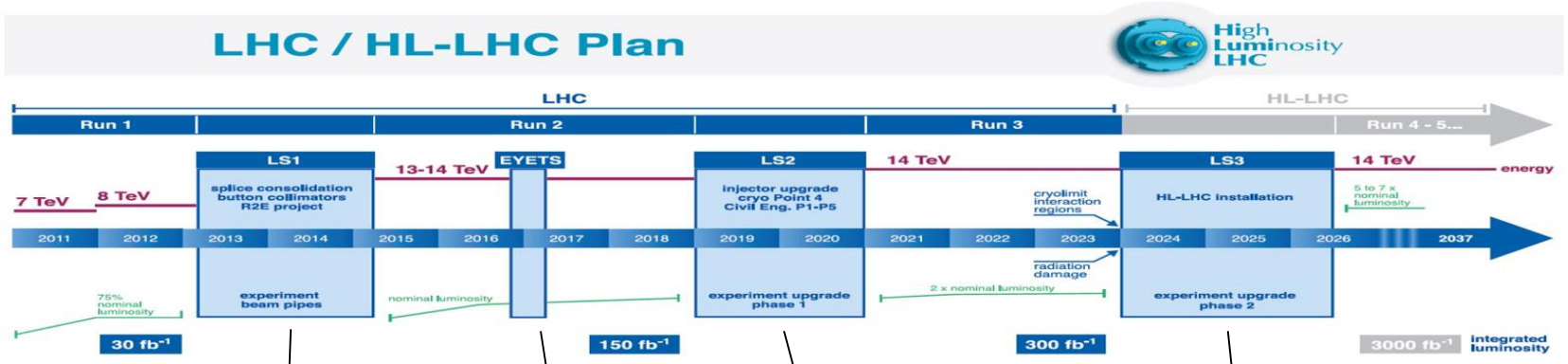
CMS DAQ (past) current and future hardware upgrades

up to post Long Shutdown 3 (LS3) times

attila.racz@cern.ch on behalf of CERN / EP-CMD group

Outline



DAQ1
in 2009

LS1 upgrade:
DAQ2

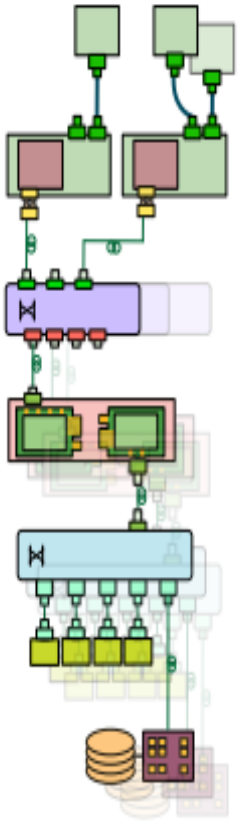
New Pixel

LS2 upgrade:
DAQ3

LS3 upgrade:
DAQ4

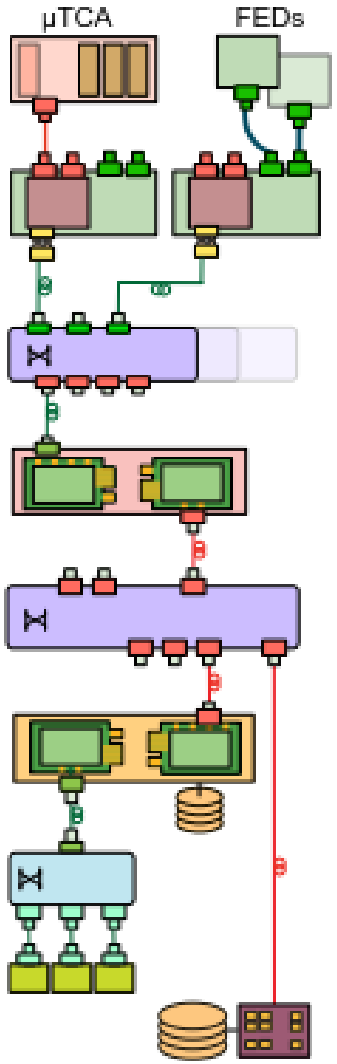
- Upgrades are necessary because of
  - Change of requirements : more data, more luminosity, more pile-up, more analysis
  - Obsolescence (failure rate)
  - Technological evolution : HS06-Gbs per Watt / $$ / m2 / U

