EPJ Web of Conferences **337**, 01136 (2025) https://doi.org/10.1051/epjconf/202533701136
CHEP 2024

Evaluating a File-Based Event Builder to enhance the Data Acquisition in the CMS Experiment

Jaafar Alawieh[1], *Kareen* Arutjunjan[1], *Miguel* Bacharov Durasov[1], *Ulf* Behrens[2], *Andrea* Bocci[1], *James* Branson[3], *Philipp Maximilian* Brummer[1], *Jan Andrzej* Bugajski[1], *Eric* Cano[1], *Sergio* Cittolin[3], *Albert* Corominas I Mariscot[1], *Georgiana-Lavinia* Darlea[4], *Christian* Deldicque[1], *Marc* Dobson[1], *Antonin* Dvorak[1], *Christos* Emmanouil[1], *Antra* Gaile[1], *Dominique* Gigi[1], *Frank* Glege[1], *Guillelmo* Gomez-Ceballos[4], *Patrycja* Gorniak[1], *Jeroen* Hegeman[1], *Guillermo* Izquierdo Moreno[1], *Thomas Owen* James[1], *Tejeswini* Jayakumar[1], *Wassef* Karimeh[1], *Rafał* Krawczyk[2], *Wei* Li[2], *Kenneth* Long[4], *Frans* Meijers[1], *Emilio* Meschi[1], *Srećko* Morović[3], *Babatunde John* Odetayo[1,5], *Luciano* Orsini[1], *Christoph* Paus[4], *Andrea* Petrucci[3,*], *Marco* Pieri[3], *Dinyar Sebastian* Rabady[1], *Attila* Racz[1], *Theodoros* Rizopoulos[1], *Hannes* Sakulin[1], *Christoph* Schwick[1], *Dainius* Šimelevičius[1,6], *Polyneikis* Tzanis[1], *Cristina* Vazquez Velez[1], and *Petr* Žejdl[1]

[1]CERN, Geneva, Switzerland
[2]Rice University, Houston, Texas, USA
[3]UCSD, San Diego, California, USA
[4]MIT, Cambridge, Massachusetts, USA
[5]University of Benin, Benin City, Nigeria
[6]Vilnius University, Vilnius, Lithuania

Abstract. The event builder in the Data Acquisition System (DAQ) of the CMS experiment at the CERN Large Hadron Collider (LHC) is responsible for assembling events at a rate of 100 kHz during the current LHC run 3, and up to750 kHz for the upcoming High Luminosity LHC, scheduled to start in 2029. Both the current and future DAQ architectures leverage on state-of-the-art network technologies, employing Ethernet switches capable of supporting RDMA over Converged Ethernet (RoCE) protocols. The DAQ Front-end hardware is custom-designed, utilizing a reduced TCP/IP protocol implemented in FPGA for reliable data transport between custom electronics and commercial computing hardware. An alternative architecture for the event builder, known as the Super-Fragment Builder (SFB), is under evaluation. The SFB comprises two separate systems: the Super-Fragment Builder and the File-based Filter Farm (F3). A super-fragment consists of the event data read by one or more Front-End Drivers and corresponding to the same L1 accept, and the SFB constructs multiple super-fragments corresponding to the number of Read-Unit (RU) machines in the DAQ system, storing them in local RAM disks. Subsequently, the F3 accesses super-fragments from all RU machines via the Network File System (NFS) over Ethernet and builds complete events within the High Level Trigger process. This paper describes the first prototype of the SFB and presents preliminary performance results obtained within the DAQ system for LHC Run 3.

