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

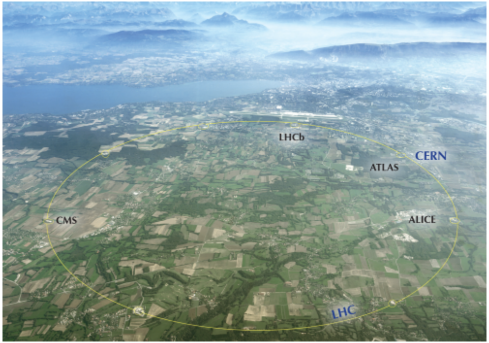
Jeroen Hegeman on behalf of the CMS DAQ project

Outline

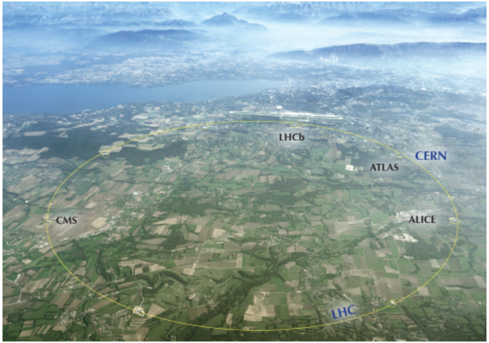
- The CMS experiment at the CERN LHC
- The CMS Phase-2 DAQ system and the DAQ and Timing Hub
- Design once, use in multiple places?

