Available on CMS information server

CMS CR -2023/241

The Compact Muon Solenoid Experiment

Conference Report

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

30 October 2023 (v2, 02 November 2023)

CMS Level-1 trigger Data Scouting firmware prototyping for LHC Run-3 and CMS Phase-2

Rocco Ardino for the CMS Collaboration

Abstract

A novel Data Acquisition (DAQ) system, known as Level-1 Data Scouting (L1DS), is being introduced as part of the Level-1 (L1) trigger of the CMS experiment. The L1DS system will receive the L1 intermediate primitives from the CMS Phase-2 L1 trigger on the DAQ-800 custom boards, designed for the Phase-2 central DAQ. Firmware is being developed for this purpose on the Xilinx VCU128 board, with features similar to one half of the DAQ-800, and validated in a demonstrator for LHC Run-3. This contribution describes the firmware development in view of the target design for the DAQ-800.

Presented at *TWEPP2023 Topical Workshop on Electronics for Particle Physics*

Prepared for submission to JINST

Topical Workshop on Electronics for Particle Physics 2023

CMS Level-1 trigger Data Scouting firmware prototyping for LHC Run-3 and CMS Phase-2

Rocco Ardino,[a,b,c,1] **Christian Deldicque,**[c] **Marc Dobson,**[c] **Dominique Gigi,**[c] **Sabrina Giorgetti,**[a,b] **Thomas Owen James,**[c] **Giovanna Lazzari Miotto,**[c] **Emilio Meschi,**[c] **Matteo Migliorini,**[a,b] **Giovanni Petrucciani,**[c] **Dinyar Rabady,**[c] **Attila Racz,**[c] **Hannes Sakulin,**[c] **Petr Zejdl**[c] **on behalf of the CMS collaboration**

[a] *Department of Physics and Astronomy "Galileo Galilei", Padova University, Via Marzolo 8, 35131 Padova, Italy*

[b] *National Institute for Nuclear Physics, Padova Division, Via Marzolo 8, 35131 Padova, Italy*

[c] *CERN, Esplanade des Particules 1, Meyrin, 1211, Switzerland*

E-mail: rocco.ardino@cern.ch

Abstract: A novel Data Acquisition (DAQ) system, known as Level-1 Data Scouting (L1DS), is being introduced as part of the Level-1 (L1) trigger of the CMS experiment. The L1DS system will receive the L1 intermediate primitives from the CMS Phase-2 L1 trigger on the DAQ-800 custom boards, designed for the Phase-2 central DAQ. Firmware is being developed for this purpose on the Xilinx VCU128 board, with features similar to one half of the DAQ-800, and validated in a demonstrator for LHC Run-3. This contribution describes the firmware development in view of the target design for the DAQ-800.

Keywords: Data acquisition, Trigger, Online data processing

[1] Corresponding author.

1 Introduction

The High-Luminosity LHC (HL-LHC) upgrade [1] will increase the peak instantaneous luminosity of the collider to $7.5 \cdot 10^{34}$ cm^{-2}s^{-1}, producing an average of 200 proton-proton collisions per bunch crossing (pileup). To cope with these extreme conditions, a significant detector upgrade is foreseen for the Compact Muon Solenoid (CMS) experiment [2, 3]. The upgraded detector, known as CMS Phase-2, will feature a new Level-1 (L1) trigger system [4], which will have access to an unprecedented level of information. Advanced reconstruction algorithms will be deployed directly on the L1 trigger FPGA-based processors, approaching the offline reconstruction resolution.

The CMS Phase-2 L1 trigger will also incorporate a Data Scouting system, designed to collect the reconstructed primitives via optical links from spare outputs of the L1 processing boards and perform a quasi-online analysis on them in a heterogeneous computing farm [4, 5]. This work presents the firmware development for the target scouting read out board, the DAQ-800, designed for the CMS Phase-2 central DAQ [6]. The firmware development and validation using the Xilinx VCU128 development board is discussed in the context of the LHC Run-3 demonstrator of the system.