*Corresponding author e-mail: Andrea.Petrucci@cern.ch

1 Introduction

At the Large Hadron Collider (LHC) [1] at CERN, one of the four major experiments is the Compact Muon Solenoid (CMS) [2]. Since 2022, CMS has been operating during LHC Run 3 at a center-of-mass energy of 13.6 TeV with a luminosity of $2\times10^{34}\text{cm}^{-2}\text{s}^{-1}$. LHC Run 3 is expected to continue until the middle of 2026, after the CMS experiment will undergo Phase-2 upgrades in preparation for the High Luminosity LHC (HL-LHC). In this phase, the collider will reach a center-of-mass energy of 14 TeV and a luminosity of $7.5\times10^{34}\text{cm}^{-2}\text{s}^{-1}$. In addition to the significant changes in the sub-detectors, the data acquisition system (DAQ) [3] will undergo a substantial upgrade to meet the new requirements of Phase-2 of CMS to accommodate the increased event size, Level-1 (L1) trigger accept rate, and High-Level Trigger (HLT). For example, these parameters will change from 2 MB, 100 kHz, and 2 kHz, respectively, to an event size of 8.4 MB, an L1 trigger accept rate of 750 kHz, and an HLT accept rate of 7.5 kHz in LHC Run 5.

This work contains preliminary studies on possible event-building architectures, a core component of the DAQ system. The event builder is responsible for assembling events, which consist of data representing a complete, detector-wide snapshot of the CMS response to a proton-proton collision. For each event accepted by the Level-1 (L1) trigger, data fragments from all sub-detectors are transmitted to the event builder, where they are combined into fully assembled events for further processing in the High-Level Trigger (HLT). To accommodate the increased data rate and size expected in future operations, a new data aggregation approach is being considered. Rather than constructing individual events, an alternative orbit-based structure is proposed. This method allows multiple events to be grouped within a single LHC orbit, corresponding to a full revolution of the LHC beams, optimizing data handling and improving scalability under the upgraded conditions.

The rest of the paper is organized as follows. Section 2 details the architecture of the DAQ system for the HL-LHC. Section 3 discusses possible alternatives for the event builder compared to the original design. Section 4 presents one of these alternative solutions, evaluated using the DAQ system of LHC Run 3. Finally, Section 5 concludes the paper and outlines future developments.

2 The CMS DAQ architecture for the HL-LHC

The CMS DAQ architecture in the CMS Phase-2 for HL-LHC is shown Figure 1 as defined in the *Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger Technical Design Report* [3]. The readout channels from the Detector Front-Ends are grouped into approximately 850 Data-to-Surface (D2S) output channels by custom DAQ boards. The Advanced Telecommunications Computing Architecture (ATCA) standard offers more space on a single board. When paired with the capabilities of modern FPGA families, it enables the integration of DAQ data aggregation from back-end boards with the Trigger, Timing, and Control (TTC) and Trigger Throttling System (TTS) functionality into a single custom board. This board, called the DAQ, Trigger, and Timing Hub (DTH), will be installed in the sub-detector back-end crates. The DAQ-800 is another DAQ ATCA board, built to support sub-detectors with higher throughput per crate. It does not include timing functionality but provides double the input connectivity. These boards deliver orbit fragments to the DAQ readout servers over separate TCP/IP streams for each D2S output channel. An orbit fragment is a collection of fragments from a single LHC orbit, averaging 67 fragments per orbit. They correspond to bunch crossings accepted by the L1 trigger within the orbit. The DTH board has five QSFP ports at 100 Gb/s connected to the Data to Surface Concentration Network (DCN), whereas

the DAQ-800 board has ten QSFP ports at 100 Gb/s. The RU/BU machines are linked to the DCN via 400 Gb/s Ethernet connection.