Initial CMS DAQ 1 features : 2009



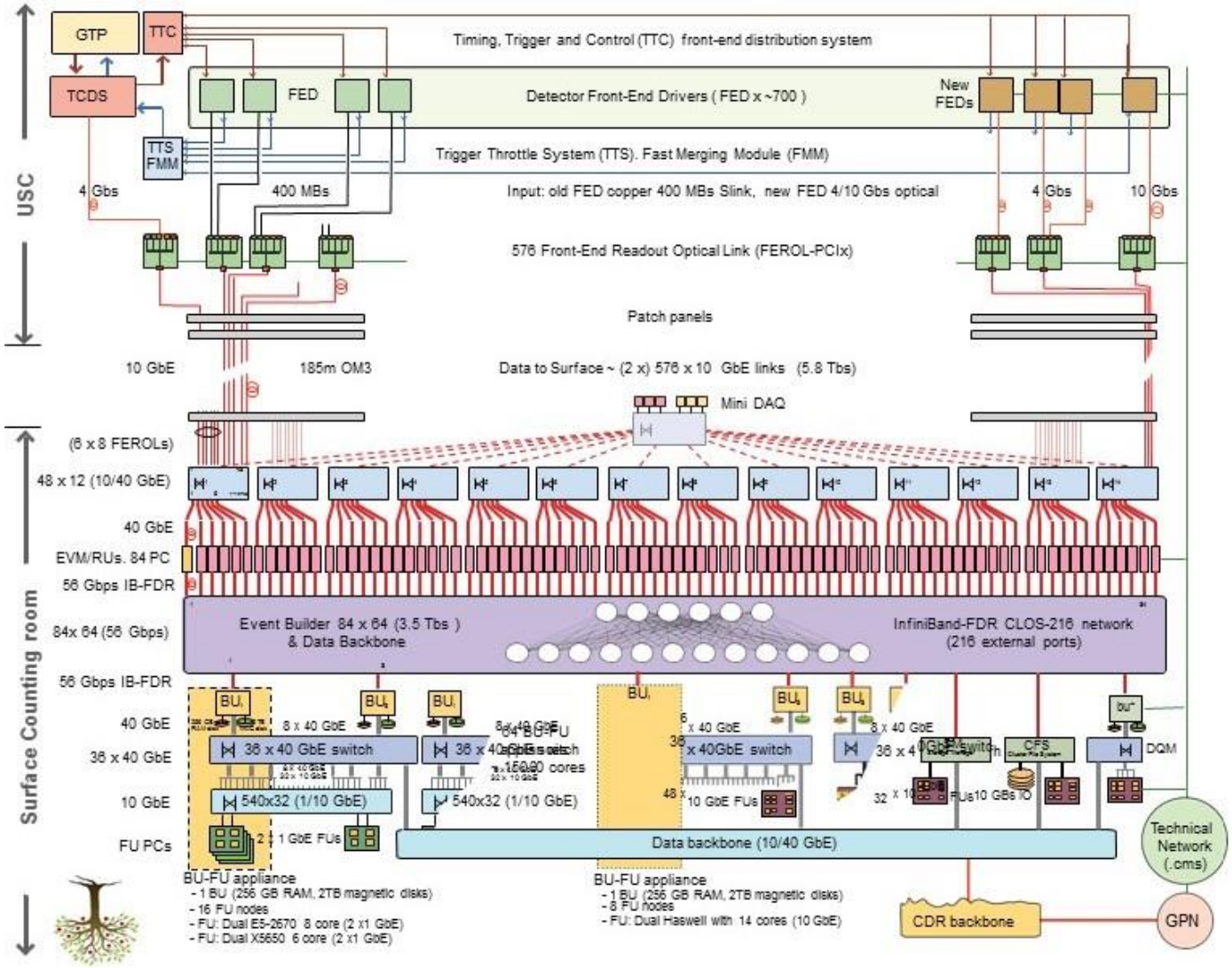
- 1MB per event, 100 kHz LV1 trigger rate, 1 Tb/s into central DAQ
- ~650 VME data sources readout by ~500 custom DAQ card (FRL)
  - max average event fragment size 2kB
- FRL input : Slink64 (parallel copper), FRL output optical Myrinet 2.5 Gb/s
  - Commercial Myrinet NIC housed and driven by FPGA on FRL
- 2 stages event builder : single stage was not viable technologically
  - Super fragment builder 12x 256-ports Myrinet switches
  - Full event builder : 8x 540-ports GbE switches (each of them being a slice @ 12.5kHz)
- High Level Triggers (online selection) performed by PC farm
  - Started with 53 kHS06 in 2009 up to 200 kHS06 in 2012, last year of Run1
- ~100-500 Hz events to permanent storage

First upgrade (2013-2015): DAQ 2

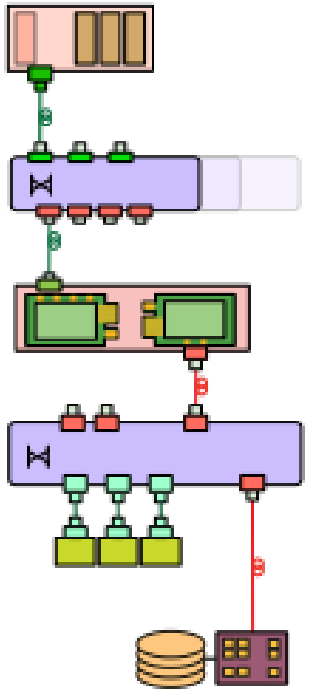


- Up to 2MB per event, same trigger rate, 2Tb/s max to cDAQ
- Replacement of Myrinet NIC on FRL by a custom design (FEROL)
  - 10 GbE TCP/IP output (full FPGA) allowing up to 4kB fragment size
  - 2x SlinkExpress 10 Gb/s inputs
  - TCP/IP allows much more flexibility for DAQ bw optimization and hardware choice
- Additional ~50 uTCA sources connected via SlinkExpress 10Gb/s
- Winter 16/17 : new Pixel detector with 112 uTCA sources
  - Ferol40 developed on purpose : 40x 10Gb/s SlinkExpress IN, 1x 40Gb/s TCP/IP OUT
- 10/40 GbE switch layer for stream aggregation into event builder
- Single stage event builder based on InfiniBand FDR techno : 56 Gb/s
  - 216 ports, made out of 18x 36-ports switches connected in a clos network
- High Level Trigger PC farm : ~350 kHS06 at restart in 2015
  - ~600 kHS06 today
- ~1000 Hz events to permanent storage

DAQ2 Block Diagram



Second upgrade (2019-2020): DAQ 3



- Little or no evolution of the detector data sources
  - No change in the "Data to Surface" (D2S) section of the DAQ
  - ~650 streams of 10GbE incoming into cDAQ with variable payload : need of a balancing layer
  - Very likely 10GbE/100GbE set of switches
- Event builder network at 100Gb/s line speed (maybe 200)
  - Candidates are InfiniBand or OmniPath
  - Chassis-based or composite TOR, we don't know : cost and reliability will be the driver
- If new CPU generation allows it (Skylake), Readout Unit and Builder Unit functions will be in a single server
  - In DAQ1 and DAQ2, they were 2 different servers
  - Fewer Event Builder ports needed, fewer servers, fewer racks, less power, etc…
  - But more demanding on I/O capability, internal data flow
  - Very delicate software optimization: core assignment, interrupts, PCIe lanes, NUMA

CMS DAQ for HL-LHC (mid 2026-….)

