



The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland



30 August 2023 (v3, 22 September 2023)

MiniDAQ-3: providing concurrent independent subdetector data-taking on CMS production DAQ resources

Vassileios Amoiridis, Ulf Behrens, Andrea Bocci, James Branson, Philipp Brummer, Eric Cano, Sergio Cittolin, Joao Da Silva Almeida Da Quintanilha, Georgiana-Lavinia Darlea, Christian Deldicque, Marc Dobson, Antonin Dvorak, Dominique Gigi, Frank Glege, Guillelmo Gomez-Ceballos, Patrycja Gorniak, Neven Gutic, Jeroen Hegeman, Guillermo Izquierdo Moreno, Thomas Owen James, Wassef Karimeh, Miltiadis Kartalas, Rafal Dominik Krawczyk, Wei Li, Kenneth Long, Frans Meijers, Emilio Meschi, Srecko Morovic, Luciano Orsini, Christoph Paus, Andrea Petrucci, Marco Pieri, Dinyar Sebastian Rabady, Attila Racz, Theodoros Rizopoulos, Hannes Sakulin, Christoph Schwick, Dainius Simelevicius, Polyneikis Tzanis, Cristina Vazquez Velez, Petr Zejdl, Yousen Zhang, Dominika Zogatova

Abstract

The data acquisition (DAQ) of the Compact Muon Solenoid (CMS) experiment at CERN, collects data for events accepted by the Level-1 Trigger from the different detector systems and assembles them in an event builder prior to making them available for further selection in the High Level Trigger, and finally storing the selected events for offline analysis. In addition to the central DAQ providing global acquisition functionality, several separate, so-called "MiniDAQ" setups allow operating independent data acquisition runs using an arbitrary subset of the CMS subdetectors. During Run 2 of the LHC, MiniDAQ setups were running their event builder and High Level Trigger applications on dedicated resources, separate from those used for the central DAQ. This cleanly separated MiniDAQ setups from the central DAQ system, but also meant limited throughput and a fixed number of possible MiniDAQ setups. In Run 3, MiniDAQ-3 setups share production resources with the new central DAQ system, allowing each setup to operate at the maximum Level-1 rate thanks to the reuse of the resources and network bandwidth. Configuration management tools had to be significantly extended to support the synchronization of the DAQ configurations needed for the various setups. We report on the new configuration management features and on the first year of operational experience with the new MiniDAQ-3 system.

MiniDAQ-3: providing concurrent independent subdetector data-taking on CMS production DAQ resources

Vassileios Amoiridis¹, Ulf Behrens², Andrea Bocci¹, James Branson³, Philipp Brummer^{1,}, Eric Cano¹, Sergio Cittolin³, Joao Da Silva Almeida Da Quintanilha¹, Georgiana-Lavinia Darlea⁴, Christian Deldicque¹, Marc Dobson¹, Antonin Dvorak¹, Dominique Gigi¹, Frank Glege¹, Guillelmo Gomez-Ceballos⁴, Patrycja Gorniak¹, Neven Gutic¹, Jeroen Hegeman¹, Guillermo Izquierdo Moreno¹, Thomas Owen James¹, Wassef Karimeh¹, Miltiadis Kartalas¹, Rafał Dominik Krawczyk², Wei Li², Kenneth Long⁴, Frans Meijers¹, Emilio Meschi¹, Srećko Morović³, Luciano Orsini¹, Christoph Paus⁴, Andrea Petrucci³, Marco Pieri³, Dinyar Sebastian Rabady¹, Attila Racz¹, Theodoros Rizopoulos¹, Hannes Sakulin^{1,**}, Christoph Schwick¹, Dainius Šimelevičius^{1,5}, Polyneikis Tzanis¹, Cristina Vazquez Velez¹, Petr Žejdl¹, Yousen Zhang², and Dominika Zogatova¹*

¹CERN, Geneva, Switzerland

²Rice University, Houston, Texas, USA

³UCSD, San Diego, California, USA

⁴MIT, Cambridge, Massachusetts, USA

⁵Vilnius University, Vilnius, Lithuania

Abstract. The data acquisition (DAQ) of the Compact Muon Solenoid (CMS) experiment at CERN, collects data for events accepted by the Level-1 Trigger from the different detector systems and assembles them in an event builder prior to making them available for further selection in the High Level Trigger, and finally storing the selected events for offline analysis. In addition to the central DAQ providing global acquisition functionality, several separate, so-called “MiniDAQ” setups allow operating independent data acquisition runs using an arbitrary subset of the CMS subdetectors.