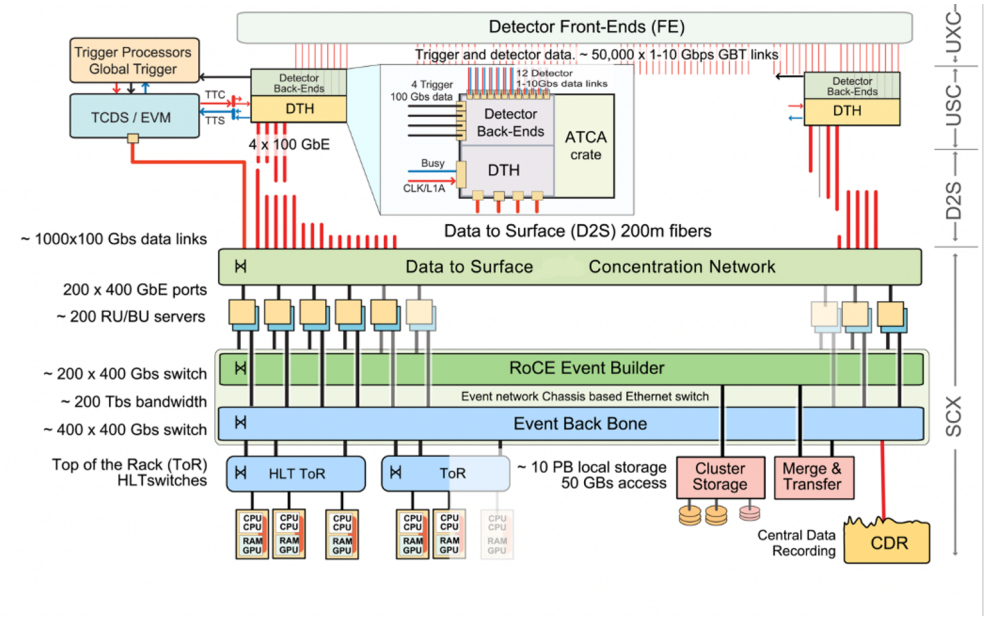

Figure 1. CMS DAQ architecture for the HL-LHC [4].

In this paper, the term *event* is used in a general sense and can mean an individual Level 1 Accept (L1A) or an aggregation of the data corresponding to several L1As. The CMS Event Builder (EVB) assembles the complete orbit by collecting all orbit fragments from CMS for a given LHC orbit. The EVB consists of three applications: Readout Unit (RU), Builder Unit (BU), and Event Manager (EVM). The RU application has two main tasks: it reads all assigned D2S output channel streams and assembles an orbit super-fragment as instructed by the EVM. The RU application then transmits the orbit super-fragments to specific BU applications as requested by the EVM. The BU application is responsible for two key tasks: it reconstructs the complete orbit by assembling all orbit super-fragments received from the RUs. Then, it writes the data to a RAM disk. The EVM assigns orbits to BU applications based on BU requests and orbit counting from the output channel of the Timing and Control Distribution System (TCDS). This mapping is forwarded to all RU applications. Each RU/BU server connects to this network using a 400 Gb/s Ethernet link with RDMA over Converged Ethernet (RoCE) v2 protocol. The RoCE Event Builder Network facilitates data exchange and control messages within the EVB.

The interface between the Event Builder and the High Level Trigger (HLT) is file-based, using the File-based Filter Farm (F3) system. Orbits are read by HLT processes on the Filter Unit (FU) servers using the Event Backbone and HLT Top-of-Rack switches. The transfer relies on Network File System (NFS) mounting of RAM disks from the RU/BU servers. Orbits are unpacked into single L1A events and HLT runs reconstruction and filtering based on physics signatures saving a fraction of event data. Selected events are stored back in the RAM disks of the RU/BU nodes before the Storage Manager system transfers them to the Lustre distributed file system. Finally, the data is sent to persistent storage outside the CMS online infrastructure for offline analysis.

3 Alternative Architectures for CMS Event Builder

Figure 2 shows the data flow of the CMS DAQ EVB system up to the filter farm nodes and the relevant networks. The process consists of four key steps where data is aggregated or transferred between nodes:

1. **Generating orbit super-fragments**: D2S output channel TCP/IP streams are read by the RU/BU machines via the D2S Concentrator Network to construct orbit super-fragments.
2. **Building orbits**: Orbit super-fragments are transferred between RU/BU nodes over the RoCE EVB network to assemble complete orbits.
3. **Reading orbits**: Filter Units access the stored orbits in the RU/BU machines using the Event Backbone network.
4. **Building events**: Filter farm processes extract events from the assembled orbits before running physics algorithms to select relevant events.