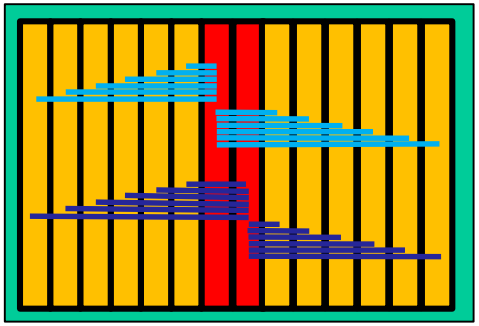
Item	DAQ1/DAQ2	Post LS3 - DAQ4	
Peak Pile-up	40 -> ~60	140	200
Level 1 rate max. (kHz)	100	500	750
Luminosity ($cm^{-2}s^{-1}$)	1 -> 2 $.10^{34}$	5 $.10^{34}$	7.5 $.10^{34}$
Event size (MB)	1 -> 1.5	5.7	7.4
Event network bw (Tb/s)	1 -> 2	23	44
HLT compute power (HS06)	53k -> 500k	4.5M	9.2M

- Many detectors replaced, but also additional detectors
- New front-end and back-end electronics (ATCA based)

- Readout requirements for cDAQ completely new

HL-LHC / DAQ4 : Data to Surface (D2S) HW

- ATCA is the hardware platform of choice for post LS3 back-end electronic



- Usually, 12 blade slots and 2 hub slots + backplane
- Each blade is connected to each hub via 2 (base) + 4 (fabric) pairs
- A pair is now specified for up to 25 Gb/s
- 100 Gb/s max from any blade to one of the hubs

- It is natural to think about a standard DAQ hub located in the hub slots of each shelf to perform :
  - Data aggregation for optimal available bandwidth usage
  - Translation to a standard switched protocol (i.e. TCP/IP)
  - Distribution of Timing and Control signals/messages to all BE cards present in the shelf
  - Collection and (pre)processing of the individual board throttling status for fast monitoring and statistics

- The name of this board is the DTH for DAQ and TCDS Hub
  - TCDS stands for Timing and Control Distribution System

DAQ4 Readout Requirements Table 1

Sub Detector	Sub Event Size (MB)	Back-End leaf-cards	Back-End crates	Av Throughput* (Tb/s)
Inner TK	1.44	24	4	8.64
Outer TK	1.15	216	18	6.60
Track Trigger	0.01	-	18	0.06
MIP Timing	0.06	16	2	0.36
ECAL Barrel	1.58	108	12	9.49
HCAL Barrel	0.24	18	2	1.45
EndCap Calo	2.00	108	9	12.00
EndCap Calo TPG	0.20	144	12	1.50
HCAL-HO	0.03	-	1	0.18
HCAL-HF	0.06	-	1	0.36
Muon DT	0.13	84	8	0.78
Muon CSC	0.20	-	2	1.20
Muon GEM	0.12	20	2	0.73
Muon RPC	-	-	-	-
Level1 Trig	0.15	120	14	0.9
Total	7.4	>858	>106	44

*at 750 kHz LV1A

DAQ4 Readout Requirements Table 2

Sub Detector	Av Throughput (Tb/s)	Av Throughput per BE crate (Tb/s)	Av Throughput per BE blade (Gb/s)
Inner TK	8.64	2.16	360
Outer TK	6.90	0.38	32
Track Trigger	0.06	-	<10
MIP Timing	0.36	0.12	12
ECAL Barrel	9.49	0.79	88
HCAL Barrel	1.45	0.73	81
EndCap Calo	12.00	1.71	111
EndCap Calo TP	1.60	0.13	11
HCAL-HO	0.12	0.12	-
HCAL-HF	0.30	0.30	-
Muon DT	0.78	0.10	8
Muon CSC	1.20	-	<100
Muon GEM	0.31	-	<100
Muon RPC	-	-	-
Level1 Trig	0.94	0.07	8
Other	0.63	0.21	31

We observe that the range for DTH inputs and outputs are large… x45 and x30

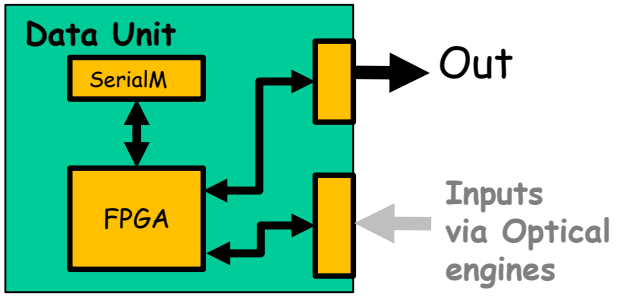
For some blades, 100Gb/s is not enough : using the backplane is not adequate

For blades with very little data, we should route all blades to a single FPGA versus few blades to multiple FPGAs when large amount of data

For global (sub+DAQ) cost optimization reasons, a single DTH design is not adequate. We have to go modular

