Design and Prototyping of the CMS Phase-2 Trigger And Timing Distribution System

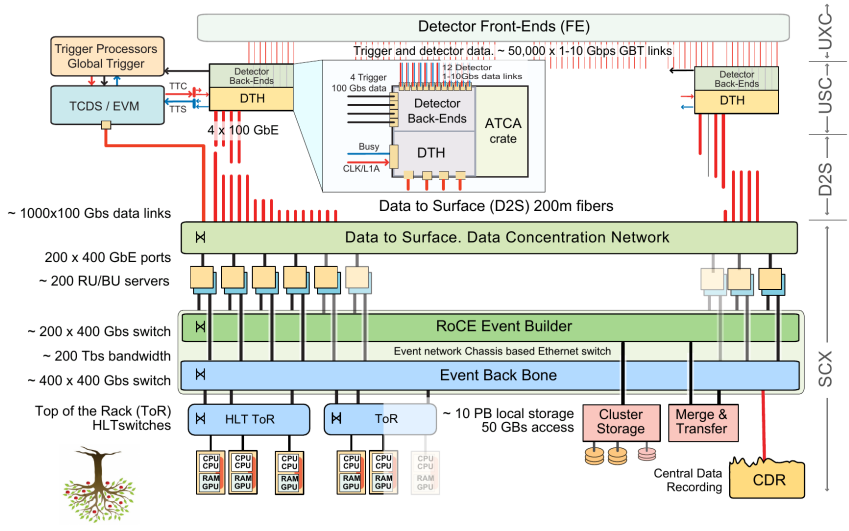
Jeroen Hegeman on behalf of the CMS DAQ project

Outline

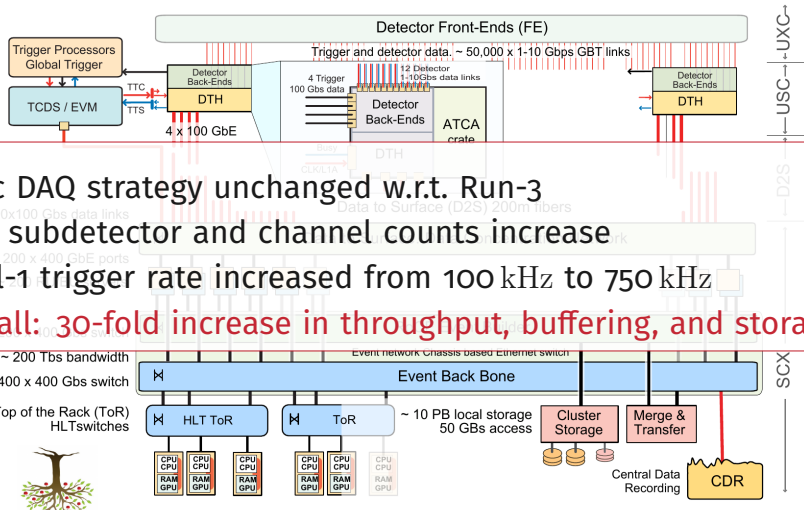
- CMS Phase-2 architecture
  - Trigger-DAQ system
  - Trigger and timing distribution
- The DTH-400 DAQ and Timing Hub and the DAQ-800 board
- TCDS2 design based on the DTH-400 and DAQ-800 boards

