PROCEEDINGS OF SCIENCE

Design and development of the DAQ and Timing Hub for CMS Phase-2

Jean-Marc André[5], **Ulf Behrens**[1], **Andrea Bocci**[2], **James Branson**[4], **Sergio Cittolin**[4], **Diego Da Silva Gomes**[2], **Georgiana-Lavinia Darlea**[6], **Christian Deldicque**[2], **Zeynep Demiragli**[6], **Marc Dobson**[2], **Nicolas Doualot**[5], **Samim Erhan**[3], **Jonathan Richard Fulcher**[2], **Dominique Gigi**[2], **Maciej Gladki**[2], **Frank Glege**[2], **Guillelmo Gomez-Ceballos**[6], **Magnus Hansen**[2], **Jeroen Hegeman**[*2] **André Holzner**[4], **Michael Lettrich**[2], **Audrius Mecionis**[5,9], **Frans Meijers**[2], **Emilio Meschi**[2], **Remigius K. Mommsen**[5], **Srecko Morovic**[5,10], **Vivian O'Dell**[5], **Samuel Johan Orn**[2], **Luciano Orsini**[2], **Ioannis Papakrivopoulos**[7], **Christoph Paus**[6], **Andrea Petrucci**[8], **Marco Pieri**[4], **Dinyar Rabady**[2], **Attila Rácz**[2], **Valdas Rapsevicius**[5,9], **Thomas Reis**[2], **Hannes Sakulin**[2], **Christoph Schwick**[2], **Dainius Šimelevičius**[2,9], **Mantas Stankevicius**[5,9], **Jan Troska**[2], **Cristina Vazquez Velez**[2], **Christian Wernet**[2], **Petr Zejdl**[5,10]

[2]*CERN, Geneva, Switzerland*
[1]*DESY, Hamburg, Germany*
[5]*FNAL, Batavia, Illinois, USA*
[6]*Massachusetts Institute of Technology, Cambridge, Massachusetts, USA*
[3]*University of California, Los Angeles, Los Angeles, California, USA*
[4]*University of California, San Diego, San Diego, California, USA*
[7]*Technical University of Athens, Athens, Greece*
[8]*Rice University, Houston, Texas, USA*
[9]*Also at Vilnius University, Vilnius, Lithuania*
[10]*Also at CERN, Geneva, Switzerland*
E-mail: jeroen.hegeman@cern.ch
This work was supported in part by the DOE and NSF (USA).

The CMS detector will undergo a major upgrade for Phase-2 of the LHC program, starting around 2026. The upgraded Level-1 hardware trigger will select events at a rate of 750 kHz. At an expected event size of 7.4 MB this corresponds to a data rate of up to 50 Tbit/s.

Optical links will carry the signals from on-detector front-end electronics to back-end electronics in ATCA crates in the service cavern. A DAQ and Timing Hub board aggregates data streams from back-end boards over point-to-point links, provides buffering and transmits the data to the commercial data-to-surface network for processing and storage. This hub board is also responsible for the distribution of timing, control and trigger signals to the back-ends.

This paper presents the current development towards the DAQ and Timing Hub and the design of the first prototype, to be used as for validation and integration with the first back-end prototypes in 2019-2020.

Topical Workshop on Electronics for Particle Physics (TWEPP2018)
17-21 September 2018
Antwerp, Belgium

*Speaker.

PoS(TWEPP2018)129

1. Introduction

The CMS detector will undergo a major upgrade for the Phase-2 of the LHC program, starting around 2026. The original CMS trigger-DAQ design [1] combines a hardware trigger (Level-1), implemented in custom electronics, with a High-Level Trigger (HLT) implemented in software and running on commodity compute nodes. This design has proven to be very flexible and scalable [2, 3], and the Phase-2 baseline DAQ design builds on this same architecture. Figure 1 shows an overview of the baseline Phase-2 DAQ design. Subdetector back-ends will transmit data on custom point-to-point links at 16 or 25 Gbit/s to a DAQ and Timing Hub in the same crate, which concentrates and balances these data for transmission to the surface counting room. The data-to-surface (D2S) network will be based on commercially available hardware, and use a standard protocol. A networked event builder will assemble all back-end event fragments into events, and transfer them to the HLT filter farm. Events accepted by the HLT are buffered locally in anticipation of transfer to the CERN computing center for permanent storage.