During Run 2 of the LHC, MiniDAQ setups were running their event builder and High Level Trigger applications on dedicated resources, separate from those used for the central DAQ. This cleanly separated MiniDAQ setups from the central DAQ system, but also meant limited throughput and a fixed number of possible MiniDAQ setups. In Run 3, MiniDAQ-3 setups share production resources with the new central DAQ system, allowing each setup to operate at the maximum Level-1 rate thanks to the reuse of the resources and network bandwidth. Configuration management tools had to be significantly extended to support the synchronization of the DAQ configurations needed for the various setups.

We report on the new configuration management features and on the first year of operational experience with the new MiniDAQ-3 system.

*e-mail: philipp.brummer@cern.ch

**e-mail: hannes.sakulin@cern.ch

1 Introduction

The Compact Muon Solenoid (CMS) is one of four main experiments at the Large Hadron Collider (LHC) at CERN [1]. The CMS experiment performs its online event selection using two trigger levels: the Level-1 trigger (L1T), implemented in custom electronics, which selects approximately 100 kHz of events, and the High Level Trigger (HLT), running on commercial computer nodes. The HLT processes fully assembled events, and selects around 2 kHz of events for persistent storage. The DAQ system needs to read out approximately 760 detector backend boards at a rate of around 100 kHz and perform event building with a throughput of about 100 GB/s.

1.1 CMS Data Acquisition for Run 3

A simplified overview of the CMS central DAQ during Run 3 [2] can be found in Figure 1. Data flow from top to bottom, starting at the detector backend boards called FEDs (FrontEnd Drivers). They send data to the FRL/FEROL boards (Frontend Readout (Optical) Link) using a custom protocol. FRLs then send data from one or multiple FEDs over the FED Builder switch to the Readout Units (RU) of the event builder using TCP/IP. FEDs are grouped into so-called FED Builders, which represent groups of readout links that send data to the same RU of the event builder. Each RU shares a physical machine with a Builder Unit (BU). The physical machine they share is referred to as a RUBU. In every run, there is one RU that takes the role of the Event Manager (EVM), steering the event building process. It receives the readout link from a partition manager in the Trigger Control and Distribution System (TCDS), which provides the reference event count. RUs can send data to any BU over the core event builder switch. Once events are built, they are made available over the event backbone network to the Filter Units (FU), machines running the HLT software on CPUs and GPUs. In the DAQ for Run 3, these FUs are grouped into so-called FU Groups, with each group consisting of 3-4 FUs which process data from the same BU.

1.2 DAQ-3 Event Builder

The DAQ-3 event builder consists of multiple RUs and BUs. Each RU can send data to any BU, as instructed by the EVM. Following the control and data flow shown in Figure 2, the event building includes the following basic steps:

1. when BUs have free event building capacity, they inform the EVM;
2. the EVM then instructs all RUs to send data from the same event to a BU with free event building capacity;
3. once all parts of the event arrived, the FUs in the FU Group connected to the BU can process it.

The event building protocol further contains optimizations, such as processing of events in batches and advanced load balancing features.