CMS Phase-2 architecture

CMS Phase-2 DAQ and trigger control overview

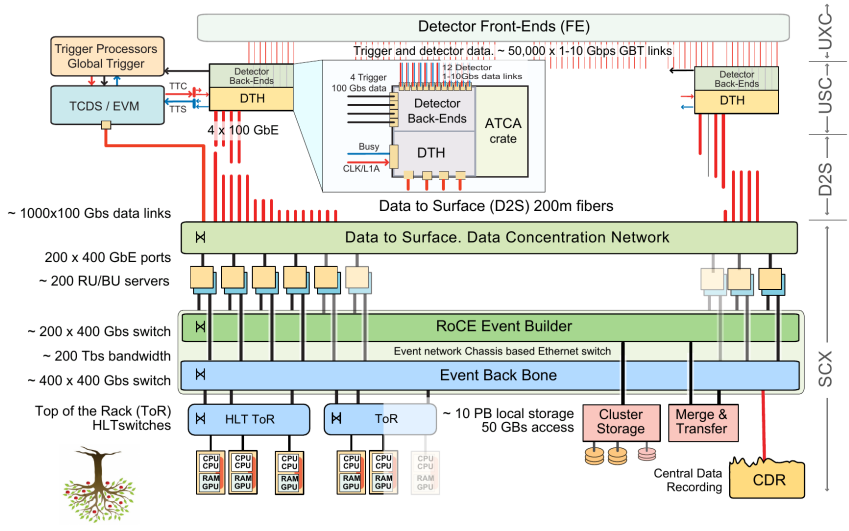


CMS Phase-2 DAQ and trigger control overview



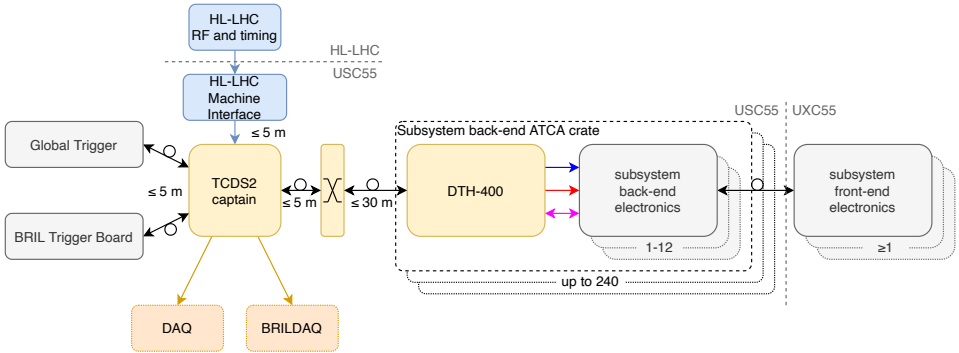
- Basic DAQ strategy unchanged w.r.t. Run-3
- Both subdetector and channel counts increase
- Level-1 trigger rate increased from 100 kHz to 750 kHz
- Overall: 30-fold increase in throughput, buffering, and storage

CMS Phase-2 DAQ and trigger control overview



CMS Phase-2 TCDS overview

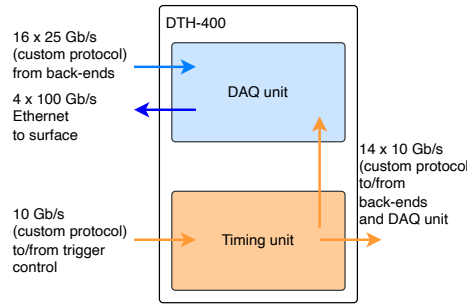
Trigger and Timing Control and Distribution System (TCDS)



The DTH-400 DAQ and Timing Hub and the DAQ-800 board

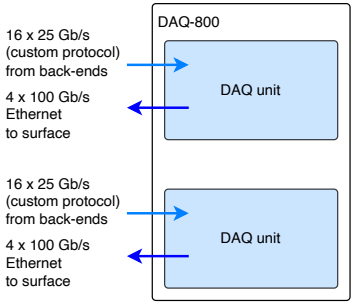
The DTH-400 DAQ and Timing Hub

- The DTH is the portal between the back-end electronics and the central DAQ, timing, and control and monitoring systems
- One DTH per back-end crate
- The DTH is equipped to drive standalone, single-crate data-taking runs for commissioning, calibration, etc.
- DTH-400 DAQ throughput: 400 $\mathrm{Gbit/s}$

