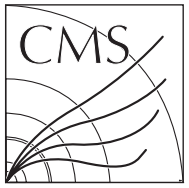
Available on CMS information server

CMS CR -2021/213



The Compact Muon Solenoid Experiment

Conference Report

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

19 October 2021 (v3, 23 October 2021)

CMS Phase-2 DAQ and Timing Hub – Prototyping results and perspectives

Vasileios Amoiridis, Ulf Behrens, Andrea Bocci, James Branson, Philipp Brummer, Sergio Cittolin, Diego Da Silva-Gomes Joao Da Silva Almeida, Georgiana-Lavinia Darlea, Christian Deldicque, Marc Dobson, Dominique Gigi, Nekija Dzemaili, Maciej Gladki, Frank Glege, Guillelmo Gomez-Ceballos, Neven Gutic, Magnus Hansen, Jeroen Hegeman, Guillermo Izquierdo Moreno, Thomas Owen James, Elias Leutgeb, Wei Li, Frans Meijers, Emilio Meschi, Srecko Morovic, Luciano Orsini, Ioannis Papakrivopoulos, Christoph Paus, Katarzyna Peron, Andrea Petrucci, Marco Pieri, Alexandros Poupakis, Ema Puljak, Dinyar Rabady, Kolyo Raychinov, Attila Racz, Hannes Sakulin, Christoph Schwick, Dainius Simelevicius, Andre Stahl, Jan Troska, Cristina Vazquez-Velez, Petr Zejdl, Vaiva Zokaite

Abstract

This paper describes recent progress on the design of the DAQ and Timing Hub, or DTH, an ATCA hub board intended for the Phase-2 upgrade of the CMS experiment. Prototyping was originally divided into multiple feature lines, spanning all different aspects of the DTH functionality. The second DTH prototype merges all R and D and development lines into a single board, which is intended to be the production candidate. Emphasis is on the process and experience in going from the first to the second DTH prototype, which included a change of the chosen FPGA as well as the integration of a commercial networking solution.

Presented at *TWEPP2021 TWEPP 2021 Topical Workshop on Electronics for Particle Physics*

Prepared for submission to JINST

TWEPP 2021 Topical Workshop on Electronics for Particle Physics
20—24 September 2021

CMS Phase-2 DAQ and Timing Hub Prototyping results and perspectives

Vasileios Amoiridis,[a] Ulf Behrens,[f] Andrea Bocci,[a] James Branson,[b] Philipp Brummer,[a] Sergio Cittolin,[b] Diego Da Silva-Gomes,[c,1] Joao Da Silva Almeida,[a] Georgiana-Lavinia Darlea,[d] Christian Deldicque,[a] Marc Dobson,[a] Dominique Gigi,[a] Nekija Dzemaili,[a] Maciej Gladki,[a] Frank Glege,[a] Guillelmo Gomez-Ceballos,[d] Neven Gutic,[a] Magnus Hansen,[a] Jeroen Hegeman,[a,2] Guillermo Izquierdo Moreno,[a] Thomas Owen James,[a] Elias Leutgeb,[a] Wei Li,[f] Frans Meijers,[a] Emilio Meschi,[a] Srecko Morovic,[b] Luciano Orsini,[a] Ioannis Papakrivopoulos,[e,1] Christoph Paus,[d] Katarzyna Peron,[a] Andrea Petrucci,[b] Marco Pieri,[b] Alexandros Poupakis,[a] Ema Puljak,[a] Dinyar Rabady,[a] Kolyo Raychinov,[a] Attila Racz,[a] Hannes Sakulin,[a] Christoph Schwick,[a] Dainius Simelevicius,[g,1] Andre Stahl,[f] Jan Troska,[a] Cristina Vazquez-Velez,[a] Petr Zejdl,[c,1] and Vaiva Zokaite,[a] on behalf of the CMS collaboration

[a] *CERN, Switzerland*
[b] *University of California, San Diego, San Diego, California, USA*
[c] *FNAL, Chicago, Illinois, USA*
[d] *Massachusetts Institute of Technology, Cambridge, Massachusetts, USA*
[e] *Technical University of Athens, Athens, Greece*
[f] *Rice University, Houston, Texas, USA*
[g] *Vilnius University, Vilnius, Lithuania*

