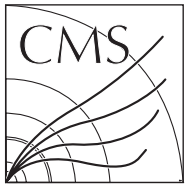
Available on CMS information server

**CMS CR -2022/130**



**The Compact Muon Solenoid Experiment**

# Conference Report

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

**18 August 2022**

# Progress in design and testing of the DAQ and data-flow control for the Phase-2 upgrade of the CMS experiment

Jeroen Hegeman for the CMS Collaboration

**Abstract**

The CMS detector will undergo a major upgrade for the Phase-2 of theLHC program the High-Luminosity LHC. The upgraded CMS detector willbe read out at an unprecedented data rate exceeding50 Tb/s, with a Level-1 trigger selecting eventsat a rate of 750 kHz, and an average event size reaching8.5 MB. The Phase-2 CMS back-end electronics will bebased on the ATCA standard, with node boards receiving the detectordata from the front-ends via custom, radiation-tolerant, opticallinks.The CMS Phase-2 data acquisition (DAQ) design tightens the integrationbetween trigger control and data flow, extending the synchronousregime of the DAQ system. At the core of the design is the DAQ andTiming Hub, a custom ATCA hub card forming the bridge between thedifferent, detector-specific, control and readout electronics and thecommon timing, trigger, and control systems.The overall synchronisation and data flow of the experiment is handledby the Trigger and Timing Control and Distribution System (TCDS). Forincreased flexibility during commissioning and calibration runs, thePhase-2 architecture breaks with the traditional distribution tree, infavour of a configurable network connecting multiple independentcontrol units to all off-detector endpoints.This paper describes the overall Phase-2 TCDS architecture, andbriefly compares it to previous CMS implementations. It then discussesthe design and prototyping experience of the DTH, and concludes withthe convergence of this prototyping process into the (pre)productionphase, starting in early 2023.

Presented at *RT2022 23rd IEEE Real Time Conference*

# Progress in design and testing of the DAQ and data-flow control for the Phase-2 upgrade of the CMS experiment

Vasileios Amoiridis*, Ulf Behrens†, Andrea Bocci*, James Branson‖, Philipp Brummer*, Eric Cano*, Sergio Cittolin‖, Diego Da Silva-Gomes‡††, Joao Da Silva Almeida*, Georgiana-Lavinia Darlea§, Christian Deldicque*, Marc Dobson*, Antra Gaile*, Dominique Gigi*, Frank Glege*, Guillelmo Gomez-Ceballos§ Neven Gutic*, Magnus Hansen*, Jeroen Hegeman* (*Member, IEEE*), Guillermo Izquierdo Moreno*, Carles Kishimoto*, Rafal Krawczyk¶, Wei Li†, Kenneth Long§ Frans Meijers*, Emilio Meschi*, Srecko Morovic‖, Luciano Orsini*, Christoph Paus§, Katarzyna Peron*, Andrea Petrucci‖, Marco Pieri‖, Alexandros Poupakis*, Dinyar Rabady*, Attila Racz*, Hannes Sakulin* (*Member, IEEE*), Christoph Schwick*, Dainius Simelevicius**††, Jan Troska*, Cristina Vazquez-Velez*, Petr Zejdl*, and Vaiva Zokaite*

***Abstract*—The CMS detector will undergo a major upgrade for the Phase-2 of the LHC program: the High-Luminosity LHC. The upgraded CMS detector will be read out at an unprecedented data rate exceeding $50\,\mathrm{Tbit/s}$, with a Level-1 trigger selecting events at a rate of $750\,\mathrm{kHz}$, and an average event size reaching $8.5\,\mathrm{MB}$. The Phase-2 CMS back-end electronics will be based on the ATCA standard, with node boards receiving the detector data from the front-ends via custom, radiation-tolerant, optical links. The CMS Phase-2 data acquisition design tightens the integration between trigger control and data flow, extending the synchronous regime of the DAQ system. At the core of the design is the DAQ and Timing Hub, a custom ATCA hub card forming the bridge between the different, detector-specific, control and readout electronics and the common timing, trigger, and control systems. The overall synchronisation and data flow of the experiment is handled by the Trigger and Timing Control and Distribution System. For increased flexibility during commissioning and calibration runs, the Phase-2 architecture breaks with the traditional distribution tree, in favour of a configurable network connecting multiple independent control units to all off-detector endpoints. In order to reduce the number of custom hardware designs required, the DAQ hardware is designed such that it can also be used to implement the Trigger and Timing Control and Distribution System.**

