DAQ Operations and Plans

Run Coordination Workshop

Torino – 24-26 Jan 2017

Frans Meijers for the DAQ group

DAQ 2016 Achievements (i)

- Developments for global DAQ
  - Repl. of HLT nodes 2nd generation (C6100) and add capacity (factor ~1.4).
  - TCDS; Consolidation, final FC7 boards, spy mode in local-DAQ (Pixel pilot)
  - Integration of new sub-systems (L1 upgrade, TOTEM, CTPPS, Pixel pilot)
  - Improvements of DAQ links to uTCA based subdet BE
    - upgrade of slink-express DAQ link from 5 Gbps to 10 Gbps
    - Multiple (2) DAQ links from AMC13 (HF since Nov)
  - Heavy Ion run p-Pb
    - Extended Lustre Storage throughput and storage by 50%. (~ 7 GB/s)
  - Optimisation of Event Building and Lustre storage.
  - Improved system in some areas taking into account operational experience

DAQ 2016 Achievements (ii)

- Operation:
  - little downtime (<1%)
    - (Almost) no problems with hardware
      - HLT nodes in DAQ2 no longer critical compared to run-1
    - Many relatively small down times (minutes)
      - DAQ shifter, “confusion” with subdet DAQ
    - Few larger ones (fraction of hour)
      - DAQ shifter
      - Confusion with DAQ configurations
      - ….
  - CMS downtime allocation takes non-negligible time by DAQ on-call

DAQ 2016 Achievements (iii)

- Developments
  - FEROL40 to accommodate new Pixel to be installed EYETS 16-17
    - Tested first prototype board successfully,
    - Production launched of 50 boards
  - Improve Test and Validation systems is ongoing
    - cDAQ test system
    - For sub-dets
      - Mini-DAQ
      - 904 is now DAQ2 + TCDS based
      - connection to pixel setup in clean room

Areas

- Migration to new OS platform and XDAQ release
- HLT replacement and possible top up
- Online cloud for offline processing
- Event Sizes
- FEROL40
- New Slink-sender mezzanine card with optical link
- Integration of new sub-systems
- Merger and storage system
- Run control and friends
- TCDS, Sysadmin, DCS and WBM (separate presentations)
- Operation
- requests for run-coordination planning

Migration to new OS platform and XDAQ release

- SLC6 and XDAQ R13
  - Small update to latest kernel for 2017
- CC7 (aka CentOS7) and XDAQ R14
  - CC7
    - Needed for WinCC (PVSS), requested by several sub-detectors
  - XDAQ R14
    - Fixes, Improvements, HAL64, ..
- cDAQ is currently testing (CC7, XDAQ, OFED driver)
  - Issues (fluctuations in EVB perf.), probably not relevant for sub-det DAQ
- Implications for sub-dets
  - Rebuild DAQ code with gcc 4.8
  - CAEN for A818 in R13-slc6 v1.5.1 and in R14-centos7 1.6
  - Revisit system services (to systemd) if you use it

HLT replacement and possible top up

- Regular replacement of “obsolete” (>5 y) PC nodes
  - For HLT replace C6220 (installed 2012) nodes
    - C6220 nodes will be relocated and used for online-cloud
  - Decided to purchase E5-2680v4 based servers and re-use IT tender
  - No replacement for 2018.
- How many nodes to purchase ?
  - Minimal: 5 racks, 180 nodes, gives factor 1.0 wrt 2016 (in HEPspec)
  - *Baseline: 8 racks, 288 nodes, gives factor 1.2 wrt 2016 (in HEPspec)*
- Schedule
  - Week 7: MB on 15 Feb for endorsement
  - Week 8: CERN purchasing
  - Week 18 (1 May): Delivery after 10 weeks (in IT contract)
  - May: installation and commissioning

HLT farm processing capacity

- HLT farm 2016

	CPU	Freq GHz	Cores/ node	HS/ node	# nodes	# cores	kHS	%
c6220	E5-2670	2.6	2*8	350	256	4096	90	18
s2600kp	E5-2680v3	2.5	2*12	538	360	8640	194	39
s2600kp	E5-2680v4	2.4	2*14	659	324	9072	214	43
					940	**21808**	**498**	

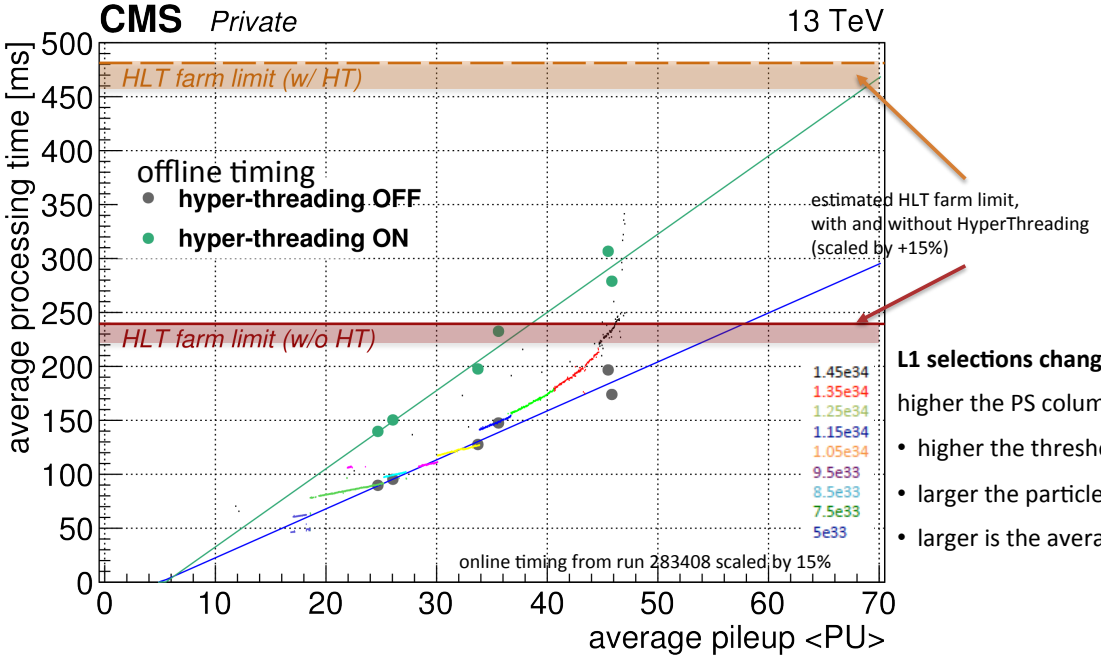
- Baseline HLT farm 2017 (up for discussion)

	CPU	Freq GHz	Cores/ node	HS/ node	# nodes	# cores	kHS	%
s2600kp	E5-2680v3	2.5	2*12	538	360	8640	194	32
s2600kp	E5-2680v4	2.4	2*14	659	324	9072	214	36
	E5-2680v4	2.4	2*14	659	288	8064	190	32
					972	**25776**	**598**	

HLT processing capacity 201672016 = 598 kHS / 498 kHS = factor 1.20

HLT timing

timing in 2016



L1 selections change event topology

higher the PS column [for ~fixed PU]

- higher the thresholds
- larger the particle multiplicity
- larger is the average HLT timing

new HF readout and new pixel detectors might have a not neglible impact on the timing in 2017 …