E-mail: jeroen.hegeman@cern.ch

Abstract: This paper describes recent progress on the design of the DAQ and Timing Hub, or DTH, an ATCA hub board intended for the Phase-2 upgrade of the CMS experiment. Prototyping was originally divided into multiple feature lines, spanning all different aspects of the DTH functionality. The second DTH prototype merges all R&D and prototyping lines into a single board, which is intended to be the production candidate. Emphasis is on the process and experience in going from the first to the second DTH prototype, which included a change of the chosen FPGA as well as the integration of a commercial networking solution.

Keywords: Data acquisition circuits, Control and monitor systems online, Trigger concepts and systems (hardware and software)

[1] Also at CERN, Geneva, Switzerland
[2] Corresponding author.

Contents

1 Introduction

The CMS detector will undergo a major upgrade for Phase-2 of the LHC program: the High-Luminosity LHC. Figure 1 shows the overall architecture of the Phase-2 CMS data acquisition (DAQ) design. On-detector front-end electronics send their data via custom, radiation-tolerant, point-to-point optical links to the respective back-ends in the CMS service cavern. Back-ends are responsible for front-end synchronisation and read-out, data integrity verification, and the forwarding of their event fragments received to the DAQ and Timing Hub (DTH), which is the entry point into the central DAQ system. This latter step uses 25 Gbit/s point-to-point optical links operating a lossless, custom protocol,

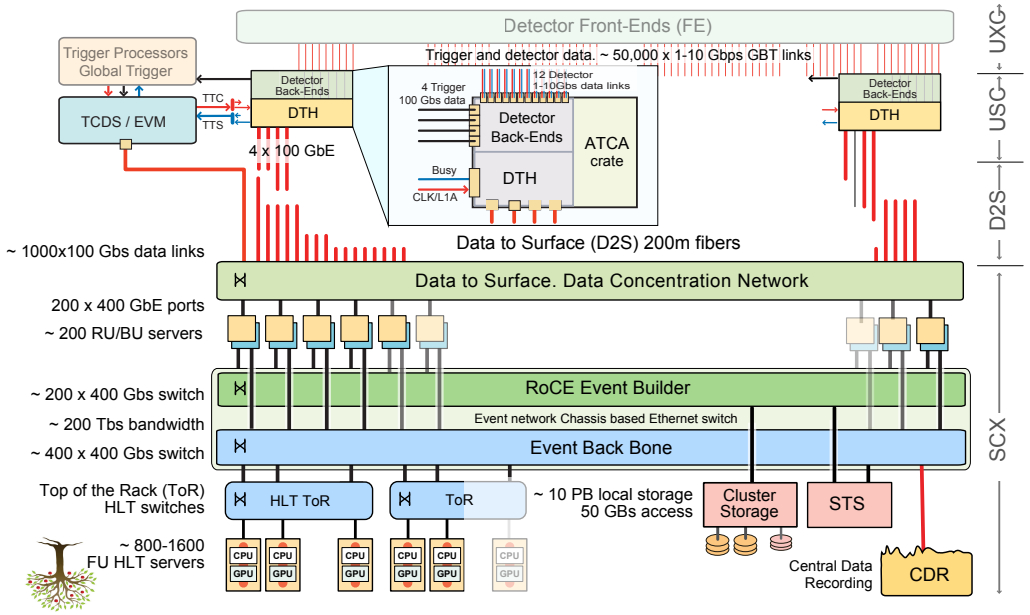


Figure 1. Architecture overview of the CMS Phase-2 data acquisition [1].

On the DTH, data are accumulated over a full LHC orbit for efficiency, and subsequently mapped into standard 100 Gbit/s Ethernet streams, before transmission by the Data-to-Surface (D2S) network connecting the underground service cavern (USC) to the online computing centre (OLC) on the surface. There, commercial compute nodes acting as combined read-out units and event builder units (RU/BUs) turn the received fragments into complete events, which are subsequently filtered for interesting physics content by the High-Level Trigger (HLT) farm. Passing events are stored locally before their transfer to the CMS central data recording (CDR), located in the CERN computing centre.