Design considerations for the Data section of DTH

- 1 large FPGA versus several "little" FPGAs on PCB ?
  - If application allows it, much more interesting cost-wise to have several FPGAs
- Using Backplane for data inputs or from Front Panel ?
  - BP limits transfer to 100Gb/s (not enough for some sub-detectors)
  - Need to use some backplane lines for timing distribution
  - Requires to have multiple DTH designs for high and low throughput
  - Signal integrity through a backplane is challenging at 25 Gb/s
  - Front panel input allows the use of optical engines (i.e. FireFly)
- Protocol translation requires large amount of fast memory
  - Currently, TCP streams require at least 10ms buffering (round-trip and congestion window)
  - We look at serial memories
- FPGAs are divided into 2 families
  - With serdes <= 16G : cheap
  - With serdes > 16G : affordable if not too big…

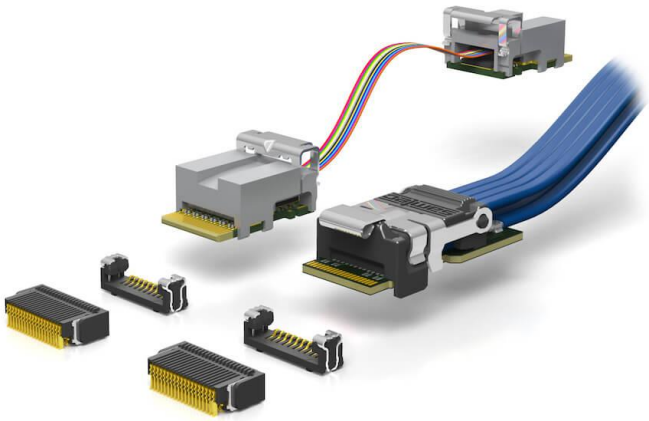


FPGA for the Data Unit

- The winner is, for us today, the KU15P-2 from Xilinx
- Usage table of the KU15P-2 serdes :

KU15P	Available	PCIGen3	TCDS	Serial Memory	4x100G Out	Serial In	Remaining
@28G	32	0	0	0	16	16	0
@16G	44	1	2	32	0	8	1

- For optical engines, we chose the FireFly technology up to 28G per link
  - We connect 6 chips of 4 Rx/Tx
  - We hope that standardization will become a reality (COBO)



- The Data Unit is hence capable of :
  - 4x 100 GbE TCP/IP out
  - 16 inputs at up to 28G and 8 inputs at up to 16G within the limit of total 400G
  - Preliminary studies shows that we can have 3 Data Units on a ATCA board
- One Data Unit : DTH400, 3 Data Units : DTH1200

DTH flexibility

- The DTH can have as little as 1x 100G output
  - 1 Data Unit equipped with only 1 optic
- As inputs, it can be anything from "almost nothing" up to the output capacity
  - FireFly sites can be left unpopulated if needed
  - Up to 24 input links per unit (8 @ 16Gb/s max and 16 @ 28Gb/s max)
  - With 16G input links only (cheap FPGA), 384 Gb/s can be reached per Data Unit (96% of the theoretical output bandwidth)
- The DTH can go as high as 1.2 Tb/s output and 72 input links at 16/28G max
  - With 3 Data Units, all 12 optics and all FireFly sites populated
  - Due to high number of fibers, 3x MPO-24 connectors per Data Unit are used on front panel
  - MPO-24 <-> 2x MPO-12 breakout cables are commonly available

attila.racz@cern.ch

DTH400/DTH1200 in DAQ4

Sub Detector	BE crates	DTH1200 Per crate	DTH400 Per crate	DTH1200 Per subdet	DTH400 Per subdet	Min. D2S links per crate	Min D2S links per sub
Inner TK	4	2	-	8	-	24	96
Outer TK	18	1	-	18	-	4	72
Track Trigger	18	-	1	-	18	1	18
MIP Timing	2	-	1	-	2	2	5
ECAL Barrel	12	1	-	12	-	9	108
HCAL Barrel	2	1	-	2	-	9	18
EndCap Calo	9	2	-	18	-	24	168
EndCap Calo TPG	12	-	1	-	12	2	24
HCAL HO	1	-	1	-	1	2	2
HCAL HF	1	-	1	-	1	4	4
Muon DT	8	-	1	-	8	1	8
Muon CSC	2	-	1	-	2	6	12
Muon GEM	3	-	1	-	3	4	12
Muon RPC	-	-	-	-	-	-	-
Level1 Trig	14	-	1	-	14	1	14
Other	6	-	1	-	6	1	6
Total	113			58	68		560