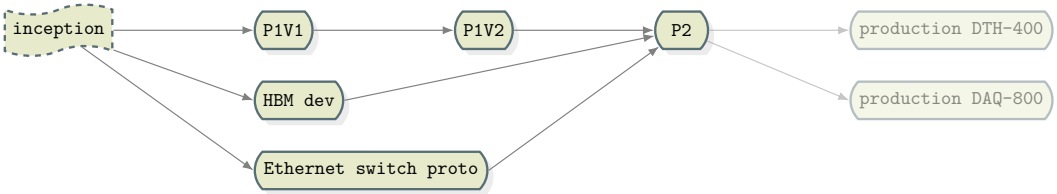


The DAQ-800 board

- Per crate, one or more DAQ-800 ‘companion boards’ can be added to increase the DAQ throughput
- DAQ-800 DAQ throughput: 800 $\mathrm{Gbit/s}$
- Can accomodate per-crate DAQ needs ranging from 10 $\mathrm{Gbit/s}$ (some muon systems) to 2.2 $\mathrm{Tbit/s}$ (inner tracker)



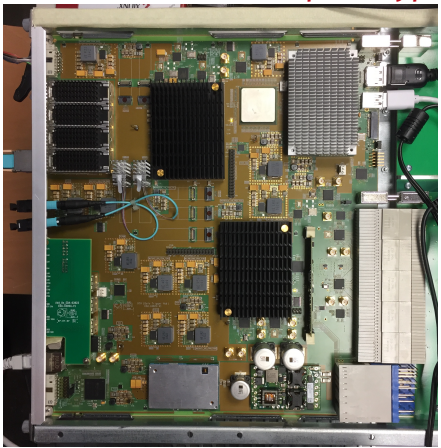
Design and prototyping of DTH-400 & DAQ-800



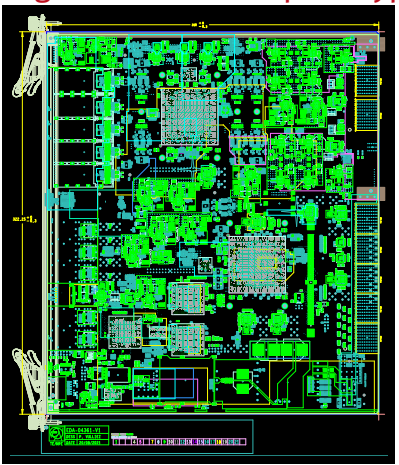
- The P1s are the main hardware validation and development platform
- 'Prototyping scatter-gather' covered all functional aspects over the last years
- The P2 merges all prototyping lines, and switches FPGAs from KU15P to VU35P, which includes 8 GB of HBM
- The P2 (with minor modifications) will become the baseline for the DTH-400 and DAQ-800 hardware production

DTH@TWEPP over the years

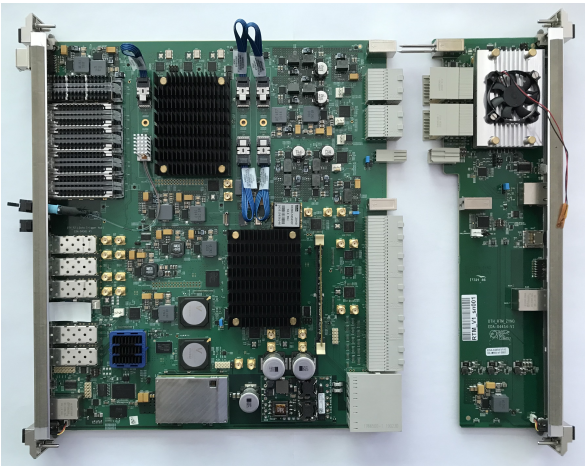
TWEPP 2019
Results from the first prototype