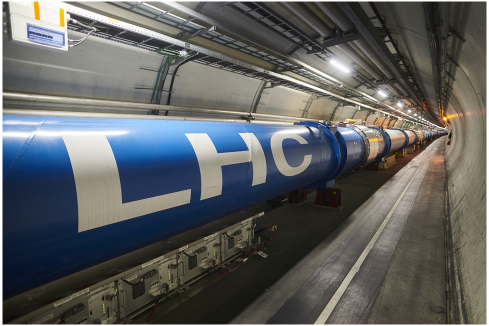
The CMS experiment at the CERN LHC

CMS is one of the experiments at the CERN LHC

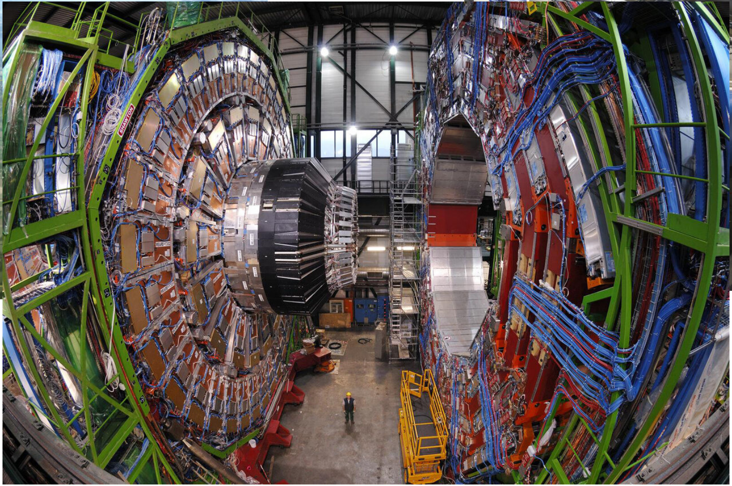


CMS is one of the experiments at the CERN LHC

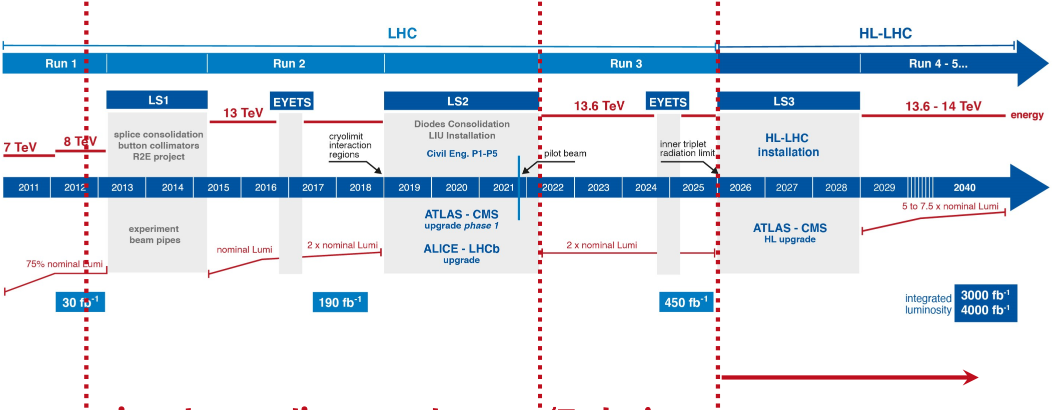




CMS is one of the experiments at the CERN LHC



The life and times of CMS and the LHC

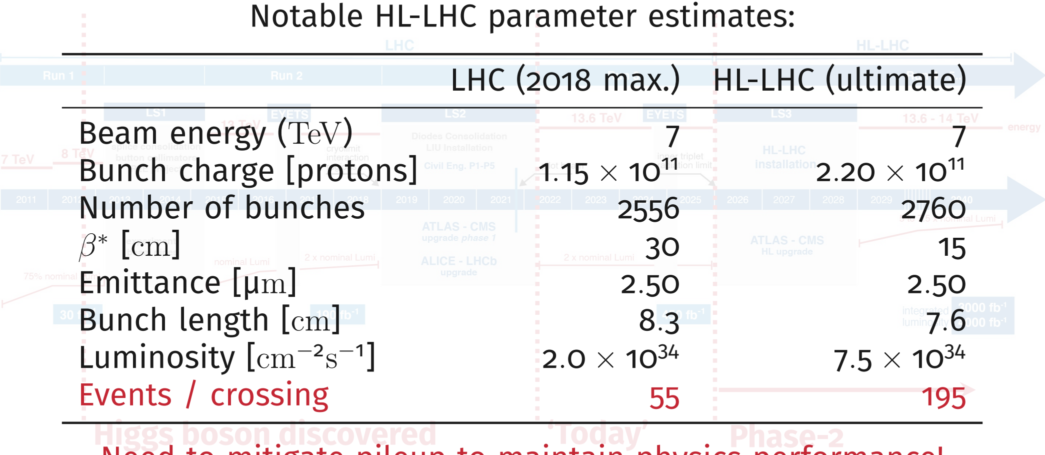


Higgs boson discovered

'Today'

Phase-2

The life and times of CMS and the LHC



Notable HL-LHC parameter estimates:

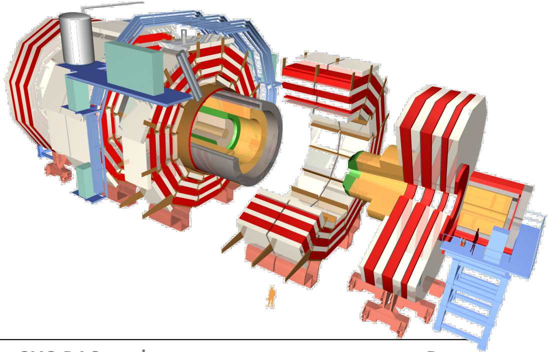
	LHC (2018 max.)	HL-LHC (ultimate)
Beam energy (TeV)	7	7
Bunch charge [protons]	1.15×10^{11}	2.20×10^{11}
Number of bunches	2556	2760
β^* [cm]	30	15
Emittance [μm]	2.50	2.50
Bunch length [cm]	8.3	7.6
Luminosity $[\mathrm{cm}^{-2}\mathrm{s}^{-1}]$	2.0×10^{34}	7.5×10^{34}
Events / crossing	55	195

Need to mitigate pileup to maintain physics performance!