2 The CMS Phase-2 Level-1 trigger upgrades

The CMS Phase-2 L1 trigger, shown in Figure 1, is designed to maintain or improve the CMS acceptance for all physics objects under the HL-LHC pileup conditions, at the price of an increase in accept rate from 100 to 750 kHz [4]. The Global Calorimeter and Muon Trigger (GCT and GMT) subsystems will provide higher granularity primitives, and a novel Global Track Trigger (GTT) will receive tracker tracks at 40 MHz from the tracker back-end to perform vertex fitting. A Correlator Trigger (CT) will combine the information from the upstream subsystems by applying a Particle Flow (PF) reconstruction directly in hardware. A Global Trigger (GT) system will select within a latency window of around 12 μs the collision events to be read out. This decision is based on a set of trigger paths consisting of selections or neural network inference on the reconstructed primitives from the upstream subsystems.

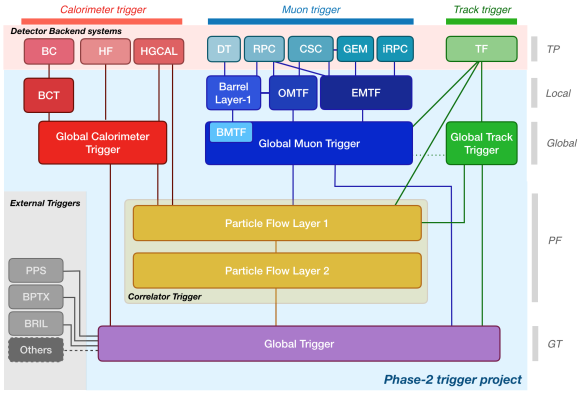

Figure 1. Architecture of the CMS Phase-2 Level-1 trigger [4]. The various levels of processing are indicated on the right: trigger primitives (TP), local and global trigger reconstruction, Particle Flow trigger reconstruction (PF), and global decision (GT).

3 Level-1 trigger Data Scouting system architecture for CMS Phase-2

The output primitives of the CMS Phase-2 L1 trigger will be collected at the full 40 MHz bunch crossing rate by the Level-1 trigger Data Scouting (L1DS) system [4, 5]. The L1DS system will provide vast amounts of data for monitoring and physics-related purposes, enabling the study of signatures too common to fit within the L1 acceptance budget or orthogonal to the standard physics trigger paths. Examples include high track multiplicity processes such as W → 3π and W → $D_s\gamma$, multiple soft jets, displaced soft leptons and long-lived charge particles spanning multiple bunch crossings. Trigger diagnostics and per-bunch luminosity measurements would also be enhanced by the data scouting approach.

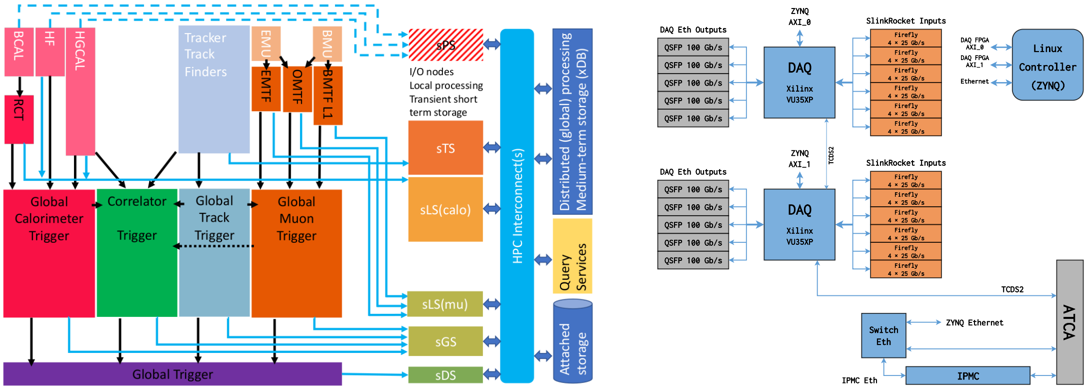

Figure 2. (Left) Architecture of the CMS Phase-2 Level-1 trigger Data Scouting system [4, 5], with the Stage 1 scouting Decision System (sDS) and scouting Global System (sGS). The system can later be extended with a scouting Local System (sLS). (Right) Simplified scheme of the DAQ-800 read out board, highlighting the two VU35P FPGAs and the input/output capabilities [6].