***Index Terms*—ATCA, Control, Data Acquisition, DAQ, Distribution, Timing, Trigger**

## I. Introduction

THE CMS experiment, one of the two general purpose particle physics experiments located in the CERN Large Hadron Collider (LHC), started its Run-1 data-taking in 2009. The data accumulated by the ATLAS and CMS experiments in Run-1 led to the discovery of the Higgs boson in 2012, leading to the award of the 2013 Nobel prize in physics to Peter Higgs and François Englert. Since then, the CMS experiment has undergone several upgrades, installed during the year-end technical stops and during Long Shutdowns 1 (2013–2015) and 2 (2019–2021).

For the second phase of the LHC program, the High-Luminosity LHC (HL-LHC), the CMS detector will again undergo a major upgrade. The Phase-2 CMS detector will be read out at a Level-1 trigger rate of $750\,\mathrm{kHz}$, with an average event size of $8.5\,\mathrm{MB}$, reaching an overall data rate exceeding $50\,\mathrm{Tbit/s}$.

From the point of view of the LHC experiments, the predominant changes in the upgrade from the LHC to the HL-LHC are 1) the increase in bunch charge from $\approx 1.15$ to $\approx 1.20$ protons, combined with 2) increased beam focusing at the interaction points ($\beta^*$ reduced from $30\,\mathrm{cm}$ to $15\,\mathrm{cm}$), and 3) a reduction in the bunch length from $8.3\,\mathrm{cm}$ to $7.6\,\mathrm{cm}$. The result is a three-fold increase in instantaneous luminosity, to (ultimately) $7.5 \times 10^{34}\,\mathrm{cm^{-2}s^{-1}}$. In addition, the shortened bunch length increases the longitudinal vertex density at the interaction point, leading to an even larger increase in pileup: from $\approx 55$ to $\approx 200$ proton-proton interactions per bunch crossing.

To maintain its excellent physics performance under HL-LHC conditions, the Phase-2 CMS upgrade addresses the mitigation of pile-up effects by 1) the redesign of the inner (pixel) and outer (pixel + strip) trackers with finer granularity and even better spatial resolution, 2) the replacement of the end-cap calorimeters with a new, high-granularity, calorimeter, 3) the addition of a timing detector layer between the outer tracker and the calorimeters, and 4) the addition of tracking information into the Level-1 trigger.

## II. CMS Phase-2 Data Acquisition Architecture

As was the case for previous upgrades of the CMS DAQ system, the Phase-2 upgrade is an evolution of the current

Manuscript submitted: August 22, 2022.
Corresponding author: Jeroen Hegeman (jeroen.hegeman@cern.ch).
This work was supported in part by the DOE and the NSF (USA).
*CERN, Geneva, Switzerland.
†DESY, Hamburg, Germany.
‡FNAL, Chicago, Illinois, USA.
§Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
¶Rice University, Houston, Texas, USA.
‖University of California, San Diego, San Diego, California, USA.
**Vilnius University, Vilnius, Lithuania.
††Also at CERN, Geneva, Switzerland.

(Run-3) system, migrating both hardware and software to current technologies, and adapting to the changing experimental requirements.

TABLE I
THE CMS PHASE-2 IN NUMBERS.

| CMS detector | Phase-1 | Phase-2 | |
|---|---|---|---|
| Peak average pileup | 60 | 140 | 200 |
| L1 accept rate (max.) [kHz] | 100 | 500 | 750 |
| Event size at HLT input [MB] | 2.0 | 7.8 | 9.9 |
| Event network throughput [Tbit/s] | 1.6 | 31 | 60 |
| Event network buffer (60 s) [TB] | 12.0 | 234 | 445 |
| HLT accept rate [kHz] | 1.0 | 5.0 | 7.5 |
| HLT compute power [MHS06] | 0.8 | 17 | 37 |
| Storage throughput [GB/s] | 2.0 | 31 | 61 |
| Storage capacity needed (1 d) [PB] | 0.2 | 2.0 | 3.9 |

From the DAQ point of view, the most influential changes (see Table I) are the increased throughput needs of some of the new subdetectors, and (due to the increased Level-1 trigger rate in combination with the increased event size) the large increase in required buffer space.