HLT Menu v4.2 on HLTPhysics Run2016H (6 skims based on both PU and PS column)
machines vocms003/004

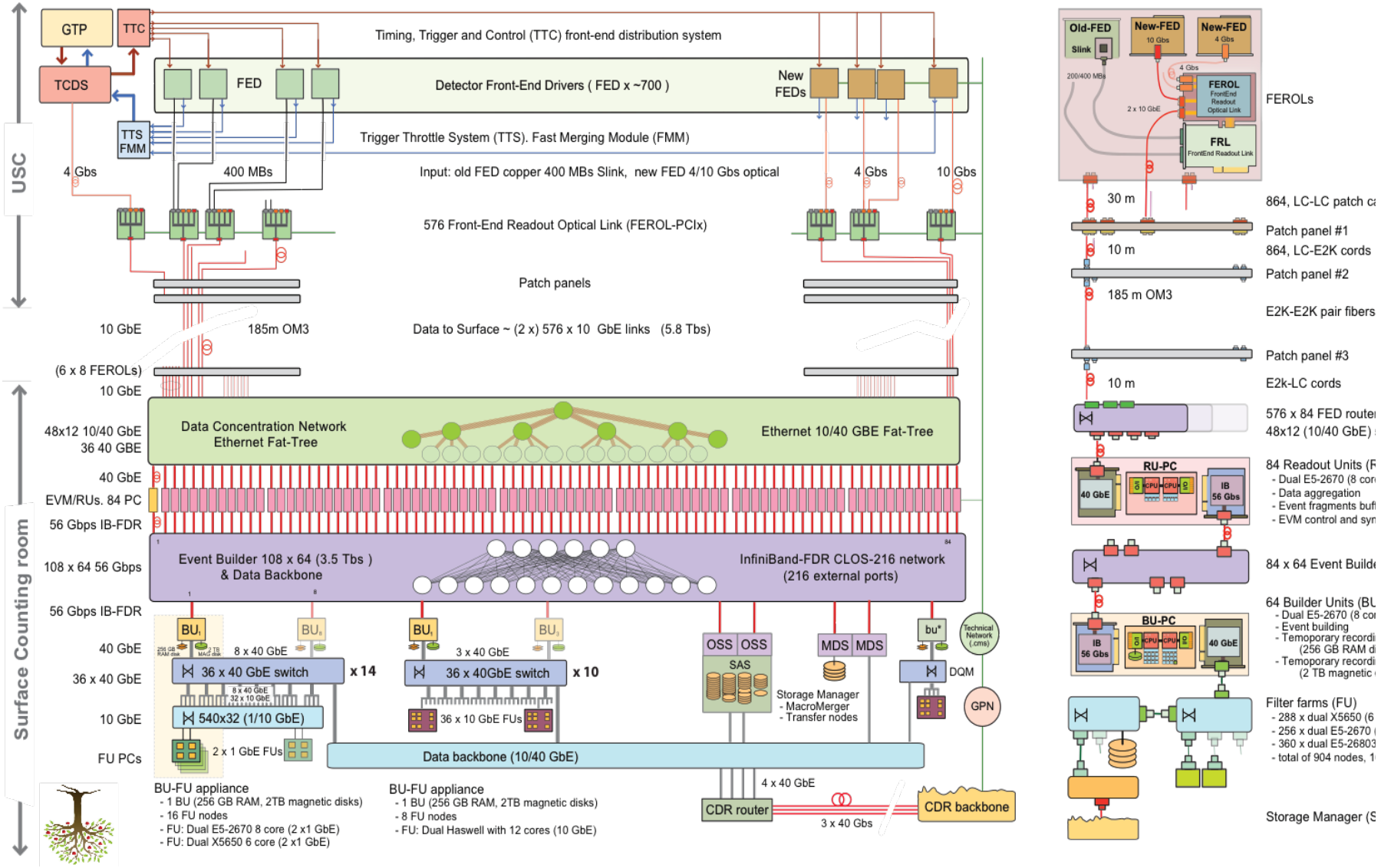
15

HLT needs for 2017

- Several Factors
  - HLT scales approx with HEPSpec
  - HLT timing
    - New Menu L1+HLT
    - Detector Change (Pixel, HCAL)
    - PU dependance
  - L1 rate close to 100 kHz ?
  - LHC operation point, peak PU
  - CMS policy on lumi-levelling
- Do we need factor 1.2 top-up ?
- If over-dimensioned
  - Can use for online cloud
  - Cuts closer to offline, could reduce HLT output rate
  - ..

Online cloud for offline processing

- Architecture and characteristics
  - Openstack infrastructure overlayed on OS
  - 4x40 Gbps link from Pt5 to IT/Meyrin
  - Can run “any” workflow if inp/out data on CERN EOS
  - Major improvement established in 2016 with Offline
    - hibernation / restart of VMs (avoiding killing of jobs)
- Modes of operation
  - Static (Used since years during TS)
  - Dynamic
    - Between fills (steered by LHC mode). Used end 2016
  - 2017
    - Want to test during fill depending on HLT CPU load (exploiting lumi-decay)