The Phase-2 upgrade of the CMS experiment

Complete overhaul of the CMS detector:

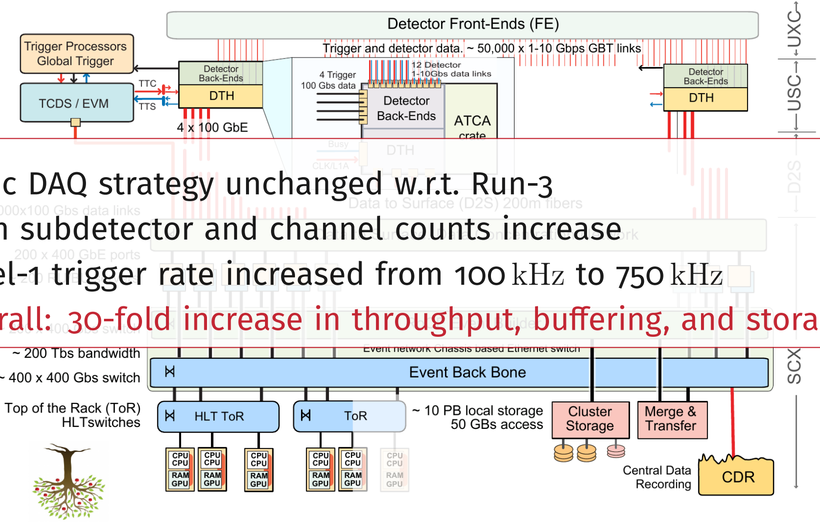
- Full redesign and rebuild of pixel and strip trackers
- Addition of MIP Timing Detector, between tracker and calorimeter
- Replacement of end-cap calorimeters with high-granularity (silicon + scintillator) ones
- Level-1 trigger latency increases from 4.3 μs to 12.4 μs
- Replacement of barrel calorimeter front-end electronics
- All muon systems receive 'minor' upgrades to stay in step with latency and technology



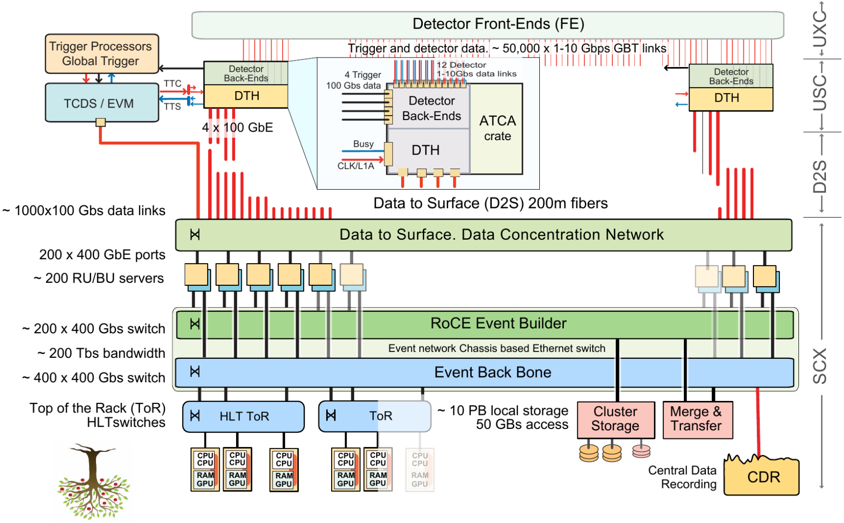
The CMS Phase-2 trigger-DAQ system and the DAQ and Timing Hub

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



Detector Front-Ends (FE)