The L1DS planned architecture is designed to be stageable and it is shown in Figure 2 (left). In the Stage 1 scouting, the decision outputs of the Global Trigger are captured by the scouting Decision System (sDS), and the output primitives of the four global subsystems are captured by the scouting Global System (sGS). This architecture can be extended in a second stage with a scouting Local System (sLS) to include local muon and regional barrel calorimeter triggers and the endcap calorimeter primitives. The L1 primitives are collected via optical links running a custom unidirectional 65b/67b encoding protocol at 25 Gbps. The DAQ-800 board, designed for the CMS Phase-2 central DAQ read out [6], has been selected for this first step. The DAQ-800 is capable of accepting up to 48 L1 input links via FireFly connectors for a total of 1.2 Tbps. The maximum theoretical output bandwidth is 1 Tbps from 10 × 100 Gb Ethernet (GbE) links from the on-board QSFPs. Thus, a moderate data reduction will be necessary to allow the board to operate at a steady output data rate of 800 Gbps. The data pre-processing logic will be implemented on the two Xilinx Ultrascale+ VU35P FGPAs mounted on the board and chosen for their built-in High-Bandwidth Memory (HBM). This is required in order to provide sufficient data buffering between the LHC-synchronous back-end and COTS switched network that relays data to standard compute servers. The L1DS will use a custom firmware implementation of the TCP/IP protocol on the DAQ-800 board FPGAs to transfer L1 primitives to a set of data servers, where they will be stored in memory prior to being sent to a computing farm for the final processing.

The production of the DAQ-800 board is foreseen for the start of 2024, with five prototypes available in the first quarter. The development and prototyping of the needed firmware is currently being performed on the Xilinx VCU128 development kit. This board is equipped with a VU37P FPGA and it can provide the same functionality as half of a DAQ-800 board. The firmware development and validation on the VCU128 kit is one goal of the LHC Run-3 demonstrator of the L1DS.

4 Firmware prototyping in L1DS demonstrator for LHC Run-3

An L1DS demonstrator has been built for the LHC Run-3 to allow the development of the system while collecting real collision data. The demonstrator, illustrated in Figure 3 (left), receives trigger primitives from the Phase-1 Global Muon Trigger (μGMT) and Calorimeter Trigger (DeMux), the Global Trigger (μGT) decision output and the input muon stubs from the Barrel Muon Track Finder (BMTF) [3, 7]. FPGA-based boards are used to receive the trigger optical links, running at 10 Gbps and with 8b/10b encoding. Simple data reduction is applied, and the pre-processed data is sent through Ethernet connection to a set of computing nodes (DSBUs and DSPUs), where further processing can be carried out. The system is heterogeneous, featuring three sets of receiver boards: the Xilinx KCU1500 and the Micron SB-852, which use DMA to a host server as output technology, and the Xilinx VCU128, which transfers the received primitives directly to the DSBU via TCP/IP.

