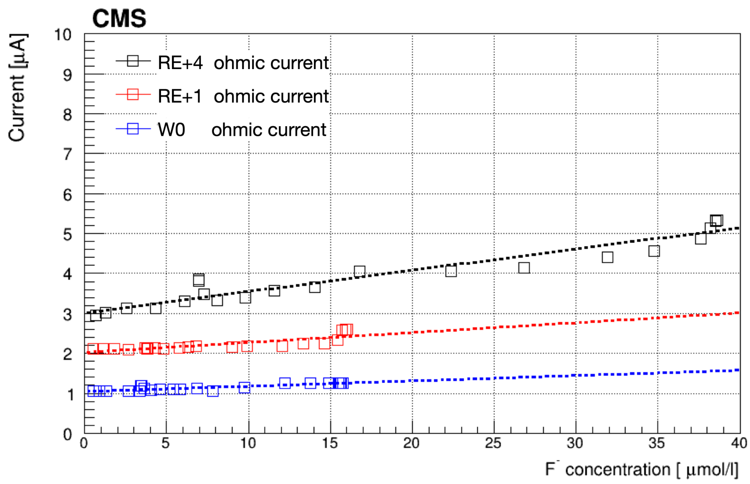


Figure 65. Ohmic current as a function of HF concentration. Reproduced from [120].

GIF++ longevity studies. The present RPC system was originally certified for ten years of LHC operation at a maximum background rate of 300 Hz/cm^2 and a total integrated charge of 50 mC/cm^2 [124, 125]. Based on the data collected in Run 2, assuming a linear dependence of the

2024 JINST 19 P05064

background rates as a function of the instantaneous luminosity, and including a safety factor of three, the expected background rates and integrated charge at the HL-LHC will be about 600 Hz/cm^2 and 840 mC/cm^2, respectively [100]. In such operating conditions, irrecoverable aging effects can appear, as the higher collision rates affect the detector properties and performance.

Therefore, since July 2016, a long-term irradiation test has been carried out at the CERN gamma irradiation facility (GIF++) [126] to study whether the present RPC detectors can survive the difficult background conditions during the HL-LHC running period [127].

Four spare RPC chambers have been irradiated, two each of type RE2/2 and RE4/2 [1, 100]. These are from the endcaps where the backgrounds are expected to be maximal [128]. Two different RPC production types have been tested, reflecting the fact that the RPC endcap production was done in two different periods, 2005 for the RE2 detectors (both RE±2) and 2012–2013 for the RE4 detectors (both RE±4). Two chambers, one from each period, are continuously being irradiated while the other two of the same type are kept as reference and are switched on only from time to time. The detector parameters, such as dark currents, noise rates, currents, and count rates for various background conditions are monitored continuously and compared with the measurements from the reference chambers.

The integrated charge is calculated as the average density current accumulated in time in the three gaps that constitute the detector, since the gamma flux, provided by the 14 TBq ^{137}Cs source at the GIF++ [126] is uniformly distributed over the detector surface. The collected integrated charge from the beginning of irradiation until September 2022 is about 813 and 478 mC/cm^2 for the RE2 and RE4 chambers, respectively, which corresponds to approximately 97 and 57% of the expected integrated charge at the HL-LHC.

Dark current and noise rate studies. Dark currents and noise rates are monitored periodically in order to spot aging effects due to irradiation. The dark-current density, i.e., the currents normalized to the surface area, for both the irradiated and reference RE2 chambers are shown in figure 66 as a function of the collected integrated charge. The dark currents were measured at 6.5 kV to determine the ohmic contribution, and at 9.6 kV to include the contribution from gas amplification.

Figure 67 (left) shows the dark-current density monitored as a function of the effective high voltage, i.e., the voltage normalized to the standard temperature of 20°C and pressure of 990 mbar [112] at different values of collected integrated charge. Since the beginning of irradiation, the dark currents have been stable in time, with only small acceptable variations. Figure 67 (right) displays the average noise rate for the irradiated and reference RE2 chambers as a function of the collected integrated charge. The average noise rate is stable with time and less than 1 Hz/cm^2.

Resistivity and current studies. Other important parameters that are measured periodically are the current in the presence of background radiation and the resistivity of the electrodes. The latter is measured several times per year, since it is a crucial performance parameter. The resistivity is measured by filling the detector with pure argon and operating it in a self-sustaining streamer mode, which occurs when the gas-quenching components such as isobutane are removed. The streamers propagate over the entire detector area, and by measuring the current at a given high voltage, the resistance, and hence the resistivity, can be calculated. The measured resistivity values are normalized to 20°C to allow comparison of the values for different temperatures [130].

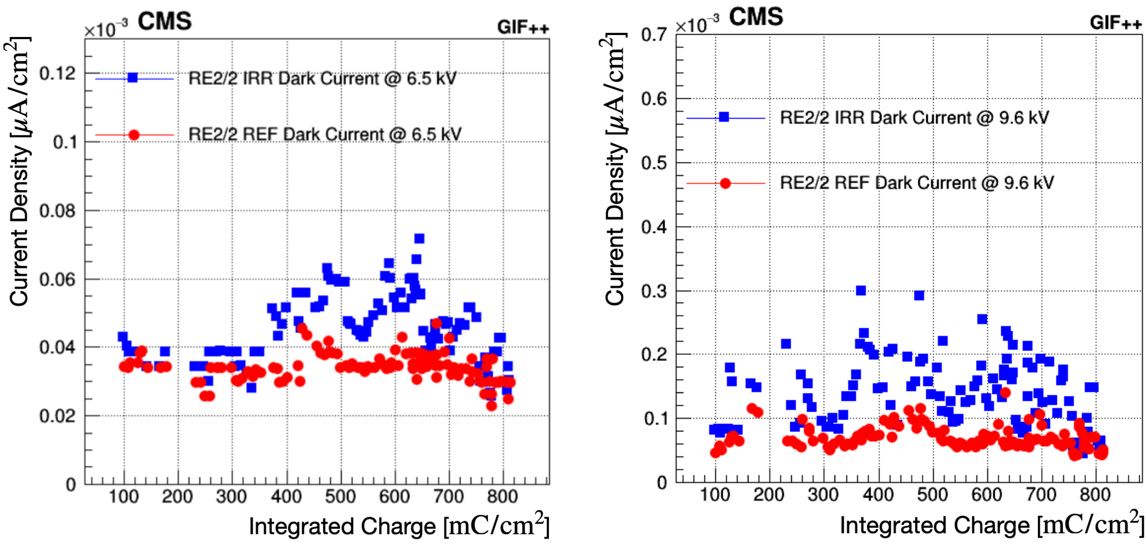


Figure 66. Dark-current density for the irradiated (blue squares) and reference (red circles) RE2 chambers as a function of the collected integrated charge at 6.5 (left) and 9.6 kV (right).

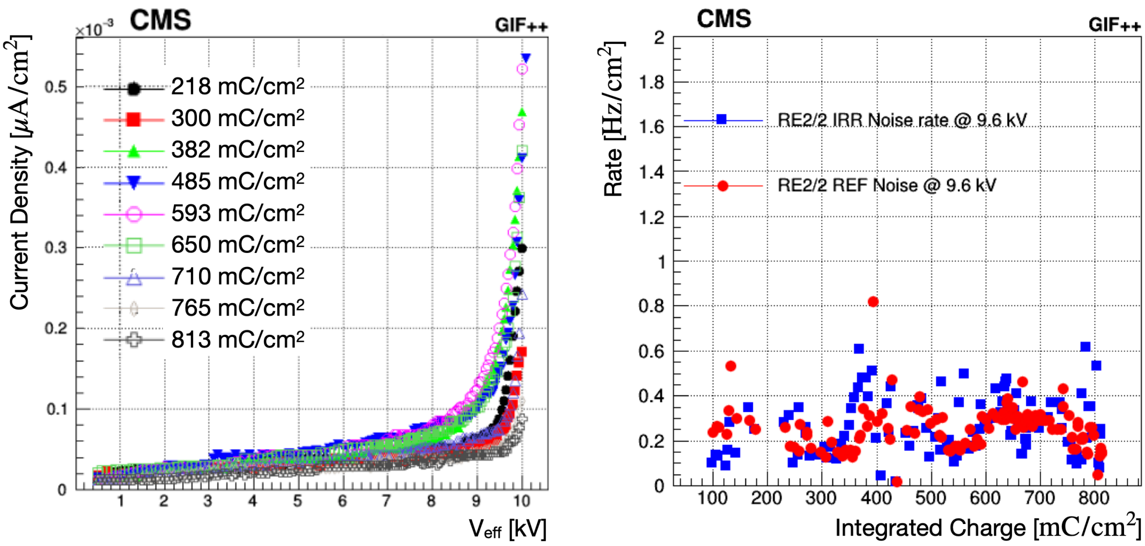


Figure 67. Left: dark-current density monitored as a function of the effective high voltage at different values of the collected integrated charge for the irradiated RE2 chamber. Right: average noise rate as a function of the collected integrated charge for the irradiated (blue squares) and reference (red circles) RE2 chambers. Reprinted from [129], Copyright (2023), with permission from Elsevier.

To exclude effects from external parameters, the ratios of resistivity and current between irradiated and reference chambers are taken for a gamma background rate of about 600 Hz/cm^2, as shown in figure 68. An increase in the resistivity is observed in the irradiated chamber during the first irradiation period, up to $\approx$300 mC/cm^2, when the detectors are operated in conditions similar to those in CMS: one gas volume exchange per hour and a relative gas humidity of 35–45%. While the RPC system operates in these conditions of humidity and gas volume exchange, they are not the optimal ones for operation at GIF++, where the high background gamma rate is about 600 Hz/cm^2, causing the HPL plates to dry

2024 JINST 19 P05064

and their resistivity to increase. During the longevity test, after an integrated charge of ≈300 mC/cm^2 was reached, the relative gas humidity was increased and maintained at about 60%, and the gas flow was increased to three gas volume exchanges per hour. With these changes, the HPL resistivity decreased and the variations were reduced. After ≈650 mC/cm^2, an increase in resistivity is observed, which is related to the return to a low gas humidity at about 40% due to sharing the same gas system with other RPCs. The increase of resistivity is confirmed by the decrease of the measured current.

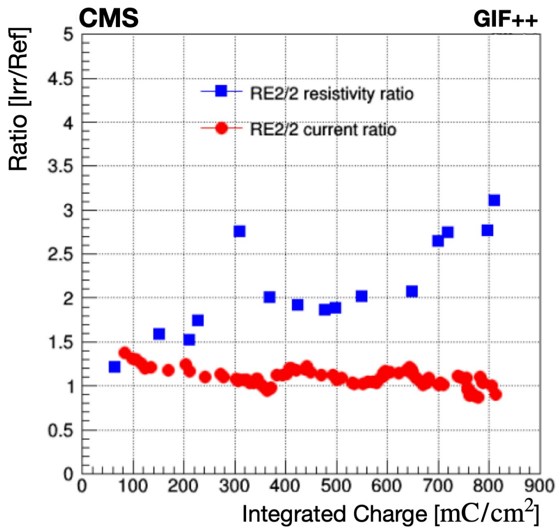


Figure 68. Resistivity ratio (blue squares) and current ratio (red circles) between the irradiated and reference RE2 chambers as a function of the collected integrated charge. Reprinted from [129], Copyright (2023), with permission from Elsevier.

Detector performance monitoring. The detector performance has been measured during test beams before irradiation and at different periods of irradiation. Comparisons between the irradiated RE2 chamber efficiency measured as a function of the effective HV without background radiation and in the presence of a background of 600 Hz/cm^2, at different values of collected integrated charge, shows that the efficiency is stable over time in the absence of background radiation, and no shift in the WP is observed [127, 131]. In the presence of background, the efficiency is stable at the WP, but a WP shift of 100 V after collecting 260 mC/cm^2 of integrated charge is introduced. This shift in the detector WP is related to the increase in the resistance (R) of the electrodes, which causes an increase of the voltage drop (RI) across them. The latter leads to a difference between the effective voltage V_{eff} applied to the electrodes and the effective voltage across the gas gap V_{gas}, which is compensated by introducing the shift [132]. A similar increase in resistivity is seen in the irradiated RE4 chamber. The quantity V_{gas} is defined as:

$$V_{\text{gas}} = V_{\text{eff}} - RI, \tag{6.3}$$

where R is the HPL resistance and I is the total current.

Since the detector operation regime is invariant with respect to V_{gas}, the efficiency as a function of V_{gas} does not depend on the HPL resistance, as shown in figure 69 (left). After removing the R-increase effect on the electrodes by introducing V_{gas} instead of V_{eff}, all the efficiency curves overlap, and no shift in the WP is observed.

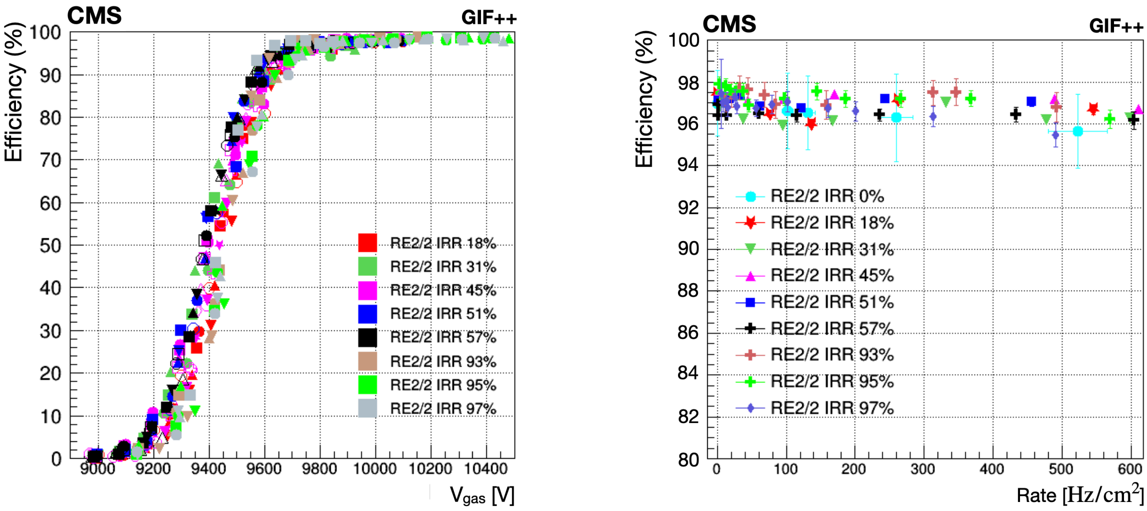


Figure 69. Left: irradiated RE2/2 chamber efficiency as a function of V_{gas} for different background irradiation rates, up to 600 Hz/cm^2, and different integrated charge values. Different marker shapes of the same color represent different background rates at the same integrated charge values. Right: irradiated RE2 chamber efficiency at the WP as a function of the background rate at different values of the collected integrated charge. Reprinted from [129], Copyright (2023), with permission from Elsevier.

The irradiated RE2 chamber efficiency at the WP is measured at different background rates and at different integrated charge values, as shown in figure 69 (right). The efficiency is stable over time up to the highest background rate expected at the HL-LHC.

Longevity studies on spare RPCs are ongoing at GIF++ under controlled conditions. Preliminary results show no evidence of aging effects. The main detector parameters and performance are stable.

6.3.4 Changes to the RPC system in LS2

Gas leak repairs and green-house gas emission strategy for Run 3. The highest priority during LS2 was the consolidation of the RPC gas system. The aim was to minimize the gas emissions and thus the environmental impact, since the standard gas mixture is composed mainly of F-gases with high global-warming potential. To accomplish this, actions were taken to minimize the number of leaks and to implement a newly developed Freon recuperation system prototype [133].

The RPC gas system is a 13 m^3 closed-loop volume with re-circulation of 7.3 m^3/h nominal mixture flow: 5 m^3/h for the barrel and 2.3 m^3/h for the endcaps. A combination of factors during production of the components and operation of the system can lead to an increase of the leak rate. Environmental conditions in the experimental cavern, such as humidity and temperature, as well as switching from Freon to N_2 and back to Freon, can accelerate the degradation process of the different components of the RPC gas system.

The gas leaks in the RPC system are mainly caused by the T-shaped or L-shaped polycarbonate gas connectors that are broken due to stress applied through the gas pipes, and the low-density polyethylene pipes that are brittle, deteriorated, or cut.

A special gas-leak repair procedure was developed to correctly identify the leaks using an endoscope, which allowed us to determine the exact location and the components that are sources of

2024 JINST 19 P05064

leaks. The RPC barrel chambers are coupled with the DT chambers and inserted into the iron yoke. Therefore, in order to reach the leak location, a partial extraction of the entire muon station (RPC and DT) by a distance of 80 cm from the back or front side is required. A surgical cut of the C-aluminum profile is then performed to gain access to the broken gas pipe or T/L connector. In figure 70, an example of the newly developed gas-leak repair procedure is shown. In one scenario, the broken pipe connecting the two chambers is removed, changing the splitting of the gas flow from internal to external (figure 70, upper center) and preserving the parallel gas distribution of the chambers. In another scenario, the repair is accomplished by gluing the T/L connector (figure 70, lower center).

Figure 70. Photographs illustrating the gas-leak repair procedures. Left: scientists working on the repair procedure. Middle left: access to broken component. Middle right: repairing of components. Right: closing and validation.

By November 2020, when all possible repairs were finalized, a total number of 88 gas leaks due to cracked or broken pipes were identified in the barrel chambers. Of those, 50 chambers were successfully repaired and returned to normal operation. This included 17 chambers that did not work during Run 2. The remaining 38 leaking chambers are either not accessible or the partial extraction technique is not applicable, either because the source of the leak is inaccessible or the source of the leak is not identified. In the endcaps, a total of 11 chambers were replaced.

The gas-leak repairs had a large impact on the chamber performance. Figure 71 shows a comparison of the efficiency of repaired chambers using cosmic ray data between the end of Run 2 before the reparation campaign and after all the repairs. This led to an overall gain in efficiency of 1.5%.