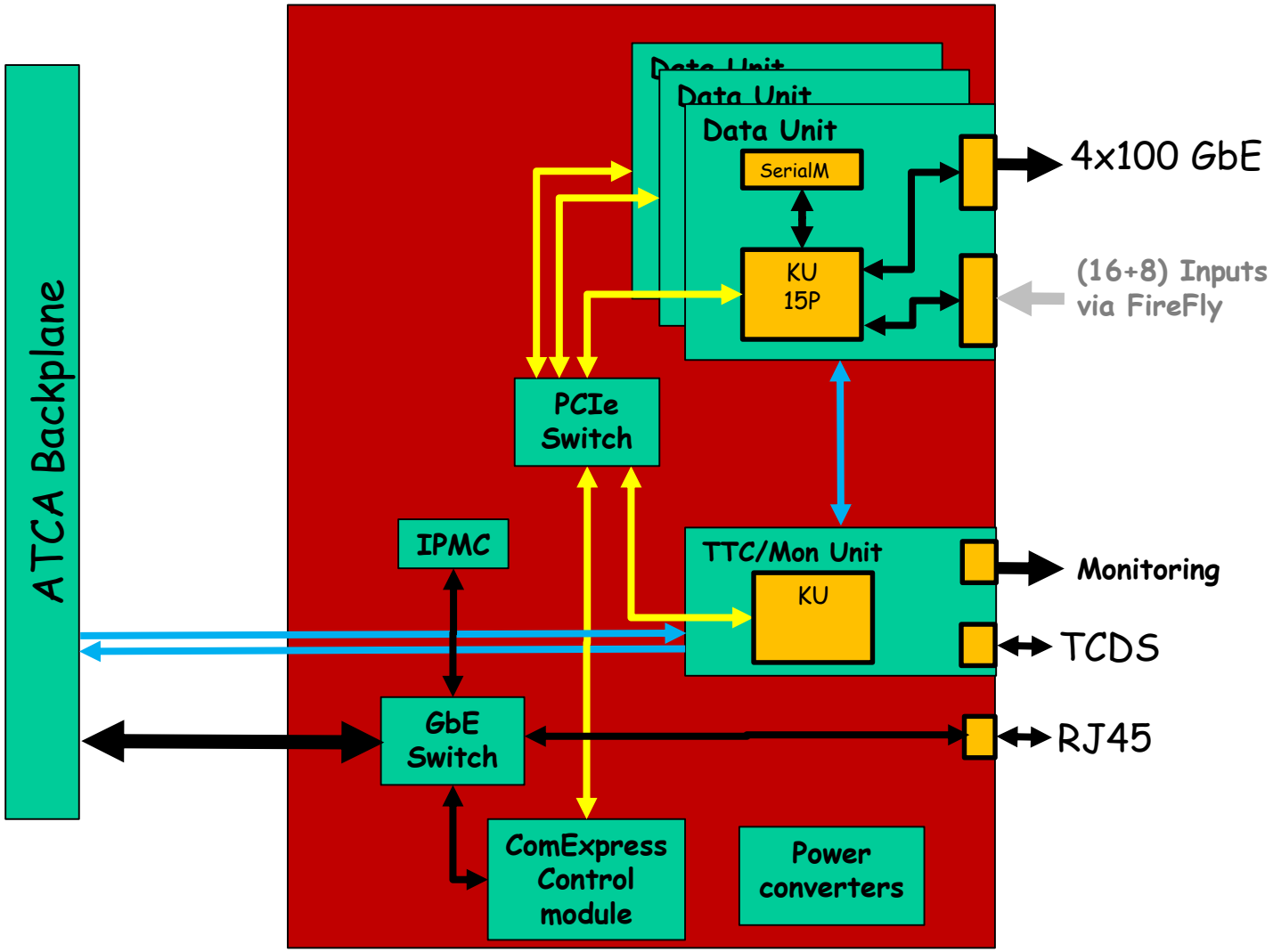
Other elements on the DTH proto

- Common ATCA and hub services
  - Power converters
  - IPMC provided by CERN (hw/sw)
  - ComExpress commercial module for remote control and operation
  - Local GbE switch for base fabric connectivity required by ATCA specs

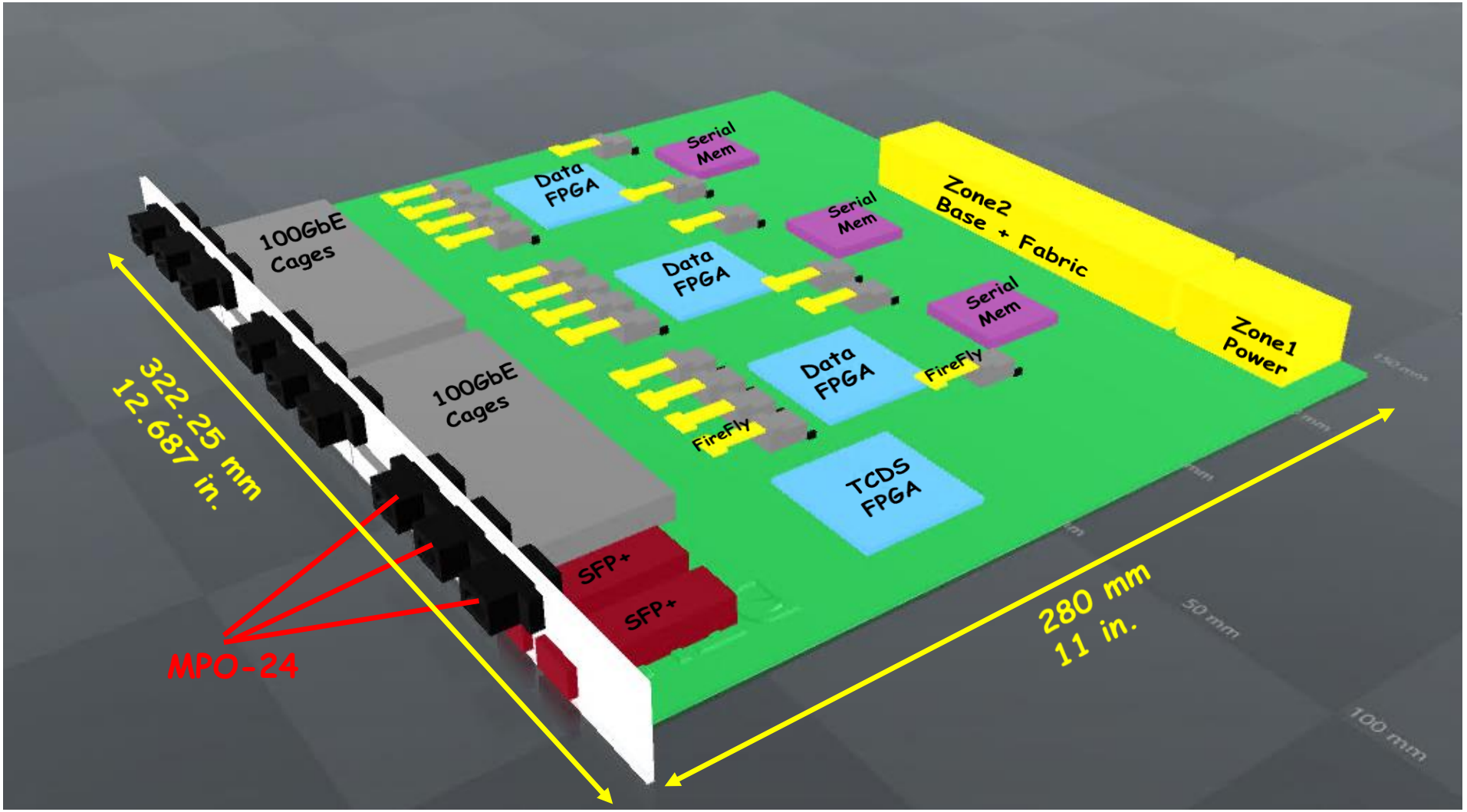
- A TCDS unit
  - Design provided by CERN TCDS group
  - Independent FPGA providing TCDS functions (evolution of current TCDS at CMS)
  - Improved monitoring capabilities (deadtime measurements, state counters, …)
  - Proto will help to measure timing performances over the ATCA backplane