2 MiniDAQ

During CMS data taking, subsystems are in the global run with the central DAQ. Between LHC fills, so-called MiniDAQs allow subsystems to run with the full DAQ chain outside of the global run. This full chain includes event building, the HLT and transfer to storage. MiniDAQ setups are commonly used by subsystems for calibration runs as well as integration

Subsystems

CMS subdetectors

FEDs - FrontEnd Drivers ~750x
detector backend boards

FRLs - Readout Links ~500x
front-end readout (optical) links
FRL/FEROL/FEROL40
receive SLINK from 1-4 FEDs
each, send TCP/IP at 10 Gbps

FED Builders - 57x
groups of readout links going to
the same RU/EVM
typically one or more per trigger
partition

RUBUs - 57x
machines running event builder
applications
one Event Manager (EVM)
Readout Units (RU)
Builder Units (BU)

100-120 kHz event building rate

AMD EPYC Rome
7502P
32 Cores
512 GB RAM

FU Groups - 57x
Filter Units - 200x

3-4 Filter Units per group
running High Level Trigger
Dual AMD Milan 7633
64 Cores
256 GB RAM
2x T4 GPUs

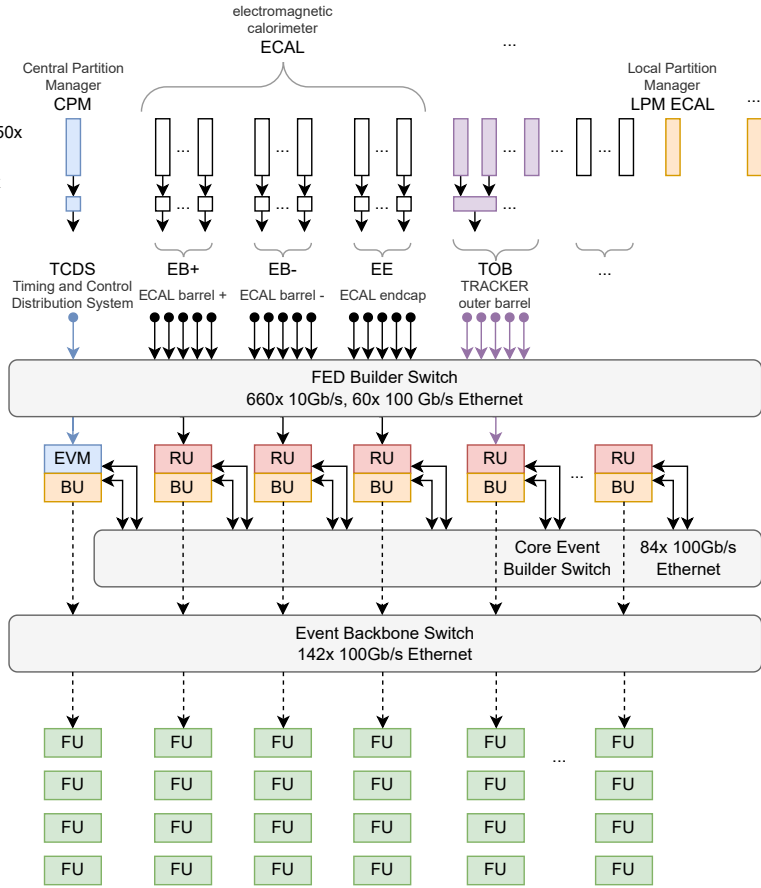


Figure 1: Overview of the CMS Central DAQ during Run 3

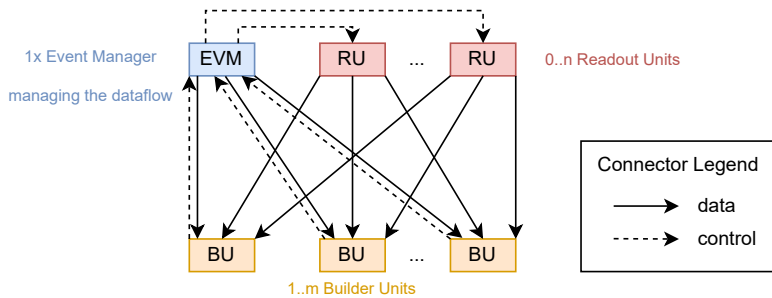


Figure 2: Overview of data and control flow in the DAQ-3 Event Builder

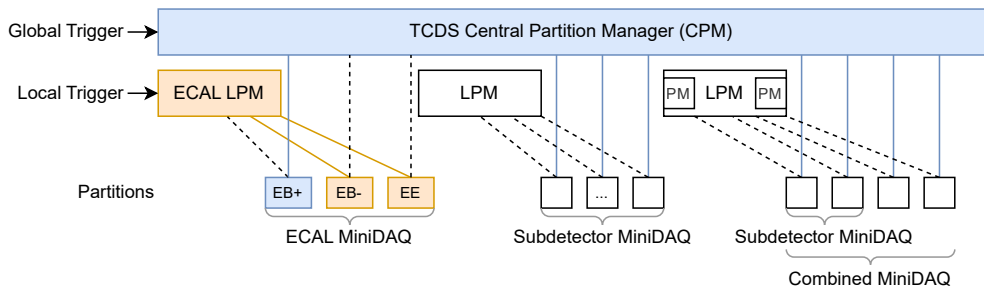


Figure 3: TCDS Central and Local Partition Managers - In the example shown, the EB+ partition is controlled by the CPM (i.e. enabled in the global run) and EB- and EE partitions are controlled by the ECAL LPM (i.e. enabled in the ECAL MiniDAQ)

tests with the DAQ and HLT. On the trigger control side, FEDs are organized into detector partitions. As illustrated in Figure 3, detector partitions can be directly controlled by the TCDS Central Partition Manager (CPM) for global data taking, using the Global Level-1 Trigger as a trigger input. Detector partitions are also associated to a Local Partition Manager (LPM) that typically controls partitions of a single sub-detector. Groups of partitions