To minimize chamber pressure variations, which are considered to be a cause of leaks, automatic regulation valves on pre-distribution racks in the underground service cavern were installed in November 2021. This operation has substantially reduced the development of new leaks in the system.

In the beginning of Run 3 in June 2022, the total number of disconnected barrel chambers was 108, corresponding to about 10% of the entire RPC system. This number includes 14 dysfunctional chambers with gas leaks, 63 operational chambers with gas leaks, and 31 gas-tight chambers that share gas channels with leaking chambers.

Another major activity was testing a Freon-recuperation prototype connected to the RPC gas system exhaust. The aim is to recuperate the $C_2H_2F_4$ from the RPC gas mixture and re-use it [133]. The $C_2H_2F_4$ and i-C_4H_{10} gases form an azeotrope, i.e., a mixture of liquids whose proportions can not be altered or changed by simple distillation, because the intramolecular force of same species is

2024 JINST 19 P05064

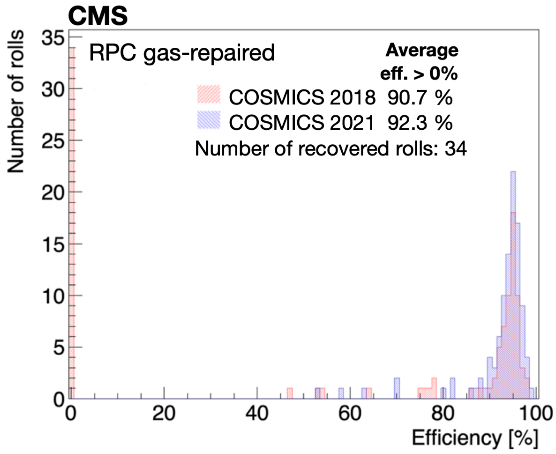


Figure 71. Efficiency comparison between 2018 (red) and 2021 (blue) cosmic ray data for chambers with repaired gas leaks.

much higher than the reciprocal attraction. The separation of the two gases is done by slowly heating the liquefied azeotrope, allowing the enrichment of the liquid R134a and the i-C_4H_{10} vapor. The CERN EP-DT gas team is finalizing the R&D on the first custom-built $C_2H_2F_4$-recuperation system with an expected efficiency of 80% [134]. To maximize the amount of gas in the exhaust line in order for efficient recuperation system operation, the procedure is to turn off and disconnect all leaking chambers. The recuperation system is expected to be ready and operational by summer 2023.

RPC commissioning during LS2. Figure 72 (left) shows the efficiencies per roll, calculated using the segment extrapolation method described in section 6.3.3. Comparing the average efficiencies for 2018 and 2021, an increase of 0.6% is observed. This increase can mainly be attributed to the recovery of chambers from single-gap to double-gap operation mode. The fraction of chambers with more than 70% efficiency rises by 6.1%, considerably raising the redundancy in the measurement of muons. These improvements are a result of the extensive maintenance and repairs during LS2.

In figure 72 (right), the cluster size distribution for barrel wheel W−2 is displayed. The average cluster size is kept around two strips, far below the possible maximum of three, in accordance with the trigger requirements. The comparison between the 2018 and 2021 cosmic ray muon data shows a stable cluster size.

The efficiency measurement using cosmic ray muon data in the endcaps can have large systematic uncertainties, related to the vertical geometry of the CMS muon system in the endcap region. Even so, the current and occupancy values at the WP during the cosmic ray muon data taking show good agreement with the expectation.

After extensive maintenance and repairs, as well as commissioning and longevity studies during LS2, the RPC detector successfully entered the data-taking period of Run 3.

6.4 Gas electron multiplier chambers

6.4.1 Motivation and general description

During the future operation of the High-Luminosity LHC (HL-LHC), the maximum hit rate in the forward region of the muon system is expected to reach 5 kHz/cm^2 in the first muon layer, with an

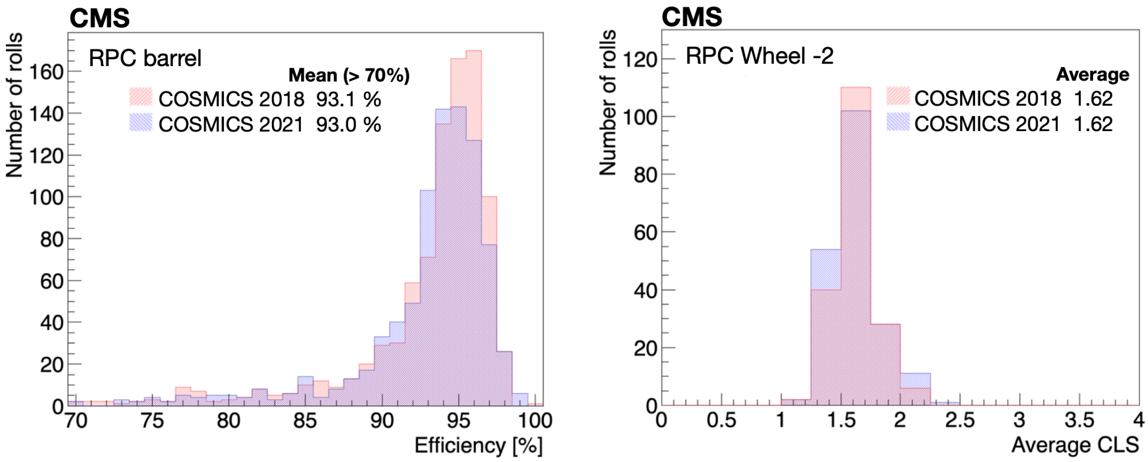


Figure 72. Left: distribution of the efficiency per roll in the RPC barrel chambers. Only chambers with rolls of efficiency greater than 70% are considered. Right: average cluster size for the RPC barrel chambers in wheel W−2. In both figures, the cosmic ray muon data from 2018 (2021) are indicated in red (blue).

integrated charge per unit area of 100 mC/cm^2 [100]. This increase compared to the present values represents a challenge to the forward muon system, which must be resistant to radiation, have a high rate capability, and maintain an adequate pattern recognition for efficient muon reconstruction, while minimizing the number of misidentified tracks and keeping the level-1 (L1) trigger rate at acceptable levels.

To enhance the track reconstruction and trigger capabilities of the endcap muon spectrometer, large-area triple-layer gas electron multiplier (GEM) detectors [135] were installed in the region covering $1.55 < |\eta| < 2.18$ of the CMS detector for the start of Run 3. This station, denoted GE1/1, is the first of three GEM rings that will be installed for the HL-LHC. Figure 39 shows a quadrant of the CMS detector with the location of the GE1/1 highlighted and outlined in red in the r-z plane. A detailed drawing is shown in figure 73.

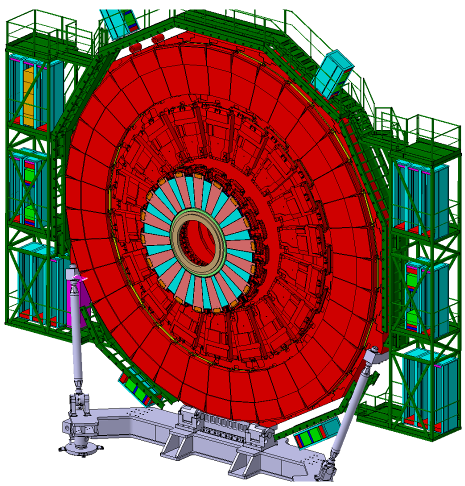


Figure 73. Sketch of GE1/1 system of one endcap indicating its location relative to the full endcap muon system, the endcap calorimeter, and the shielding elements. Reprinted from [136], Copyright (2019), with permission from Elsevier.

2024 JINST 19 P05064

Muons experience the largest bending in the muon spectrometer at the position of the first muon station as the magnetic field lines bend around the endcap flux return. Due to the increasing background rates at large η and the reduction in the magnetic field in the first muon station, the trigger rate in this region is large and difficult to mitigate. The insertion of the GE1/1 chambers increases the lever arm traversed by muons by a factor of 2.4–3.5, relative to ME1/1 alone, leading to a significant improvement in the muon trigger momentum resolution and a large reduction in the L1 trigger rate. Figure 74 shows the expected trigger rates in the forward muon spectrometer with and without the GE1/1 upgrade. This indicates that the upgrade will lower the trigger rates by a factor of 3–10 depending on the p_T threshold. The muon trigger is described in detail in section 10.2.

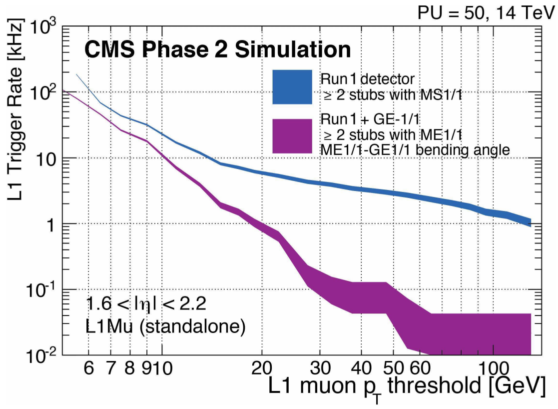


Figure 74. L1 muon trigger rate with and without the GE1/1 upgrade, assuming an instantaneous luminosity of 2×10^{34} cm^{-2} s^{-1}, where MS1/1 indicates the first muon station [100].

Each endcap GE1/1 detector consists of 36 double-layered triple GEM chambers located just in front of the first CSC station, labeled ME1/1, each covering a 10° sector in azimuth. The chambers provide full coverage in ϕ and were constructed in two sizes, the odd-numbered GE1/1 are slightly longer in order to maximize the pseudorapidity coverage while fitting in the available space constrained by the support structure, as shown in figure 75. The GEM detector technology can withstand rates up to 1 MHz/cm^2 and has excellent spatial and timing resolution of approximately 250–500 μm and <10 ns per layer, respectively. The combined GE1/1 station spatial resolution is on the order of 100 μm. The GE1/1 detector is placed, in global CMS coordinates, in z between 566 and 574 cm, and at a radius between 145 and 230 cm.

6.4.2 Technical design

The CMS triple GEM detector is a micro-pattern gas detector that comprises four gas gaps separated by three GEM foils, as shown in figure 76. It has an active area of 990 × (220–455) mm^2 with a 3/1/2/1 mm wide drift/transfer-1/transfer-2/induction field gap configuration [136]. The bottom of the GEM assembly is a printed circuit board that holds the drift electrode and voltage divider, while the top of the assembly is the readout board consisting of radially oriented readout strips along the long side of the chamber. The strip pitch ranges from 0.6 to 1.2 mm, and the readout board is segmented in up to 10 × 3 η-ϕ partitions, each with 128 strips. The triple GEM arrangement allows for a charge amplification factor of up to a factor of several 10^5, while limiting the probability of electrical breakdown or discharge. The amplified charge induces a measurable signal on the readout

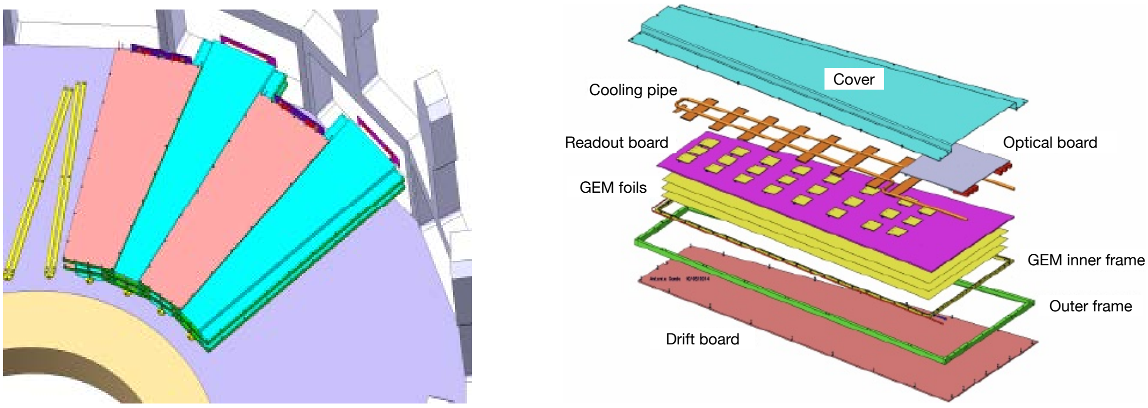


Figure 75. Left: layout of the GE1/1 chambers along the endcap ring, indicating how the short and long chambers fit in the existing volume. Reproduced from [137]. CC BY 4.0. Right: blowup of the trapezoidal detector, GEM foils, and readout planes, indicating the geometry and main elements of the GEM detectors, from ref. [138] John Wiley & Sons. © 2017 Society of Plastics Engineers.

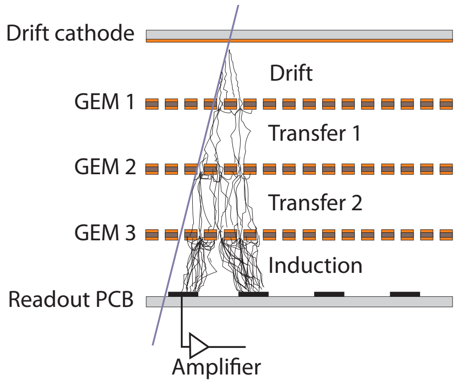


Figure 76. Sketch of a triple GEM detector showing the three foils, cathode, readout PCB and amplification. Reproduced from [100]. CC BY 4.0.

electrode, which is segmented to provide positional information. The gas mixture was chosen to be $Ar+CO_2$ in a 70:30 proportion.

In the GE1/1 station shown in figure 73, pairs of triple GEM detectors are matched to form a “super-chamber”, providing two measurement planes and maximizing the detection efficiency for the station. Thus, if each chamber operates at 98.0% efficiency, the logical OR of the two signals from the super-chamber provides a 99.9% efficiency. The structure of a GE1/1 chamber is shown in figures 75 (right) and 76, while the dimensions and specifications of the two chamber types, “short” and “long”, are given in table 10.

The main elements of an individual chamber are: the drift board holding the drift electrode, the GEM foils that amplify the ionization signal, the readout board where induced charge is read out on segmented strips, the internal and external frames, and a gas distribution system. A more complete technical description of the layout and assembly of the GEM chambers can be found in ref. [136].

The drift board is a trapezoidally shaped printed circuit board (PCB) inside the active gas volume and coated with a copper layer that serves as the chamber’s drift electrode. Outside the

2024 JINST 19 P05064

Table 10. Dimensions and specifications of the GE1/1 short and long chambers, from ref. [136].

Specification	Short	Long
Chamber length [cm]	113.5	128.5
Chamber width [cm]	28–48.4	26.6–51.2
Chamber thickness [cm]	1.42	1.42
Active readout area [cm^2]	3787	4550
Active chamber volume [liters]	2.6	3
Geometric acceptance in η	1.61–2.18	1.55–2.18

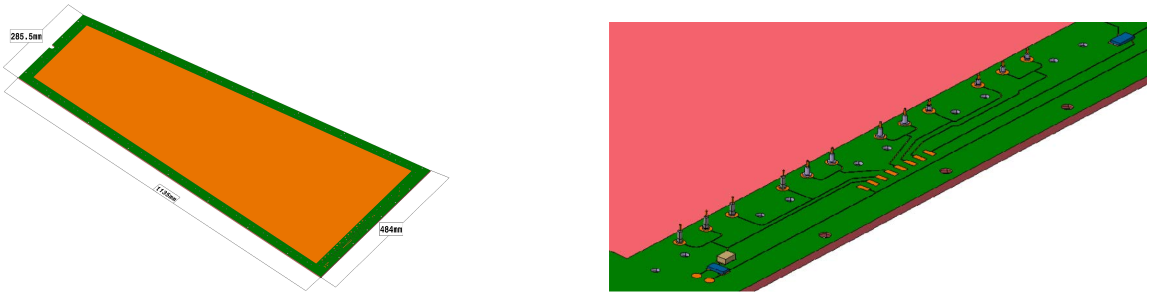


Figure 77. GE1/1 drift board (left) and a magnified view of the drift board (right) showing the HV pins and the resistor and capacitor network connecting to the chamber HV supply. Reprinted from. [136], Copyright (2019), with permission from Elsevier.

active gas volume, a 100 kΩ resistor and 330 pF capacitor are installed on pads on the PCB to limit the current from the high-voltage (HV) power supply and decouple the signal from the HV. Twelve pins that carry the HV to the GEM foils are mounted as shown in figure 77.

Four internal frames composed of halogen-free glass epoxy with thicknesses of 3/1/2/1 mm are coated with polyurethane varnish and are used to define the spacing between the drift board, the three GEM foils, and the readout board. The GEM foils are attached to the frames by screws that pass through the entire chamber assembly to hold it in place.

The GE1/1 detector uses three identical GEM foils shown in figure 78, which are thin polyimide foil clad with a thin copper layer containing micro-pattern holes etched in a periodic grid. A voltage up to 400 V is applied across the copper-clad surface producing a strong electric field of between 60 and 100 kV/cm. The triplet structure allows for an amplification of around 10^5 when moderately high voltages are applied. The side of the foil facing the readout boards is a continuous conductor, while the strips facing the drift board are segmented into sectors of approximately equal area of 100 cm^2. The segmentation ensures that, in the extreme case of a large discharge creating a short circuit between the GEM electrodes of a single sector, the affected dead area is limited to the 100 cm^2 of a single sector instead of deactivating the entire detector. Each sector is connected separately to the HV supply via a 10 MΩ resistor to limit currents from the HV supply and to quench any possible discharge.