Figure 1 shows the overall architecture of the Phase-2 CMS data acquisition (DAQ) system. On-detector front-end electronics located in the experimental cavern (UXC) send their data via custom, radiation-tolerant, point-to-point optical links to the respective back-ends in the (radiation-free) service cavern (USC). Back-ends are responsible for front-end control and, upon instruction from the Level-1 hardware trigger, for forwarding their event data to the DAQ and Timing Hub (DTH), which provides the connection to the central DAQ system. This latter step uses $25\,\mathrm{Gbit/s}$ point-to-point optical links operating a lossless, custom protocol.

On the DTH, data are accumulated to increase transfer efficiency, mapped into standard $100\,\mathrm{Gbit/s}$ Ethernet streams, and transmitted via the Data-to-Surface (D2S) network connecting the underground service cavern to the online computing centre (OLC) on the surface. There, commercial compute nodes assemble the received fragments into complete events. These events are subsequently filtered for interesting physics content by the High-Level Trigger farm. Events that pass this second selection are stored locally before their transfer to the CERN computing centre for full reconstruction and permanent storage.

A more detailed description of the Phase-2 CMS DAQ system can be found in the Technical Design Report [1].

## III. CMS PHASE-2 TRIGGER AND TIMING DISTRIBUTION ARCHITECTURE

The CMS Trigger and Timing Control and Distribution System (TCDS) is the hardware system tasked with the low-level control and synchronisation of CMS data-taking. In addition to a high precision accelerator-synchronous clock, it distributes fast timing, synchronisation, and recovery commands to all back-end electronics.

Previous CMS fast-command systems were all based on a 'distribution tree' (Fig. 2, top), in which the cabling determined the connectivity between, and hence the possible grouping of, subdetectors. The Phase-2 TCDS (Fig. 3), instead, will use a configurable 'switching network' (Fig. 2, bottom) to allow any subset of detector back-ends to be operated independently, with the possibility of operating multiple data-taking, commissioning, and/or calibration runs in parallel. A complicating factor in the definition of this network is the fact that, in addition to timing and trigger information flowing from the run controller to the back-ends, back-end readiness information has to be collected and transmitted to the run controllers, in opposite direction, on the same links. So whereas for the timing information this network implements configurable switching (which is a passive task), for the readiness information it implements configurable merging, making the 'switching' an active task.

An important characteristic of the design presented here is the fact that the switching network is distributed across multiple nodes. This reflects the expectation that no single FPGA/device will be able to contain the full network, and adds a level of scalability. The same holds true for the layer of run controllers: the distribution of these across multiple physical devices provides the design with a certain amount of scalability. Combined, these two considerations impose the need for a full-mesh network connecting the run controller nodes with the switch nodes. This, as well as the added complexity in the configuration of the system, is the price to pay for the added grouping flexibility compared to the traditional distribution tree.

More detail about the design of the CMS Phase-2 TCDS can be found in [1].

## IV. DESIGN AND PROTOTYPING 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 shelf 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: 1) an ATCA-compliant baseboard, housing 2) a data acquisition unit, 3) a timing and fast-command unit, and 4) a managed Ethernet switch.

The DAQ unit uses up to six 4-channel mid-board optical engines operating at $16\,\mathrm{Gbit/s}$ or $25\,\mathrm{Gbit/s}$ 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 $400\,\mathrm{Gbit/s}$ towards the data-to-surface network.