TWEPP 2021
Design of the second prototype

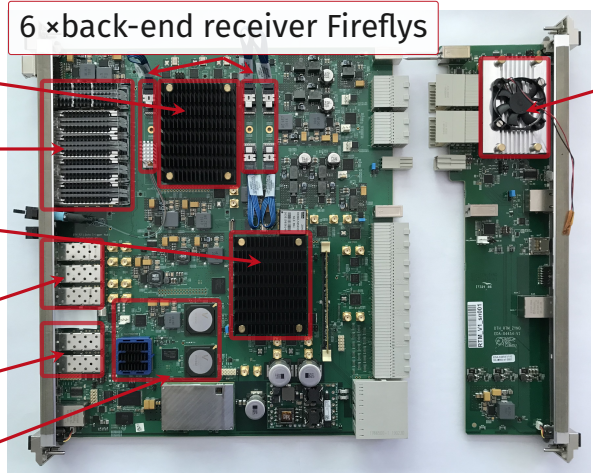


Current state-of-the-art: the DTH-P2



Expected to comfortably meet all clock quality and DAQ throughput requirements for Phase-2 CMS

Current state-of-the-art: the DTH-P2



Expected to comfortably meet all clock quality and DAQ throughput requirements for Phase-2 CMS

TCDS2 design
based on the DTH-400 and DAQ-800 boards

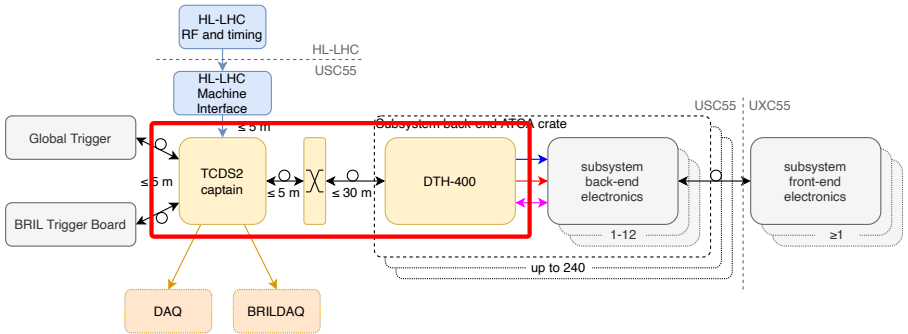
Can DAQ (hardware) build a timing system?

Challenge: (re)design the DTH-400 and DAQ-800 such that they can *also* be used to implement the central part of the Phase-2 TCDS

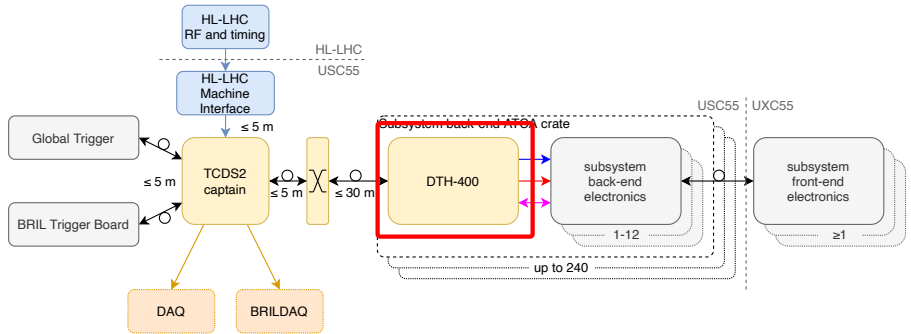
This would reduce the number of different board designs by one or two, and hence

- reduce design effort,
- reduce maintenance effort, and
- reduce engineering and prototyping cost.

CMS Phase-2 trigger control architecture



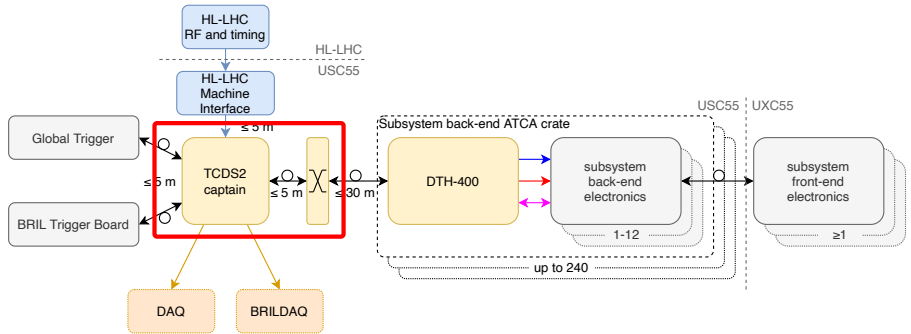
CMS Phase-2 trigger control architecture