The readout board is a printed circuit board with 3072 radially oriented readout strips on the side facing the interior of the chamber. The strips are connected by metalized vias to the outer side of the board, which are routed to readout pads on the exterior-facing side of the board. The board is segmented into 8×3 η-ϕ partitions with three sets of 128 strips in ϕ. Each set of 128 strips is read

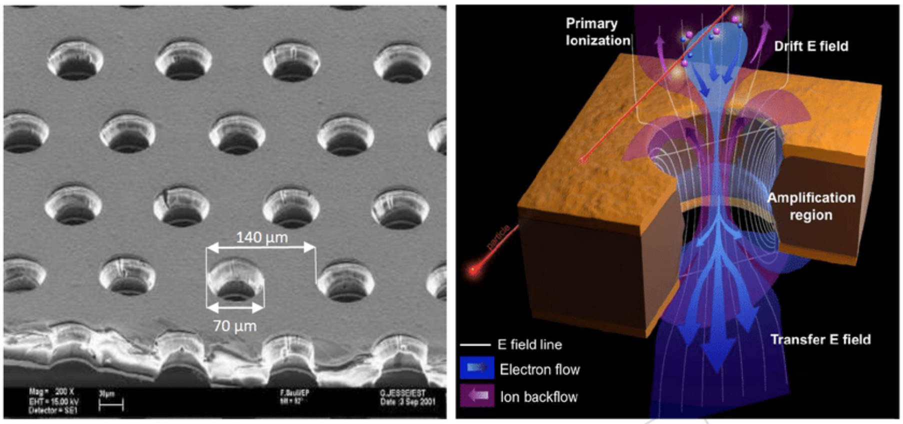


Figure 78. Electron scanning microscope image of a GE1/1 foil (left) and a diagram of the multiplication principle (right). Reproduced from [100]. CC BY 4.0.

out by the first stage of the frontend electronics, as described below. The strip pitch varies between 0.6 and 1.2 mm on the shorter and longer ends of the trapezoidally shaped board.

6.4.3 Gas system

During operation, the chambers are run using an Ar:CO_2 gas mixture with a ratio of 70/30. A gas mixer is installed in the surface gas barrack (SGX) where the argon and carbon dioxide gases are mixed to the required percentages. The mixer has input and output lines that go to the underground gas room (UGX). Four racks in the UGX are used to pump and control the gas from the surface using input and output lines to a crate on the periphery of the first muon ring on each endcaps. The rack on the periphery uses twelve gas lines, each of which provides the gas mixture to six consecutive single chambers. Each chamber in turn has a single inlet and outlet through which the gas enters and exits the chamber. In figure 79 a diagram of the periphery crate (left) and the location of the gas lines (center) are shown. The entire system is monitored at different stages (mixer, pre-distribution, and distribution), and the pressure and flow-rate values are supervised by the gas control system (GCS).

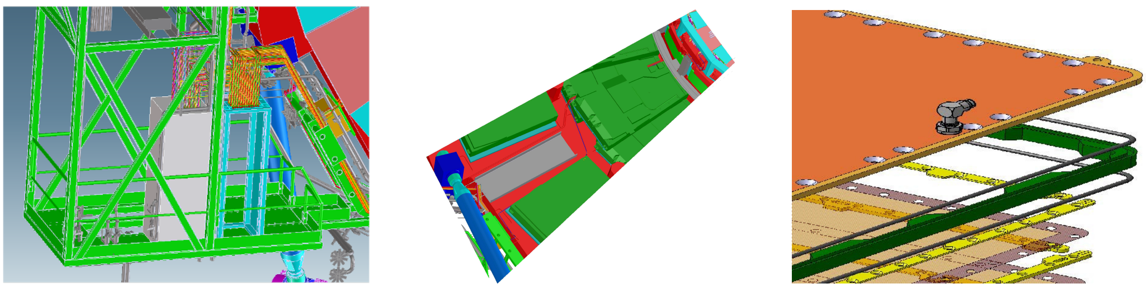


Figure 79. Left: periphery GE1/1 gas rack. Center: location of the gas lines feeding the GE1/1 chambers. Right: exploded view of a GE1/1 chamber showing the gas plug attached to the readout board. Reprinted from [136], Copyright (2019), with permission from Elsevier.

Each chamber is connected to the gas system via a single inlet and single outlet attached to the readout board, as shown in figure 79 (right). The gas flows from an inlet placed at one corner of the readout board across to the outlet placed at the diagonally opposite corner of the trapezoid and through holes in the internal frames.

6.4.4 Electronics

The GEM readout electronics are divided into frontend components that are mounted on the detector itself, and backend electronics that are outside the experimental cavern. A schematic overview of the different components is shown in figure 80. The readout is organized according to 128 readout strips by the VFAT3 chip [139], which is the ASIC on the readout board responsible for the digitization of the induced signals. The signals from individual VFAT3 chips are then sent via traces on a large chamber-sized PCB called the GEM electronics board (GEB) [140]. The VFAT3 and opto-hybrid [141] boards are attached to the GEB, which also routes high and low voltages to the chambers from external off-chamber power supplies. The opto-hybrid board transmits the data via optical links to the off-chamber backend readout and also sends data to the CSC trigger board for processing at L1. The optical links are routed from the opto-hybrid boards on the chambers to patch panels on the periphery crates and then to off-detector electronics in a μTCA crate [142]. The backend boards are CTP7 μTCA boards, which are described in ref. [143]. They contain bidirectional links to communicate with the frontend chamber electronics and are connected via AMC13 [25] cards to the CMS L1 trigger and DAQ systems [144], as discussed in section 9.

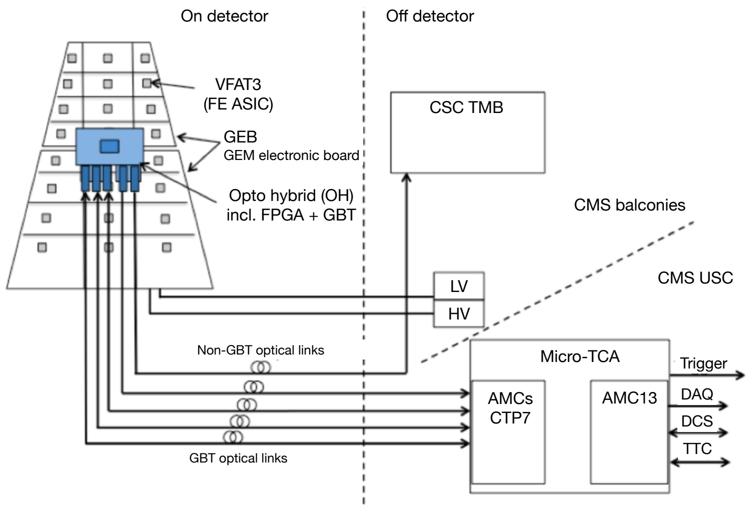


Figure 80. GE1/1 electronics overview showing the frontend electronics, VFAT, GEB, and opto-hybrid boards, as well as the optical links and backend readout with connections to the L1 trigger and DAQ.

The VFAT3 chip was developed from an existing frontend readout chip, the VFAT2 [145] (very forward ATLAS and TOTEM) ASIC, developed for the ATLAS and TOTEM readout. The VFAT3 design was modified for application in the CMS GEM to provide readout in a high-rate environment, collect the total charge produced by particles crossing the GEM volume, and be radiation hard. A picture and a high-level schematic of a VFAT3 chip are shown in figure 81.

The VFAT3 must collect, amplify, shape, digitize, and buffer the signal produced on each of the 128 channels connected to the readout strips on the GEM readout board. The preamplifier and shaper are charge sensitive and are followed by a constant-fraction discriminator that digitizes the incoming

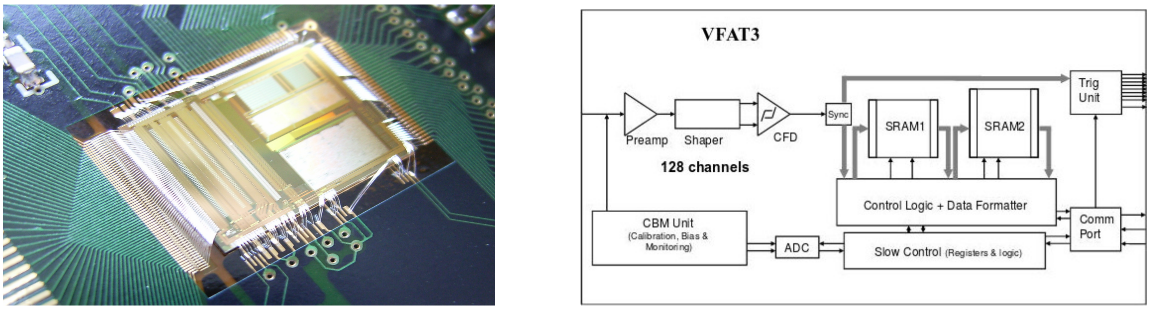


Figure 81. Picture of the VFAT3 ASIC (left) and a high-level schematic (right). Reproduced from [146].

charge pulses. Following the discriminator, the binary comparator results are synchronized with the LHC clock in the synchronization unit. A programmable threshold allows a channel-by-channel calibration to optimize the signal-to-noise ratio, and is set to obtain approximately 98% efficiency per single hit so that the logical OR of two layers has a 99.9% efficiency.

The data are then split into two paths. A trigger signal, with fixed latency, is sent to the trigger unit, while an asynchronous tracking signal is sent to a buffer for full granularity readout via the communications port. The signal is accompanied by a time stamp, the bunch crossing, for further processing.

The signal can last up to approximately 60 ns depending on the gas mixture, and the VFAT3 chip allows the shaping time to be adjusted to fully integrate the charge and maximize the signal-to-noise. The timing resolution is optimized by the use of a constant-fraction discriminator and a gain with a programmable dynamic range. The buffer for the tracking data is implemented by a SRAM memory that operates as a circular buffer 128 channels wide and 1024 bunch crossings deep. It continuously samples all channels in every cycle of the LHC clock. After 1024 bunch crossings, the buffer overwrites the first entry.

Since the L1 trigger operates with fixed latency, there is a fixed time between the level-1-accept (L1A) trigger signal and the data corresponding to the bunch crossing of interest. Upon receiving an L1A, the appropriate data are transferred from the SRAM1 buffer to the SRAM2 buffer, which is 512 bunch crossings deep and stores the data until they are transmitted off the chip. The chip supports a latency of up to 12.5 μs with an L1A rate of up to 1 MHz. The data can either be sent in a lossless mode or zero suppressed.

The fixed latency, or trigger path, is used to provide fast hit information synchronous with the LHC clock, which can be put in coincidence with other detectors to decide if the event will generate an L1 trigger. The logical OR of two adjacent channels is sent to the trigger unit.

The slow-control signals are used to send calibration, bias, monitoring, and control information via the communication port and to control and monitor the settings and status of the VFAT3 chip.

The output of each VFAT3 chip is routed along the GEM electronics board (GEB), which is mounted directly on top of the readout board. The GEB is divided into two sections that cover the entire chamber. The GEB routes the signals between the 24 VFAT3s on each chamber and the low voltage from the external power supplies to the chamber and frontend electronics. CAEN power supplies provide voltage to the FEAST boards mounted on the GEB. The FEAST boards perform the DC-DC conversion to provide the proper low voltages for the readout electronics, as well as the HV for the operation of the chamber.

The opto-hybrid board serves as the chamber hub for data communication, transfer, and control. It receives data from all the VFAT3 ASICs, formats it for the entire chamber, and sends the data using the GBT protocol and optical links to the backend electronics for final readout and use in the L1 muon trigger. Additionally, each opto-hybrid board contains three GBTx chipsets [147], one Virtex-6 FPGA [148], three versatile link transceivers (VTRx), and two versatile link transmitters (VTTx) [68]. Trigger data are sent over the VTTx via optical fibers to both the backend readout and the CSC trigger motherboard, while the three VTRx transmit tracking data to the backend readout (CTP7 card). Each GBT can handle up to ten frontend chips at a transfer rate of 320 MHz, and the tracking data are transferred at 4.8 Gb/s through the VTX. The trigger data are transferred at 3.2 Gb/s using 10b/8b encoding. In operation, the trigger data are formatted within the Virtex-6 FPGA, while the tracking data flow directly from the frontend by the GBT VTRx links to the backend. Calibration, and trigger, timing, and control (TTC) signals are sent to the opto-hybrid board and distributed to the frontend via the GEB. Irradiation tests have shown that the single- and double-error rates from single-event upsets (SEUs) are well within acceptable margins for the total ionizing dose (TID) expected over the lifetime of the HL-LHC project. A photo of the opto-hybrid board and the measured error rates are shown in figures 82 and 83 .

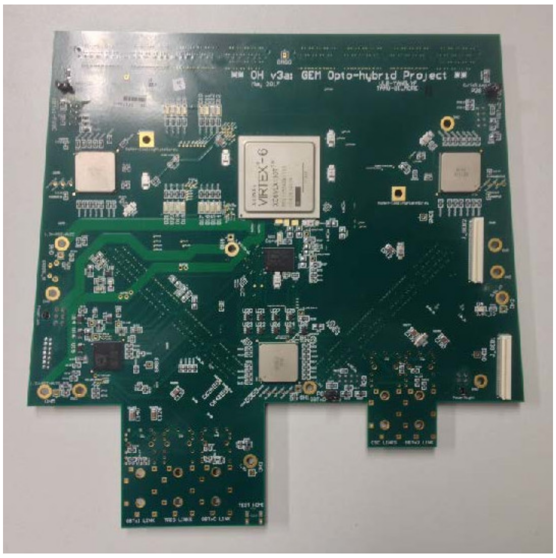


Figure 82. Photo of the GE1/1 opto-hybrid board with the Xilinx Virtex-6 FPGA at the center.

The backend electronics are μTCA cards that were originally developed for a calorimeter trigger Phase 1 upgrade project [143]. The CTP7 has 67 optical receivers, 48 optical transmitters, a Xilinx Virtex-7 FPGA, and a Zync processor. The firmware was adapted for the GEM detector with 36 GBT cores servicing 12 triple GEM detectors with one link to a CMS μTCA AMC-13 card operating the standard DAQ and trigger links. Using this hardware, all of the GE1/1 can be read out with one μTCA crate hosting six CTP7 cards and one AMC-13 card.

6.4.5 CSC/GEM trigger for Run 3

Figure 84 shows a view in the global CMS transverse (ϕ-z) plane indicating the location of the GE1/1 and the first CSC station, ME1/1. The diagram indicates the relative position of GE1/1 and ME1/1

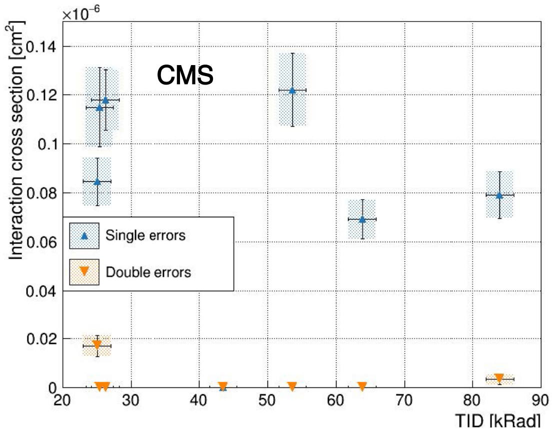


Figure 83. Single- and double-error cross sections measured in irradiation tests as a function of the TID.

in the global z coordinate and shows two muon trajectories as they traverse the stations' volume. The deflection of muons with a p_T of 5 and 20 GeV is approximately 3 and 12 mm, respectively, in the local x (global ϕ) direction. By reconstructing the ϕ direction in both the entrance to the GE1/1 station and the exit from the ME1/1 station, an estimate of the track p_T can be made and used for the trigger decision in the L1 muon trigger system. The GE1/1 chambers add two new independent position measurements, which improves the redundancy, lowers the muon misidentification rate, increases the robustness of segments found in the muon spectrometer, and raises the efficiency of the stub-finding algorithm from 90 to 96%. Finally, the joint segments are utilized in the endcap muon track finder (EMTF) to optimize the final track reconstruction and resolution of the L1 trigger, as described more fully in section 10.2.

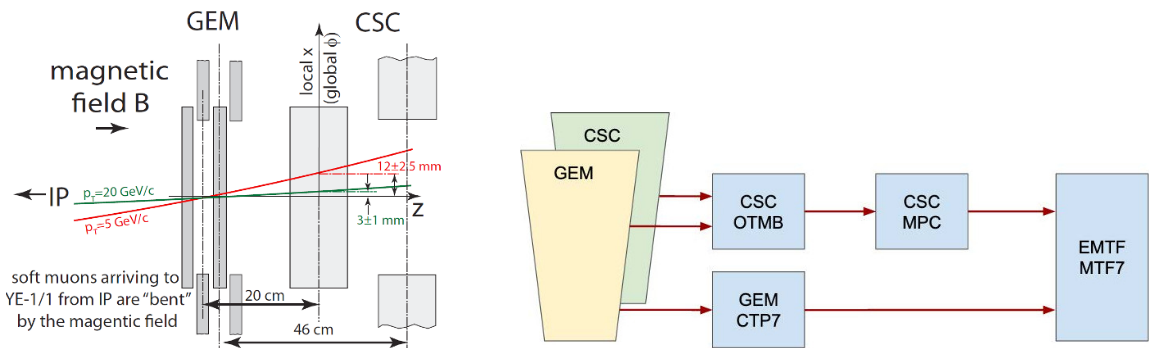


Figure 84. Side view of the GEM-CSC trigger coincidence (left) and schematic overview of the data flow in the GEM and CSC trigger processor (right). The addition of the GE1/1 station significantly increases (by a factor of 3 to 5) the lever arm of the distance traveled in z by a muon originating from the interaction point. The bending angle between the entrance of the muon to the GE1/1 station and the exit from the ME1/1 CSC station can be used to estimate the momentum of the muon. Reproduced with permission from [149].

The data flow from the GEM chambers to the EMTF via two paths is shown in figure 84. First, trigger data are sent through the opto-hybrid board to the CSC optical trigger motherboard (OTMB) [150] that uses hits in the GE1/1 detector and the CSC ME1/1 to form stubs from the combined stations. These stubs are in turn sent to the CSC muon port card that sorts the stubs from multiple chambers by sector and sends them to the appropriate EMTF card where the muon momentum is estimated from hits in the CSC, RPC, and GEM using a neural network algorithm [151], as described in section 10.2. In figure 85, simulated distributions of the muon rate in the endcap as a function of p_T and η are shown [100].