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 shelf slots, and collects back-end readiness and status information in the 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,

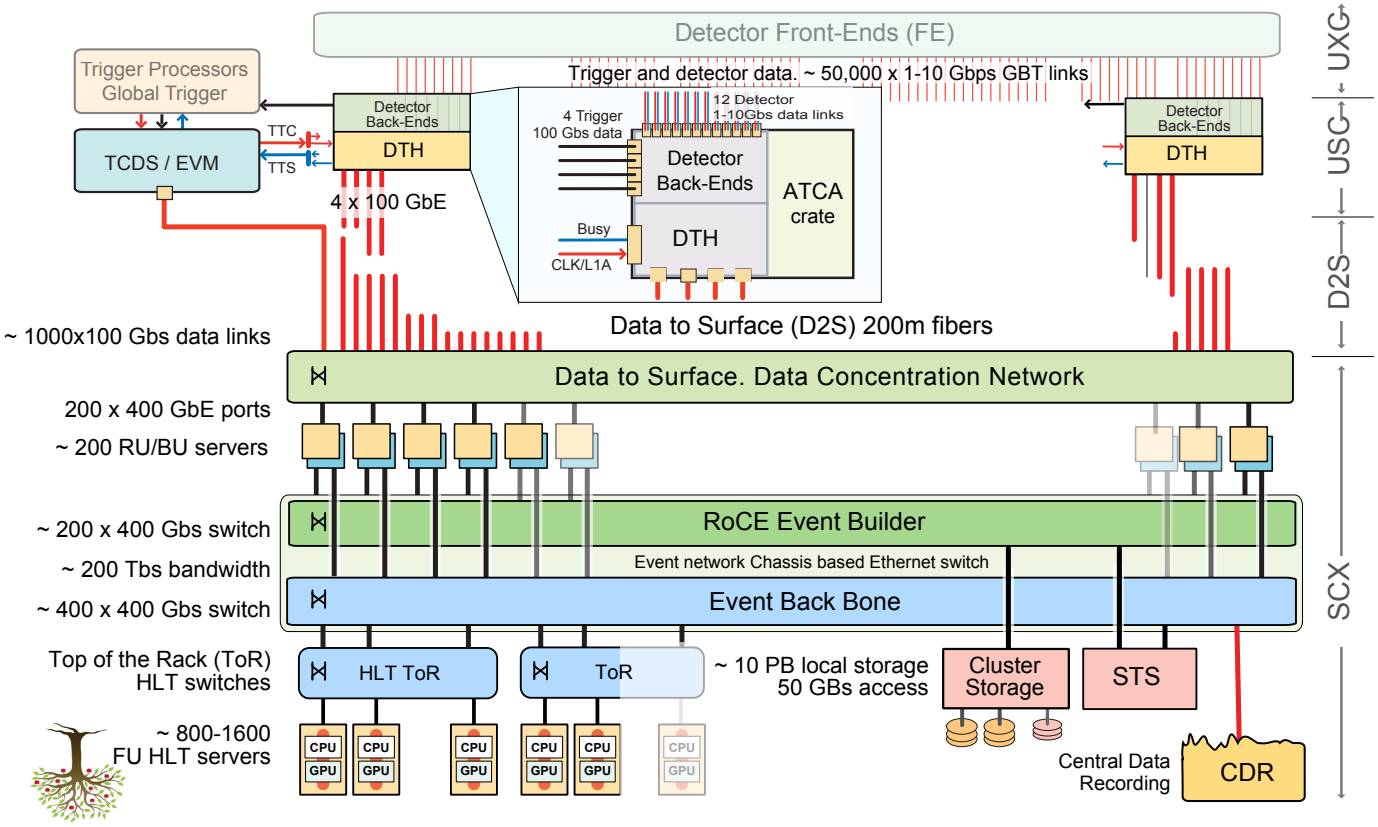



Fig. 1. Overall architecture of the CMS Phase-2 data acquisition system. Physics data flow from the on-detector front-ends in the underground experimental cavern (UXC), via detector-specific back-end electronics in the radiation-safe environment of the Underground Service Cavern (USC) to DAQ and Timing Hubs. The DTHs aggregate the data for efficient transmission via the data-to-surface network to the online computing centre on the surface (SCX). On the surface, read-out units (RU) and event builder units (BU), interconnected on a RoCE-based network, assemble the individual subdetector event fragments into coherent events, and transport these to the High Level Trigger (HLT) for filtering. Events passing the HLT are buffered in local storage before transfer to the CERN central computing centre.