associated to an LPM can be independently controlled by one of two partition manager modules inside the LPM. In this case a local trigger input to the LPM is used. As long as there is no overlap in enabled FED Builders or detector partitions, MiniDAQs can run in parallel to the central DAQ and to each other. Separate MiniDAQ setups are provided for most of the CMS subdetectors. The readout link of the corresponding LPM module is used as an input to the the MiniDAQ's EVM.

3 MiniDAQs for Run 2

Run 2 of the LHC took place from 2015 to 2018. MiniDAQs were provided for most CMS subdetectors, for a total of ten MiniDAQ setups. During Run 2, MiniDAQ setups were using resources separate from that of the central DAQ, with each MiniDAQ having one RU, one BU and one FU machine. Unlike during Run 3, RUs and BUs did not share physical machines. The limited resources led to MiniDAQs having limited event building capacity and HLT performance compared to the global system, where most subdetectors span multiple RUs and all BUs and FUs were available. With these constraints, during Run 2, most subdetectors were unable to take data at the full event rate in their MiniDAQ. As resources were separate from those of the central DAQ, the synchronization effort between the global and MiniDAQ setups was minimal, however, additional machines were needed for every setup.

Figure 4 illustrates the differences in available resources between the central DAQ and MiniDAQ for the Electromagnetic Calorimeter (ECAL) subdetector of CMS. In the central DAQ (Figure 4a), ECAL readout links are split into the ECAL barrel at the positive (EB+) and negative (EB-) ends of the detector, and the ECAL endcap (EE). The ECAL FED Builders are spread over three RUs and the DAQ has many BUs and FUs available. In contrast, the ECAL MiniDAQ (Figure 4b), has only one RU available, which in this case also takes the role of the EVM. This means that there can only be one FED Builder which consists of all ECAL readout links.