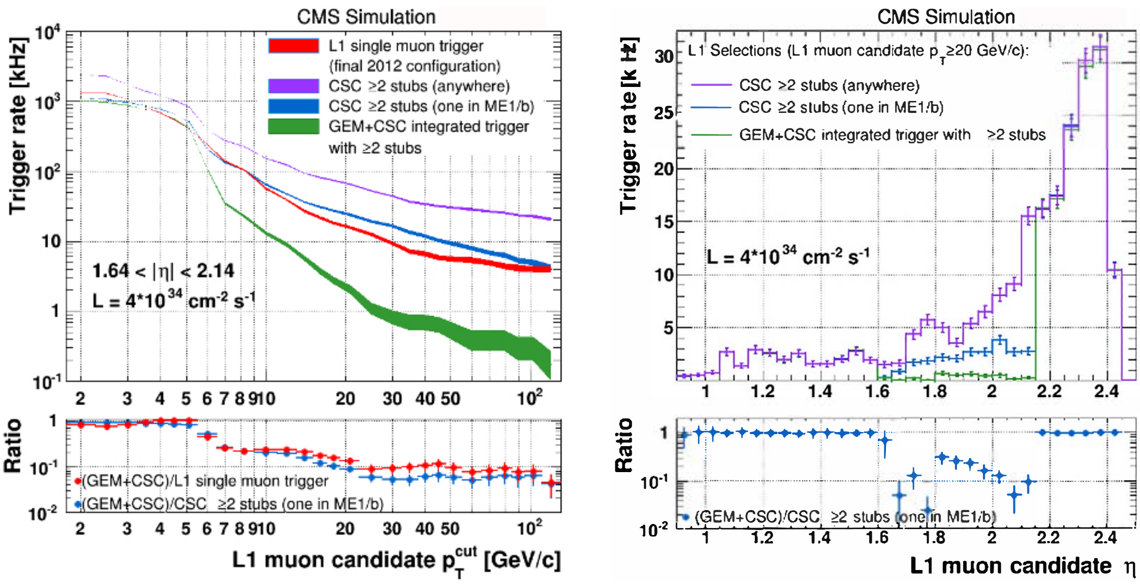


Figure 85. Muon rate from simulation as a function of p_T (left) and η (right) with and without the integration of the GEM chambers under various assumptions. Reproduced from [100]. CC BY 4.0.

6.4.6 Detector control system

The CMS detector control system (DCS) [73] is based on the Siemens Simatic WinCC Open Architecture (WinCC OA), previously known as Prozess-Visualisierungs- und Steuerungssystem (PVSS), and is described in ref. [152]. The DCS [153] was designed to control and monitor the high and low voltage and to monitor the gas conditions.

Even though the gas system is controlled by the central CERN gas group, the DCS maintains the function of monitoring the mixture composition (fraction of Ar and CO_2), as well as the pressure and flow rate to the chambers. In particular, the DCS reflects the structure of the gas system, which in CMS is divided into three main elements: mixer, rack, and flowcells.

The DCS system also controls both the high voltage used to operate the GEM foils and the low voltage used to power the frontend electronics on each chamber. The system allows turning on and off either the high or low voltage for every component, as well as monitoring the voltages and currents in the system.

6.4.7 Chamber assembly and installation

To ensure high-quality performance of all the components, a detailed quality control process was put into place to assure the quality of every element. The ten-step process, described in greater detail in ref. [154], can be summarized as follows:

- quality control of the components for production (QC1–2),

2024 JINST 19 P05064

- assembly and commissioning of the GE1/1 chambers at the production sites (QC3–5),
- assembly and commissioning at CERN before installation in CMS (QC6–10).

First, all components were cleaned using ultrasonic baths, baked, sand-blasted, and visually inspected for faults. The components were then optically and electrically inspected including verification measurements, specifically measuring the *I-V* curves and checking for shorts between readout strips on the drift board. The foils were optically inspected, and checked that the leakage currents were less than 30 nA when 500 V was applied between the two sides of the GEM foils. Each chamber was pressurized by dry nitrogen at 30 mbar and checked for gas leakage. The chambers were then flushed for several hours with Ar+CO_2 and monitored for output at moderate HV. The uniformity of the gas gain was verified by placing an X-ray source 1 m away from a chamber enclosed in a copper protective box, which simultaneously illuminated the full chamber.

Once shipped to CERN, the gas leakage and HV tests were repeated for every chamber, and the frontend electronics and chambers were mounted into super-chambers with cooling plates. Electrical conductivity tests, gas leak tests, and electronic noise measurements were repeated with the cooling on. Finally, the super-chambers were mounted on a cosmic ray test stand where efficiency, noise, and tracking studies were done to confirm the correct and uniform operation of each chamber. While in storage and before installation into CMS, the gas-leak and HV stability were tested again and monitored over one month before each GEM was installed and declared ready for commissioning in the CMS detector.

6.4.8 Preliminary commissioning results

Before installation in the CMS cavern, a GE1/1 chamber was tested at the CERN H4 beam, which was extracted from the SPS [155]. A secondary beam of pions was produced when the proton beam struck a beryllium target. Finally, that secondary beam was filtered by collimators to produce a beam of 150 GeV muons from the decay of the pions in the secondary beam.

Several key performance parameters were measured in the test beam. Two are the single-hit efficiency and the time resolution of the GEM chambers. A set of scintillators read out by photomultiplier tubes (PMTs) were placed upstream of the chamber under test. Events were selected requiring a triple coincidence in each of the three layers and the efficiency was estimated by counting the number of hits in the GE1/1 tracking chamber in that region. Time resolution information was obtained by measuring the standard deviation of the distribution of the measured time between the trigger and GE1/1 detector signal. As can be seen in figure 86, an efficiency greater than 98% and a time resolution less than 10 ns were obtained. Other results, such as discharge rate and rate capacity as a function of voltages and gas mixture are given in ref. [155].

During LS2, the GE1/1 chambers were installed in CMS, connected to all services (LV/HV, gas, and optical fiber connection for readout), and the chambers were commissioned in situ. Latency scans were done for all GE1/1 chambers on both endcaps to account for the variable cable and optical fiber lengths, and HV values were optimized for signal-to-noise ratio using cosmic ray muons. It should be noted that, unlike in LHC collisions, cosmic ray muons do not arrive at a fixed and known time. Thus, the latency and other operational parameters cannot be fully optimized for LHC operations until all chambers are exposed to both muons originating from the interaction point and the LHC background conditions.

2024 JINST 19 P05064

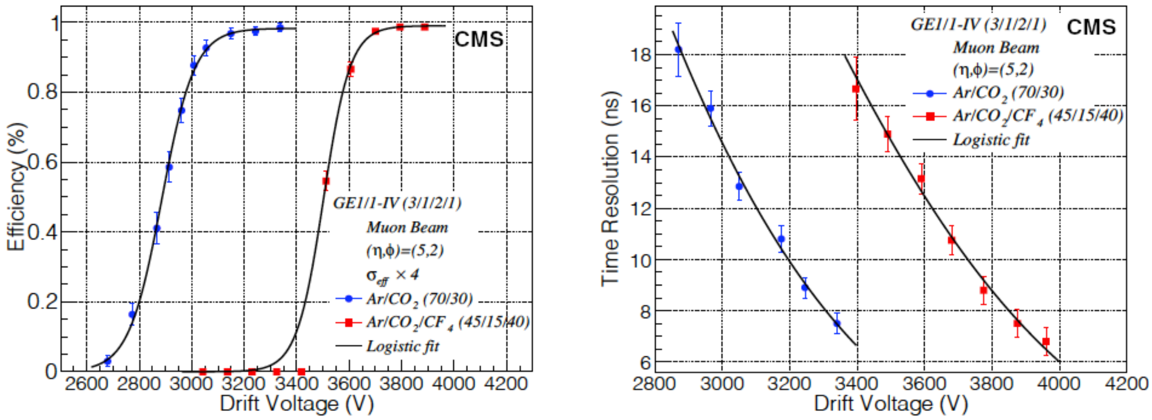


Figure 86. GEM test beam results showing the efficiency (left) and time resolution (right) as a function of the drift voltage placed on the GEM foils. The chosen gas mixture Ar/CO_2 (70/30) has very similar properties to $Ar/CO_2/CF_4$ (45/15/40) and was selected as CF_4, a greenhouse gas, is being phased out of industry. Reprinted from [155], Copyright (2020), with permission from Elsevier.

Figure 87 shows the occupancy in both GE1/1 endcap rings. Not all chambers were included in the data taking with cosmic rays due to minor temporary commissioning issues. This explains the missing sectors in the $-z$ endcap. Furthermore, since electronic noise was being addressed, some chambers show higher-than-average occupancy. The GE1/1 system was fully commissioned with cosmic rays and has been successfully operated since the beginning of LHC Run 3 in summer 2022.

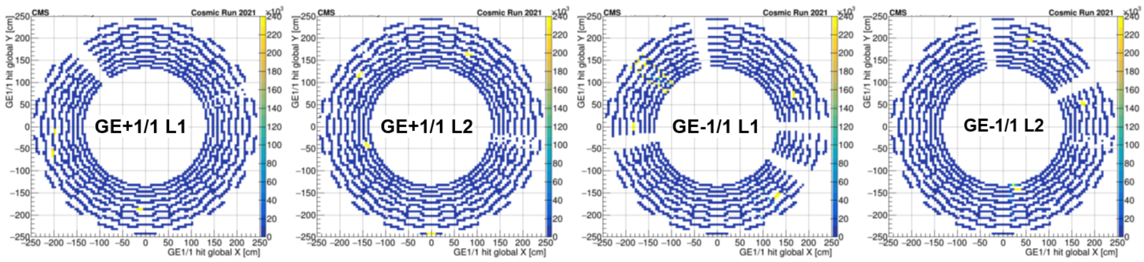


Figure 87. GE1/1 occupancy in cosmic ray muon data recorded in 2021.

7 Precision proton spectrometer

The precision proton spectrometer (PPS) [156] is a set of near-beam detectors, located in the LHC tunnel at a distance of about 200 m from the CMS IP, on both sides. Originally conceived and developed as a joint project of the CMS and TOTEM collaborations (CT-PPS), it then evolved into a standard CMS subdetector following a Memorandum of Understanding between CERN, CMS, and TOTEM in 2018. For the sake of clarity, only the PPS acronym is used throughout this section. During Run 2, PPS took part in the CMS data taking, with various configurations between the years 2016 and 2018, and recorded data corresponding to an overall integrated luminosity of about 110 fb^{-1}. Following approval of the extension of the PPS experimental program to the LHC Run 3, a broad upgrade plan has been launched, involving replacement of all PPS detectors.

2024 JINST 19 P05064

7.1 Forward protons and the roman pot system

The measurement of very forward ("leading") protons [156] implies the detection of such protons at large distance from the IP and in proximity of the LHC beam. The main variable characterizing the proton kinematics is its fractional momentum loss, defined as $\xi = (|\vec{p}_\mathrm{i}| - |\vec{p}_\mathrm{f}|)/|\vec{p}_\mathrm{i}|$, where $\vec{p}_\mathrm{i}$ and $\vec{p}_\mathrm{f}$ are the initial and final proton momentum vectors, respectively. The coverage in ξ at the PPS location is limited by the LHC magnet and collimator lattice in Run 2 and Run 3 to a range below 0.2. In order to extend it as much as possible in the low-ξ region, scattered protons must be detected as close as few (2–3) mm from the beam. This can be accomplished by means of special movable beam pockets, the so-called roman pots (RPs), which host the particle detectors such that they can be moved close to the beam during stable beam operations. The PPS experimental setup includes both tracking detectors, to measure ξ from reconstructed proton tracks, and timing detectors, to suppress the contribution of protons originating from different primary interactions in the same LHC bunch crossing (pileup). Tracking and timing detectors are hosted in separate RPs.

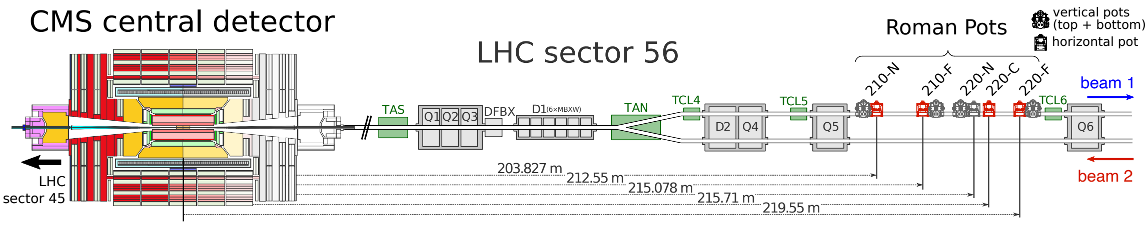


Figure 88. Schematic layout of the beam line between the interaction point and the RP locations in LHC sector 56, corresponding to the negative z direction in the CMS coordinate system. The accelerator magnets are indicated in gray; the collimator system elements in green. The RP units marked in red are those used by PPS during Run 2; the dark gray ones are part of the TOTEM experiment. In Run 3 the "220 near" horizontal unit is used in place of the "210 near". Reproduced from [12]. The Author(s). CC BY 4.0.

The TOTEM experiment [157] has been using an extensive system of box-shaped roman pots to conduct its physics program since the start of LHC operations in dedicated, low-luminosity runs. The system has been conceived as a two-arm spectrometer, composed of two pairs of stations on each side of the CMS IP, at about 147 and 220 m along the beam line, each station comprising two RPs approaching the beam vertically, and one approaching it horizontally. Within each pair, the station installed closer to the IP is called "near", and the other "far". For Run 2, the layout of the spectrometer has been modified as sketched in figure 88 for one of the two arms. The stations at 147 m have been relocated in proximity of those at 220 m and identified as "210 m". Following these conventions, the four stations are usually named 210-N, 210-F, 220-N, 220-F, where N stands for near, and F for far. In order to allow operations at standard LHC luminosity with PPS, some of the horizontal RPs have been equipped with cylindrical ferrite shields to reduce their radio-frequency impedance [156] and have been reserved for the PPS tracking detectors (section 7.2). Moreover, an additional horizontal RP with modified geometry (cylindrical RP) has been installed in each arm between the near and far stations at 220 m to host timing detectors; it is referred to as 220-C (section 7.3). All vertical RPs remain reserved for the TOTEM physics program. However, their role is crucial for PPS operations at the beginning of each data-taking period, when the tracking detectors they host are used in special runs, in conjunction with those in the horizontal RPs, to determine the global alignment parameters [12].

2024 JINST 19 P05064

Figure 89 shows a view of the LHC tunnel, in the region where the RPs are installed in one of the two sectors. The detector housings are made of 2 mm thick stainless steel, with a useful depth to contain the detectors of about 110 mm; the box-shaped pots have a rectangular cross section of $124 \times 50\,\mathrm{mm}^2$, while the cylindrical pots have a diameter of 141 mm. In order for detectors to approach the LHC beam as closely as possible, and to minimize the amount of material traversed by scattered protons, the thickness of the walls is reduced to 200 (300) μm around the active sensor region in the box-shaped (cylindrical) RPs: this section is called “thin window”. Figure 90 shows the structure of a box-shaped and a cylindrical RP, as well as the insertion system installed on a section of the LHC beam pipe.

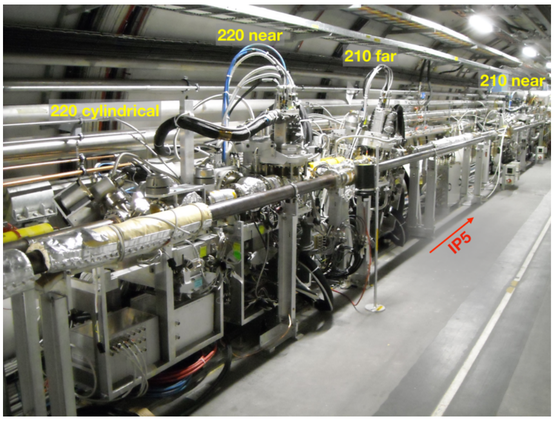


Figure 89. View of a section of the LHC tunnel in sector 45, with part of the PPS and TOTEM RP stations along the beam line.

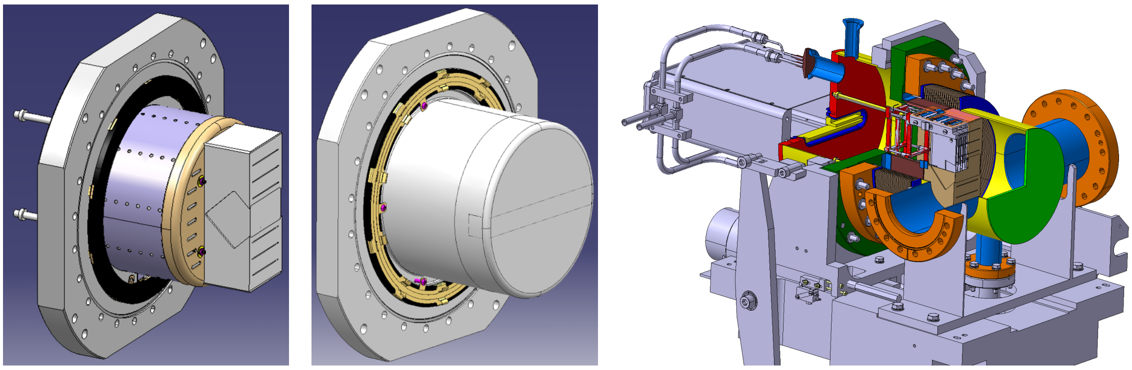


Figure 90. Sketches of the horizontal roman pots. Left: box-shaped pot, with the ferrite RF-shield in place. Center: cylindrical pot where the ferrite RF-shield is integrated as a ring in the flange. The thin window shapes, different for the two cases, are visible in the rightmost, vertically centered part of the pots. Right: overview of the insertion system of a box-shaped pot in a section of the LHC beam pipe. Reproduced with permission from [158].