- (Re)-use as much as possible existing hardware and synergies from CMS subsystems and related projects

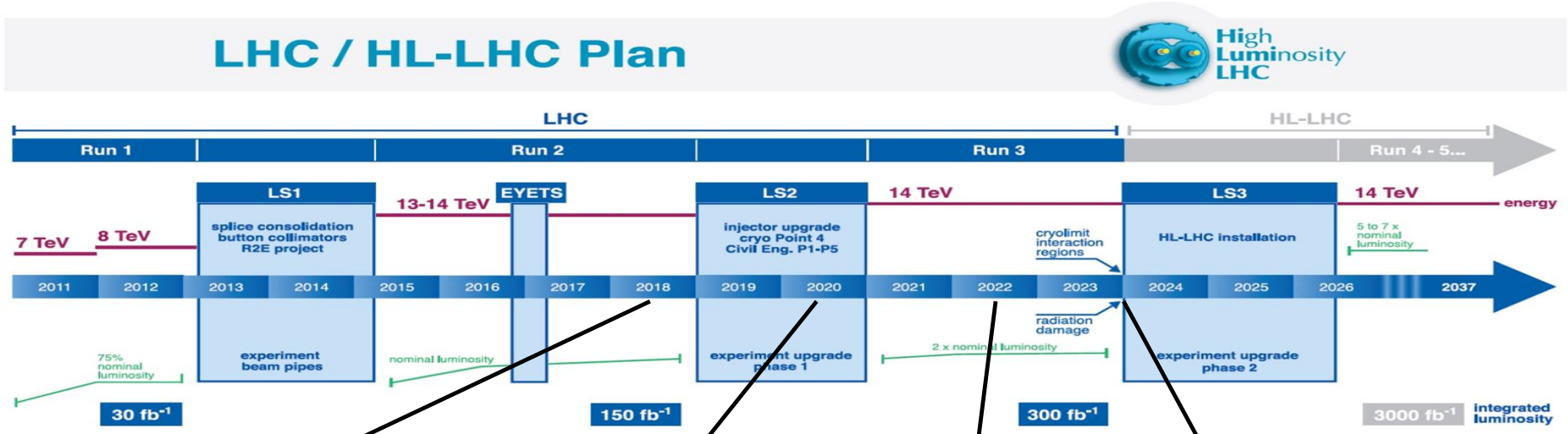
CMS DAQ DTH block diagram



Artistic view of the DTH (9 MPO-24 version)



DTH tentative prototyping program



P1 mid 2018

- Will implement one Data Unit
- Generic TCDS Unit
- Common services

P2 mid 2020

- Will implement one/more Data Unit
- Full Function TCDS
- To be used for Software dev