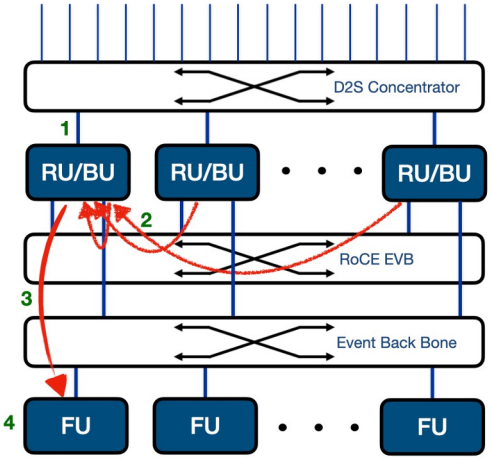

Figure 2. This diagram illustrates the data flow of the CMS DAQ system up to the filter farm nodes.

As described above, the main goal is to make the orbit available in the FU nodes. In the architecture described in DAQ TDR[4], the orbit is assembled in the RU/BU machines and then transferred to the FU servers. However, if the orbit were built directly in the FU nodes, the intermediate step of transferring super-fragments between RU/BU machines would be removed. The following are potential benefits of building orbits within the filter farm process:

- **Eliminating the RoCE EVB network** reduces costs associated with network line cards and NICs, simplifying network configuration, monitoring, and maintenance.
- **Storing complete orbits in local HLT process memory on filter farm PCs** reduces the total number of memory copies in the data flow, improving efficiency. "
- **Reduced memory and I/O performance demands** for the final orbit-building, as the workload is distributed across a larger number of HLT nodes."

The DTH400 and DAQ-800 boards are designed to perform critical data consistency checks (e.g. detection of event ID mismatch) and this simplifies the event builder code by requiring fewer tasks.

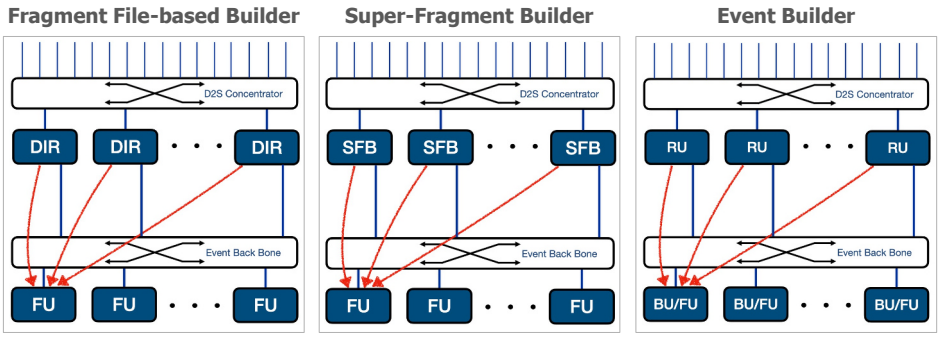

Figure 3. These tree diagrams present the architecture options for the CMS DAQ in the HL-LHC.

Figure 3 presents three different approaches for the CMS event builder in the HL-LHC, where orbits are constructed in the FU nodes. Each option reflects a different distribution of event building tasks between the EVB applications and the filter farm application. The first option delegates most tasks to the filter farm application, while the third option follows the canonical event builder model as describe in the TDR, but with the BU application running on the FU node. the second Option represents an intermediate approach, balancing responsibilities between the two components.

3.1 Fragment File-based Builder

In the Fragment File-based Builder, the DTH Input Receiver (DIR) running on the RU/BU node reads multiple TCP D2S output streams and writes a separate file into RAM disk for each stream. Therefore, approximately 850 files for the same orbit range are distributed across approximately 200 DIR machines. The FUs in the HLT cluster then access these files via NFS, parsing the full orbit data to assemble individual events. This approach is expected to have significant file handling and network I/O overhead due to the large number of files being read from distributed RAM disks.