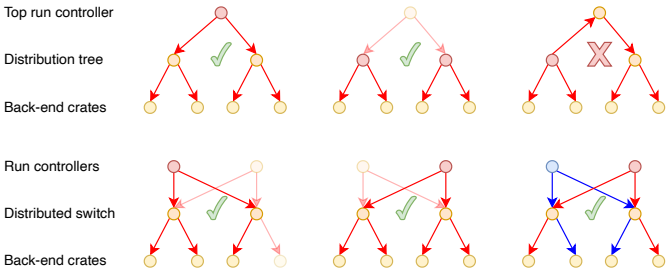



Fig. 2. Top: Diagram of a traditional trigger distribution tree, as implemented in past and current CMS. Firmware 'run controller' entities govern data-taking runs. Independent run controllers are present at the top of the tree and in all intermediate nodes. Data-taking with the full detector requires the top-level run controller, since this is the only controller with access to all branches of the distribution tree. Intermediate-level run controllers can be used to drive all subsystems connected to their branch, but cannot access subsystems in neighbouring branches. Bottom: A 'switched network' approach to trigger distribution. All run controllers are now implemented at the top level, and the intermediate nodes form a distributed, configurable network. By careful configuration, any of the top-level run controllers can be given access to any subset of the subdetector back-end electronics.

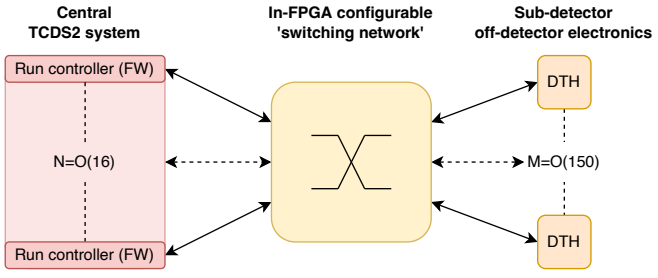



Fig. 3. The Phase-2 TCDS does away with the traditional distribution tree, in which the cabling configuration defines the possible combinations of subdetectors that can operate together. Instead, a configurable 'switching network' is used, allowing any run controller (on the left of the diagram) to command any subset of the DTH endpoints in the experiment (on the right of the diagram).

as part of the standard ATCA fabric. The connection to the control network is comprised of dual 10GbE uplinks implemented using commercial SFP+ devices.

### A. The DAQ-800 companion to the DTH-400

To accommodate back-end shelves requiring more than 400 Gbit/s of DAQ throughput, a DAQ-800 companion board to the DTH has been designed. This DAQ-800 will contain two DAQ units, for a total throughput of 800 Gbit/s, but no timing or Ethernet switch functionality. (See Fig. 4.)

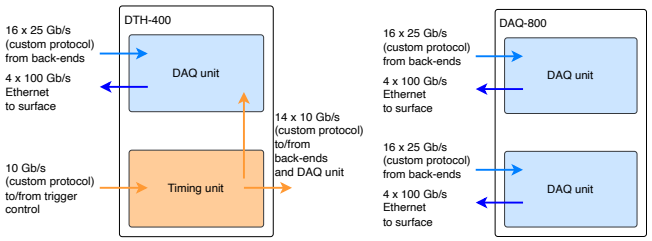



Fig. 4. Simplified diagrams of the DTH-400 (left) and the DAQ-800 (right) boards. The DTH contains a timing unit and a DAQ unit (as well as an Ethernet switch). The timing unit connects to the central TCDS via front panel optics, and to the back-end electronics in the shelf via the backplane. The DAQ unit receives event data from the back-end electronics in the shelf via custom front panel links, and transmits those, after aggregation, on standard $100\,\mathrm{Gbit/s}$ Ethernet connections, again on the front panel. The DAQ-800 replaces the timing unit and the Ethernet switch with a second copy of the DAQ unit, doubling the DAQ bandwidth to at least $800\,\mathrm{Gbit/s}$.