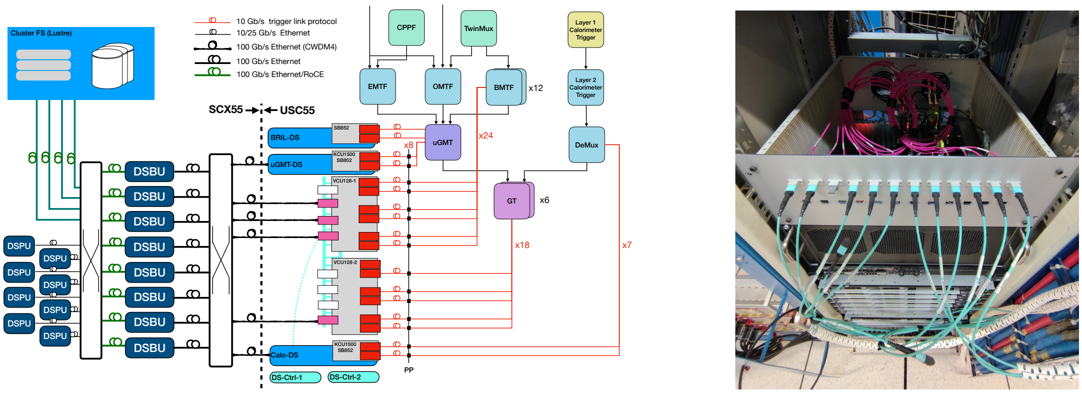

Figure 3. (Left) L1DS Run-3 demonstrator architecture [3]. It captures the trigger primitives from the Global Muon Trigger (μGMT) and the Calorimeter Trigger (DeMux), the Global Trigger (μGT) decision output and the input stubs to the Barrel Muon Track Finder (BMTF). After basic zero suppression, the received data is propagated via Ethernet to compute nodes (DSBUs and DSPUs) for further processing and subsequently sent to long-term storage. (Right) PCIe crate in the Run-3 demonstrator with two VCU128 boards.

Two VCU128 boards are installed in the demonstrator and housed in a PCIe crate from *One Stop Systems*, shown in Figure 3 (right). The monitoring and control of the boards is handled via AXI protocol communication by a server connected to the PCIe bus of the crate. For each board, up to 24 × 10 Gbps trigger input links can be connected to additional QSFP slots available through a *HT-Global* mezzanine, while the four on-board QSFPs are reserved for 100 GbE output links to the CMS surface computing room. The firmware to exploit these resources has been developed in three parts: the input data receiver, the trigger data processing pipeline and the output core.

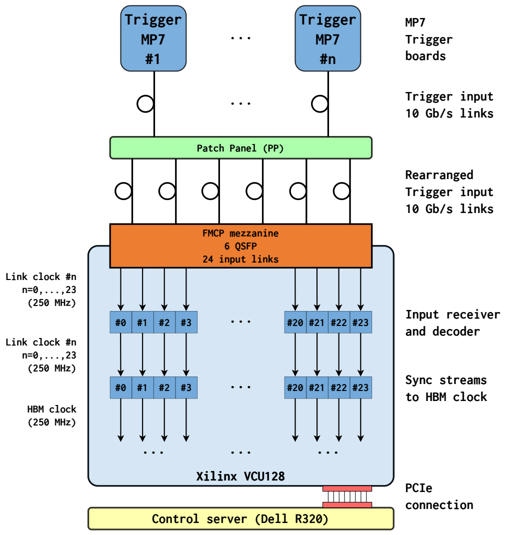
Trigger MP7 #1
Trigger MP7 #n
MP7 Trigger boards
Trigger input 10 Gb/s links
Patch Panel (PP)
Rearranged Trigger input 10 Gb/s links
FMCP mezzanine 6 QSFP 24 input links
Link clock #n n=0,...,23 (250 MHz)
#0 #1 #2 #3
#20 #21 #22 #23
Input receiver and decoder
Link clock #n n=0,...,23 (250 MHz)
Sync streams to HBM clock
HBM clock (250 MHz)
Xilinx VCU128
PCIe connection
Control server (Dell R320)

Figure 4. Design of the input receiver of the L1DS firmware for the VCU128 and the DAQ-800.