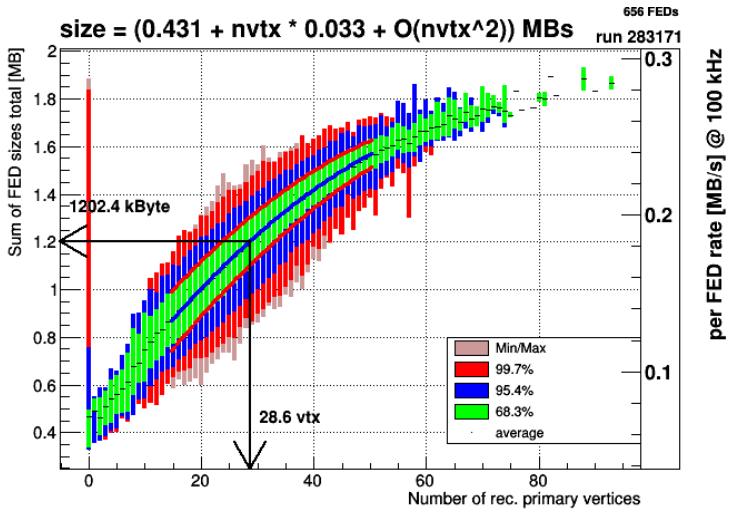

USC
Surface Counting room
GTP
TTC
TCDS
TTS FMM
Timing, Trigger and Control (TTC) front-end distribution system
FED
Detector Front-End Drivers (FED x ~700)
New FEDs
Trigger Throttle System (TTS). Fast Merging Module (FMM)
4 Gbs
400 MBs
Input: old FED copper 400 MBs Slink, new FED 4/10 Gbs optical
4 Gbs
10 Gbs
576 Front-End Readout Optical Link (FEROL-PCIx)
Patch panels
10 GbE
185m OM3
Data to Surface ~ (2 x) 576 x 10 GbE links (5.8 Tbs)
(6 x 8 FEROLs)
10 GbE
48x12 10/40 GbE
36 40 GBE
Data Concentration Network
Ethernet Fat-Tree
Ethernet 10/40 GBE Fat-Tree
40 GbE
EVM/RUs. 84 PC
56 Gbps IB-FDR
108 x 64 56 Gbps
Event Builder 108 x 64 (3.5 Tbs)
& Data Backbone
InfiniBand-FDR CLOS-216 network
(216 external ports)
56 Gbps IB-FDR
40 GbE
36 x 40 GbE
10 GbE
FU PCs
BU
8 x 40 GbE
36 x 40 GbE switch
x 14
8 x 40 GbE
32 x 10 GbE
540x32 (1/10 GbE)
2 x 1 GbE FUs
3 x 40 GbE
36 x 40GbE switch
x 10
36 x 10 GbE FUs
OSS OSS
SAS
MDS MDS
Storage Manager
- MacroMerger
- Transfer nodes
bu*
DQM
GPN
Technical Network (.cms)
Data backbone (10/40 GbE)
4 x 40 GbE
CDR router
3 x 40 Gbs
CDR backbone
BU-FU appliance
- 1 BU (256 GB RAM, 2TB magnetic disks)
- 16 FU nodes
- FU: Dual E5-2670 8 core (2 x1 GbE)
- FU: Dual X5650 6 core (2 x1 GbE)
BU-FU appliance
- 1 BU (256 GB RAM, 2TB magnetic disks)
- 8 FU nodes
- FU: Dual Haswell with 12 cores (10 GbE)

Old-FED
Slink
New-FED
10 Gbs
New-FED
4 Gbs
200/400 MBs
4 Gbs
FEROL
FrontEnd Readout Optical Link
2 x 10 GbE
FRL
FrontEnd Readout Link
FEROLs
30 m
864, LC-LC patch ca
Patch panel #1
10 m
864, LC-E2K cords
Patch panel #2
185 m OM3
E2K-E2K pair fibers
Patch panel #3
10 m
E2k-LC cords
576 x 84 FED router
48x12 (10/40 GbE)
RU-PC
40 GbE
IB 56 Gbs
84 Readout Units (F
- Dual E5-2670 (8 cor
- Data aggregation
- Event fragments buf
- EVM control and syn
84 x 64 Event Builde
BU-PC
IB 56 Gbs
40 GbE
64 Builder Units (BU
- Dual E5-2670 (8 cor
- Event building
- Temporary recordi
(256 GB RAM di
- Temporary recordi
(2 TB magnetic
Filter farms (FU)
- 288 x dual X5650 (6
- 256 x dual E5-2670
- 360 x dual E5-2680
- total of 904 nodes, 1
Storage Manager (S

Event Sizes

- Limitations on total event size, FED size, FEDbld size
- In 2016
  - Event size of ~1.2 MB when TK had extra diagnostics
  - Optimised EVB performance: ~< 2 MB @ 100 kHz

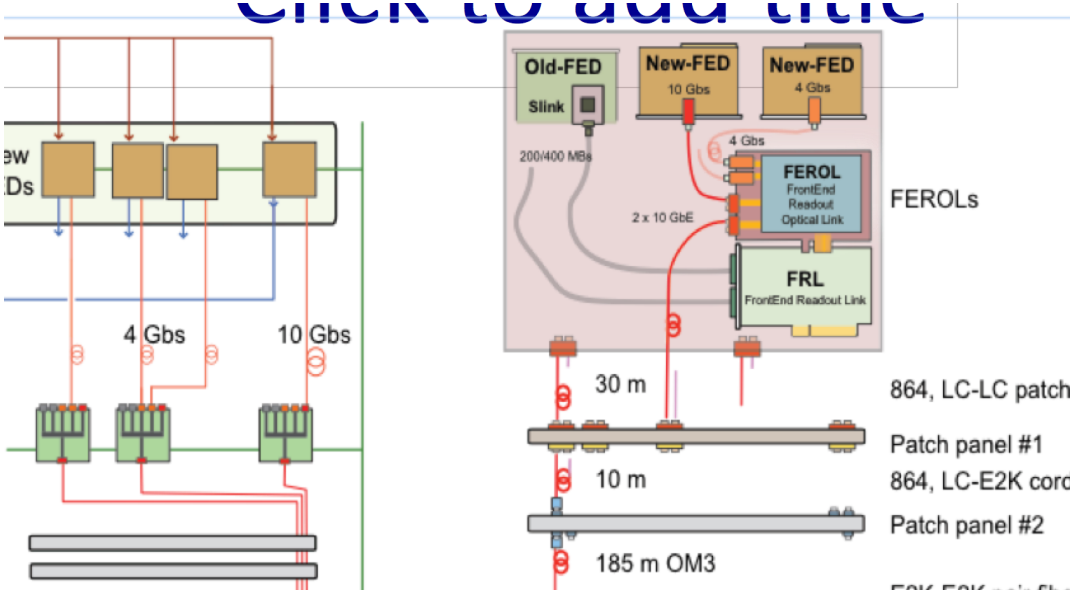
standard pileup



- 2017
  - New pixel detector (sub-evt size x ~4), HF

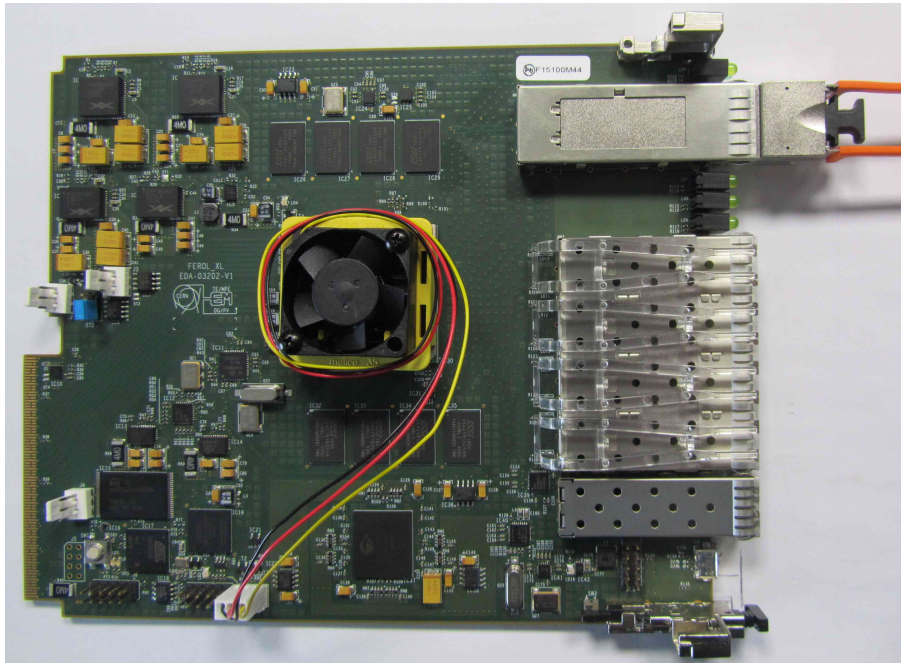
FRL/FEROL/FEROL40

- FEROL is housed in FRL modules (from run-1)
  - 1 or 2 Legacy Slinks (copper cable)
    - With 50 MHz clock, 400 MB/s = 3.2 Gbps
  - 1 or 2 Optical slink-express 5 Gbps (8/10b encoded)
    - Used for TCDS
  - 1 Optical slink-express 10 Gbps (64/66b encoded)
    - All uTCA BE sources (AMC13)



FEROL40

- Motivation
  - Pixel upgrade has 108 FEDs (FC7)
  - Need sufficient FEROL(40) spares till LS3
- FEROL40
  - uTCA based module (not PCI
  - 4 slink-express 10 Gbps inputs
  - 4x10 Gbps Ethernet output
  - Can see it as 4 FEROL in 1 module