At a pileup of 200 proton-proton collisions, the upgraded CMS detector will produce an estimated event size of ≈ 7.4 MB. At the design Level-1 accept rate of 750 kHz this corresponds to an event-builder throughput of ≈ 44 Tbit/s. The HLT is supposed to reduce this to an output rate of 7.5 kHz to storage.

A more in-depth description of the CMS Phase-2 DAQ upgrade can be found in [4, 5].

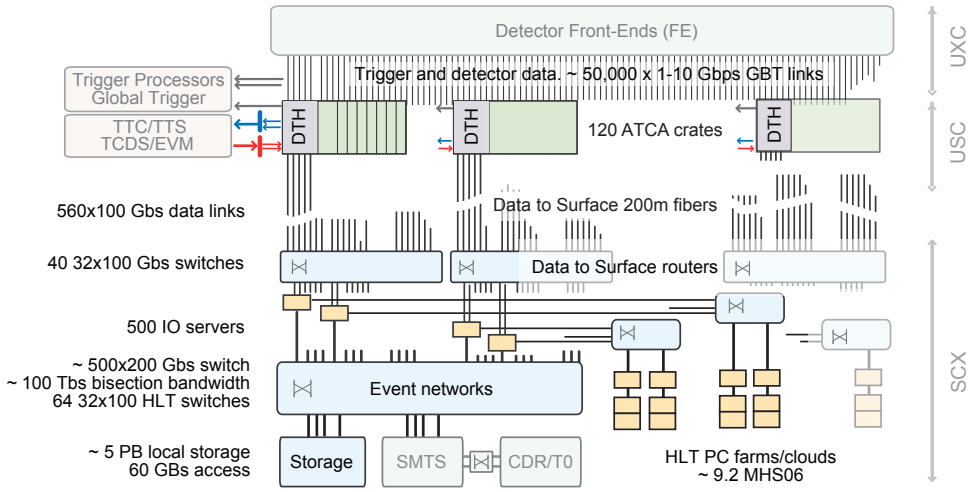

Figure 1: Overview of the CMS Phase-2 DAQ and Trigger system. The upgrade design still follows the original two-level design with a hardware trigger (Level-1) and a software trigger (High-Level Trigger). The main changes lie in the interconnect architecture and the integration of the DAQ and Timing Hub in the subsystem back-end crates. (Color version available online.)

2. The DAQ and Timing Hub

Optical links coming from the detector front-ends are aggregated in detector-dependent ATCA-based back-end boards. A DAQ and Timing Hub (DTH) aggregates several data streams from multiple back-end boards over point-to-point links. (See Fig. 2.) These links use a custom protocol,

an evolution of the CERN S-link protocol family [6], with datagrams corresponding to one event transmitted to the DTH and flow-control in the reverse direction. The DTH combines these streams to feed commercial high-speed optical links forming the data-to-surface (D2S) network, each with 100 Gbit/s or larger bandwidth. The D2S links carry the data to the surface, connecting the DTH output via network to I/O servers for event building. At the same time the DTH is responsible for distributing timing and trigger control signals (TTC) to the back-end electronics from where they are distributed to the front-ends. These TTC signals are received by the DTH from the TCDS (Trigger and Timing Control and Distribution System) master over an optical link and distributed through the backplane to all node slots. Each node slot sends its data-taking readiness status along with monitoring data to the DTH over the backplane. The DTH, in turn, elaborates a global status to provoke trigger throttling as necessary. Finally, the DTH provides monitoring of DAQ and TTC/TTS functions, and emulation of data sources, both self-triggered and externally triggered, for testing purposes.