Moderate vacuum conditions have to be kept inside the detector housings, in order to minimize the pressure gradient between the detector volume and the ultra-high vacuum of the LHC beam pipe. The exact value of the pressure depends on the detector technology, but does not exceed 100 mbar. Several RPs are serviced by the same vacuum system.

The solid-state detectors equipping PPS need to operate at low temperatures (−20 to 5°C, depending on the detector) to mitigate the effects of radiation damage. This is accomplished by means of a fluorocarbon evaporative system [159]: the cooling fluid is transported to the RPs, where suitable evaporation circuits, thermally coupled to the detectors, supply the needed cooling power [156, 157].

A more detailed description of the TOTEM and PPS roman pot system and its operations can be found in refs. [156, 157] and references therein.

7.2 Tracking detectors

The energy loss of protons can be measured, through detailed knowledge of the LHC optics parameters and dedicated calibrations [12], from the parameters of their reconstructed tracks. In order to obtain the needed accuracy in ξ, a resolution of few tens of µm is required on the coordinates of the track impact point. The hit rate distribution is highly nonuniform in the region covered by the tracking detectors, with peak particle flux values of about 3×10^9 p/(cm^{2}s) and a range spanning over three orders of magnitude. This reflects in the high radiation dose the detectors have to withstand, with reference fluence values in the most irradiated areas of 1–3×10^{15} $\mathrm{n_{eq}/cm^2}$ for an integrated luminosity of $100\,\mathrm{fb^{-1}}$.

7.2.1 Detector units

Proton tracks are reconstructed in PPS by two tracking stations per arm. During Run 2, the roman pots hosting the trackers and the technology chosen for the detectors have changed over time.

In 2016, in order to advance data collection by one year with respect to the original plan, two of the TOTEM tracking stations, equipped with silicon strip detectors [160, 161], were installed in the horizontal RPs at the 210-N and 210-F locations. The TOTEM strip detectors were originally designed to operate in special LHC runs at very low luminosity, up to about $10^{30}\,\mathrm{cm^{-2}\,s^{-1}}$. Within PPS, they have been operated successfully well beyond design specifications, providing excellent position measurements. However, they suffered from severe limitations related to the high hit rate. Firstly, since the hit position in the detector planes is determined by the coincidence of pairs of perpendicular strips, the frequent presence of multiple proton tracks in the same bunch crossing leads to ambiguities that cannot be resolved. Secondly, the radiation dose in the regions with the highest hit rate induced damages that caused an early loss of efficiency: for those regions, the efficiency dropped to zero after $\mathcal{O}(10\,\mathrm{fb^{-1}})$ of integrated luminosity.

In 2017, new silicon pixel detectors, based on the 3D technology [162], were installed in the RPs at the 220-F locations, replacing the strip trackers at 210-N. These detectors were specifically developed for PPS, with much improved radiation tolerance and rate capability with respect to the strip detectors, allowing for the reconstruction of multiple tracks per bunch crossing. Finally, in 2018, both stations at 210-F and 220-F were equipped with pixel detectors.

Despite the improved resistance to radiation, the dose accumulated by the pixel sensors and readout chips during Run 2 has been such to degrade significantly the detector performance, as briefly described in the following. This, and the lack of enough replacement parts, called for the construction of new detector modules; with the aim of mitigating the adverse effects of radiation damage, a new design for the support mechanics has been developed for Run 3, which in turn implied a redesign of the frontend electronics. Here, the main differences of the Run 3 design from that used in Run 2 are outlined. A more complete description of the Run 2 setup can be found in ref. [163].

2024 JINST 19 P05064

Each pixel tracking station consists of six detector planes, oriented at an angle of 70° with respect to the thin window, i.e., 20° with respect to the plane perpendicular to the beam axis (it was 18.4° in Run 2). This orientation removes the geometrical inefficiency associated to the junction and ohmic columns of the sensor, and increases the probability of charge sharing between nearby pixels for particle signals. The sensor size is driven by that of the readout chip (ROC): for Run 3 it is $16.20 \times 16.65\,\text{mm}^2$, corresponding to a 2×2 ROC matrix; in Run 2 most sensors were longer, $20.40 \times 16.65\,\text{mm}^2$, and read out by a 3×2 ROC matrix. The pixel size is $150 \times 100\,\mu\text{m}^2$, also adapted to that of ROC cells; the pixels located at the horizontal or vertical edge of a ROC have double size in the corresponding coordinate, except for those at the sensor edge. The details of the sensor module geometry and the arrangement in the tracking station are shown in figure 91.

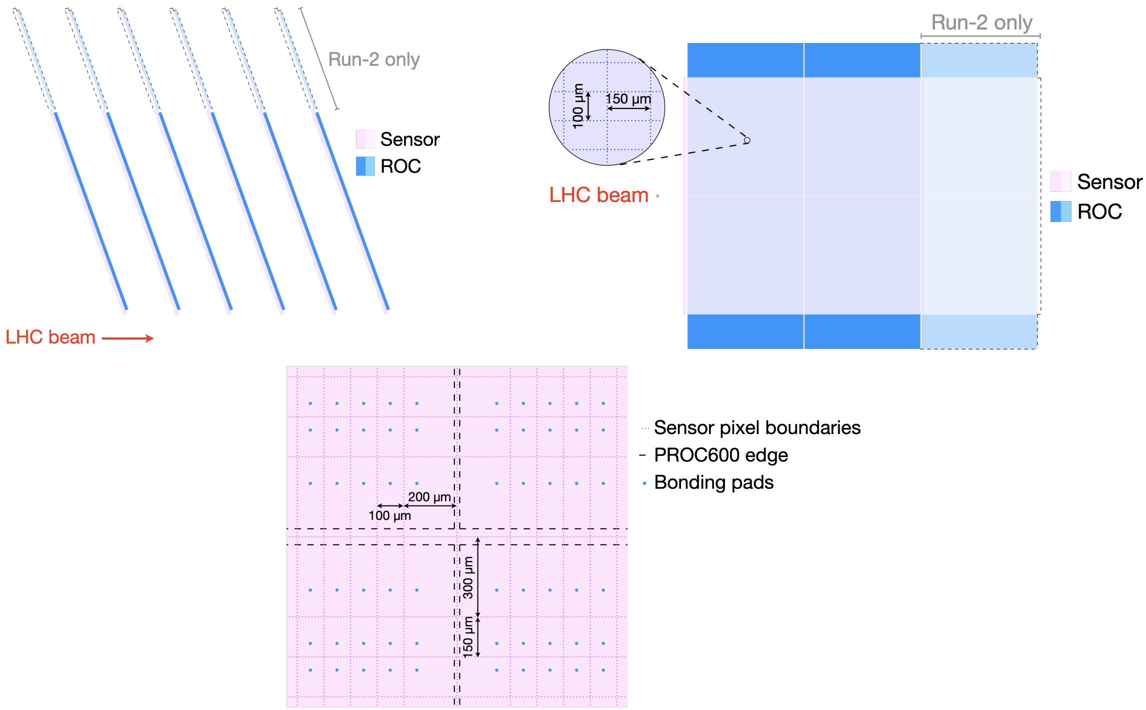


Figure 91. Geometry and arrangement of the pixel sensor modules. Upper left: arrangement of the sensor modules in a tracking station, relative to the LHC beam. Upper right: sensor, ROCs, and pixel geometry. The shaded areas with dashed contours refer to the larger, 3×2 modules used in Run 2. Lower: detail of pixels at internal ROC edges.

The 3D sensors have been produced by the Foundation Bruno Kessler (FBK) in Trento, Italy [164]. They are realized with "single-side" technology, where both the junction and ohmic columns are etched from the same side of the silicon wafer; the thickness of the active volume is $150\,\mu$m. A so-called 2E configuration is used, where each pixel has two readout (junction) columns, thus improving the radiation resistance characteristics. In Run 2, the sensors, produced by Centro Nacional de Microelectrónica (CNM) in Barcelona, Spain [165], were realized with "double-side" technology, with the two types of columns etched on opposite sides of the wafer, with $230\,\mu$m active thickness, in both 1E and 2E configurations. Figure 92 shows schematically a cross section of the column layout for the two productions. Due to the 3D configuration, bulk depletion occurs at inverse polarization values of a few Volts. Currents below 1 µA are required at the working point, but are generally much lower (a looser requirement was set on Run 2 sensors); the breakdown tension must exceed 60 V.

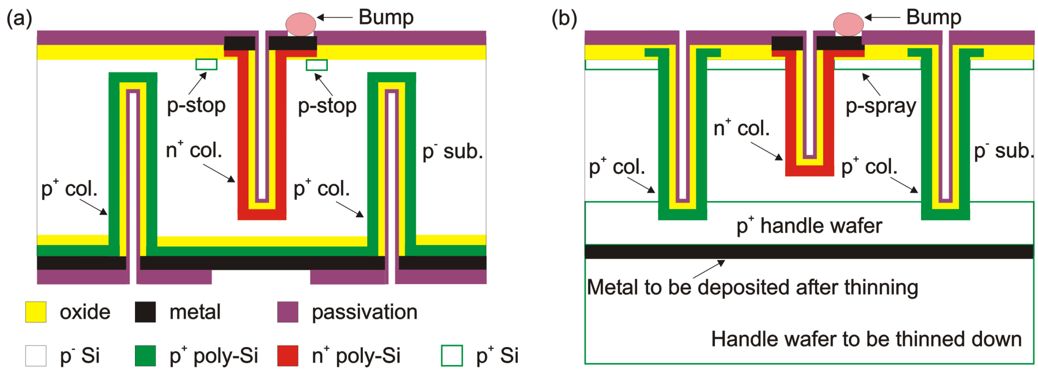


Figure 92. Schematic layout of the columns of 3D pixels. Left: double-sided technology, used in Run 2. Right: single-sided technology, used in Run 3. Metal bumps on top are used for interconnection with the readout chips; p-type surface implants, such as p-stop or p-spray, provide isolation between contiguous pixels. Reproduced with permission from [166].

7.2.2 Readout electronics

The readout of pixel detectors is based on the frontend electronics developed for the Phase 1 upgrade of the CMS pixel tracker [15]. A schematic view of the readout chain is shown in figure 93.

Signals collected from the sensor electrodes are processed by the PROC600 chips developed for the layer 1 of the barrel pixel detector, described in section 3.1.4: four chips are connected to each sensor via bump bonds. In Run 2, the PSI46dig chips [167] were used. A flexible printed circuit board is glued on top of the sensor; the ROCs are connected to the board via wire bonds. A TBM10d chip (token bit manager), mounted on the board, distributes clock, fast commands and configuration instructions to the ROCs, and collects data through a token ring architecture; only one of the two TBM cores is used, resulting in a single output data stream. In Run 2, the flexible board was mounted around the sensor module, and carried one TBM08c chip. The two versions of the flexible board are shown in figure 94.

The six detector modules of a tracking station are connected, through connectors on the flexible boards, to a concentrator card, called PortCard (figure 95). The connections are made on a section of the PortCard that enters the secondary vacuum of the RP through a vacuum-sealed feed-through. The rest of the board, operating at atmospheric pressure, contains all the active electronics needed to perform the board tasks:

- manage the configuration of the integrated circuits equipping the board itself via I^2C commands sent on the slow-control ring;
- send the clock signal, fast commands, and configuration instructions to the frontend readout (via the TBM) through an optical link connection;
- collect and send, through individual optical links, the data received from the detector modules;

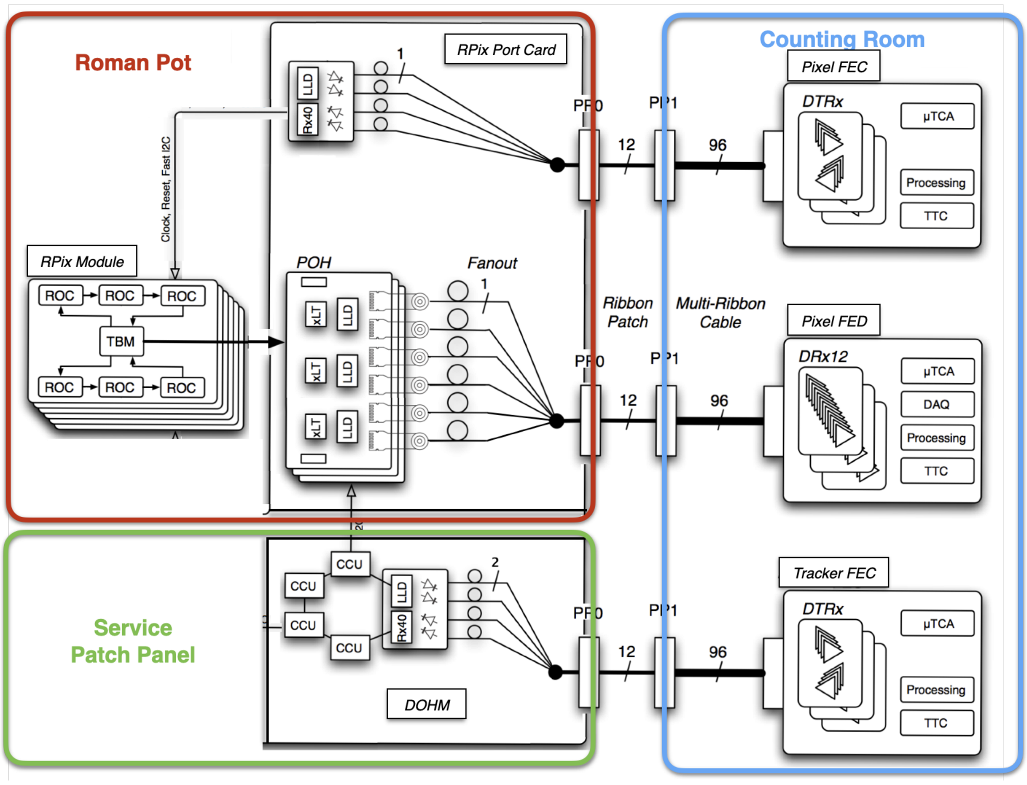


Figure 93. Schematic diagram of the PPS pixel tracker readout chain. The various components shown are described in ref. [15]. Reproduced from [15]. CC BY 3.0.

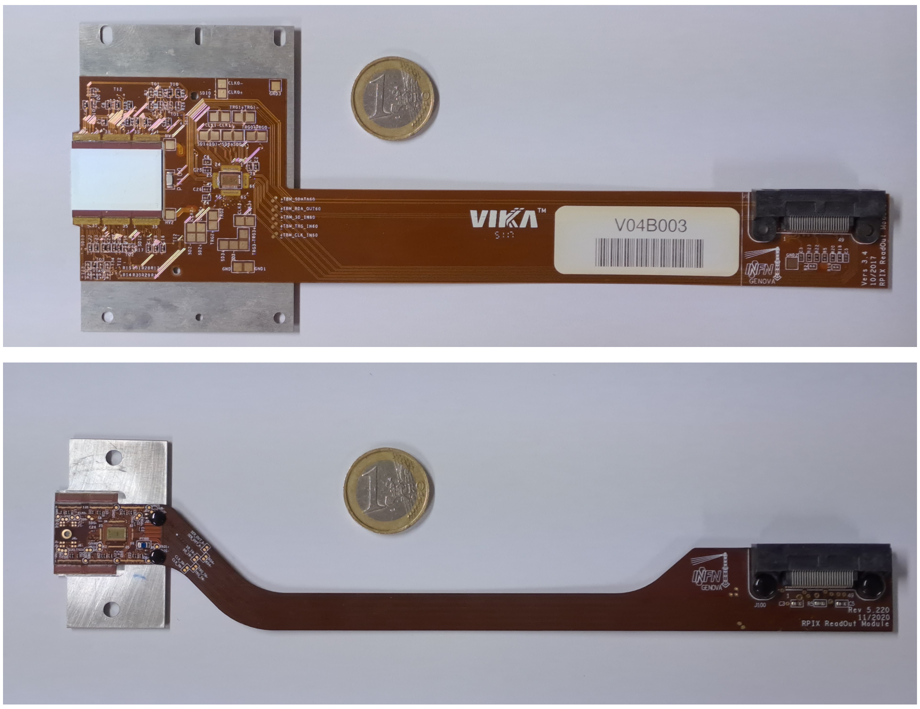


Figure 94. The detector module of the silicon pixel tracker, in its Run 2 (upper) and Run 3 (lower) versions.

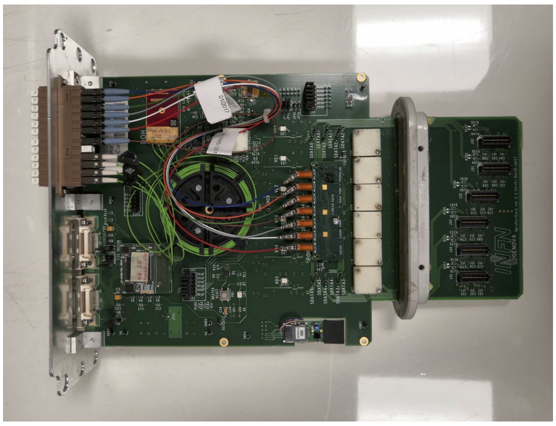


Figure 95. The pixel PortCard for Run 3; the small flange near the bottom acts as a feed-through for the board. The section below is housed inside the secondary vacuum volume of the pot.

- route lines needed to read out environmental parameters in the detector housing and control the internal movement system (described later).

The PortCard can also control, through a parallel port register, the selective powering of the detector planes.

7.2.3 Support structure and internal motion system

The sensor modules and the flexible boards are glued, using a flowable silicon-based sealant, to aluminum support plates, ensuring thermal connection to the cooling circuit on the sides of the detector package (figure 95). Precision holes in the plates are used to control the position of the sensors with respect to the thin window of the detector housing. In Run 2, larger support plates were employed, realized in thermal pyrolytic graphite (TPG) enclosed in thin aluminum sheets.