Trigger and detector data. ~ 50,000 x 1-10 Gbps GBT links

Trigger Processors
Global Trigger

TCDS / EVM

TTC

TTS

Detector Back-Ends

DTH

4 x 100 GbE

4 Trigger
100 Gbs data

12 Detector
1-10Gbs data links

Detector Back-Ends

ATCA crate

DTH

Busy

CLK/L1A

Data to Surface (D2S) 200m fibers

UXC

USC

D2S

SCX

~ 1000x100 Gbs data links

Data to Surface. Data Concentration Network

200 x 400 GbE ports

~ 200 RU/BU servers

~ 200 x 400 Gbs switch

~ 200 Tbs bandwidth

~ 400 x 400 Gbs switch

RoCE Event Builder

Event network Chassis based Ethernet switch

Event Back Bone

Top of the Rack (ToR)
HLTswitches

HLT ToR

ToR

CPU CPU RAM GPU

~ 10 PB local storage
50 GBs access

Cluster Storage

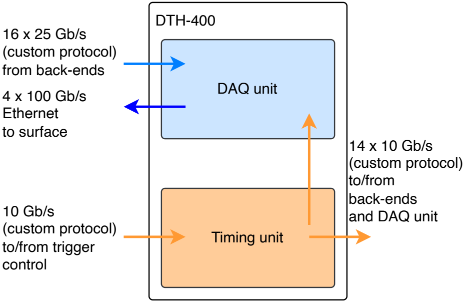
Merge & Transfer

Central Data Recording

CDR

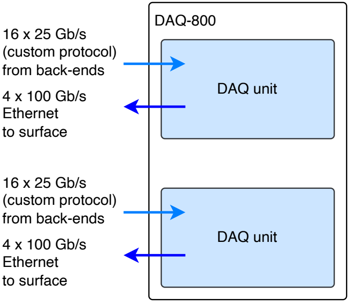
The DTH-400 DAQ and Timing Hub

- 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 $\mathrm{Gbit/s}$


DTH-400
16 x 25 Gb/s (custom protocol) from back-ends
DAQ unit
4 x 100 Gb/s Ethernet to surface
14 x 10 Gb/s (custom protocol) to/from back-ends and DAQ unit
10 Gb/s (custom protocol) to/from trigger control
Timing unit

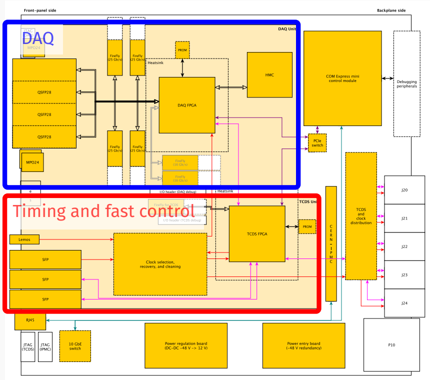
The DAQ-800 node board

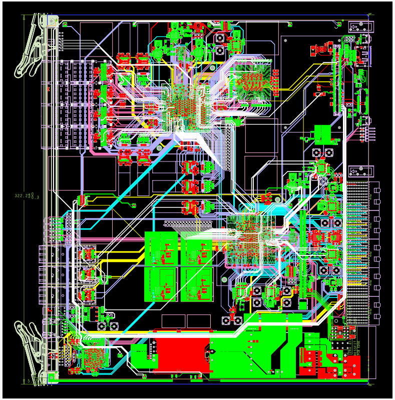
- Per crate, one or more DAQ-800 ‘companion boards’ can be added to increase the DAQ throughput
- DAQ-800 DAQ throughput: 800 $\mathrm{Gbit/s}$
- Can accomodate per-crate DAQ needs ranging from 10 $\mathrm{Gbit/s}$ (some muon systems) to 2.2 $\mathrm{Tbit/s}$ (inner tracker)


DAQ-800
16 x 25 Gb/s (custom protocol) from back-ends
DAQ unit
4 x 100 Gb/s Ethernet to surface
16 x 25 Gb/s (custom protocol) from back-ends
DAQ unit
4 x 100 Gb/s Ethernet to surface

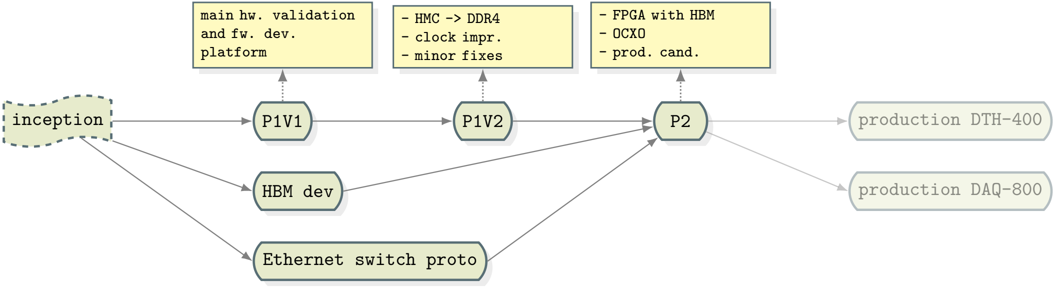
Flashback to Real Time 2018

DAQ and Timing Hub (DTH)