Pre-prod mid 2022

- Full perf, full functions
- Distributed to sub-detectors groups

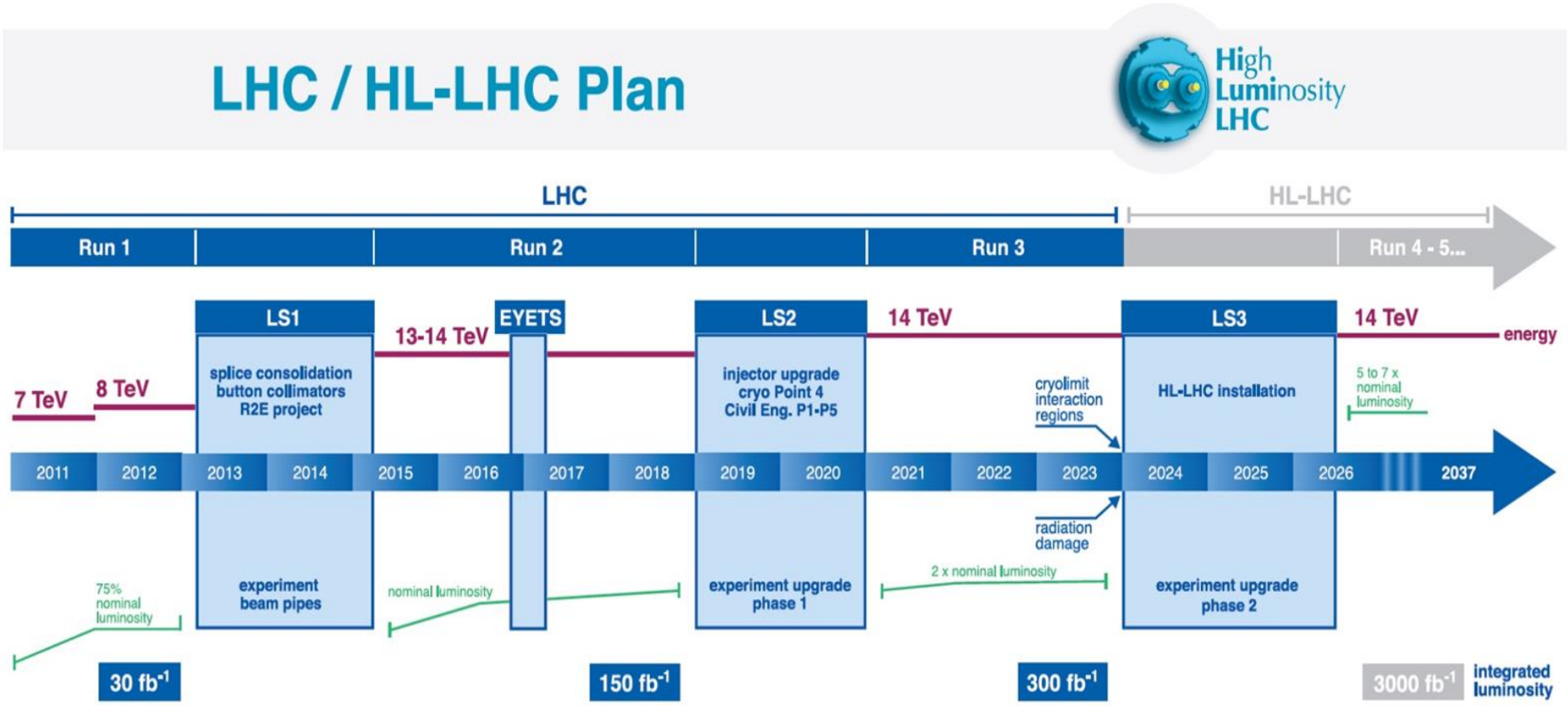
Full prod end 2023

Summary/Conclusions

- Since 2009, CMS DAQ adapts itself to an evolutive detector and evolutive LHC conditions
- Intermediate hardware layer (FRL) allows
  - Decoupling between P2P data transfer and standard packet switched networks
  - Decoupling between sub-detector readout systems and central DAQ
- TCP/IP direct from FPGA enabled the use of efficient balancing layers and different event builder technologies
- Seen from sub-systems, interface to cDAQ is still the same and this will continue
- Continuous improvements in max event size, HLT processing
- For post LS3 DAQ, new dectector, new accelerator, bandwidth x40 wrt DAQ1
- A key hardware element for data collection into the BE crates has been described
- Many more on all the other aspects of post HL-LHC DAQ in the DAQ interim document recently submitted to LHC : CERN-LHCC-2017-014 / CMS-TDR-17-005
  - Be aware that some numbers in tables can be different due to very last corrections ☺