3.2 Super-fragment Builder

In the Super-Fragment Builder (SFB), the first stage of event aggregation in the EVB is preserved, while the second stage is delegated to the filter farm application. The SFB applications construct multiple orbit super-fragments, with the number of super-fragments corresponding to the number of SFB machines in the DAQ system. These super-fragments are stored in local RAM disks. The HLT filter farm nodes then retrieve approximately 200 orbit super-fragment files per orbit range, significantly reducing file I/O operations compared to the Fragment File-based approach. This aggregation at the SFB level improves data transfer efficiency and event reconstruction.

3.3 Event Builder

The Event Builder approach eliminates the need for NFS file access in the DAQ system by fully constructing the orbit within the FU nodes. The RU applications assemble orbit super-fragments, which are then sent to the Build Units (BUs) running alongside the filter farm processes for final orbit assembly. The fully built orbits are stored in RAM disks, enabling the HLT filter farm to read complete orbits directly from local memory. This reduces network traffic, increases processing speed, and minimizes memory copies but requires configuring RoCE on the Event Backbone and HLT top-of-rack switches.

4 Evaluating the Super-Fragment Builder in DAQ for LHC Run 3

The section 3 describes alternative architectures for the CMS event builder in HL-LHC. The next step is to prototype these options to evaluate their feasibility. Currently, the only avail-

able hardware for prototyping is the CMS DAQ system used during LHC Run 3[5]. The *Fragment File-based Builder* option can not be tested on the current hardware because the custom DAQ hardware lacks the necessary checks required for the correct operation of the DAQ system, which are implemented in the current event builder. The *Event Builder* approach, which runs the BU application on the FU machine, can also not be tested in the current DAQ system during the year-end technical stop. The main challenge is configuring of the RoCE network in the Event Backbone and HLT top-of-rack switches. The existing RoCE implementation in the EVB network operates with a single hop, while bringing RoCE to the FU nodes would introduce two network hops, requiring further study. The only feasible option for testing in the current DAQ system is the *Super-Fragment Builder*, as it can run on the LHC Run 3 hardware with modifications to the existing EVB and filter farm software.

4.1 Super-fragment Builder Architecture for LHC Run 3

The Super-Fragment Builder (SFB) constructs multiple super-fragments corresponding to the number of SFB machines in the DAQ system and stores them in local RAM disks. An event super-fragment consists of data read from one or more Front-Ends corresponding to the same L1 accept. The SFB is composed of three applications: Readout Unit (RU), Super-Fragment Worker (SFW), and Event Manager (EVM). The RU application reads data from one or more FEDs via TCP streams and assembles the super-fragments. The SFW receives these super-fragments from the RU application and writes them to the RAM disk. The EVM schedules SFB operations based on requests from SFWs and event counting in the TCDS FED. Figure 4 shows the SFB protocol between SFB applications. In the first step,

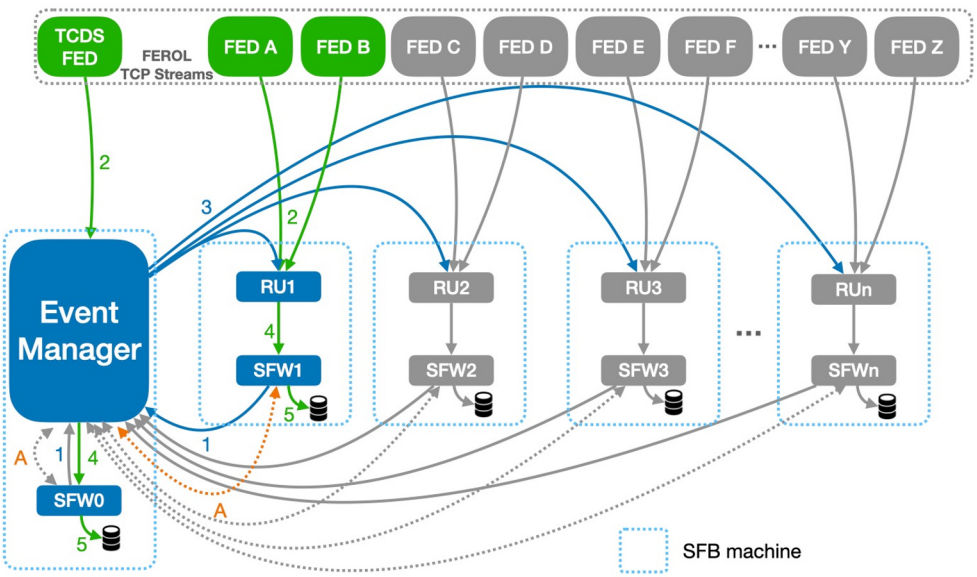

Figure 4. This drawing describes the SFB protocol between SFB applications.