An aluminum support structure carries the six detector modules, ensuring their precise positioning and coupling them to the evaporators of the cooling system. Six slits on the sides of the support structure host the aluminum plates of the modules, and pins on both sides keep them in place. Three adjustable transfer-ball feet provide contact with the floor of the detector housing; on the opposite side, two spring-loaded frames allow the correct sealing of the upper flange of the RP and the transverse movement of the detector package, as described in the following. In Run 2, planes were arranged in pairs in separate sections of the support structure, eventually assembled together and to the rest of the package mechanics. Feet were static and no internal movement was allowed. Figure 96 shows the assembled detector packages, as well as the connections to the readout and services outside the secondary vacuum volume.

Both the sensors and the readout chips used in Run 2 were designed to withstand fluence values up to 2–3$\times10^{15}$ n_{eq}/cm^2 [162, 168]. However, the highly nonuniform distribution of the hit rate (figure 97) implied localized radiation damage in the readout chip structures that could not be compensated by changing global configuration registers of the chip itself. In particular, the position of the time window to accept hits (WBC register) could not be optimized for all pixels simultaneously: after a data-taking period corresponding to about 20 fb^{-1} of integrated luminosity, the regions with the highest rate could not be read out.

2024 JINST 19 P05064

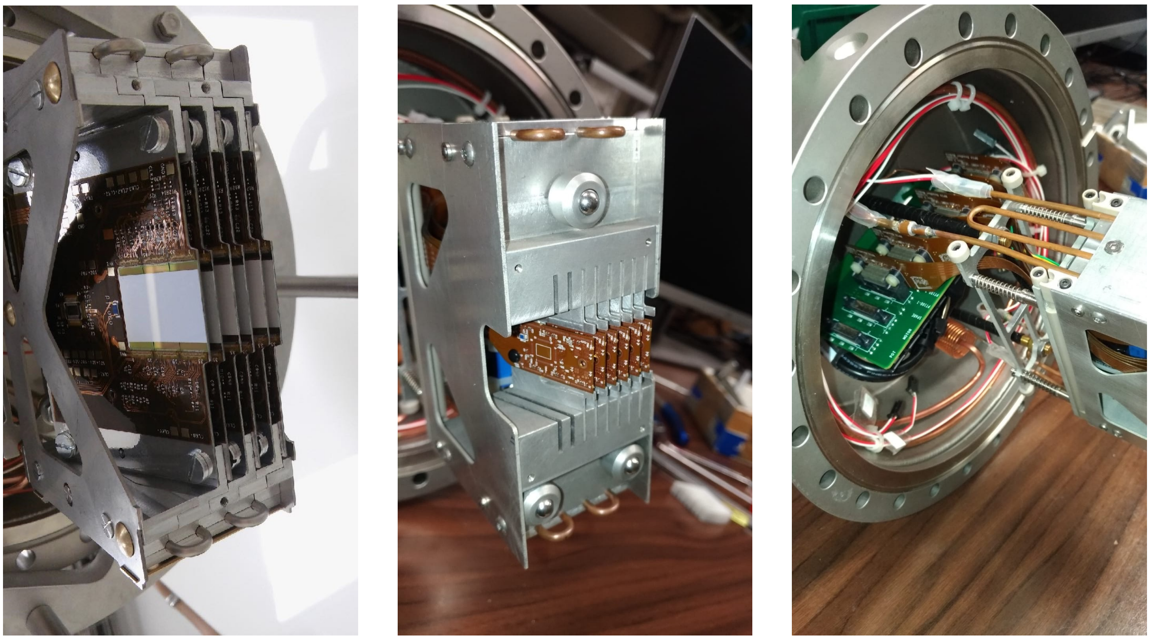


Figure 96. Details of the pixel detector package. Left: Run 2 version. Center: Run 3 version. Right: mechanical, cooling, and electrical connections to the upper part of the RP in Run 3: the vacuum section of the PortCard is visible, as well as the cooling circuit and the connections to the environmental sensors.

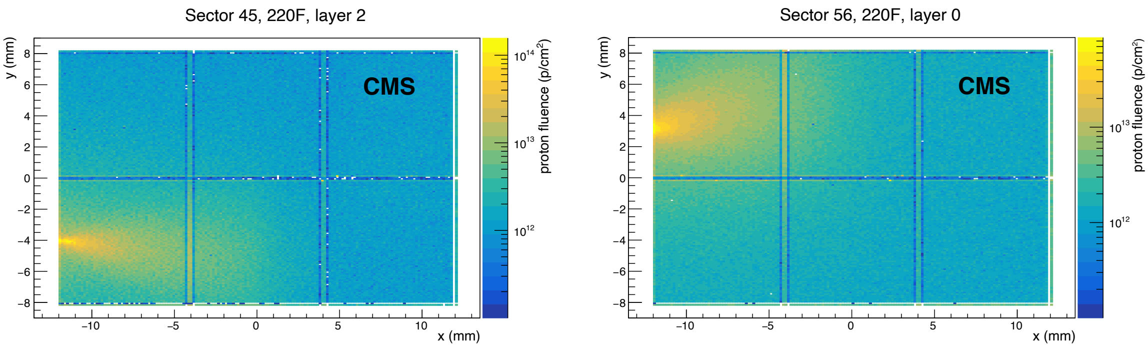


Figure 97. Proton fluence in two planes (one per arm) of the PPS pixel tracker in a sample run in 2017, normalized to an integrated luminosity of 1 fb^{-1}. The horizontal and vertical lines in the distributions are an artifact due to the different pixel size of the sensors in those regions.

With the aim of extending the lifetime of the detectors, in Run 2 the support of the corresponding RP was raised with respect to the floor of the LHC tunnel, in two steps of 0.5 mm, during technical stops. Because the inefficient area was so localized, these small shifts were sufficient to move it to a region with much smaller hit rate, and thus almost restore the overall initial efficiency. However, the range for this kind of movements cannot be further extended; moreover, the operation is onerous and not without risk.

In order to perform the same kind of movement in a more practical and safe fashion, an internal (vertical) movement system for the pixel detector package has been developed. The very design of the new detector modules has been driven by this requirement. Their width has been minimized (by

2024 JINST 19 P05064

placing the flexible board on top of the sensor) so as to exploit all the available space in front of the RP thin window. The width of the thin window is 29 mm, while that of the sensor module is just below 22 mm. A movement range of 5 mm for the detectors has been foreseen, including some safety margin. This will allow to shift vertically the sensor modules, during the data-taking period, in 11, 500 μm-wide steps, thus distributing the radiation damage.

The task is accomplished by means of a miniaturized stepping-motor linear actuator (Zaber LAC10A-T4A), mounted inside the detector package (figure 98). The actuator has an excursion of 10 mm and a step size of about 23.8 nm. Its body is fastened to the lower frame of the spring-loaded structure; the head of the actuator is screwed into one of the side support parts. These two parts can slide on each other thanks to low-friction elements realized in polyether ether ketone (PEEK). Their relative position is measured by a simple linear motion potentiometer (Bourns 3048L-5-103), mounted in a similar fashion as the actuator, whose resistance can be read out externally. Extensive tests have been performed on the fully assembled system, in particular performing repeated movements inside an exact reproduction of the RP and in conditions similar to the working ones ($T \approx -20^{\circ}$C, $P \approx 15$ mbar). Some results are shown in figure 98. The tests have demonstrated that the measured variation in resistance is consistent with the expected shift, showing the adequacy of the position measurement, where only a moderate precision is required. Some hysteresis effect is observed, most likely due to the mechanics of the potentiometer, when the shift direction is inverted; however, this effect is reproducible over several movement cycles.

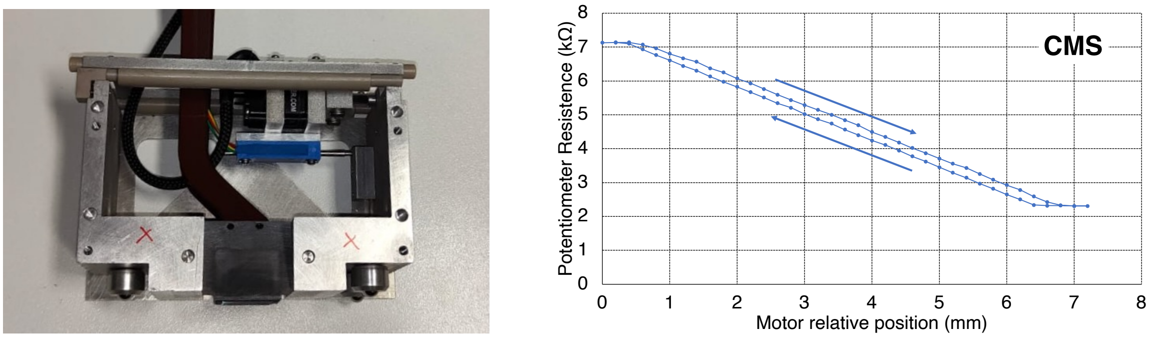


Figure 98. The system for internal motion of the pixel detector package. Left: detail of a partially assembled detector package: the stepper motor is the black object on top; the blue object below is the potentiometer used to monitor the position; both of them have their body tied up to the sliding slit on top, while their mobile tip is screwed to the support structure for the modules. Right: results of a motion test inside a RP at standard working conditions (-20°C and about 5 mbar), with measured versus nominal position. Two sets of points can be identified for forward and backward movements, revealing a hysteresis effect.

The linear actuators are controlled by dedicated two-channel controllers (Zaber X-MCC2). One controller per arm is installed inside the RR53 and RR57 shielded areas (“alcoves”) along the tunnel. They are connected to Raspberry PI microcomputers via USB link, which also reads out the detector position through an external ADC board; the microcomputers can be accessed from outside via the 4G network in the tunnel. This kind of setup is adequate in view of the noncontinuous operation of the motors. In fact, only a limited number of detector shifts is performed over the whole data-taking time, in interfill periods.

7.3 Timing detectors

In the presence of multiple proton-proton interactions in the same bunch crossing (pileup), the separation of overlapping events in CMS relies on the reconstruction of multiple primary interaction vertices. However, because the proton tracks reconstructed by PPS have scattering angles very close to zero, the tracking system described in the previous section cannot associate them to CMS primary vertices. For events with two protons detected on opposite sides, this limitation can be overcome if a precise measurement of the arrival time of protons is available: from the difference in time $\Delta t = t_+ - t_-$, where t_+ and t_- are the time measurements in the positive and negative arm of PPS, respectively, the position in z of the pp vertex can be inferred from $z_{\text{pp}} = c\Delta t/2$. Dedicated studies [156] have shown that a time resolution of $O(10\,\text{ps})$ is needed to achieve the correct association of forward protons to the events reconstructed centrally by CMS. For pileup conditions corresponding to $\mu = 50$, a time measurement with 10 (30) ps resolution could reject the combinatorial background from random combination of uncorrelated protons by about a factor 30 (20) while keeping about 60 (50)% of the signal. However, larger time resolution values, within 100 ps, can still help in reducing the background. As in the case of the tracking system, timing detectors must be able to operate with large and highly nonuniform particle fluxes, to tolerate high radiation doses and to work in a vacuum. Moreover, the material thickness must be limited.

During Run 2, one roman pot on each arm was used for timing, with four planes per RP. For Run 3, instrumenting an additional station at the 220-N location will allow a second timing RP on each arm, giving a total of eight planes per arm.

7.3.1 Detector modules

The PPS timing detectors [169–171] are based on synthetic single crystal chemical vapor deposit (scCVD) diamonds. The detectors combine good time resolution, extreme hardness against large and nonuniform radiation, low material budget, and fine segmentation near the beam.

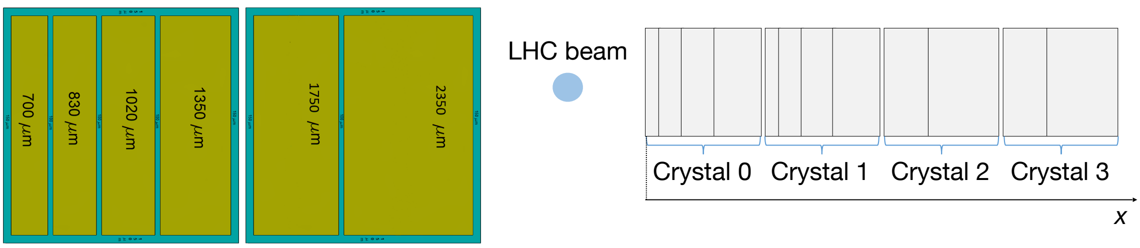


Figure 99. Left: details of the four-pad and two-pad segmentation of the diamond sensors used in the Run 3 modules. Right: arrangement of the four crystals in a Run 3 module, where the position of the beam is indicated by the spot on the left. Reproduced with permission from [172].

At the conclusion of Run 2, half of the planes were based on an improved double-diamond architecture, in which two diamonds are connected to the same amplification channel. This nearly doubles the signal, while keeping a similar noise level, resulting in a significant improvement of the timing resolution compared to a single-diamond design. In Run 3, all timing planes are of the double-diamond design. The crystals have dimensions of $4.5 \times 4.5\,\text{mm}^2$, with a total active surface area of approximately $20 \times 4.5\,\text{mm}^2$ per plane. The final segmentation is achieved during the metalization process, resulting in a total of 12 channels per plane. Due to the highly nonuniform flux of particles, the segmentation varies with the distance from the beam in the x direction. Several configurations of pad dimensions and sensor layout have been used in Run 2; in Run 3, sensors with two-pad and four-pad segmentation are used, with dimensions as detailed in figure 99, where the layout of four of these sensors in a detector module is also shown. Pads close to the beam position have smaller size (as small as 0.55 mm), while a coarser segmentation is used farther from the beam. This results in a more uniform occupancy, with low inefficiency due to multiple hits in the same channel. Because of the horizontal crossing of the LHC beams at the IP, hits are more widely distributed along the x axis. For this reason, the detectors are only segmented horizontally.

The diamonds for Run 3 include newly produced crystals, and crystals that were previously used in Run 2, cleaned and remetalized. Samples of the latter were first studied in beam tests after dismounting, and found to maintain high efficiency and achieve single plane time resolutions of $\sim$ 80–95 ps, after being exposed to fluences as large as 5×10^{15} p/cm^2 [173]. Compared to the nominal resolution of about 50 ps per plane for a new double-diamond detector, the decrease in resolution was found to be largely consistent with radiation damage to the preamplification electronics.

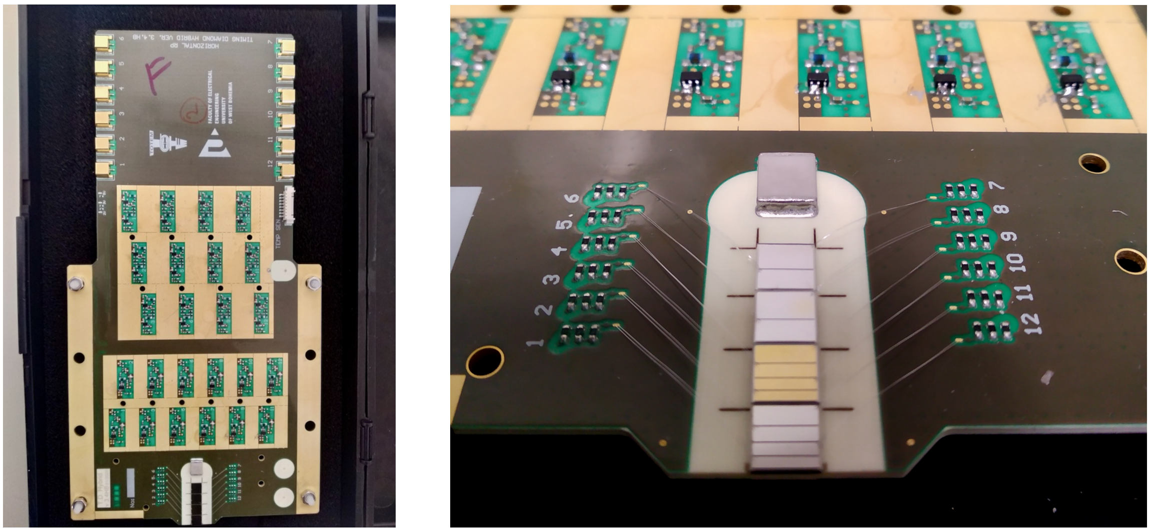


Figure 100. Left: the hybrid board for the Run 3 timing detectors' readout; the lower, wider section hosts the sensors and is housed inside the secondary vacuum volume of the pot. Right: detail of the diamond sensors on one side, connected via wire bonds to the frontend electronics.

The diamonds are glued to a hybrid board (figure 100), containing both the sensors and a multistage amplification chain for each of the 12 channels. As mentioned above, radiation damage to the amplification chain was identified as a limiting factor for the timing performance in Run 2. Therefore, a revision of the hybrid board was designed for Run 3, with a modified layout to mitigate the exposure of the preamplifiers to radiation. Additional modifications to the design were made to improve the high-voltage isolation and the stability against RF noise pickup. Finally, remote control of the amplifier gains was implemented, giving the opportunity to better fine tune the settings and compensate for any degradation during data taking.

7.3.2 Readout electronics

The signals from the hybrid boards are transmitted by individual coaxial cables to custom “NINO boards”, mounted in a mini-crate about 1 m above the LHC beam pipe (figure 101). The main data path for reading all channels at the full trigger rate is based on the fast, low-power NINO ASIC [174], with four chips per board. The NINO performs discrimination of the input signals above an adjustable threshold. The width of the output signal is proportional to the input charge, and is then stretched by a constant value for compatibility with the next piece in the readout chain. This serves as a proxy for the signal amplitude, allowing for time-walk corrections to be derived from the data. In Run 3, a revision of the NINO board also allows for signals from a subset of channels to be split off to record the full waveform information, at reduced rate.