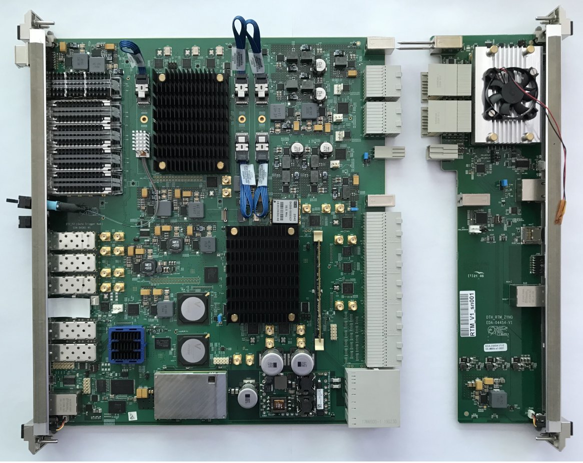


Design and prototyping of DTH-400 & DAQ-800



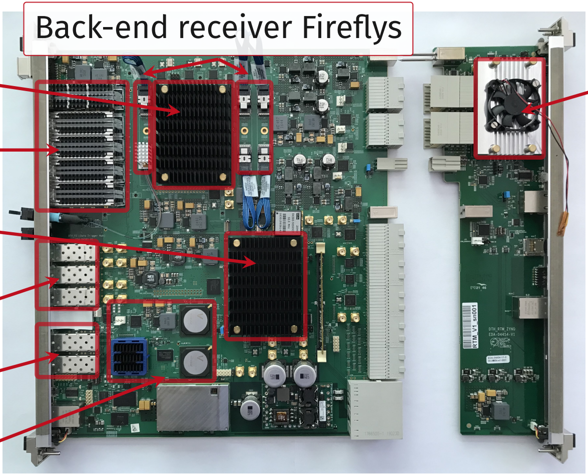
- The P2 merges all prototyping lines, and switches FPGAs from KU15P to VU35P
- Adopted in-FPGA High-Bandwidth Memory for Ethernet buffering
- The DAQ-800 is a ‘creative copy-paste’ of the DTH-400

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



Comfortably meets clock quality and
DAQ throughput requirements for Phase-2 CMS

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



Back-end receiver Fireflys

DAQ FPGA

5 × DAQ QSFP28

Timing FPGA

3 × timing I/O SFP+

2 × 10 GbE (SFP+)

Ethernet switch

Zynq-based controller

Comfortably meets clock quality and DAQ throughput requirements for Phase-2 CMS

Design once, use in multiple places?

A new kind of optimisation challenge

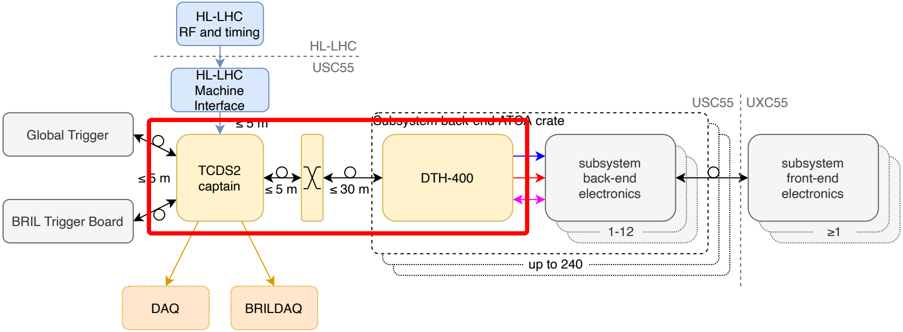
Driven by a wish to

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

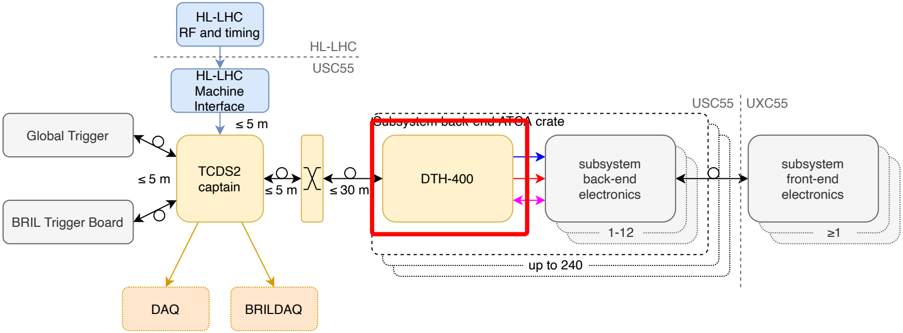
we were prompted to consider designing the Phase-2 DAQ hardware such that it could also serve for the Trigger and Timing Control and Distribution System (TCDS).