The DTHs:

- Connect all CMS back-end crates to the central trigger, DAQ, and control systems

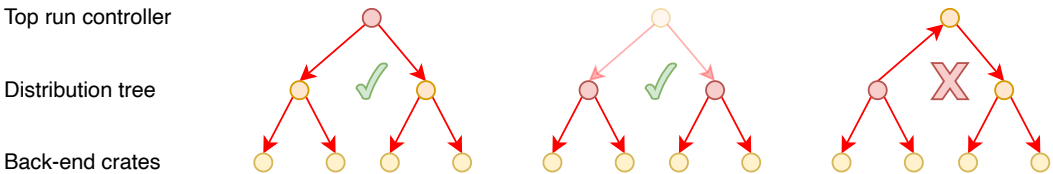
CMS Phase-2 trigger control architecture



The TCDS2 captain:

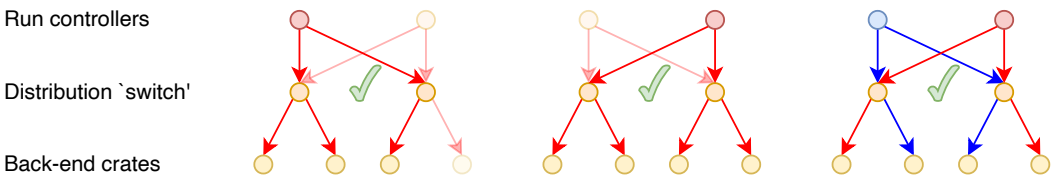
- Houses several firmware ‘run controllers’ to drive data-taking runs
- Contains a configurable ‘switch’ to assign groups of CMS back-ends to these runs

A switch or a tree?



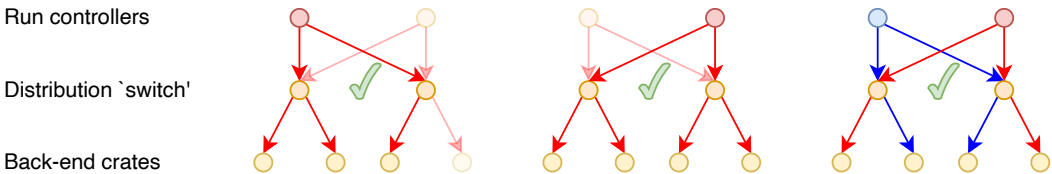
- Simultaneous runs with different subdetectors are necessary for commissioning, calibration, etc.
- Only the top-level run controller can reach all end-points
- Each sub-level run controller can reach a *fixed* subset of end-points
- Ad hoc changes in subsets require recabling

A switch or a tree?



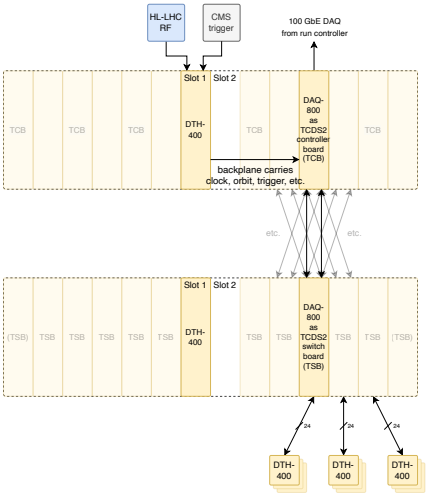
- Flexibility increases by relocating run controllers outside the distribution layer
- Each top-level run controller can reach all end-points, in any arbitrary combination
- Subset assignment is now ‘just configuration’
- This achieves full flexibility for many simultaneous data-taking runs

A switch or a tree?