### *B. The DTH prototyping and development process*

The DTH R&D was divided into three independent branches, each focused on a particular part of the DTH functionality. The main branch focused on the development of a dual-FPGA ATCA hub board, including all required connectivity to the front panel and the backplane connections for the distribution of clocks and fast timing commands. 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 P1V1, a second version, the P1V2, was developed, which fixed a few design mistakes and improved the layout in order to further reduce phase noise. Both the P1V1 and the P1V2 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 $\sigma_{\mathrm{RJ}} < 4\,\mathrm{ps}$ [2], [3], where the most stringent subdetector requirement is $\sigma_{\mathrm{RJ}} < 10\,\mathrm{ps}$ [1].

In 2019, about 25 P1V2 boards were produced, which are since then 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 is implemented based on a trio of commercial switch ASICs. The main functionality is provided by a single-chip managed Ethernet switch, which provides two $10\,\mathrm{Gbit/s}$ 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 $5\,\mathrm{Gbit/s}$ QSGMII lanes, providing Gigabit Ethernet connections to all shelf 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. The development of the software and configuration for the control of the more advanced networking features in the main switch required several iterations with the vendor. This process clearly demonstrated that even though these advanced commercial components can greatly expand the possibilities of our custom hardware, the required engineering effort for successful integration remains significant.

The DTH DAQ unit needs a significant amount of buffer memory to absorb fluctuations in the D2S network throughput, or those due to 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 the first prototype was developed. Two alternatives were investigated: the P1V2 included several DDR4 RAM banks replacing the buffer memory component, and in parallel a proof-of-concept was developed based on an FPGA with built-in high-bandwidth memory (HBM). The latter was chosen for the final design, mainly due to the ease of integration with the FPGA vendor toolkit and the absence of timing-critical routing between the FPGA and multiple RAM banks. 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 $25\,\mathrm{Gbit/s}$. For efficiency, the fragments are aggregated, 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. 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 $25\,\mathrm{Gbit/s}$ streams to a single 100GbE link has shown stable and error-free throughput 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 $107\,\mathrm{Gbit/s}$. A first design implementing all 24 input links and all 5 100GbE outputs, targeting the DTH FPGA, indicates an approximate resource usage of $70\,\%$.

### *C. The DTH – Prototype 2*

All branches prototyping the different DTH functionality groups combine into the second overall DTH prototype: the DTH P2 (Fig. 5). This board includes all P1V2 fixes and improvements, changes the FPGA from a Xilinx KU15P to a Xilinx VU35P (speed grade 1) which includes $8\,\mathrm{GB}$ of HBM, and incorporates the Ethernet switch ASIC trio. In addition, a Zynq-based on-board controller replaces the original COM Express x86-based module.

It turned out to be impossible to fit all required functionality for a full-fledged DTH on a single ATCA front-board PCB. Therefore, the on-board controller was moved to a rear transition module (RTM). One advantage of this approach is that the

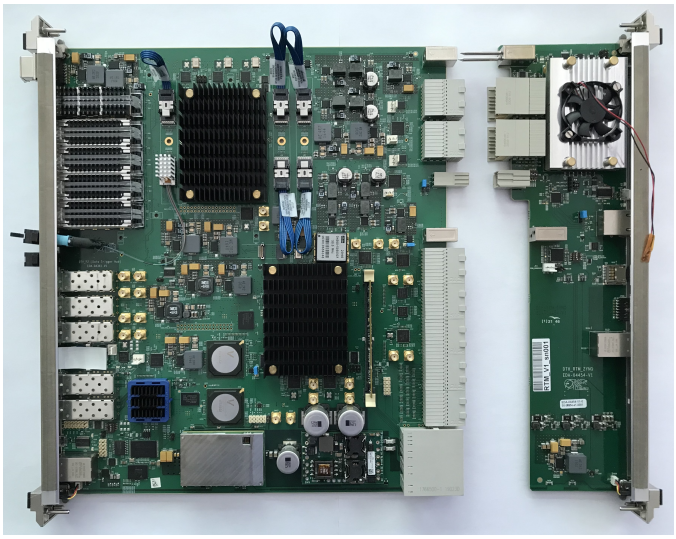


Fig. 5. The first assembled P2 DTH prototype. The DAQ and TCDS FPGAs (under the top and bottom heatsinks respectively) are clearly visible. On the front panel (left-hand) side one recognises the QSFP28 cages for the data-to-surface connections at the top, and the SFP+ cages for the TCDS and network uplinks at the bottom. The on-board controller has been moved to a Rear Transition Module (RTM), allowing the controller to be upgraded independently, as well as making it a common item between the DTH-400 and DAQ-800 boards.