FEROL40 Module production

- Need 32 modules
  - 12 for barrel +, 12 for barrel 1 and 8 for FPIX
- Production
  - Produced 41 cards
    - 2 have electrical problem
    - 2 have problem after replacing voltage regulator (out of 5)
    - 37 pass the tests
  - 9 remaining cards are on the way to be assemble with components
- Re-discovered issue with voltage regulator (Altera EN2392QI)
  - High prob of dying when power cut, orderly shutdown OK
  - Appeared that heating board for exchange of regulator risky
  - Designing small circuit board with other regulator chip
  - UPS in USC
    - Need shutdown mechanism and/or diesel backup

FEROL40 for Pixel

- Installed fibres USC-SCX FEROL40 to switches
- Software developed, tested in DAQVal
- Incorporated in FED-builder switch configuration
- Tested 1 module in uTCA crate in USC to global DAQ
- To be done:
  - Install all 32 FEROL40 modules
  - fibre cables Pixel to FEROL40

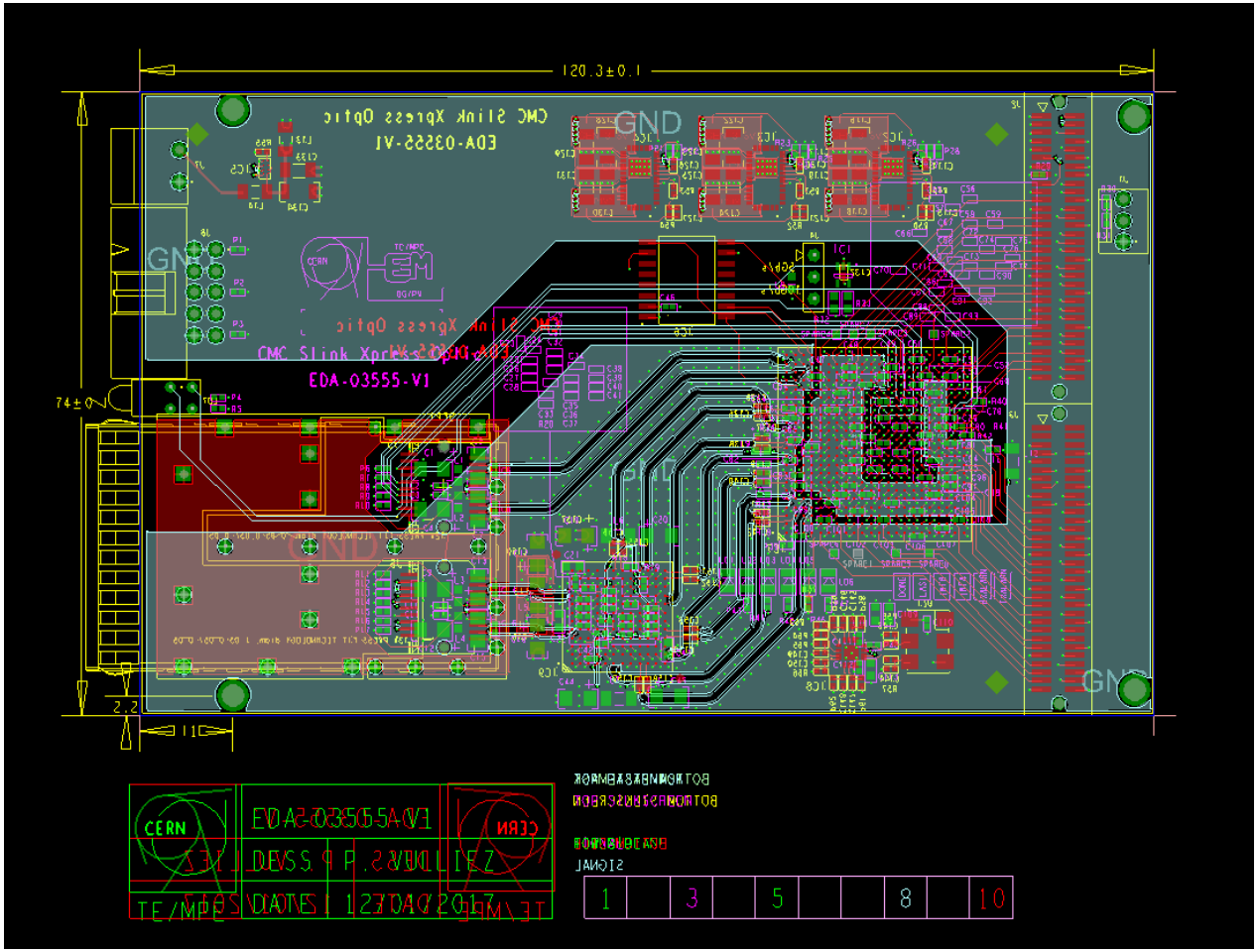
New Slink-sender mezzanine card with optical link (i)

- Slink-sender mezzanine on legacy FED
  - FED clock
    - subdet 40 MHz (Hence 320 MB/s effectively)
    - Some (eg ECAL DCC) 80 MHz
  - Small FIFO Buffer of 2 kB (XOFF at 1536 B)
  - Drained over s-link cable with 50 MHz (400 MB/s), so 4 kB @100 kHz
  - Consequence for FEDs with clock >50 MHz
    - frequent instantaneous slink-XOFF, not-trivial to interpret FED busy
  - Was assumed for CMS Phase-0 that FED itself has sufficient buffer space to absorb bursty event arrival

New Slink-sender mezzanine card with optical link (ii)

- Slink-sender mezzanine on legacy FED
- Project for optical slink sender mezzanine
  - Not a critical item
  - Main use case is ECAL for time being
    - Could reduce dead-time slightly (especially FED dealing with 100 Hz laser R/O)
  - Design
    - FPGA with 1 MB buffer
    - Two SFP cages (5 Gbps, 10 Gbps), selectable with jumper
    - With 5 Gbps links, could merge by 2 in FEROL
  - 70 pieces
  - schedule
    - Jan PCB layout
    - 2 card mid-March (ECAL can test in 904)
    - 70 cards for 56 needed, mid-April
    - switch over in may in USC, needs also pull fibre cables ECAL feds to ferols.

New Optical CMC (5/10Gbs)



Integration of new sub-systems

- Major (in terms of channels) new Pixel
  - Installation FEROL40 etc : Finish by end-Jan
  - Commissioning by central DAQ with emulator
    - By mid-Feb
  - Commissioning of Pixel BE with Central DAQ
    - Involves DAQ link, TCDS, run-control
    - When there is possibility in Feb – Mar
    - Integrated running by MWGR3 (15 April) ?
- Minor (in terms of channels)
  - CTPPS uTCA back-end
  - GEM pilot in uTCA
  - RPC Endcap via L1 ?