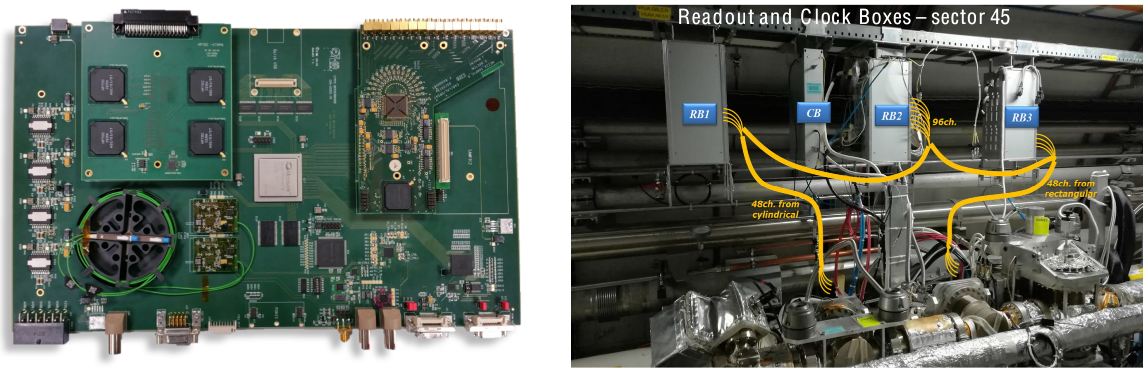


Figure 101. Left: the digital readout unit. In this example both an HPTDC mezzanine (upper left) and a SAMPIC mezzanine (upper right) are mounted on the motherboard. Right: location of PPS mini-crates installed above the LHC beam pipe, for both the cylindrical timing RP already used in Run 2, and the rectangular 220-N RP newly equipped for timing in Run 3. The readout boxes (labeled “RB”) contain the NINO boards and DRUs as described in the text. The box labeled “CB” holds components for the precise reference clock.

From the NINO boards, the signals are transmitted to the digital readout units (DRUs), mounted in the same mini-crate above the beam pipe. Each DRU consists of a digitizer motherboard supporting two types of mezzanine for timing measurements. The first, used in Run 2 and for all channels in Run 3, includes four high-performance time-to-digital converter (HPTDC) chips [175]. The HPDTC for PPS is used in very high resolution mode, where an on-chip four-point sampling and interpolation of a delay-locked loop is performed to improve the time resolution. This results in a TDC with a binning of about 25 ps, at the cost of limiting the number of output channels per chip to eight. The HPTDC records the time of both the leading and trailing edges of the signal, allowing for time-walk corrections to be performed as a function of the reconstructed input charge value. In Run 3, additional DRUs are also equipped with a second type of mezzanine, based on the SAMPIC [176] fast waveform sampler. This allows the readout of the full analog signal shape, for calibration purposes, on a fraction of the data, for trigger rates up to 100 kHz. Each digitizer motherboard includes a radiation-hard MicroSemi SmartFusion2 FPGA, which formats the data from the HPTDC (or SAMPIC) mezzanine into the final data frames. The data from each board are then transmitted over optical fibers to the backend using two GOH opto hybrid mezzanines.

In addition to its role in the data acquisition, the DRU is also responsible for receiving and distributing slow-control commands, clock, and fast commands. It receives both the standard clock

2024 JINST 19 P05064

and the dedicated precise reference clock [169] that is used for the timing measurement. Along with the configuring of components on the DRU itself, it can propagate commands via I^2C interface to set the thresholds on the NINO board, and, for Run 3, the low voltage settings on the diamond hybrid. The ability to remotely set the low voltage will allow for better fine tuning of the operational settings, and provide the ability to compensate for possible radiation damage effects.

A schematic view of the timing control and readout chain components installed in the LHC tunnel is shown in figure 102.

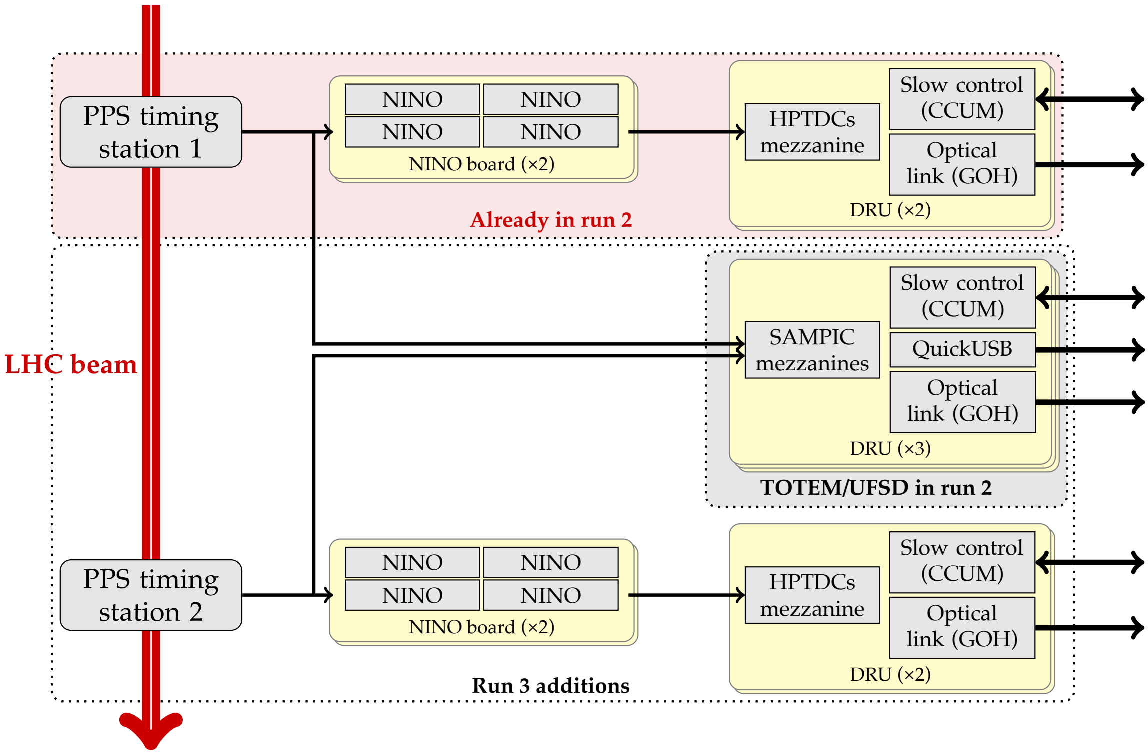


Figure 102. Schematic diagram of the full readout chain for the PPS timing detectors in Run 3, also showing the configuration used in Run 2. The arrows on the right represent electrical connections to the slow-control ring and optical connections to components in the underground service cavern of CMS, described in section 7.4. The central set of DRUs (3 units) was already present in Run 2, but was employed for the diamond detectors equipping the TOTEM experiment at that time, and for one layer of UFSD detectors used in 2017 for the PPS timing.

7.3.3 Reference clock distribution

For precise timing measurements, a clock distribution system that provides time information at points separated by large distances, with a picosecond precision, is needed. This requires a system capable of the highest precision and of the utmost time stability.

PPS employs the clock distribution system developed by the TOTEM Collaboration [177], an adaptation of the universal picosecond timing system [178, 179], developed for FAIR (Facility for Antiproton and Ion Research) at GSI, Darmstadt. The system has been installed and used during the TOTEM and PPS runs since 2017.

The system can be logically subdivided into four major blocks: the transmission unit, the distribution unit, the measurement unit, and the receiving unit. Receiving units are installed in the LHC tunnel, as close as possible to each timing detector, while the transmission, distribution, and measurement units are located in the IP5 counting room. A block diagram of the entire system is shown in figure 103. The system leverages the use of a dense wavelength division multiplex (DWDM) technique, exploiting the transmission of multiple signals of different wavelengths over a common single-mode fiber. This allows the use of standard telecommunication modules compliant with ITU (International Telecommunications Union) standards.

The transmission unit optically modulates two reference clock signals using two different DWDM wavelength carriers: λ_1 and λ_2. These optical signals are multiplexed into a single fiber and transmitted to the distribution unit, where they are split to be distributed to all receiving units located in the tunnel on both sides. The multiplexed signal is then optically amplified via an erbium-doped fiber amplifier (EDFA) to compensate for the attenuation due to the multiplexing and splitting steps.

These signals are further multiplexed with a third one, of wavelength λ_M, with the goal of measuring the transmission delay over each fiber. This is done in the measurement unit, where a network analyzer drives the optical modulation of this third DWDM signal, which is then sent to and reflected back by the receiving unit. In this way, the signal delays can be determined and possible drifts can be monitored.

In the tunnel, receiving units separate and convert the multiplexed optical signals generated in the transmission unit back to electrical signals, and reflect, via a fiber Bragg grating reflector, the one generated in the measurement unit. The electric signals are then routed to frontend and readout electronics.

Measurements performed with the installed system have shown an additional contribution to jitter of slightly less than 1 ps, mainly due to the inherent jitter of the clock source signal, the noise added by the optical components, and the bandwidth of the transmission system itself.

7.4 Data acquisition and detector control

The backend data acquisition systems for the various detectors used in PPS have been developed independently and are based on different hardware. The strip trackers have maintained the original structure developed for TOTEM, and are integrated into the CMS data acquisition (DAQ) system, described in section 9. A similar architecture is used for the timing detectors. The pixel tracker system is based on the DAQ scheme employed by the Phase 1 upgrade of the CMS pixel detector, described in section 3.1.4.

7.4.1 Pixel detectors

The pixel DAQ is based on μTCA FC7 carrier boards, following the design illustrated in figure 93, and described in more detail in ref. [21]. The boards are employed as either FEDs or FECs, depending on the type of attached mezzanine card and the firmware deployed. Two FC7 boards are equipped as FEDs, and receive data transmitted by POHs over optical fibers from the pixel detectors on the two arms of the PPS spectrometer. The typical data volume in Run 2 was about 0.5 kB/event or less, depending on the instantaneous luminosity and other LHC conditions, and is similar in Run 3. Additional FC7s are equipped with FEC mezzanines, transmitting signals over fibers to the frontend boards in the LHC tunnel. Depending on the firmware, the FECs are used as either PxFECs, sending

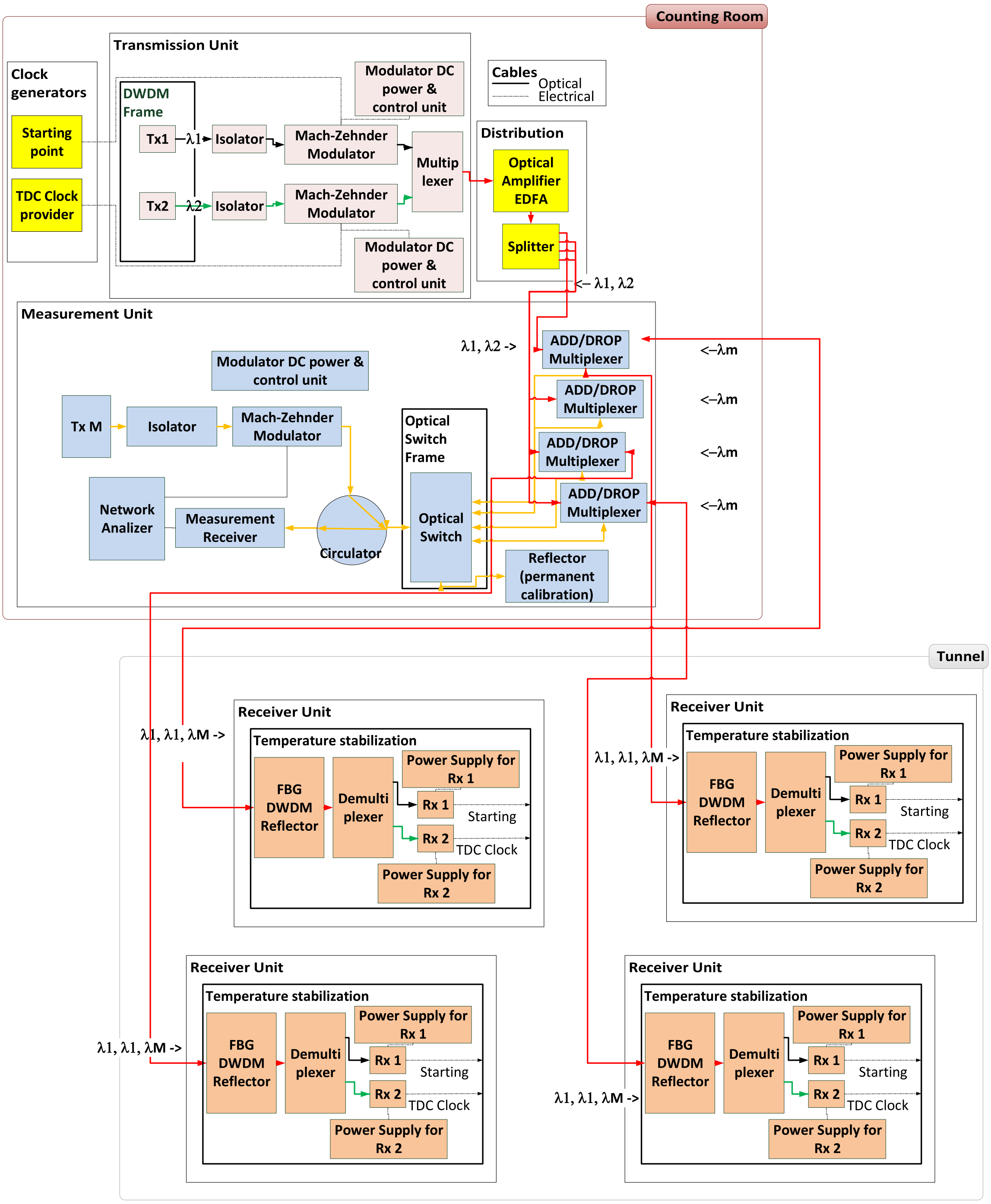


Figure 103. A block scheme of the clock distribution system for the PPS timing detectors. The four receiver units correspond each to a timing station in the tunnel; the remaining elements of the system are all located in the counting room. Different colors of the fiber lines represent the different wavelength carriers used by the system, λ_1, λ_2, λ_M (black, green, and yellow, respectively), and the multiplexed signals (red).

the clock and trigger, or TkFECs, responsible for sending slow-control commands. The entire system, shown in figure 104 (left), is housed in a single μTCA crate, containing one AMC13 for clock and trigger distribution, and one MCH crate controller.

Figure 104. Left: the PPS pixel DAQ crate. On the left are two FC7 FECs, sending clock, trigger and slow-control commands; in the center-right are two FC7 FEDs, receiving data from the two arms of the PPS spectrometer; on the far left is the MCH crate controller, on the far right is the AMC13. Right: the timing and strips DAQ crate. The SLinks are placed in the backplane of the OptoRx boards delivering data to upstream CMS DAQ.

7.4.2 Strip and timing detectors

The DAQ for the PPS timing detectors, and for the strip detectors used in special alignment fills, was developed by the TOTEM Collaboration, based on a scalable readout system (SRS) [180]. Optical receiver mezzanines (OptoRx) are connected to a custom designed board, which further connects to a frontend concentrator board. The OptoRx receives data transmitted over fibers from GOHs on the timing digital readout units. A scalable readout unit (SRU) is responsible for receiving and distributing the clock, trigger, and fast commands within the SRS crate (figure 104 right). The system was designed for full compatibility with the CMS DAQ. Data are transmitted via SLink to the FRLs of the central CMS DAQ, where they are treated in the same way as those from other FEDs. During Run 2, a total of two VME-based FEDs were used for the timing detectors in normal data taking, while four additional such FEDs were used to read out the strip detectors in the vertical RPs for special alignment runs. In Run 3, a total of six timing FEDs are used in standard runs, to support the addition of a second timing detector station, and the readout of a limited amount of data with a SAMPIC waveform digitizer.

The SRS firmware was originally designed to read data from the VFAT chip [145] used by the TOTEM strip detectors. In order to use this system with minimal changes, the timing data from the HPTDCs are packed into a VFAT-like frame, with no zero-suppression. In Run 2, the data volume from the full timing system was around 2 kB/event. In Run 3, this will more than double with the addition of a second timing station and the SAMPIC readout option. The system is capable of sustaining the CMS trigger rate of more than 100 kHz.

The distribution of (non-reference) clock, trigger, and slow-control commands for the timing electronics is handled by a VME FEC, of the same type and configuration as used in the electromagnetic calorimeter preshower (section 4). The commands are transmitted over fibers to the frontend boards of the timing system. One control loop is dedicated to the strip detectors, while another

2024 JINST 19 P05064

is dedicated to the timing detectors. In addition, firmware for the FPGA of the timing digitizer motherboard can be deployed over the slow-control loop, using a new software development that takes advantage of the JTAG master of the CCU25 communication and control unit [29]. This allows firmware updates to be made remotely, without requiring access to the LHC tunnel.

7.4.3 Detector control system

The PPS detector control system (DCS) is a legacy from the TOTEM experiment built with the industrial WinCC OA software (PVSS) used to control the low and high voltage of detector packages, to monitor the radiation dose, pressure, and temperature sensors, as well as the roman pots movements and the status of interlocks of the detector safety system (DSS). The TOTEM DCS system includes the automatic matrix actions to manage the power-supply state of the roman pots according to the LHC beam status. The system for Run 3 is integrated within the CMS DCS framework, and runs with WinCC OA version 3.16.