- ! The number of end-points (O(160)), plus the full-configurability requirement, requires the switch be distributed across multiple nodes
- ! This architecture requires a full mesh network connecting all controller nodes to all switch nodes
- ! The actual implementation also needs to gather end-point status information and deliver that to the corresponding run controllers

Using the DAQ-800 to implement the TCDS



- Two layers of DAQ-800: one with run controllers, one as ‘distributed switch’
- Use the ‘back-end data’ Fireflys to mesh-interconnect the controller boards and the switch boards
- Use the ‘DAQ QSFPs’ to connect the switch to the DTHs
- Number of run controllers scales with the number of controller boards
- The number of end-points scales with the number of switch boards

The determining scale factor appears to be the FPGA resources required to implement each N × M (sub)switch

Using the DAQ-800 to implement the TCDS

The current Phase-2 TCDS design aims to:

- implement the run controllers on DAQ-800 boards
- connect the run controllers to the subsystem DTHs via a distributed switch implemented on DAQ-800 boards
- connect the run controllers and the switch nodes to the LHC RF and the CMS trigger using DTHs with dedicated firmware

Some ingenuity is needed for the ‘dual use’ of the DAQ-800

- The four-fold DAQ-optimised optical connectivity will be adapted to the many-to-many TCDS mesh network using custom shuffle fibres
- The FPGA choice is driven by the DAQ need for the buffer HBM, with less need for basic logic
  TCDS firmware may have to adapt to the available (types of) resources

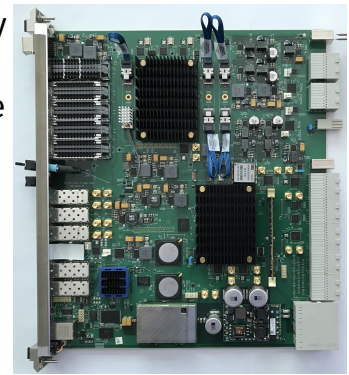
Using the DAQ-800 to implement the TCDS

An ongoing study, using the first DTH-P2 board, should soon point the way to the optimal implementation for the TCDS2 captain

- Baseline implementation of run controllers
- Number of switch nodes
- Maximum number of end-points
- Etc.

Closing words

- The design of the CMS central DAQ hardware, both the DTH-400 and DAQ-800, is approaching the final production designs
- All initial DTH-400 prototypes meet clock quality DAQ throughput requirements
- First studies look promising for the re-use of the DAQ hardware for the implementation of the trigger control system
  - Greatly reduces the engineering effort, as well as the engineering and development cost
  - Does require some small un-DAQ-like additions
  - Will involve some level of compromise on the TCDS side. Studies should show how much.



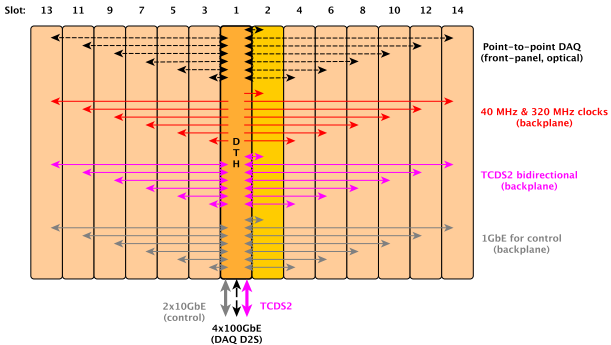
Phase-2 CMS DAQ in numbers

Bottom line: high rate and enormous throughput

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

CMS Phase-2 DAQ and Timing Hub (DTH)

- ATCA baseboard handling power, IPMC, etc., including on-board controller
- Managed Ethernet switch to all node slots and both shelf managers
- Timing and control unit handling clock recovery, cleaning, and distribution
- DAQ unit converting from custom back-end links to commercial Ethernet



Using the DAQ-800 to implement the TCDS