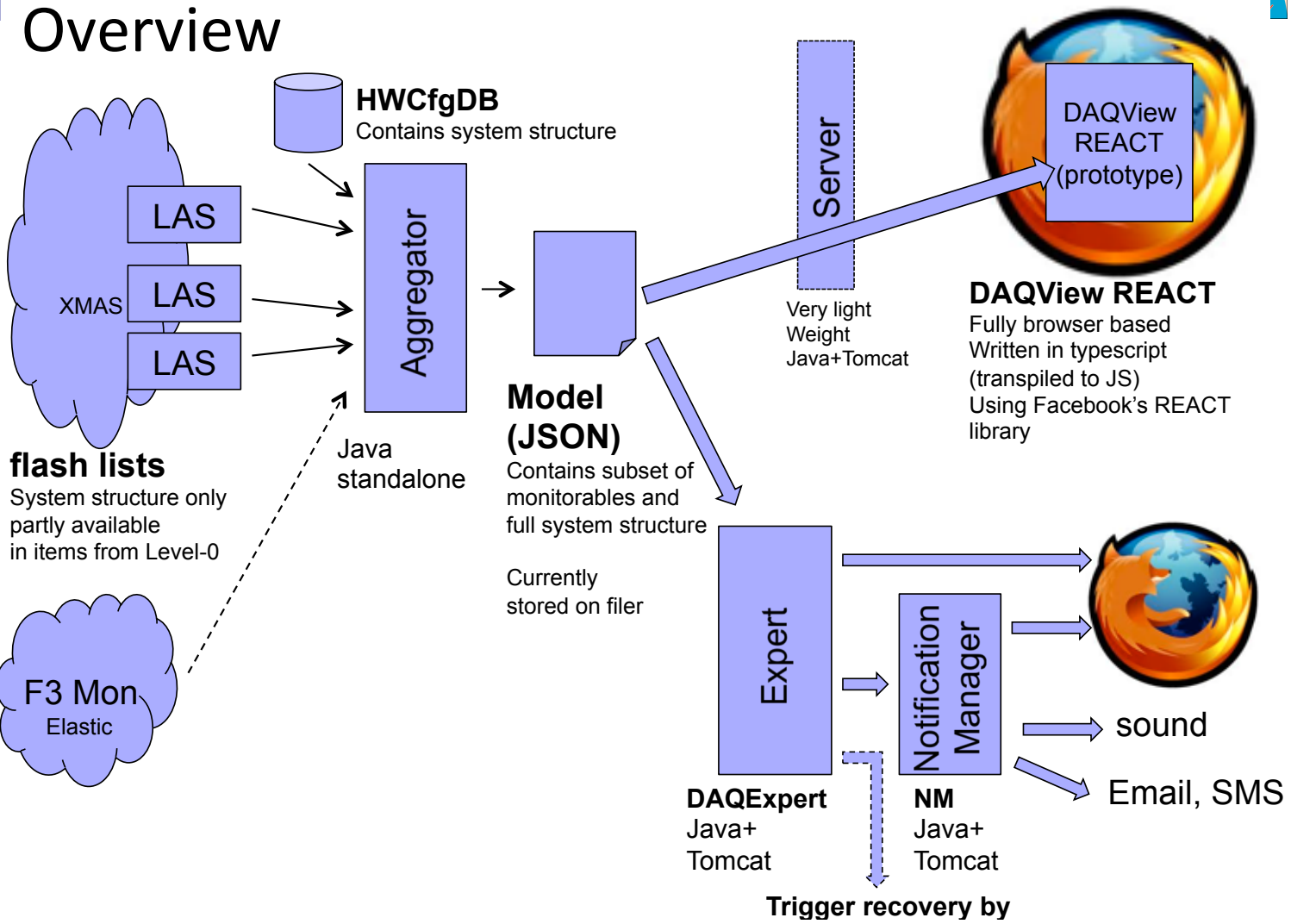
Merger and storage system

- 2016
  - Extended Storage HW by 50%
  - Optimisations of Merger, Storage
    - For example: more nodes for transfer to deal with EOS
- 2017
  - Update and uniformise Lustre versions
  - New transfer system
    - python based, communication via dbase
    - First iteration tested in daqval
      - Transfer of files works,
      - T0 wants modification needs for re-packer communication

Run control and friends

- Run control:
  - Level of automation increased throughout the years
- Re-designed chain of monitoring
  - Level0 timeline
  - DAQ aggregator
    - Store (important part of) monitoring persistently to allow playback
    - Clients: new DAQview and new Expert System
  - Expert System
    - Guidance for DAQ operator
    - Use cases
      - DAQ flow chart
      - Relevant ones from old expert system, so that this can be obsoleted
      - Interface to sound system
    - Suggestion to use for guiding change of L1 / HLT prescale
  - Considering how to use it to help downtime assignment

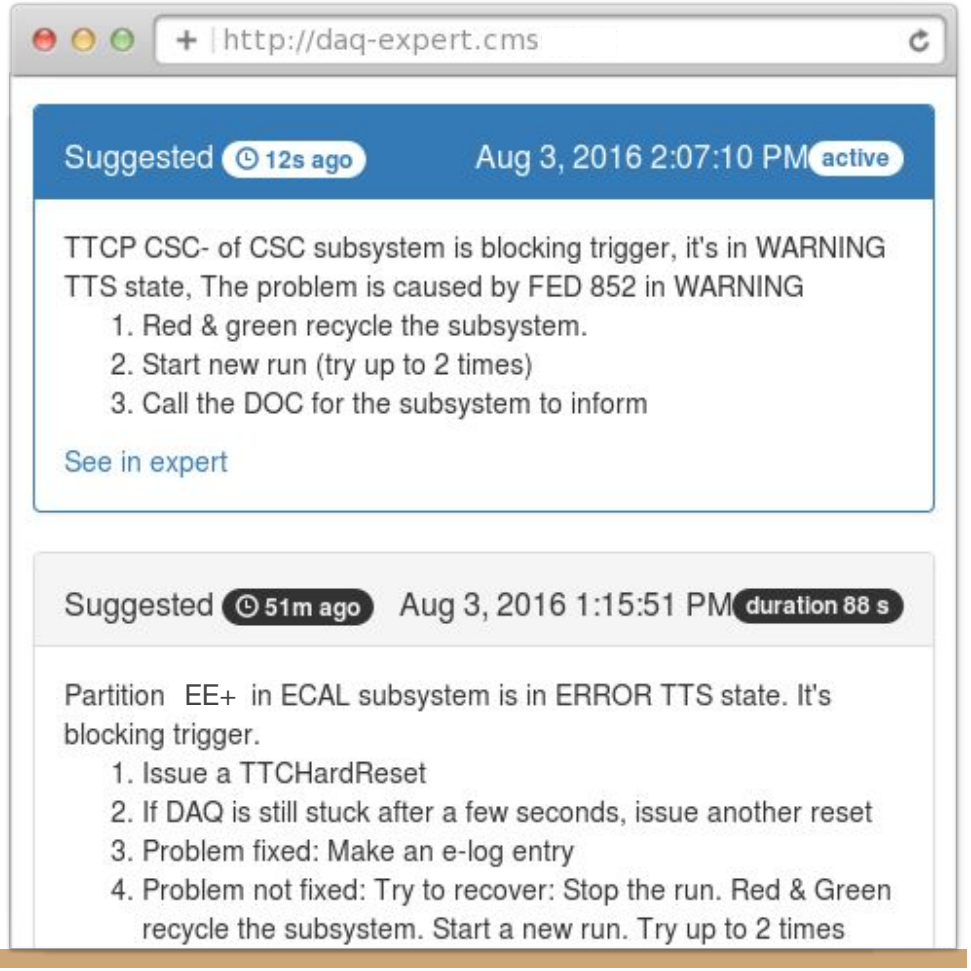
Overview


HWCfgDB
Contains system structure
LAS
XMAS
LAS
LAS
Aggregator
Server
DAQView
REACT
(prototype)
DAQView REACT
Fully browser based
Written in typescript
(transpiled to JS)
Using Facebook's REACT
library
Very light
Weight
Java+Tomcat
flash lists
System structure only
partly available
in items from Level-0
Java
standalone
Model
(JSON)
Contains subset of
monitorables and
full system structure
Currently
stored on filer
F3 Mon
Elastic
Expert
Notification
Manager
sound
Email, SMS
DAQExpert
Java+
Tomcat
NM
Java+
Tomcat
Trigger recovery by