To note: this challenge was posed at the right time,
i.e., during the design phase

CMS Phase-2 trigger control architecture



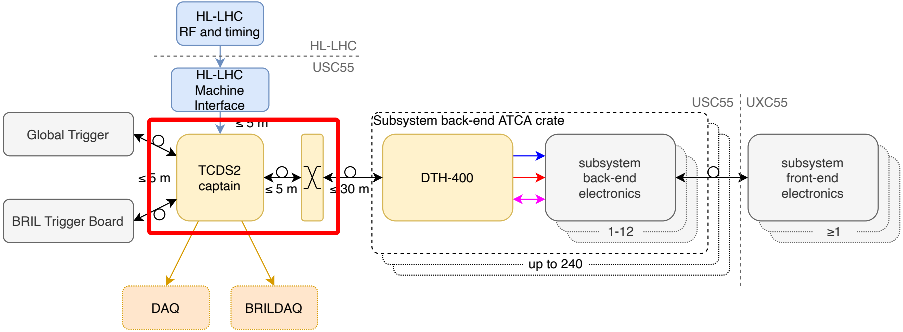
CMS Phase-2 trigger control architecture



The DTHs:

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

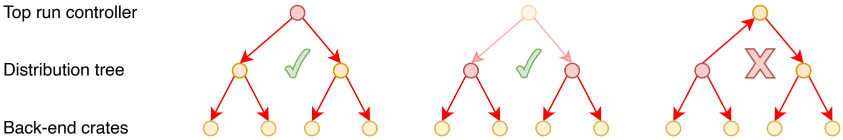
CMS Phase-2 trigger control architecture



The TCDS2 captain:

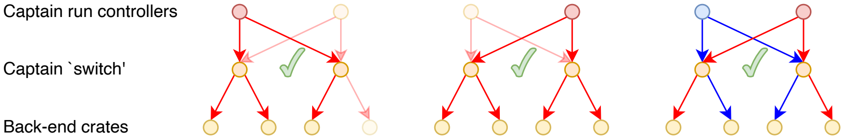
- Houses several firmware ‘run controllers’ to drive data-taking runs
- Contains a configurable ‘switch’ to assign groups of CMS back-ends to these runs

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 controllers can reach a *fixed* subset of end-points
- Ad hoc changes in subsets require recabling

A switch or a tree?



Captain run controllers

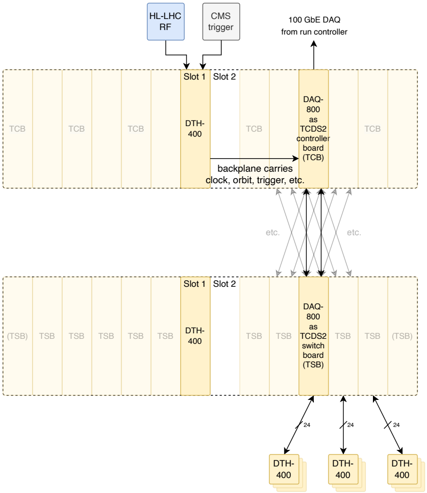
Captain `switch'

Back-end crates

- 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

So a switch it is, then!

Using the DAQ-800 to implement the TCDS


HL-LHC RF
CMS trigger
100 GbE DAQ from run controller
Slot 1
Slot 2
TCB
TCB
TCB
DTH-400
TCB
DAQ-800 as TCDS2 controller board (TCB)
TCB
backplane carries clock, orbit, trigger, etc.
etc.
etc.
Slot 1
Slot 2
(TSB)
TSB
TSB
TSB
TSB
TSB
DTH-400
TSB
TSB
DAQ-800 as TCDS2 switch board (TSB)
TSB
TSB
(TSB)
24
24
24
DTH-400
DTH-400
DTH-400

- 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 good (which is beyond question)

Removes the need for a separate design, production, spares, etc.