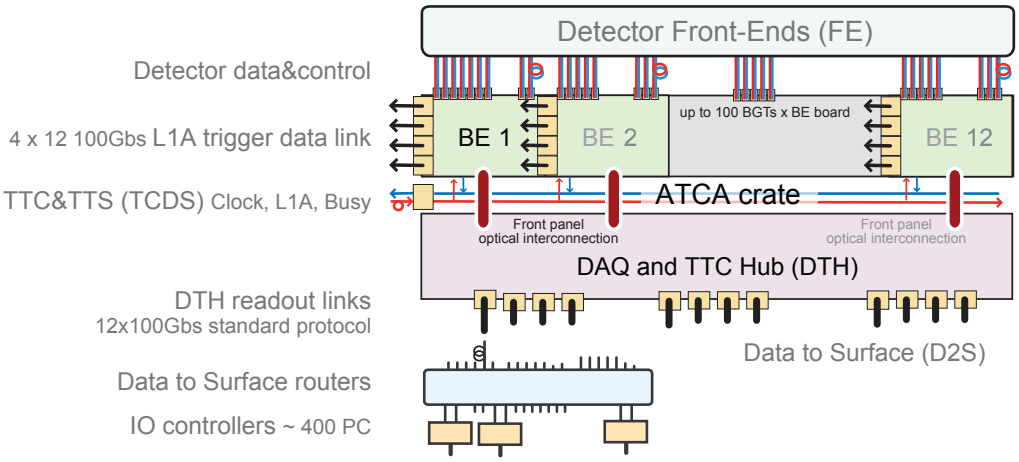

Figure 2: A DAQ and Timing Hub (DTH) in a back-end ATCA crate. Front-end to back-end connectivity is point-to-point. Level-1 trigger information is diverted by the back-end boards, and the event data are sent to the DTH on point-to-point optical links. The DTH balances and concentrates these data, and sends them on to the commercial data-to-surface network. (Color version available online.)

In consideration of the structure of the ATCA crate, a standardized DTH board will be located in the first hub slot. Because of the wide variety of required per-crate and per-board throughput, a modular design was adopted for the DTH, consisting of a timing unit for clock distribution and synchronization tasks, and one or more DAQ units handling the data streams. Data from a subset of the leaf boards in a crate can be handled by one FPGA operating independently as a DAQ unit. The baseline DTH design consists of one timing unit and one or more DAQ units. In cases where the required crate throughput may exceed the maximum bandwidth of a single DTH board, it will be possible to install one or more additional DTH boards with DAQ units in other slots.

The first DTH prototype (P1), targeted for end 2018/early 2019, will include a first version of one timing unit and one DAQ unit. The Xilinx Kintex UltraScale+ KU15P-2 FPGA on the DAQ unit provides enough high-speed transceivers for a DAQ unit capable of 400 Gbit/s including 24×16 Gbit/s or 16×25 Gbit/s input pairs via Samtec FireFly mid-board optics, 16 serial link

pairs to a Micron HMC (Hybrid Memory Cube) memory and 16 25 Gbit/s output pairs to four QFSP28 100GbE cages.

3. PCB design and power integrity simulation

The DTH P1 PCB has been designed focusing on the signal integrity of all high-speed serial and clock signals. Special care has been taken to 'sandwich' all high-speed layers between two ground layers. The result is a 20-layer stack-up in Isola I-Tera MT40, chosen for its dielectric constant of $Dk = 3.45$ up to 25 Gbit/s. As part of the PCB design process, the integrity and performance of all power planes has been simulated using Mentor HyperLynx. Based on the initial simulation results the PCB design was successfully adapted to remove current hot-spots and to minimize voltage drops across power supply planes.

4. Thermal modeling and cooling performance

For cost reasons, i.e., to reduce the space and the number of boards required, a high data throughput per DTH is preferred. This implies the presence of multiple high-power FPGAs per board, as well as a large number of optical transceivers. Especially for the latter, it is important to maintain the package/heatsink temperature below $\approx 50\,^{\circ}$C in order not to impact their longevity [7].

Based on the PCB layout of the DTH P1, a simplified thermal model has been developed. This model only includes those components that either:

- significantly contribute to the heat load, based on their power consumption of ≥ 3 W, or
- significantly affect the airflow based on their dimensions.

Figure 3 shows the resulting simulation model, together with preliminary airflow and temperature estimates. At the moment the aim is to develop the required simulation experience in collaboration with other developers and experts in CMS and at participating institutes. The DTH P1 will be instrumented with temperature sensors on the front and back sides of the board. These temperature readings, together with direct component temperatures where available, can be used to improve the thermal model. The intention is to use the resulting model to optimize the thermal properties of further generations of the DTH design.