Tools (dashboard)

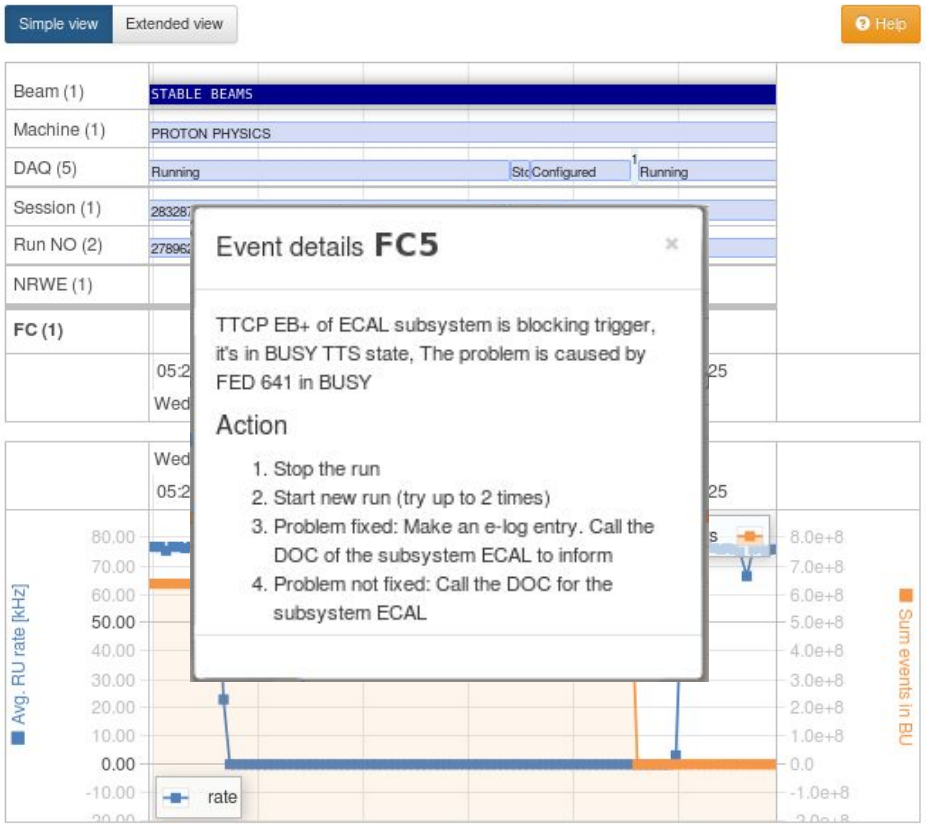
- Main view for CR
- RT suggestions to shifters:
  - Reduce reaction time
  - Avoid wrong decisions
- Suggestion format:

  what's the problem +
  what's the best action to take



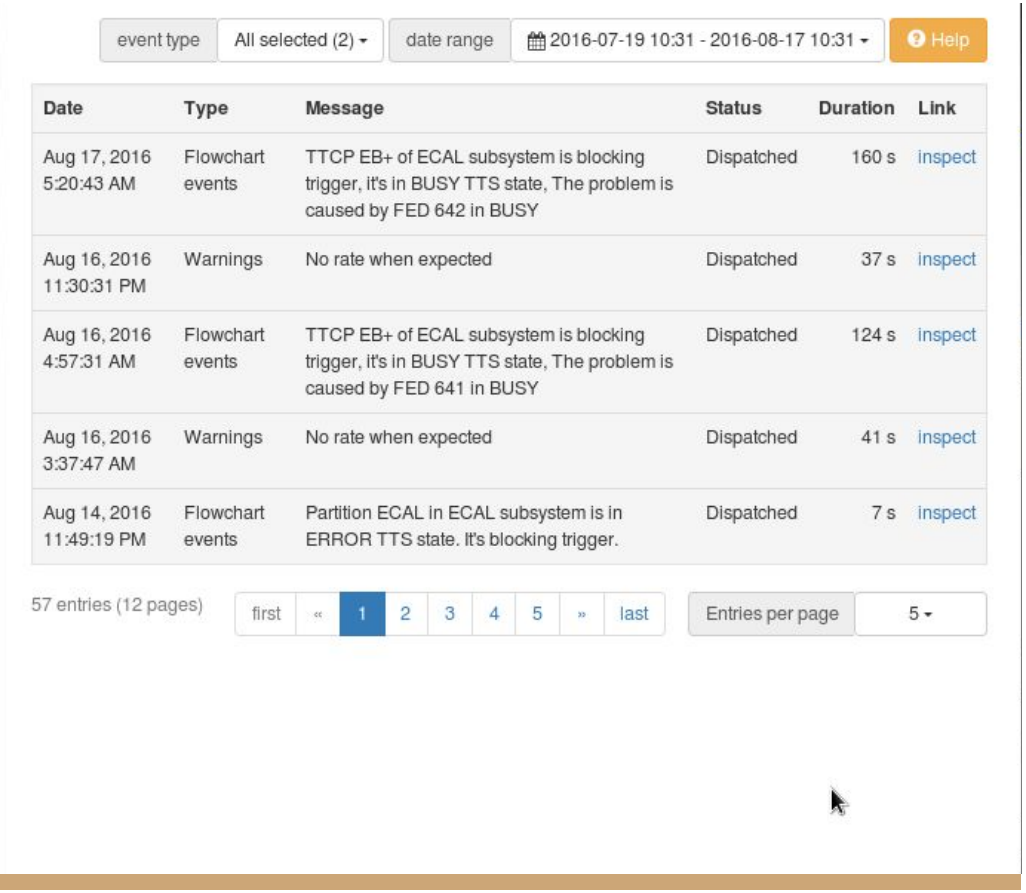
Tools (browser)

- Goal: post-mortem analysis
- Visualizes analysis in time
- Analysis panel
  - 1 Row - 1 Logic Module*
- Raw data panel
  - Parameters from snapshots
  - Raw snapshot popup (JSON)
- Freely move and zoom in time
- simple/extended view + experimental mode



Tools (notifications)

- Goal: browse all generated notifications
- Inspect link
- Filter by type
- Date range picker