RTM becomes a common component between the DTH-400 and DA-800 boards, which can be easily upgraded to a more powerful controller at a later time. It also allows all debug interfaces to be neatly located at the rear, without sacrificing any front-panel space.

### *D. Powering and PCB design*

The change to a larger FPGA with HBM significantly increases the DTH power requirements. For each of the two FPGAs on the DTH, the HBM alone may draw 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 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. Since the total PCB thickness now approaches the upper limit of the ATCA specification of $(2.4 \pm 0.2)\,\mathrm{mm}$, two grooves are machined along the top and bottom edges to locally reduce the thickness to $2.4\,\mathrm{mm}$.

## V. Design of the Trigger Control System Using DAQ Hardware

Based on the desire to reduce both engineering effort and prototyping cost, and in an attempt to reduce the number of custom hardware designs, the CMS central DAQ project was challenged to investigate if the upgraded central trigger control and timing system could be implemented using the DTH-400 and DAQ-800 boards already under development.

Looking back at Section III, it becomes clear that the basic building block for both the TCDS run controllers and for the switch nodes is the same: an FPGA equipped with a number of high-speed optical links. The size of the FPGA is relevant both for the run controllers and for the switch nodes. The run controllers have to be able to monitor many independent end-points (estimated to be O(150) at the time of writing), and perform the rate and deadtime accounting of the whole experiment, involving different trigger types, deadtime sources, etc. In the switch nodes the multiplexing of the $10.26\,\mathrm{Gbit/s}$ streams dominates the FPGA resource use.

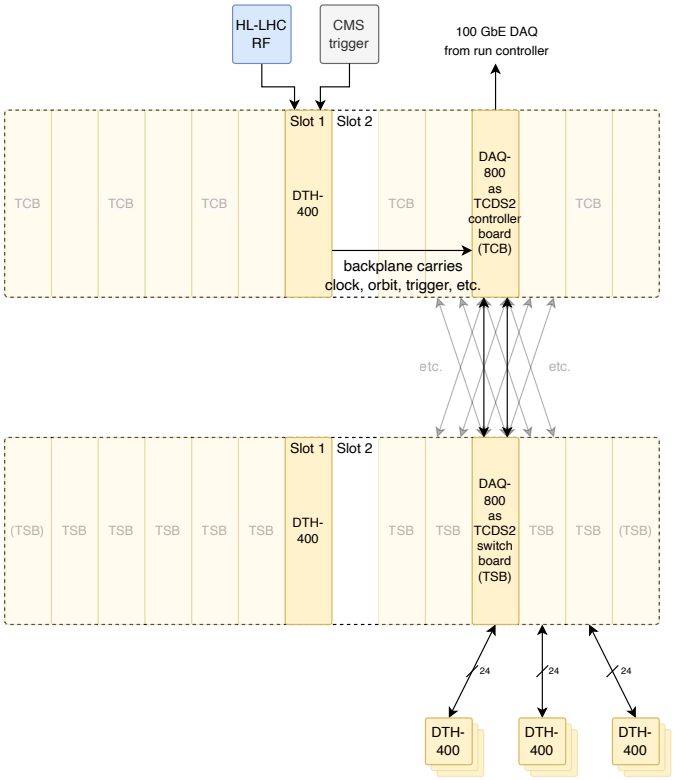



Fig. 6. Architectural diagram of the CMS Phase-2 Trigger and Timing Control and Distribution System. All three layers of the bottom diagram of Fig. 2 are represented here. The run controllers (top level) will be implemented in the FPGAs of the DAQ-800 board in the top ATCA shelf. The distributed switching network (middle layer) will be spread over the FPGAs of the DAQ-800 boards in the bottom ATCA shelf. The DTH-400s at the bottom of the figure form the connection between the central system and the individual subdetector back-end shelves.