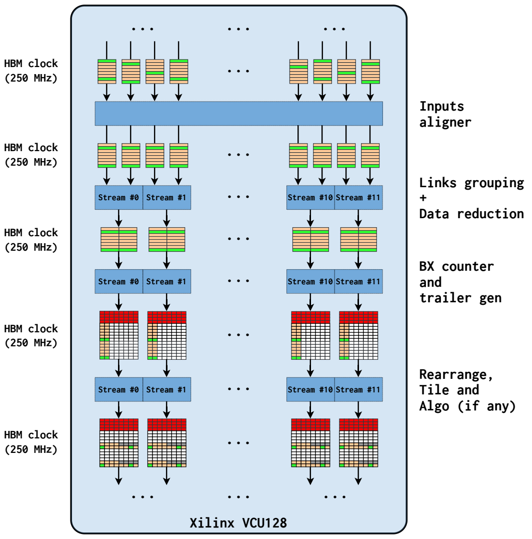

Figure 5. Design of the processing pipeline of the L1DS firmware for the VCU128 and the DAQ-800.

Input data receiver and decoder The Phase-1 trigger processors send bunch crossing (BX) data at 250 MHz, where a single BX block is a fixed size record of 6 × 32-bit frames per input link. Multiple BX blocks are grouped into orbits, sent at a rate of approximately 11.2 kHz. The input logic to receive these data is shown in Figure 4. An input receiver core decodes the data and key words from the input links, recovering a clock independently for each stream. Thus, a synchronization stage is required to move the streams to the HBM clock domain.

Processing pipeline A BX-wise alignment of the HBM clock synchronous streams is needed before performing any processing step in the pipeline, schematized in Figure 5. The aligned output streams are then arranged into groups. A simple data reduction condition is applied to every BX record in a group, suppressing for instance blocks without valid trigger primitives inside. The grouped streams are subsequently padded to accommodate the 256-bit alignment of the HBM banks and at the same time the BX blocks in an orbit are numbered. A trailer with a 3564-bit field carrying the information of the non-suppressed BXs is appended at the end of the orbit. The last stage of the processing pipeline applies a specific algorithm on the trigger primitives and tiles the output in a compact format. Examples include neural network inference for L1 muons recalibration [8] or muon stubs fitting, which can be deployed directly on FPGA through `hls4ml` [9, 10].

Output logic and Ethernet core The final stage is schematized in Figure 6. The pre-processed trigger data is encapsulated in "scouting blocks" with a programmable number of orbits. The packager logic is aware of the instantaneous occupancy of the HBM and handles the backpressure by discarding orbits. Prior to writing the scouting data blocks to the HBM, a 256-bit frame is reserved for the scouting block header. After the block payload and trailer have been written, a set of counters to place in the header is fully determined and the HBM writing logic places the frame in the reserved space. The TCP/IP logic controls the HBM reading and it allows to multiplex 4, 6 or 12 TCP/IP streams into a single 100 GbE core, which transmits the streams to the compute nodes.

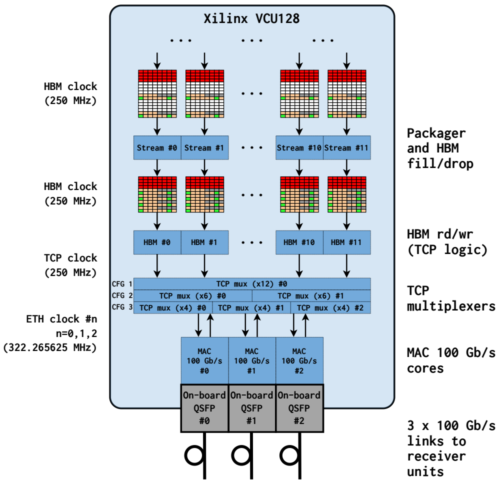

Figure 6. Design of the orbit packager, HBM read/write logic and output core of the L1DS firmware for the VCU128 and the DAQ-800.