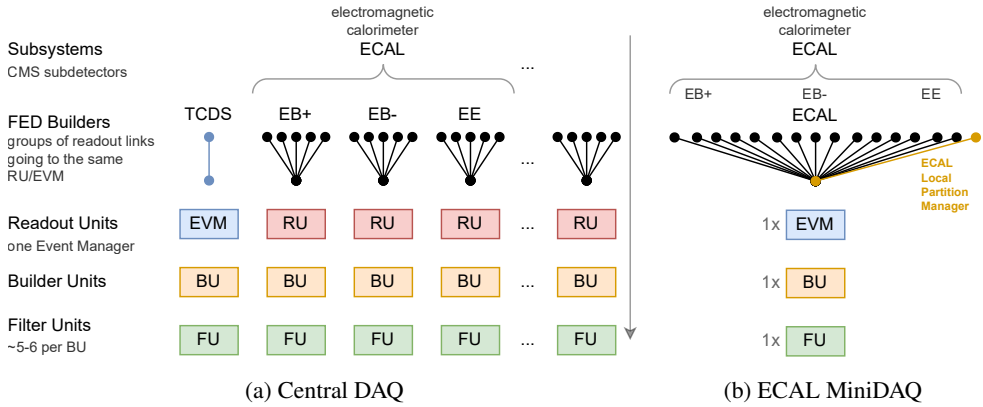


Figure 4: Central DAQ and ECAL MiniDAQ during Run 2

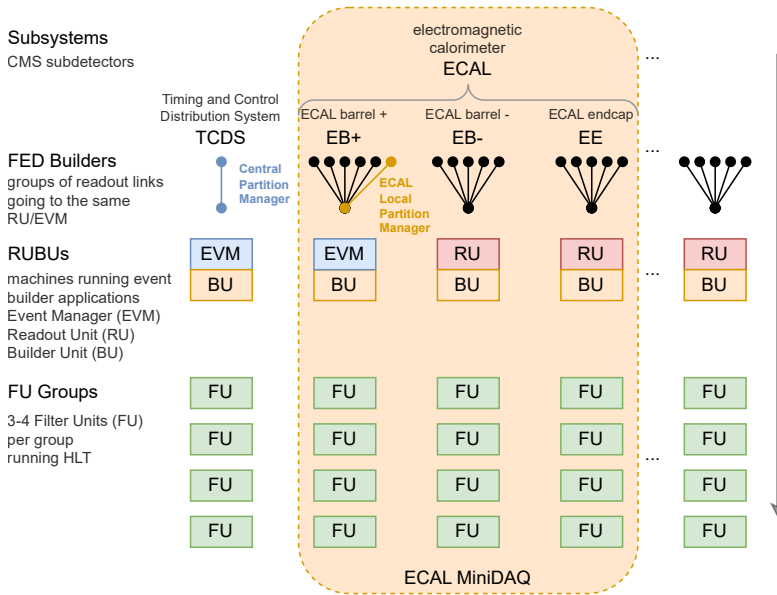


Figure 5: Central DAQ-3 and ECAL MiniDAQ-3 with shared resources

4 MiniDAQs for Run 3

Run 3 of the LHC started in June of 2022. In the CMS DAQ for Run 3, MiniDAQ setups share event building and HLT resources with the central DAQ. To facilitate this, FED Builders, RUs, BUs and FU Groups (3-4 FU machines each) are assigned to physical RUBU machines. A physical RUBU machine has a RU and BU application assigned in a way that the RU receives data from a specific FED Builder and the BU is connected to a specific FU Group. When a FED Builder is masked out from the global data taking, the assigned DAQ resources (RUs, BUs, FUs) are masked out from the central DAQ as well. These resources are then available to be used in the subdetector's MiniDAQ.