A comprehensive description of the Phase-2 upgrade of the DAQ and HLT can be found in the Technical Design Report [1].

2 The design of the DAQ and Timing Hub

With all back-end electronics implemented following the ATCA standard, the DTH is designed as an ATCA hub board. In addition to its DAQ functionality, it also provides the connections between an ATCA back-end crate and the central trigger and timing systems, as well as with the CMS online control network. Four main components can be identified in the DTH design, corresponding to its main tasks: an ATCA-compliant baseboard, housing a data acquisition unit, a timing and fast-command unit, and a managed Ethernet switch.

The DAQ unit uses up to six 4-channel mid-board optical engines operating at 16 Gbit/s or 25 Gbit/s to receive event fragment data from back-end boards, and up to five standard QSFP28 optics connected to the data network to guarantee a throughput of at least 400 Gbit/s.

The timing and fast-command unit uses standard QSFP+ optics to connect to the central Trigger and Timing Control and Distribution (TCDS) system, and distributes the LHC bunch clock and a higher-frequency precision clock on the backplane. In addition, it distributes fast-command and synchronisation information to all crate slots, and collects back-end readiness and status information in opposite direction, for transmission to the central TCDS.

As a true hub board, the DTH provides Gigabit Ethernet to all slots, as well as Fast Ethernet to both shelf managers, as part of the standard ATCA fabric. The connection to the control network is comprised of dual 10GbE uplinks implemented using commercial SFP+ devices.

To accommodate back-end crates requiring more than 400 Gbit/s of DAQ throughput, a DAQ-800 companion board to the DTH will be designed, containing two DAQ units, but without timing or switch functionality.

3 The DTH prototyping and development process

Following the division into functional units described in the previous section, the DTH R&D was divided into three independent branches. The main branch focused on the development of a dual-FPGA ATCA hub board, including all required connectivity to the front panel (two times SFP+ for the central TCDS connections, and four times QSFP28 connecting to the D2S network) and the backplane connections for the distribution of clocks and fast timing commands to all crate slots. This board, dubbed the P1V1 (prototype 1, version 1), was also used to gain experience with ATCA-related mechanical, powering, and cooling aspects. Based on initial results with the

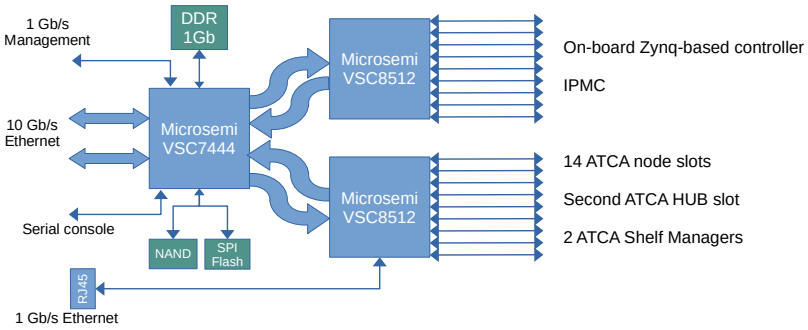


Figure 2. Overview of the network connectivity as implemented on the switch-board prototype, and as planned for the DTH P2 [1]. A commercial, managed switch ASIC provides the core functionality, two 10GbE uplinks to the CMS online control network, and management interfaces both via Gigabit Ethernet and via a serial console. Two additional 12-channel ASICs from the same manufacturer provide Gigabit Ethernet PHYs for all backplane slots, the shelf managers, and several on-board end-points.

P1V1 [2], a second version, the P1V2, was developed, which fixed a few design mistakes and improved the layout surrounding the various jitter cleaners on the board in order to further reduce phase noise. Both the P1V1 and -V2 satisfy the requirements for clock distribution of the Phase-2 CMS detectors. The typical performance achieved on a mock-up clock distribution chain using DTH prototypes achieves a front-end recovered bunch clock with a jitter of $\sigma_{\mathrm{RJ}} < 4\,\mathrm{ps}$, where the most stringent subdetector requirement is $\sigma_{\mathrm{RJ}} < 10\,\mathrm{ps}$. About 25 P1V2 boards were produced and are successfully used for subsystem development and integration.