5. R&D on input and output technologies

In anticipation of the production of the first DTH P1 boards, several parallel efforts are ongoing using evaluation boards containing FPGAs comparable to the one targeted for the DTH.

Data from sub-detector back-ends arrive to the DTH on point-to-point optical links running a custom protocol. This protocol, called SlinkRocket, will be an evolution of the S-link and Slink-Express protocols [6] used already in CMS, adapted for wider counters, larger data words, and expanded to support several additional features in the Phase-2 DAQ and TCDS design. A prototype SlinkRocket sender IP core has been developed and shown to operate at near line speed in a simple loop-back test. Current studies focus on reducing the resource utilization, and the addition of a event fragment generator for use in connection tests and for DAQ system optimization purposes.

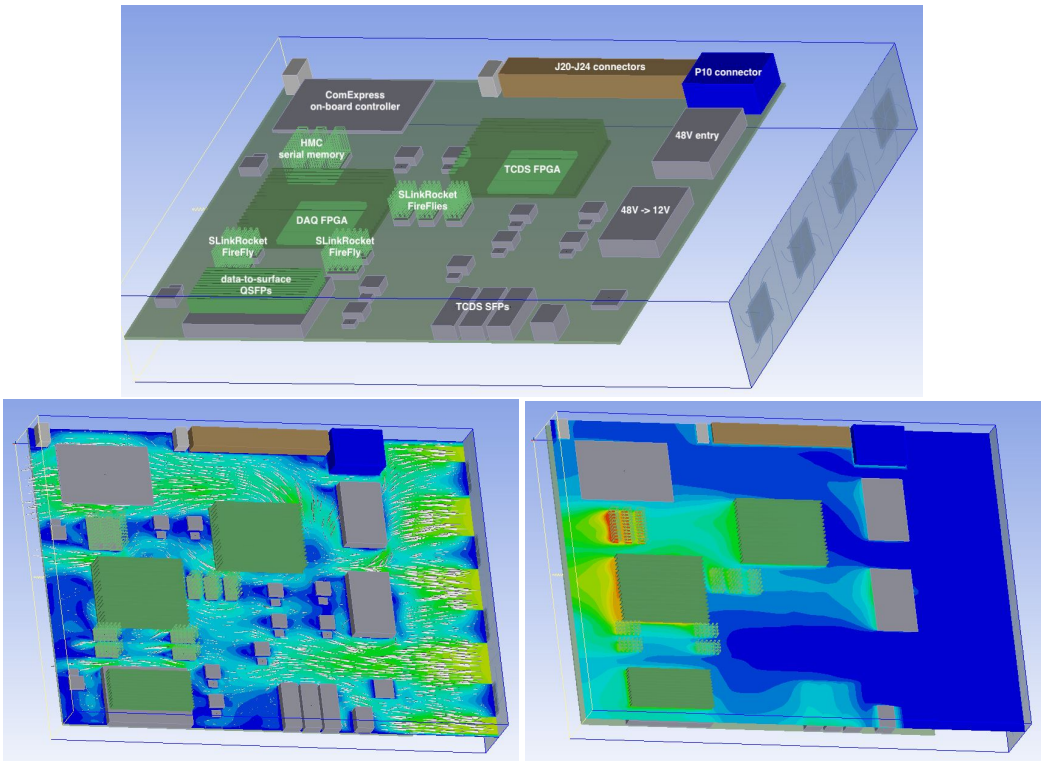

Figure 3: Top: the simplified DTH P1 thermal model includes only those components shown: the ones that contribute at least 3 W of thermal load, and those that significantly affect the airflow. The top of the board is oriented to the left. The right-hand side shows the ATCA crate cooling fans. Bottom left: the resulting airflow pattern. Bottom right: the resulting temperature distribution. Assuming an intake air temperature of $\approx 20\,^{\circ}\mathrm{C}$, the hot-spot at the top of the HMC shows a clear opportunity for improvement. (Color version available online.)