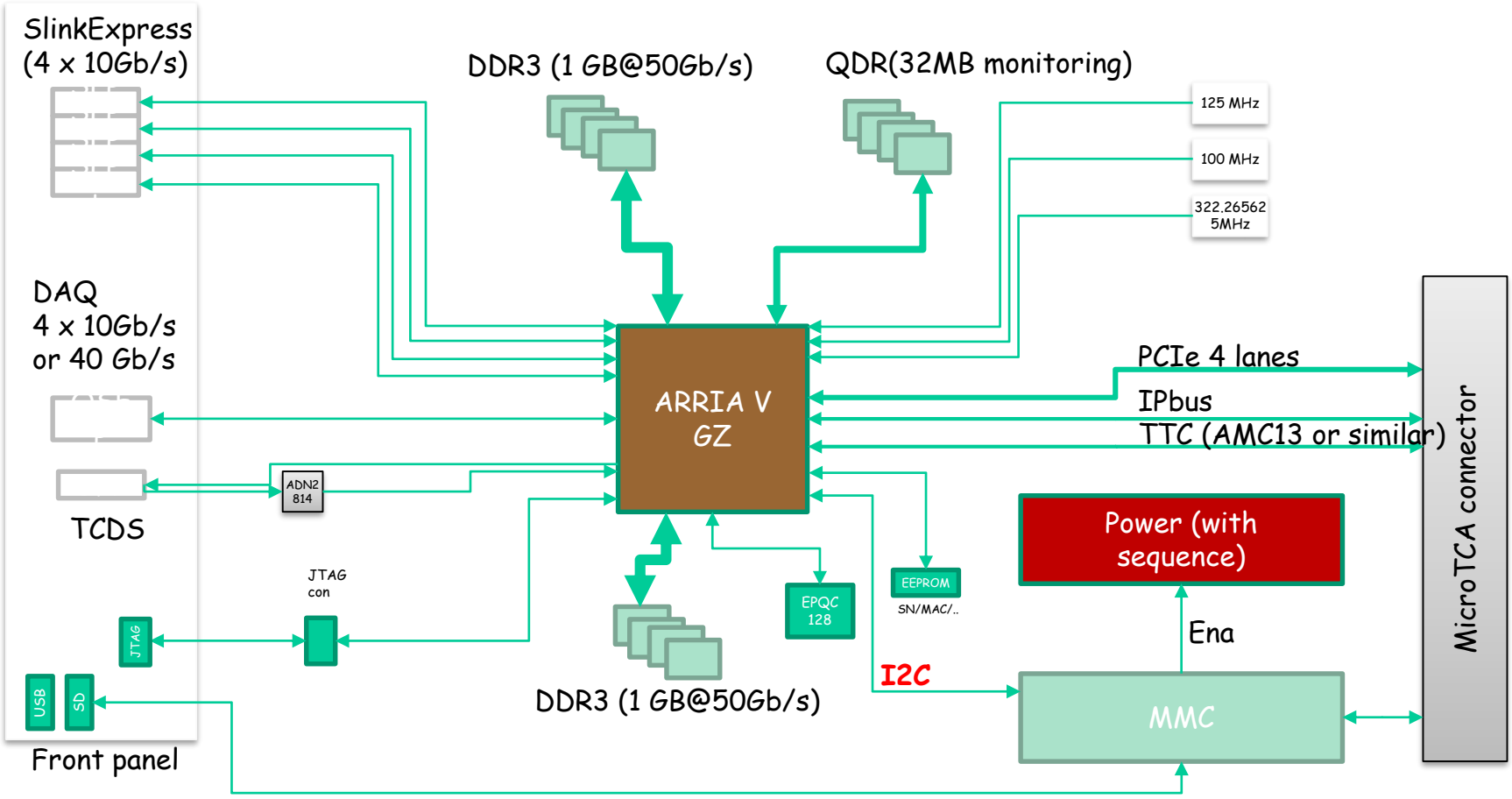
Extra Slides

LHC operation schedule

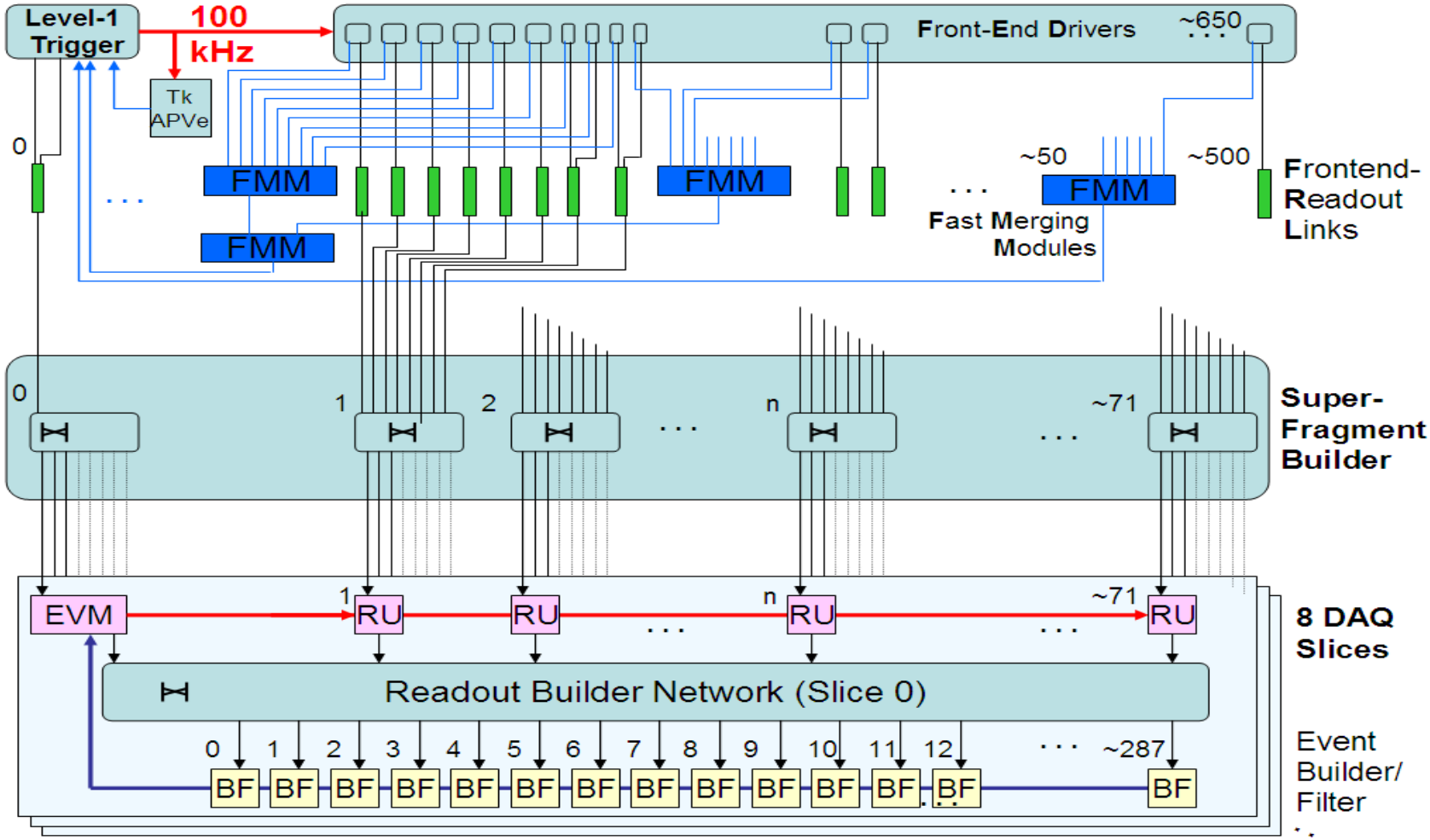


attila.racz@cern.ch

FEROL40 Block Diagram



DAQ1 Block Diagram



attila.racz@cern.ch

FEROL40 picture