A separate ATCA carrier was developed, using the same board infrastructure as the P1, to prototype the Ethernet switch, and to evaluate the switch management software. The switch functionality was implemented based on a trio of commercial switch ASICs, as outlined in figure 2. The main functionality is provided by a single-chip managed Ethernet switch, which provides two 10 Gbit/s uplinks, as well as several management interfaces. This main switch is backed by a pair of 12-port 10/100/1000BASE-T PHYs, each connected to the main switch by three 5 Gbit/s QSGMII lanes, providing Gigabit Ethernet connections to all crate slots and to the on-board controller and IPMC, as well as Fast Ethernet connections to both ATCA shelf managers. In terms of hardware development this design was rather straightforward. Software-wise, however, the effort required to confirm the necessary functionality turned out to be significant.[1] The moral here is that while commercial, off-the-shelf components enable access to advanced functionality, their integration is by no means a zero-effort endeavour.

The DTH DAQ unit needs a significant amount of buffer memory to ensure data-taking stability in the presence of D2S network throughput fluctuations as well as the non-real-time nature of the receiving PCs. In the original design this buffer memory was implemented in the form of a separate high-bandwidth memory component, which was retracted from the market around the time

[1]Packet corruption issues that only presented on certain ports, and only at Gigabit Ethernet speeds, were only resolved after several iterations with the switch ASIC manufacturer. This issue was eventually tracked down to the accidental enabling of a second reference clock input, the effects of which were not immediately clear from the available documentation. It should be noted that up to this point all evaluation of the switch hardware and software was based on the manufacturer's evaluation kit and reference design. It could be argued that some of the debugging might have been easier had the full software suite been purchased at the start of evaluation.

the first prototype was developed. At that time, two independent alternatives were investigated: the P1V2 included several DDR4 RAM banks replacing the buffer memory component, and in parallel a proof-of-concept was developed on an evaluation kit containing an FPGA with built-in high-bandwidth memory (HBM). The latter approach is the more convenient one, mainly due to the tight integration with the FPGA vendor toolkit and the absence of timing-critical routing between the FPGA and multiple RAM banks. Figure 3 shows a diagrammatic representation of the data flow through the DTH DAQ unit. Event fragments from subdetector back-end boards arrive at the DTH by up to six 4-channel mid-board optics devices, using a custom 'SLinkRocket' protocol operating at 25 Gbit/s. For efficiency, the fragments are aggregated on an orbit-by-orbit basis, before being stored in the in-FPGA high-bandwidth memory (HBM). On the D2S side, the aggregated data are read from the HBM, mapped to TCP/IP streams, and transmitted to the D2S network using the built-in 100GbE transceivers.[2] In order to reduce resource usage, the mapping of data sources to Ethernet streams is limited to five predetermined configurations, instead of a fully configurable routing implementation. This design purposely passes all data through the buffer memory, effectively decoupling throughput fluctuations between the input and output sides, in order to avoid switching between explicit 'read-out' and 'buffer' modes. Firmware prototyping for the DAQ functionality was performed using an evaluation kit containing an FPGA comparable to the one chosen for the DTH. A preliminary design transferring data from four SLinkRocket streams to a single 100GbE link has shown stable and error-free throughput (approximately 3 PB transmitted over three days) at over 88 Gbit/s to a commercial network interface in a Linux PC. In this setup the receiving host is believed to be the bottleneck. The occupancy of the firmware busses into and out of the HBM shows no saturation yet, and indicates a theoretical throughput ceiling of approximately 107 Gbit/s. A first design implementing all 24 input links and all 5 100GbE outputs, targeting the DTH FPGA, indicates an approximate resource usage of 70 %.

4 The DTH – Prototype 2

All branches prototyping the different DTH functionality groups combine into the second overall DTH prototype: the DTH P2. This board includes all P1V2 fixes and the layout improvements related to the jitter cleaners, changes the FPGA to a model with built-in high-bandwidth memory, and incorporates the Ethernet switch ASIC trio. To align with the majority of CMS back-end boards, the COMExpress on-board controller on the P2 is replaced by a (commercial) Zynq-based module.