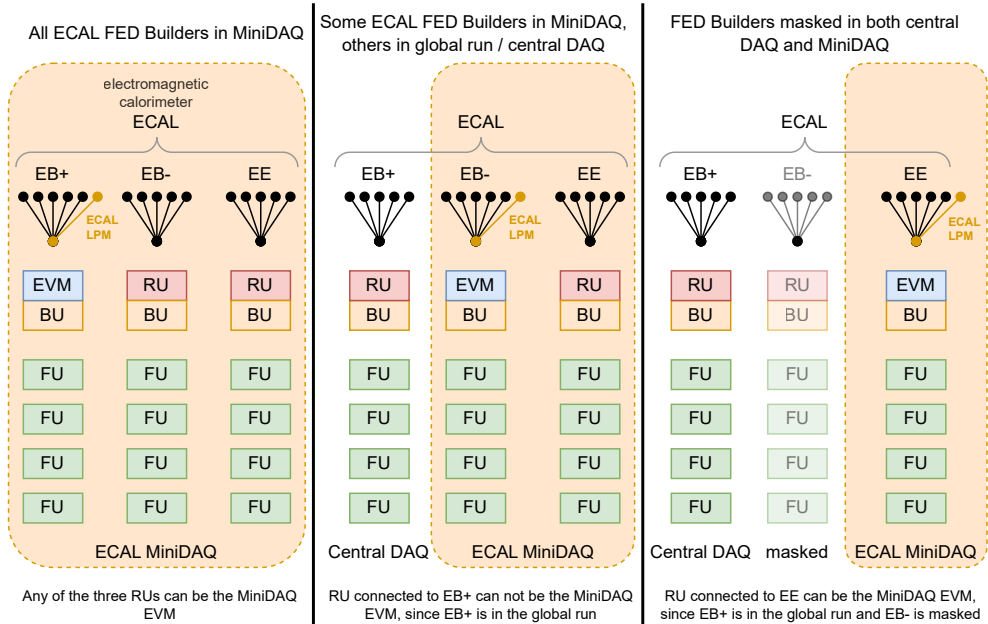


Figure 6: MiniDAQ-3: Event Manager Assignment

In Figure 5, an example for the ECAL subdetector is shown. The EB+, EB- and EE FED Builders are removed from the global data taking and can instead be used in the ECAL MiniDAQ, with event building and the HLT using the same DAQ resources when active in either setup. Every setup requires an EVM to coordinate the event building. In the global run, the EVM is the RU that reads out the CPM. It is included in every run. In a MiniDAQ, the EVM reads out the corresponding LPM in addition to a FED Builder.

4.1 Event Manager Assignment

While MiniDAQs for Run 2 had only one RU per setup, so only one machine to place the EVM on, in Run 3, MiniDAQs typically span multiple RUs. However, not all FED Builders of a subdetector need to be enabled in a MiniDAQ at the same time: typically, MiniDAQs span multiple detector partitions that can be operated independently, so that that some FED Builders may remain in the global run. Also, FED Builders may simply be disabled. The MiniDAQ's EVM then needs to be moved to one of the active RUBU machines, as shown in Figure 6. This means that masking a FED Builder in a MiniDAQ might require a change in the MiniDAQ's configuration, leading to a synchronization effort that did not exist during Run 2.

4.2 Sharing Resources with the Production System

Sharing resources between the central DAQ and MiniDAQ setups requires synchronization between the setups. This subsection focuses on the implemented mechanisms to enable concurrent independent data-taking.

Run Control

The CMS Run Control [3] is used by operators to control the data-taking of the experiment. It is based on a hierarchical tree structure of functional nodes called Function Managers (FMs). Parent nodes control the lifecycle of child nodes and send them control commands along with configuration data. Child nodes send notifications to their parent nodes, informing them of state changes and error conditions.

The operator controls the hierarchy using the web interface of the top-level node, the so-called Level Zero FM. Nodes on the first level represent the various subsystems of CMS, among them TCDS, DAQ and subdetectors like ECAL. The DAQ node configures and controls the DAQ resources, including the event builder (EVM, RUs, BUs) and the readout links.

The operator can use the Level Zero FM's web interface to mask readout links between runs. If all readout links in a FED Builder are masked out, the associated RU and BU are masked out as well and can instead be configured in a MiniDAQ setup. Each MiniDAQ setup has their own Run Control hierarchy and control interface.

Resource Assignment

As described in section 4, RUs and BUs are assigned to physical RUBU machines. This assignment is stored and versioned in a database and can be modified by experts using a tool called the DAQ Configurator [4]. The assignment is commonly modified when a physical RUBU machine needs to be replaced with a hot spare. The DAQ Configurator follows an assignment algorithm designed to minimize changes to unaffected parts of the system, in such a way as to not disturb ongoing independent runs. An alternative assignment algorithm for test runs is available. Additionally, an editor is provided for experts to modify generated resource assignments or create custom ones. After changes to the assignment, the DAQ Configurator is used to generate a new DAQ configuration for the central DAQ. This configuration is stored in a different database, the resource service, where it is loaded from by the Run Control. Each MiniDAQ has its own resource service and DAQ configuration, however, all of them follow the assignment that is used to create the central DAQ's configuration.