7.5 Roman pot insertion and running scenarios

The position of the roman pots with respect to the LHC beam is controlled by the LHC operators through standardized procedures. During the injection, acceleration and luminosity tuning stages, the RPs are kept in a retracted ("garage") position, at about 40 mm from the proton beam. When "Stable Beams" are declared, the RPs are moved closer to the protons, at a distance depending on the beam width at that location, $\sigma_{x,\mathrm{XRP}}$. This distance, d_{XRP}, was given in Run 2 by the following rule from machine protection:

$$d_{\mathrm{XRP}} = \max\left[(n_{\mathrm{TCT}} + 3)\sigma_{x,\mathrm{XRP}} + 0.3\,\mathrm{mm}, 1.5\,\mathrm{mm}\right], \tag{7.1}$$

where n_{TCT} is the distance of the tertiary collimator (TCT) from the beam center in units of the beam width $\sigma_{x,\mathrm{TCT}}$ at its location. The 3σ retraction ensures that the RPs stay in the shadow of the TCT, while the additional 0.3 mm margin protects against accidental beam orbit deviations. From arguments of mechanical rigidity of the RP thin window, which could bulge towards the beam in case of a secondary vacuum loss, an absolute lower limit of 1.5 mm was imposed. In Run 2, with $n_{\mathrm{TCT}} = 8.5$, the resulting d_{XRP} ranged between 2.2 mm (210-F) and 1.5 mm (220-F).

Operations in Run 3 are characterized by a far more complex luminosity-leveling scheme with concurrent changes in the crossing angle, called α in this section, and the beta function at the IP, β^*. The most recent concept of the leveling scheme can be represented by the trajectories shown in figure 105 for the years 2022 and 2023.

Both α and β^*, as well as the collimation scheme, have a decisive impact on operation and performance of the RP spectrometer.

7.5.1 TCT collimator and roman pot insertion scheme

The RP approach distance d_{XRP}, as given in eq. (7.1), depends on β^* through the beam width, $\sigma_{x,\mathrm{XRP}}$, and through the TCT distance, $n_{\mathrm{TCT}} = d_{\mathrm{TCT}}/\sigma_{x,\mathrm{TCT}}$. The function $d_{\mathrm{TCT}}(\beta^*)$ characterizes the TCT collimation scheme.

In the old standard collimation scheme, used until the end of 2022, the TCT did not move during stable beams, i.e., $d_{\mathrm{TCT}}(\beta^*) = \mathrm{const}$, so the nominal distance d_{XRP} was entirely determined by the

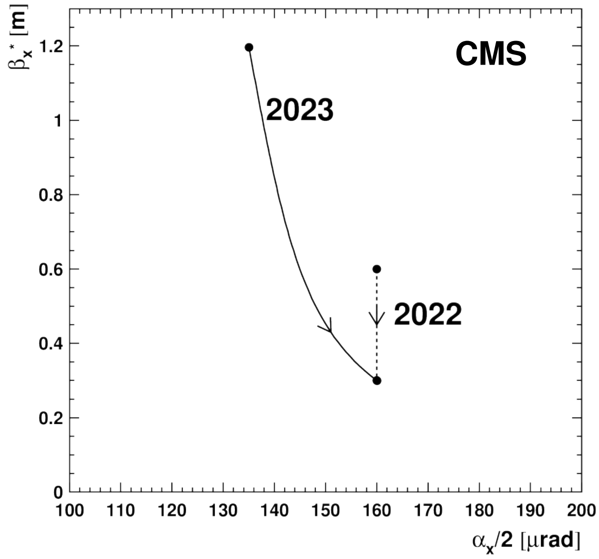


Figure 105. Luminosity-leveling trajectories for typical LHC fills in the $(\alpha/2, \beta^*)$ plane for 2022 (dashed line) and 2023 (continuous line).

evolution of the beam widths, $\sigma_{x,\mathrm{XRP}}$ and $\sigma_{x,\mathrm{TCT}}$, with β^*. With β^*-leveling, d_{XRP} changes during the fill. However, an automated RP movement synchronized with the β^* evolution during the fill is not foreseen in the control system and would be difficult to implement, in particular because also the position limits in the interlock system would have to follow this movement. In Run 2, this was not a problem, since the range of β^* (25–30 cm in 2018) was so small that the variation of d_{XRP} was negligible. The RPs were kept fixed at the maximum distances along their nominal trajectories. In Run 3, the wider β^* range complicates the situation. The evolution of the RP distance with β^*, for the old collimation scheme with fixed TCT positions, is shown in figure 106, for two example RPs.

In 2022, where the β^* leveling covered the range from 0.6 to 0.3 m, the RP with the widest nominal distance range was 210-F. Ideally, it would have had to move by 0.22 mm. The RP 220-F, on the other hand, could stay fixed at the limit of 1.5 mm throughout the fill. The situation of the two other pots, 220-N and 220-C, was between these extremes, with ideal movement ranges smaller than 0.2 mm.

In 2023, the much wider β^* range (1.2 to 0.3 m) would have led to movement amplitudes up to 1 mm. To judge whether it is acceptable to keep the RPs fixed at the most distant points along their trajectories like in Run 2, the impact of the RP distance on the mass acceptance has to be considered. The minimum accepted mass M_{min} of centrally produced states X from the process $\mathrm{pp} \to \mathrm{pXp}$ with double proton detection is given by:

$$M_{\mathrm{min}} = \sqrt{s}\,\frac{d_{\mathrm{XRP}} + \delta}{D_{x,\mathrm{XRP}}}, \tag{7.2}$$

where $\delta \approx 0.5$ mm is the insensitive margin from the outer RP window to the point of full efficiency in the detector, and $\sqrt{s} = 13.6$ TeV. The horizontal dispersion $D_{x,\mathrm{XRP}}$ of the LHC from IP5 to the RP depends linearly on the crossing angle, according to:

$$D_{x,\mathrm{XRP}}(\alpha) = D_{x,\mathrm{XRP}}(0) - D'_x\alpha, \tag{7.3}$$

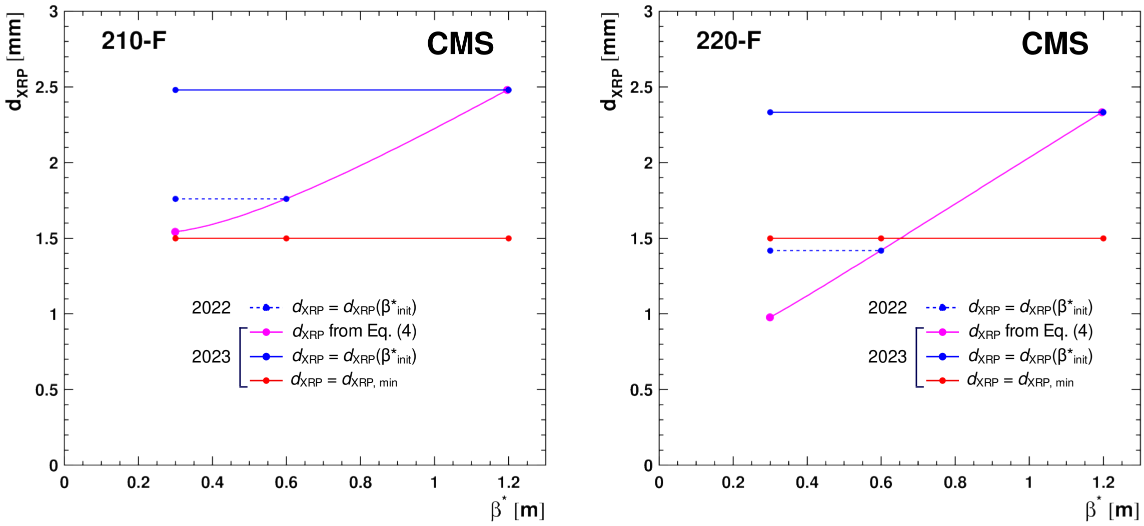


Figure 106. RP insertion distance d_{XRP} for the leveling trajectories of 2022 (dashed lines) and 2023 (continuous lines) from figure 105, evaluated for the pots 210-F (left) and 220-F (right) in the case of the TCT fixed to $d_{\mathrm{TCT}} = 8.5\sigma_{x,\mathrm{TCT}}(\beta^* = 30\,\mathrm{cm})$. The magenta line shows the nominal distance according to eq. (7.1). The blue lines show the most conservative constant RP distance, i.e., never closer to the beam than the nominal distance. If this blue distance is smaller than the limit of 1.5 mm (red line), which is the case for 220-F, the pot has to stay at that limit. The fill evolves from the right to the left.

with a constant $D'_x > 0$, implying that smaller crossing angles yield a larger dispersion and hence a better low-mass acceptance.

Figure 107 shows the evolution of M_{min} during β^* leveling in 2022 and 2023 for the old collimation scheme with fixed TCTs. The differences in M_{min} between 2022 (dashed) and 2023 (continuous) for a given β^* are caused by the concurrent crossing-angle leveling in 2023. The magenta lines represent the case of the RPs moving along their nominal d_{XRP}. The blue lines show the effect of keeping the RPs fixed at the most distant position along their trajectories. If a blue line crosses the limiting red line (1.5 mm distance), the RP has to be fixed at 1.5 mm, and its M_{min} follows the red line.

For 2022, the loss in M_{min} at the end of the fill, at a β^* of 30 cm, by not following the nominal trajectory, at most 35 GeV for 210-F, was considered to be still acceptable. Therefore, the run in 2022 proceeded along the old scenario. Despite the slight loss from omitting the movement, the acceptance for a given $(\alpha/2, \beta^*)$ point was better than in Run 2.

For 2023, on the other hand, the M_{min} acceptance loss would have amounted to 150 GeV. To mitigate this performance deterioration, while maintaining fixed RP positions throughout the fill, the collimation working group has elaborated an alternative scheme where the TCTs, having a more sophisticated movement control system, adapt their positions to the β^* evolution. In this scheme, operationally implemented in 2023, the constant RP positions are all close to the 1.5 mm limit and thus the M_{min} acceptance limits (figure 108) near the safely achievable optimum.

A quantitative mass acceptance overview for 2018, 2022, and 2023 is given in table 11, anticipating the upper mass limits, M_{max}, discussed in section 7.5.2. The higher dispersion in Run 3 leads to generally lower values for both M_{min} and M_{max} than in Run 2.

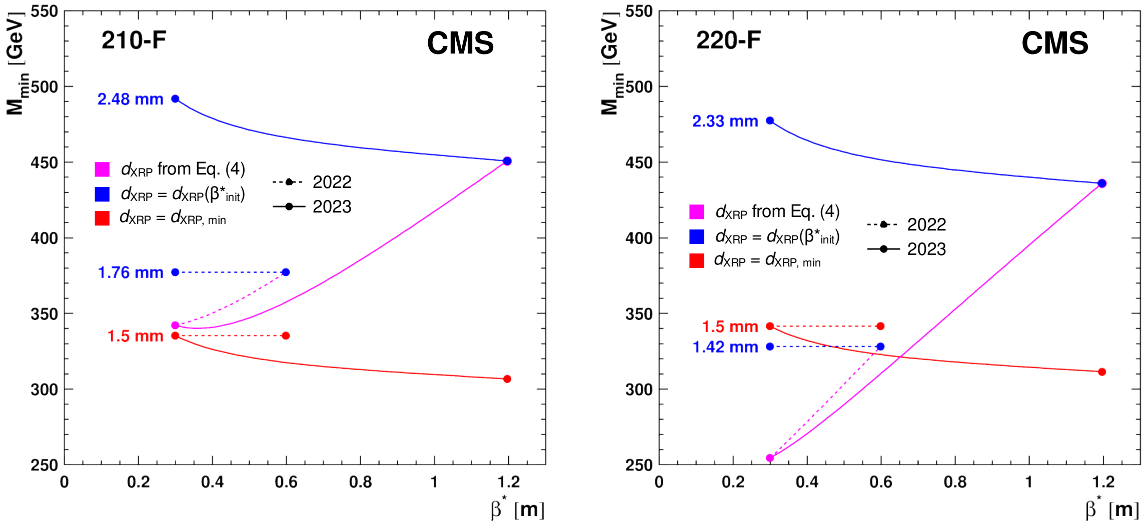


Figure 107. Minimum accepted central mass $M_{\rm min}$ in the RPs 210-F (left) and 220-F (right) for the old collimation scheme with the TCTs fixed at $d_{\rm TCT} = 8.5\sigma_{x,\rm TCT}(\beta^* = 30\,\rm cm)$ and two cases for the RP positions. Magenta lines: RPs moving according to eq. (7.1) and figure 106. Blue lines: RP positions fixed on the most distant point of the nominal trajectory. The red lines correspond to the 1.5 mm distance limit. The fill evolves from the right to the left.

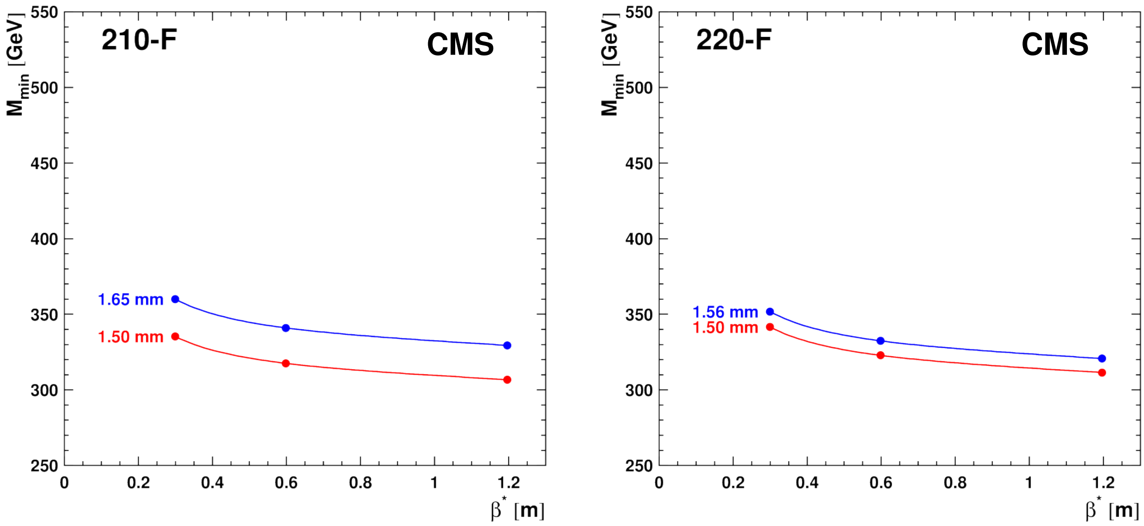


Figure 108. Minimum accepted central mass $M_{\rm min}$ in the RPs 210-F (left) and 220-F (right) for the new collimation scheme with moving TCTs and fixed RP positions. The red lines correspond to the 1.5 mm distance limit. The fill evolves from the right to the left.

7.5.2 TCL collimator insertion scheme

The debris collimators, TCL4 and TCL5, in the outgoing beams upstream of the RPs determine the high-mass acceptance limit $M_{\rm max}$ of PPS. A TCL collimator inserted to the distance $d_{\rm TCL}$ creates a

Table 11. Accepted central mass range, $[M_{\text{min}}, M_{\text{max}}]$ (in GeV), for each RP at the beginning and at the end of the levelling trajectory in 2018 ($\sqrt{s}$ = 13 TeV), and 2022 and 2023 ($\sqrt{s}$ = 13.6 TeV), for the collimation schemes laid out in the text. Due to the coincidence requirement, the RP with the highest M_{min} (typeset in bold face) defines the spectrometer acceptance.

Year	2018	2022	2023	
$\alpha/2$ [µrad]	160	160	135	160
β^* [cm]	30–25	60–30	120	30
210-F	[**449**, 2485]	[**377**, 2086]	[**329**, 1968]	[**360**, 2086]
220-N	[407, 2485]	[355, 2086]	[324, 1968]	[355, 2086]
220-C	[397, 2485]	[350, 2086]	[323, 1968]	[354, 2086]
220-F	[347, 2485]	[342, 2086]	[321, 1968]	[352, 2086]

mass cut-off:

$$M_{\text{max}} = \sqrt{s}\,\frac{d_{\text{TCL}}}{D_{x,\text{TCL}}}, \tag{7.4}$$

where $D_{x,\text{TCL}}$ is the horizontal dispersion at the location of the TCL. Throughout Run 2, TCL4 was inserted to a fixed distance of typically $15\sigma_{x,\text{TCL4}}(\beta^* = 30\,\text{cm})$ from the beam center. The distance of TCL5 was chosen such that it did not introduce any tighter cut: $d_{\text{TCL5}} = 35\sigma_{x,\text{TCL5}}(\beta^* = 30\,\text{cm})$.

In Run 3, the distance of TCL4 is unchanged with respect to Run 2. Due to a slightly different optics, it now corresponds to $17\sigma_{x,\text{TCL4}}(\beta^* = 30\,\text{cm})$. The TCL5 collimator is placed at $42\sigma_{x,\text{TCL5}}(\beta^* = 30\,\text{cm})$, again a position mostly in the shadow of TCL4. Because of the crossing-angle dependence of the dispersion, M_{max} varies throughout the fill as shown in figure 109. The M_{max} values for 2018, 2022, and 2023 are given in table 11.

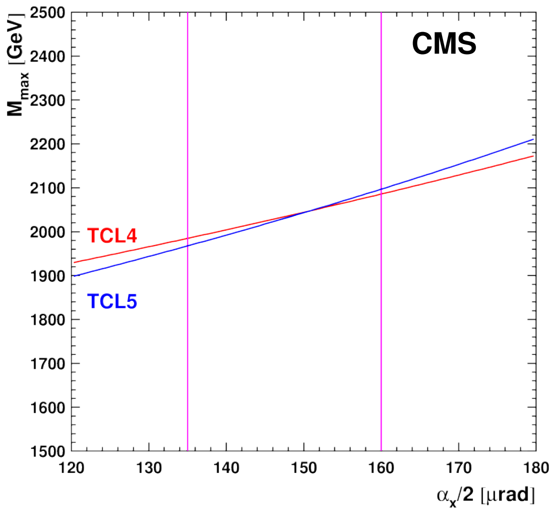


Figure 109. Upper mass cut-off in Run 3 caused by the debris collimators TCL4 and TCL5 for the settings explained in the text as a function of the crossing angle. The fill evolves from the left to the right between the magenta lines.