It turned out to be unfeasible to design a full-featured DTH fitting on only the ATCA front-board PCB. Therefore, the on-board controller was moved to a rear transition module (RTM). This RTM shares its hot-swap handle with the front board, effectively making the DTH a 'two-piece board'. The advantage of this approach is that the same RTM design can be shared between the DTH and the DAQ-800 board. It also allows all debug interfaces to be neatly located at the rear, without sacrificing any scarce front-panel space.

[2] The original DTH design was based on an FPGA with four 100GbE transceivers, and could be configured with either six 4-channel inputs at 16 Gbit/s or with four 4-channel inputs at 25 Gbit/s. The current FPGA allows one additional 100GbE output, adding a comfortable amount of headroom.

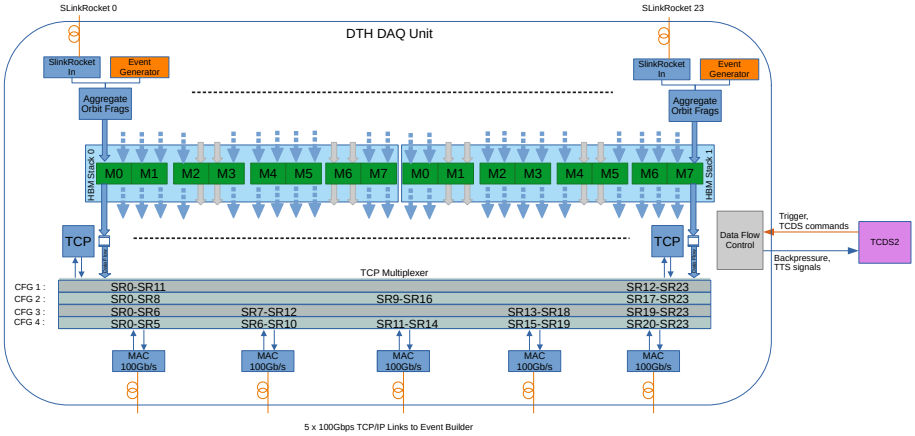


Figure 3. Schematic diagram of the flow of event fragment data from the back-ends through the DTH DAQ unit to the D2S network [1]. Note that all data flows through the buffer memory (implemented in the in-FPGA High-Bandwidth Memory (HBM)) in order to avoid mode switches between transmitting and buffering triggered by changes in D2S network throughput.

The change to a larger FPGA, and especially the use of an FPGA with HBM, significantly increases the DTH power requirements. The HBM alone may consume up to 20 A at 1.2 V, and the main core voltage, $V_{\mathrm{cc,\,int}}$, up to 100 A at 0.85 V. For the DTH P2, $V_{\mathrm{cc,\,int}}$ is carried by two 70 µm (i.e., a '4 oz layer') power planes for each FPGA. The shape of these planes, as well as the placement of the power converters with respect to the FPGAs, was optimised based on simulations, avoiding sharp corners and narrow lanes. The areas of the high-current power planes that carry no current were enlarged on purpose, to help with the distribution of heat, thereby improving thermal stability and cooling performance. With the doubling of the $V_{\mathrm{cc,\,int}}$ power planes the stack-up of the DTH P2 printed circuit board spans a total of 22 layers. In order to reduce cost, blind vias were avoided, although the use of back-drill vias was unavoidable.

5 Summary and outlook

The division of the DTH prototyping efforts into different functionality groups has paid off, and has led to the successful completion of proof-of-concepts for all required features. A small series of the main prototype has been produced and distributed for development use by subsystems. The next step is the production of a second DTH prototype, the DTH P2, combining all development lines. The design, layout, and simulation-based optimisation have been finalised in Q3 2021. After verification using a first few boards, expected in early 2022, this DTH P2 should serve as the candidate for production in 2024.

References

[1] The CMS collaboration, *The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger*, Tech. Rep. CERN-LHCC-2021-007, CMS-TDR-022, CERN, Geneva (Mar, 2021).

[2] J. Hegeman, R. Blažek, U. Behrens, J. Branson, P. Brummer, S. Cittolin et al., *First measurements with the CMS DAQ and Timing Hub prototype-1*, *PoS* **TWEPP2019** (2020) 111. 5 p.