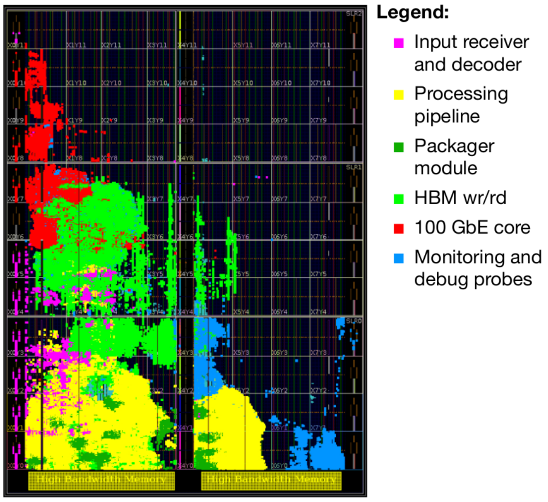

Figure 7. Floorplanning on a VU37P FPGA of a design with a 24 input links receiver, 12 HBM banks, 12 TCP/IP streams and 3 × 100 GbE cores.

Resource utilization and test The floorplanning on a VU37P chip for a design with a 24 trigger links receiver, 12 HBM banks, 12 TCP/IP streams and 3 × 100 GbE cores is shown in Figure 7. The resource utilization and the extrapolation to a VU35P chip is reported in Table 1. The neural network inference module is not included and it would allocate $O(10^2$ DSP units). The design has been tested by injecting test patterns into the scouting processing pipeline. A received throughput of around 98 Gbps from 12 TCP streams was processed without backpressure by an Intel TBB-based DAQ software on an AMD EPYC 7502P 32-Core CPU, writing the processed data to a ramdisk.

	VU37P			VU35P extrapolation	
Resource	**Available**	**Utilization**	**Utilization [%]**	**Available**	**Utilization [%]**
LUT	1303680	199478	15.30	871680	22.88
FF	2607360	328338	12.59	1743360	18.83
BRAM	2016	490	24.31	1344	36.45
URAM	960	48	5.00	640	7.50
DSP	9024	6	0.07	5952	0.10

Table 1. Target firmware resource utilization for a VU37P FPGA and extrapolation to a VU35P chip.

5 Conclusions and future plans

The development and validation of the L1DS firmware for the DAQ-800 board with the Xilinx VCU128 development kit in the Run-3 demonstrator has been discussed. The implemented firmware satisfies the timing closure for a VU37P FPGA. The extrapolation to a VU35P chip shows a resource utilization lower than 30%, allowing a timing-safe allocation of additional HBM banks and 100 GbE cores, if necessary. Future plans for CMS Phase-2 include the test of new hardware technologies, e.g. producing a "DAQ-1200" board with a Versal HBM chip. New ideas for output link protocols are also under study, such as RDMA over Converged Ethernet from the FPGA to compute nodes.

References

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

[2] CMS Collaboration, "The CMS experiment at the CERN LHC. The Compact Muon Solenoid experiment", *JINST* **3** (2008) S08004.

[3] CMS Collaboration, "Development of the CMS detector for the CERN LHC Run 3", *sub. to JINST* (2023) [2309.05466].

[4] CMS Collaboration, "The Phase-2 Upgrade of the CMS Level-1 Trigger", CMS-TDR-021, 2020.

[5] D.S. Rabady et al., "A 40 MHz Level-1 trigger scouting system for the CMS Phase-2 upgrade", https://doi.org/10.1016/j.nima.2022.167805, 2023.

[6] CMS Collaboration, "The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger", CMS-TDR-022, 2021.

[7] CMS Collaboration, "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} =$ 13 TeV", *Journal of Instrumentation* **15** (2020) P10017.

[8] CMS Collaboration, "40 MHz Scouting with Deep Learning in CMS", CMS-DP-2022-066, 2022.

[9] FastML Team, "fastmachinelearning/hls4ml", https://doi.org/10.5281/zenodo.1201549, 2023.

[10] J. Duarte et al., "Fast inference of deep neural networks in FPGAs for particle physics", *JINST* **13** (2018) P07027 [1804.06913].