Resource Locking

In order to avoid the central DAQ and MiniDAQs using the same resources, locking was implemented. Resource locking happens when a resource is configured in a setup and is implemented for multiple resource types. Like in Run 2, the FRL hardware is locked on configuration. In Run 3, locking of the RU, EVM and BU applications was added. As a consequence, if a FED Builder is enabled (at least one readout link is not masked) and configured in the central DAQ, it can not be configured in a MiniDAQ at the same time and the other way around. If a resource is attempted to be configured while it is already locked, the Run Control operator is informed of the conflict in the Level Zero FM web interface.

Concurrent Independent Data-Taking

Resources that are not locked in one setup can be configured in another. As long as there is no overlap in enabled FED Builders and detector partitions, the central DAQ and MiniDAQs can be configured at the same time, allowing for concurrent independent data-taking on shared resources.

4.3 Tooling for Automated Configuration

Keeping the resource assignment of all nine MiniDAQ-3 setups synchronized with that of the central DAQ would require a lot of effort if done manually. Instead, an automated procedure

was implemented and works as explained by the following example. For illustration, we assume that the physical RUBU machine receiving data from the ECAL EB+ FED Builder becomes dysfunctional. Since the data from this FED Builder still needs to be read out, the RU and BU on the RUBU need to be moved to a spare machine. To do this, a DAQ expert disables the broken machine using the DAQ Configurator tool, which places the RU and BU applications connected to EB+ on a spare machine and updates the assignment database, before creating a new DAQ configuration for the central DAQ. However, since the EB+ FED Builder is also used in the ECAL MiniDAQ, the MiniDAQ configuration needs to be updated to use the changed resource assignment. Instead of requiring an expert to do this manually, the MiniDAQ Run Control updates the MiniDAQ configuration automatically, by calling the DAQ Configurator tool when configuring the DAQ Run Control node. The tool then uses the latest assignment to create a new MiniDAQ configuration compatible with that of the central DAQ. This automated compatibility check also covers other cases where a different MiniDAQ configuration is needed, for example when a MiniDAQ's EVM needs to be moved to a different RUBU machine because a FED Builder was masked (see section 4.1).

4.4 Operational Experience

MiniDAQ-3 was deployed at the start of Run 3 of the LHC in June of 2022 and has since been used by subdetectors for calibration and test runs. Nine MiniDAQ-3 setups were deployed, with between one and ten FED Builders in each setup, meaning up to ten different EVM placements and DAQ Configuration variations for the same resource assignment. As intended, the automated configuration updates resulted in a greatly reduced workload for DAQ experts. One side effect of the resource sharing has been observed during high-rate DAQ tests in inter-fills with only very few FED Builders included in the global run. As the HLT capacity is reduced proportional to the number of included FED Builders, back-pressure from the DAQ may be observed due to limited HLT resources when running with a full HLT menu. This effect is not observed in the MiniDAQ setups, where a simplified HLT menu is used. To counter this effect, such high rate tests are now performed with additional FED Builders (and thus BUs and FUs) included from one of the larger sub-detectors.

4.5 Summary

MiniDAQs allow CMS subdetectors to take independent runs with the full DAQ chain. With MiniDAQ-3, these setups share event building and HLT resources with the global DAQ, meaning that data taking is possible at a much higher rate compared to what was possible during Run 2. In many cases, the achievable event rate in MiniDAQ-3 setups is comparable to that in the central DAQ. Additionally, sharing resources saves costs, as no additional machines are needed for the MiniDAQ setups. The added complexity due to the need to synchronize the shared resources is covered by the automated configuration update procedures.

References

- [1] S. Chatrchyan et al. (CMS), *JINST* **3**, S08004 (2008)
- [2] The CMS Collaboration, *Development of the CMS detector for the CERN LHC Run 3* (2023), CERN-EP-2023-136, submitted to *JINST*
- [3] G. Bauer et al., *IEEE Transactions on Nuclear Science* **59**, 1597 (2012)
- [4] G. Bauer et al., *J. Phys. Conf. Ser.* **219**, 022003 (2010)