Replacement of DAQDoctor

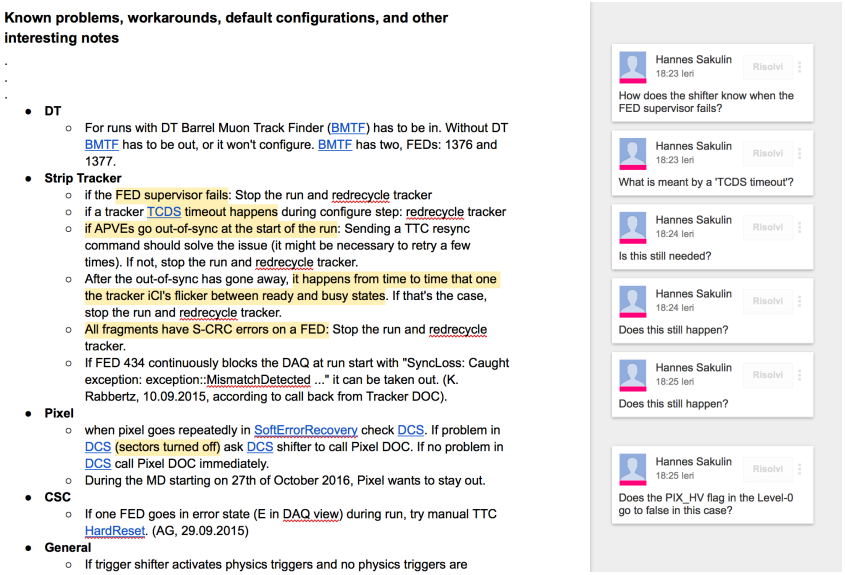
Check	D1	D2	E
Beam mode transition	x	x	x
Machine mode transition	x	x	x
Session transition	x	x	x
Run number transition	x	x	x
DAQ transition	x	x	x
L0 transition	x	x	x
L1 rate out of range	x	x	x
HLT rate out of range	x	x	
HLT bandwidth			
Global deadtime	x	?	x

Partition deadtime	x	x	x
FED deadtime	x	x	x
Subsystem running degraded			x
Subsystem error	x	x	x
Subsystem soft error			x
Cpu usage	x	?	
Temperature	x	x	
DQM latency			
DQM file size			
Check BU output disk usage			
Check BU ram disk usage			

Abnormal transitions	x		
Flowchart cases			x
Stale monitoring data	x	x	
Subsys stuck in fixing soft error			
Crash detection	x		
Other infrastructure monitoring			
Resync storms	x		
Alignment check	x		
Splash	x	x	
Warn FED CRC error	x		
Warn Slink CRC error	x		
Myrinet getting stuck	x	-	-
Monitor health of transceivers	x		

DAQ shifter instruction etc

- Need a review of sub-det related actions
  - DAQ shifter instruction has grown organically through the year
  - Lead to in-efficiency because complex and error-prone
  - Maybe some rules can be coded as guidance in new expert
- Sub-detectors please provide feedback on
  - https://docs.google.com/document/d/1qAHj2a6BN6Z-AC30to6vb3ojQpTulyVrHQgo6dqT-c0/edit?usp=sharing



Known problems, workarounds, default configurations, and other interesting notes

- **DT**
  - For runs with DT Barrel Muon Track Finder (BMTF) has to be in. Without DT BMTF has to be out, or it won't configure. BMTF has two, FEDs: 1376 and 1377.
- **Strip Tracker**
  - if the FED supervisor fails: Stop the run and redrecycle tracker
  - if a tracker TCDS timeout happens during configure step: redrecycle tracker
  - if APVEs go out-of-sync at the start of the run: Sending a TTC resync command should solve the issue (it might be necessary to retry a few times). If not, stop the run and redrecycle tracker.
  - After the out-of-sync has gone away, it happens from time to time that one the tracker iCI's flicker between ready and busy states. If that's the case, stop the run and redrecycle tracker.
  - All fragments have S-CRC errors on a FED: Stop the run and redrecycle tracker.
  - If FED 434 continuously blocks the DAQ at run start with "SyncLoss: Caught exception: exception::MismatchDetected ..." it can be taken out. (K. Rabbertz, 10.09.2015, according to call back from Tracker DOC).
- **Pixel**
  - when pixel goes repeatedly in SoftErrorRecovery check DCS. If problem in DCS (sectors turned off) ask DCS shifter to call Pixel DOC. If no problem in DCS call Pixel DOC immediately.
  - During the MD starting on 27th of October 2016, Pixel wants to stay out.
- **CSC**
  - If one FED goes in error state (E in DAQ view) during run, try manual TTC HardReset. (AG, 29.09.2015)
- **General**
  - If trigger shifter activates physics triggers and no physics triggers are

Operation

- Sysadmin on-call
- DCS on-call (too few people)
- DAQ on-call
  - In 2016 shared by 9+2 new newcomers heads
  - 2+ are leaving in 2017
- DAQ Shifts 2017
  - Shifts start 27 Mar
  - First half-year opened
    - Group-1 (people who have already done shifts)
    - Asked to leave holes for newcomers
    - Identified 10 newcomers from group2+3
  - Need to organise a tutorial

Summary for Run coordination planning

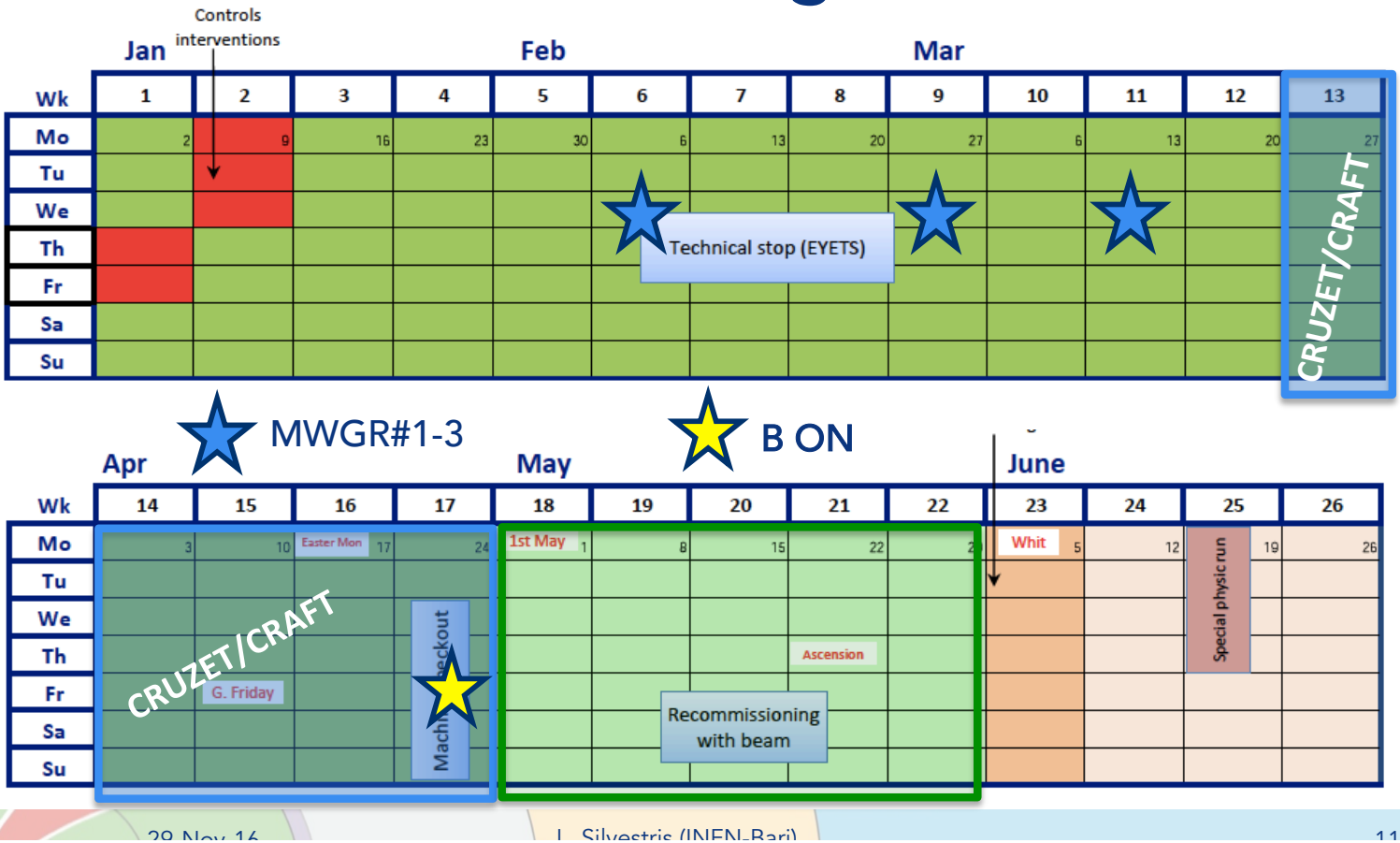
- forthcoming (to be confirmed)
  - recommendation to sub-det to migrate to CC7 and XDAQ R14
- need feed-back on capacity of HLT farm (small top up?) before 13 Feb
- need review by sub-det on DAQ operations procedures