The ‘bad’ (which complicates life)

The needs of a DAQ system are largely orthogonal to those of a timing/control system

- The DAQ functionality hinges on the High-Bandwidth Memory, a control system benefits more from logic resources
- The DAQ profits from high-density optics, e.g., CWDM QSFP28s, and the architecture of a timing distribution system is all single point-to-point links

Reusing back-end or trigger boards has similar trade-offs

The ugly (which makes it possible)

- Optics connectivity can be addressed with break-out fibres
- Firmware can be written with narrow(er) counters, latching and using the HBM to buffer, and the software can gather and post-process

Closing words

- The CMS central DAQ hardware, both the DTH-400 and DAQ-800, is well on its way towards Phase-2
- The DTH-400 prototypes meet clock quality and 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.

http://cms.cern.ch

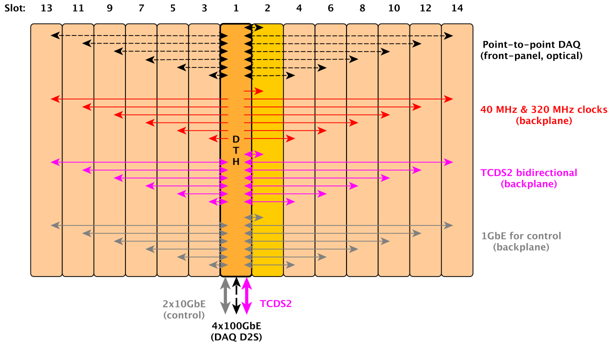
Phase-2 CMS DAQ in numbers

Bottom line: high rate and enormous throughput

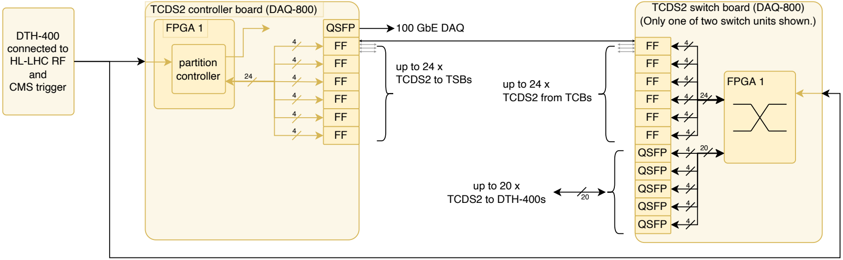
CMS detector	Phase-1	Phase-2	
Peak average pileup	60	140	200
L1 accept rate (max.)	100 kHz	500 kHz	750 kHz
Event size at HLT input	2.0 MB	7.8 MB	9.9 MB
Event network throughput	1.6 Tbit/s	31 Tbit/s	60 Tbit/s
Event network buffer (60 s)	12.0 TB	234 TB	445 TB
HLT accept rate	1.0 kHz	5.0 kHz	7.5 kHz
HLT 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


Slot: 13 11 9 7 5 3 1 2 4 6 8 10 12 14
D T H
Point-to-point DAQ (front-panel, optical)
40 MHz & 320 MHz clocks (backplane)
TCDS2 bidirectional (backplane)
1GbE for control (backplane)
2x10GbE (control)
TCDS2
4x100GbE (DAQ D2S)

Using the DAQ-800 to implement the TCDS


DTH-400 connected to HL-LHC RF and CMS trigger
TCDS2 controller board (DAQ-800)
FPGA 1
partition controller
24
4
QSFP
FF
FF
FF
FF
FF
FF
100 GbE DAQ
up to 24 x TCDS2 to TSBs
up to 24 x TCDS2 from TCBs
up to 20 x TCDS2 to DTH-400s
20
TCDS2 switch board (DAQ-800) (Only one of two switch units shown.)
FF
FF
FF
FF
FF
FF
QSFP
QSFP
QSFP
QSFP
QSFP
FPGA 1
24
20
4