the SFW applications request N resources from the EVM application to process the super-fragments. While the SFB is running, the event fragments are asynchronously transmitted from FEROLs to EVM and RUs (step 2). Once the EVM collects enough event fragments from the TCDS FED to fulfill a request for N resources from all SFWs, the step 3 is triggered. In this step, the EVM sends messages to each RU that includes the L1A number of the events, the SFW ID, and the RU ID. In step 4, the EVM and RUs construct the super-fragments for the requested events and send them to their respective SFW. Upon receiving a super-fragment,

the SFW writes it to the RAM disk (step 5). The SFWs can retrieve the number of events per lumisection from the EVM at anytime (step A). Figure 5 presents the event-building

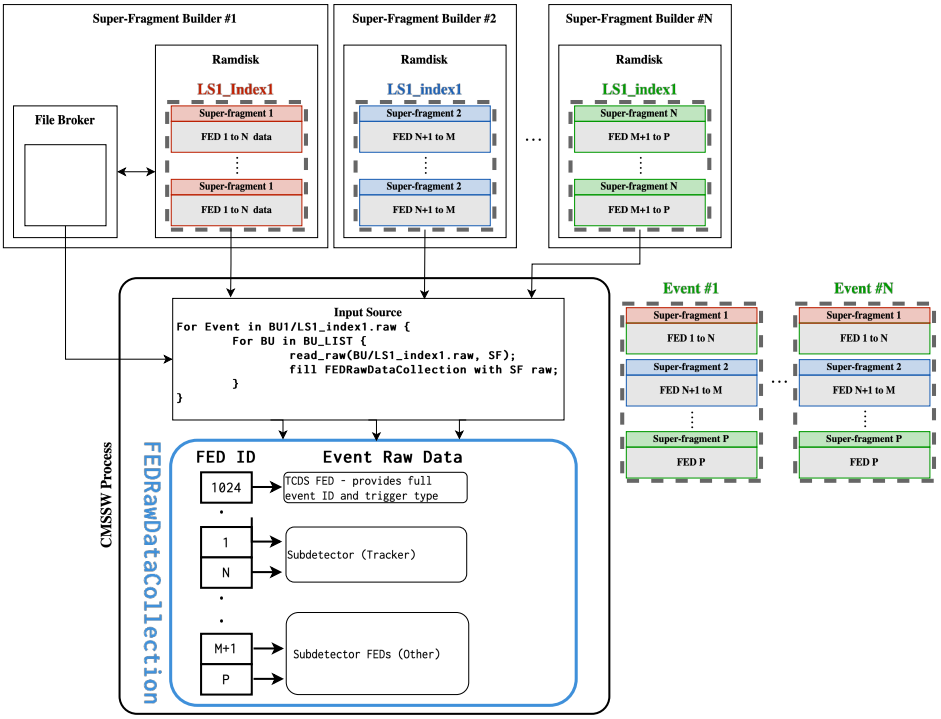

Figure 5. This diagram draws the architecture how the filter farm process constructs the event.

architecture in the filter farm process using super-fragments from the SFB machines. The File Broker application is responsible for assigning super-fragment stream files from all SFBs, within the same event range, to a filter farm application (CMSSW process). After combining all fragments into fully build events, the Input Source of the CMSSW process unpacks them into single-event raw data further used by the HLT event processing, and writes bookkeeping metadata. The Output Module of the CMSSW process writes output streams and metadata to the local RAM disk. In the FU node, the High-Level Trigger Daemon (HLTD) merges the output streams from the filter farm processes and transfers them to the Storage and Transfer System (STS). The HLTD monitors and schedules CMSSW processes in the FU node.