Summary for Run coordination planning

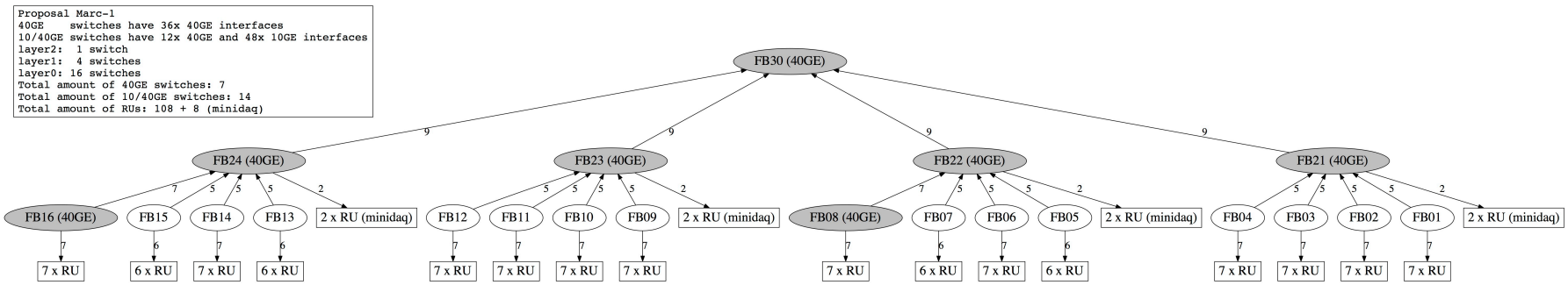
- Jan - end Mar (before CRUZET/CRAFT)
  - need dedicated test time for
    - cdaq internal
    - integration new sub-systems (pixel, ct-pps, ecal)
    - integration of new transfer system with T0
- End April
  - Optical slink sender mezzanine for ECAL
- May
  - integration of new HLT nodes
- Lumi ramp-up and beyond
  - online cloud interfill mode
  - suggest to test running at L1=100kHz

ADDITIONAL MATERIAL

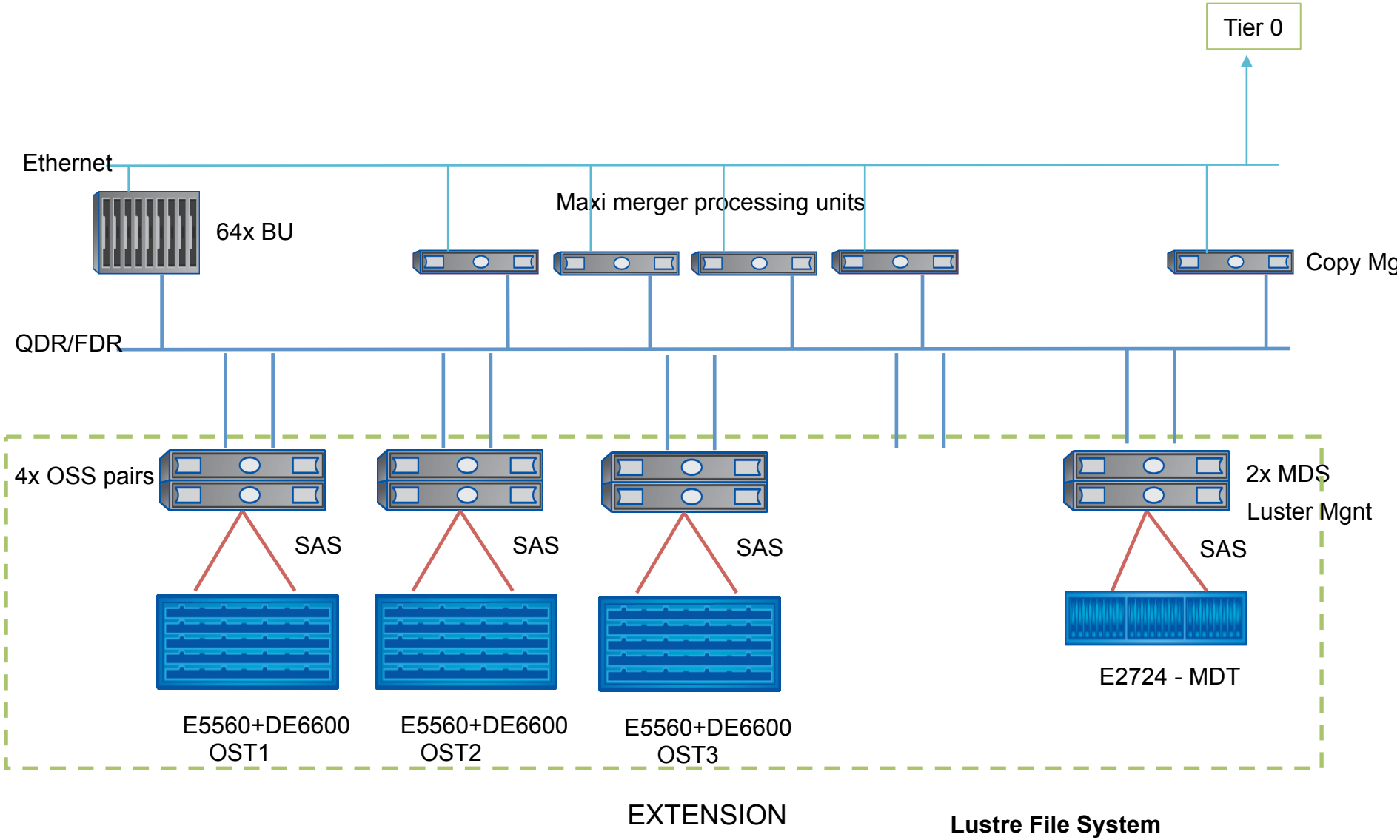
2017 LHC draft Schedule and CMS Commissioning



Ethernet Fattree network



Storage E5560/DE6600 Lustre



DAQExpert data flow & tasks

- 3 sub projects + DAQView
- Michail joined DAQAggregator Jul 13