One example of resource optimization efforts is the data integrity verification approach to take. While a 32-bit CRC is the most efficient for the CPU-based end-point on the event-building side (due to the presence of dedicated instructions for this purpose), a 16-bit CRC would occupy significantly fewer FPGA resources on the sending side: O(860) LUTs and O(540) Flip-Flops vs. O(2030) LUTs and O(570) Flip-Flops. Based on the fact that the number of back-end SlinkRocket senders is much larger than the number of DTHs sending data to the D2S network, the current baseline is to use a 16-bit CRC between the back-ends and the DTH, and to add a 32-bit CRC in the DTH DAQ unit for the transmission on the data-to-surface network.

As was the case in previous generations of the CMS DAQ system, the data-to-surface network will be based on commercial hardware and standard protocols. The choice of the exact protocol and network technology is still under evaluation. The CMS DAQ group has experience with TCP/IP implementations in various FPGA families, and at multiple speeds [8, 9]. The baseline D2S technology at the time of writing is 100 Gbit/s Ethernet (four 25 Gbit/s lanes) using 100GBASE-CWDM4 interfaces to handle the approximately 250 m cable length to the computing center on the surface. Proof-of-principle has been shown in a development setup, transmitting ≈ 55 Gbit/s in a single TCP/IP stream from a Virtex UltraScale+ evaluation board to a PC, both directly and via a

network switch. The throughput in this case is limited by the receiving CPU (core). In a test with Ethernet flow control disabled and TCP/IP congestion control disabled, the outgoing throughput achieved (i.e., including retransmissions) approached 100 Gbit/s, indicating that the firmware implementation can handle the required transmission rate. A PC-to-PC test shows that the receiving PC can handle 100 Gbit/s when divided into two TCP/IP streams and handled by two CPU cores.

The DTH baseline throughput of 400 Gbit/s in a single FPGA requires this FPGA to be accompanied by a relatively large, high-bandwidth buffer memory. The DTH P1 design is based on the Hybrid Memory Cube from Micron: a high-speed serial memory with built-in memory controller. Recent changes in the Micron roadmap make this component obsolete in the near future, requiring a fundamental redesign of the DTH. Current investigations focus on comparing the design and financial consequences of HBM vs. DDR4-based solutions.

6. Summary and outlook

The DAQ and Timing Hub plays a central role in the design of the Phase-2 CMS trigger and DAQ systems, as well as in the distribution of a high-quality sampling clock to all back-end electronics.

A first prototype of the DTH should be available for testing and validation by the end of 2018/early 2019, to be followed by initial integration tests with the current generation of back-end electronics prototypes by summer 2019.

The upcoming discontinuation of the Hybrid Memory Cube forces the DTH design to be fundamentally reconsidered, in search of other high-bandwidth memory solutions.

References

[1] CMS Collaboration , S. Cittolin et al., *CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger. CMS trigger and data-acquisition project*, tech. rep., Geneva, 2002.

[2] G. Bauer et al., *Operational experience with the CMS Data Acquisition System*, Tech. Rep. CMS-CR-2012-138, CERN, Geneva, Jun, 2012.

[3] J.-M. André et al., *Performance of the new DAQ system of the CMS experiment for run-2*, tech. rep., 2016. 10.1109/RTC.2016.7543164.

[4] CMS Collaboration, *The Phase-2 Upgrade of the CMS DAQ Interim Technical Design Report*, Tech. Rep. CERN-LHCC-2017-014. CMS-TDR-018, CERN, Geneva, Sep, 2017.

[5] CMS Collaboration , J. G. Hegeman, *The CMS Data Acquisition System for the Phase-2 Upgrade*, Tech. Rep. CMS-CR-2018-099, CERN, Geneva, Jun, 2018.

[6] CERN, *S-LINK homepage*. http://hsi.web.cern.ch/HSI/s-link/.

[7] F. Vasey, *Versatile Plus Status and Plans (at ACES 2018)*. https://indico.cern.ch/event/681247/contributions/2928993/.

[8] G. Bauer et al., *10 Gbps TCP/IP streams from the FPGA for High Energy Physics*, Tech. Rep. CMS-CR-2013-402, CERN, Geneva, Nov, 2013.

[9] D. Gigi et al., *The FEROL40, a microTCA card interfacing custom point-to-point links and standard TCP/IP*, *PoS* **TWEPP-17** (2017) 075. 5 p.