Figure 6 shows the proposed implementation of the central TCDS using DTH-400 and DAQ-800 boards, all equipped with dedicated firmware for this purpose. One ATCA shelf contains a number of DAQ-800 boards, on which each FPGA implements a single run controller. A second ATCA shelf contains DAQ-800 boards, on which each FPGA implements a node of the distributed switch connecting the run controllers to the subsystem DTHs. The Firefly connections that normally receive back-end event data are used to interconnect the run controllers and the switch nodes. The QSFP28s on the switch nodes (which would normally connect to the data-to-surface

network) are used for the connections to the DTHs in the subsystem back-end shelves. On the run controller side only one of the five QSFP28s is used, in this case to connect the run controller event data directly to the data-to-surface network.

Both the run controller shelf and the switch node shelf also house a DTH-400, in its usual hub slot. This board interfaces these shelves to the HL-LHC RF system, and to the central CMS Level-1 trigger. For this purpose, the boards will operate another, dedicated, firmware version.

The architecture proposed above contains a certain amount of flexibility 1) in the number of run controllers that can be implemented, 2) in the number of DTHs connected to each switch node, and 3) in the number of nodes in the distributed switch.

The two main constraints to the design are related to the number of nodes in the switch layer. Given a fixed number of run controllers and a fixed number of DTH end-points, the number of switch nodes determines the amount of FPGA resources required to realise each node's contribution. This sets a lower limit on the number of switch nodes. The requirement of a full mesh network connecting run controllers to switch nodes, combined with the number of available optical connections on each run controller, sets an upper limit on the number of possible switch nodes. (Conversely, the number of optical connections per switch node limits the maximum number of run controllers.)

### A. How using common hardware affects the DAQ and TCDS

Both the DTH-400 and the DAQ-800 are primarily designed as DAQ aggregator/adaptor boards. Some minor modifications were made to these boards to enable their use in a synchronous trigger and timing system. These modifications are mostly related to the routing to, and use of, the LHC clock in the DAQ optics, which are normally operated asynchronously, and do not affect the performance of these boards for DAQ purposes.

Because all DAQ connectivity is implemented in 'quad form' using 4-channel Fireflys and QSFPs, the connections between the run controllers and the switch nodes requires custom 'shuffle fibres' to break-out, shuffle, and fan-in the fourfold fibre groups.

Preliminary *in Vivado* studies indicate that it will be possible to implement the architecture proposed above using the current design of the DTH-400 and DAQ-800 boards, with the help of custom 'shuffle fibres' to adapt the optical connectivity. The intention is to demonstrate proof-of-principle using the (DAQ units on the) first DTH-P2 boards, before committing to the final architecture of the Phase-2 TCDS.

## VI. Summary and outlook

The design and development of the Phase-2 CMS custom DAQ hardware is well under way.

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 first prototype has been produced and distributed for development use by subsystems. The second prototype consolidates the experience of all development branches, and is expected to become the production design. A companion board, the DAQ-800, derived from the DTH-400 design, has been designed for subsystems requiring additional event data throughput in a single back-end shelf.

The DTH-400 and DAQ-800 boards have been designed such that they can also be used to implement the Phase-2 CMS Trigger and Timing Control and Distribution System. Whereas from a purely technical point of view this may look like a compromise, the reduction in the number of custom board designs makes this a worthwhile effort. Preliminary studies indicate that sufficient interconnections and FPGA resources should be available on the DAQ-800 boards to accomplish the TCDS design proposed in the TDR. Proof-of-principle will be demonstrated using the first DTH-400 preproduction boards, well before the expected installation date of the system in 2027.

## References

[1] CMS Collaboration, "The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger," CERN, Geneva, Tech. Rep., Mar. 2021. [Online]. Available: https://cds.cern.ch/record/2759072.

[2] J. G. Hegeman, "First measurements with the CMS DAQ and Timing Hub prototype-1," CERN, Geneva, Tech. Rep., Oct. 2019. [Online]. Available: https://cds.cern.ch/record/2797789.

[3] V. Amoiridis, U. Behrens, A. Bocci, *et al.*, "CMS Phase-2 DAQ and Timing Hub – Prototyping results and perspectives," CERN, Geneva, Tech. Rep., Oct. 2021. [Online]. Available: https://cds.cern.ch/record/2797680.