4.2 Preliminary test of SFB with DAQ LHC run 3 Emulator Runs

The DAQ LHC Run-3 software (EVB/CMSSW modules) was modified to implement the SFB. To test the SFB, the DAQ Emulator was used to generate events in the DAQ hardware, simulating fragment sizes like those produced by each sub-system's front ends, targeting a pile-up of 65 in CMS. Several optimizations were applied to the SFB machines, including memory and core pinning for super-fragment building, as well as NFS tuning. Multiple DAQ Emulator runs were performed, including long-duration tests. For instance, an overnight run of 10 hours with an average event size of 1924 kB and an HLT output of approximately 10 GB/s, the SFB maintained the required L1A rate after deadtime at 114 kHz, with a negligible DAQ TTS deadtime of 0.163%. This confirms that the SFB can operate under real conditions.

Another key aspect of the test was evaluating NFS performance over TCP/IP for reading out super-fragments from HLT processes in an all-to-all connection between 58 SFB machines and 216 FU nodes. Figure 6 presents the throughput of an SFB node with the highest

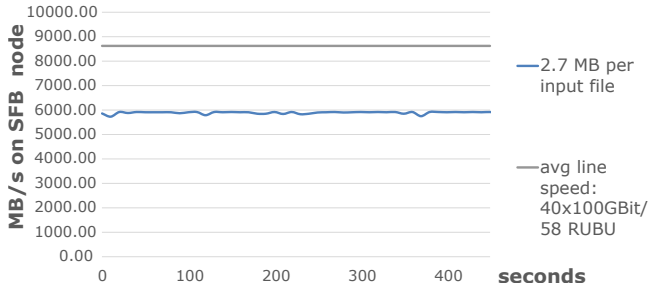

Figure 6. This diagram shows the throughput on SFB node with the highest super-fragment size in the SFB DAQ Emulator test with 58 SFBs x 216 FUs.

super-fragment building scenario of the SFB DAQ Emulator test, with 58 SFBs connected to 216 FUs. The total throughput was limited by the 40 x 100 GbE (8 links per rack) top-of-rack switch uplinks which provided an bandwidth of 8620 MB/s averaged per 100 GbE uplink and the highest super-fragment throughput reached 6000 MB/s. An important finding from the test was the time required for automatically mounting all 58 SFB RAM disks on all 216 FUs. This process took approximately 1 minute and 40 seconds with the current system, without any specific improvements to the NFS server for scalability.

5 Conclusions and Future Developments

The upgrade of the CMS data acquisition system for the High-Luminosity LHC requires handling larger event sizes, higher L1 trigger accept rates, and increased high-level trigger accept rates than the current Run 3 CMS DAQ system. This paper describes the CMS DAQ team's efforts to explore alternative architectures that eliminate the need for a dedicated Event Builder network, improve resource utilization, and simplify the event builder software. The evaluation of the Super-Fragment Builder demonstrated a viable alternative to traditional event-building architectures.

Future evaluations will focus on configuring the RoCE network in the Event Backbone and HLT top-of-rack switches to enable NFS over RoCE and to run the BU application on FU nodes. During LHC Long Shutdown 3, a small-scale setup will be tested with increased network speed (400 Gb/s per link), and the three alternative DAQ system architectures proposed in this paper will be evaluated.

References

[1] L. Evans (ed.) and P. Bryant (ed.), LHC Machine, 2008 JINST 3 S08001

[2] CMS Collaboration, The CMS experiment at the CERN LHC, 2008 JINST 3 S08004

[3] CMS Collaboration, Development of the CMS detector for the CERN LHC Run 3, Journal of Instrumentation, Volume 19, May 2024.

[4] CMS Collaboration, The Phase 2 Upgrade of the CMS Data Acquisition and High Level Trigger, Technical Design Report CERN-LHCC-2021-007, CMS-TDR-022, CERN, 17 June 2021.

[5] CMS Collaboration, Development of the CMS detector for the CERN LHC Run 3, Journal of Instrumentation, Volume 19, May 2024.