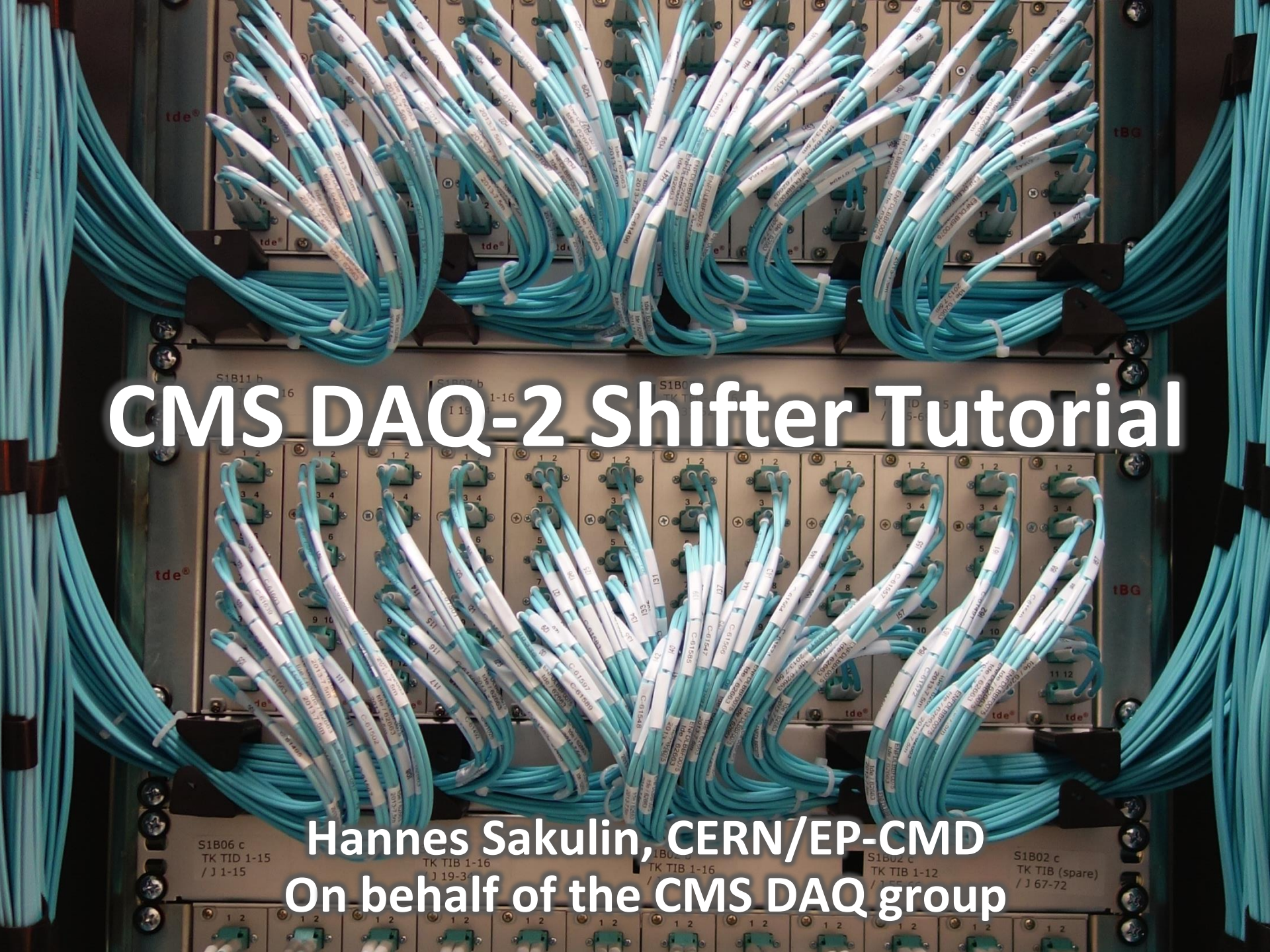


CMS DAQ-2 Shifter Tutorial

Hannes Sakulin, CERN/EP-CMD
On behalf of the CMS DAQ group

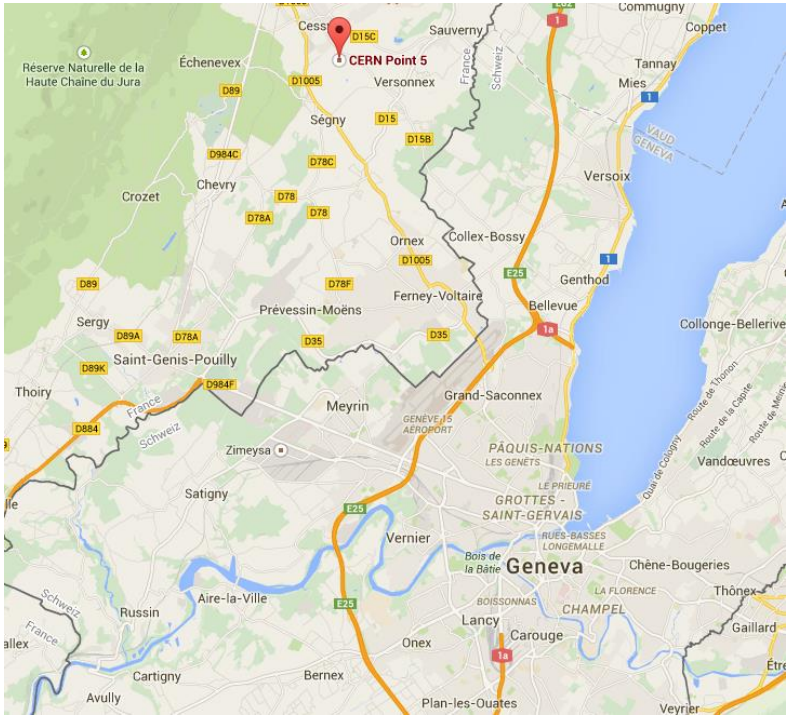
DAQ2 Tutorial Outline

- Part 1: Your tasks as a DAQ shifter
- Part 2: Overview of the DAQ-2 system
  - Change from DAQ-1 to DAQ-2
  - DAQ-2 hardware and data flow from the detector to storage / Tier 0
  - Flow control
  - Software
- Part 3: Controlling data taking through Run Control
- Part 4: DAQ monitoring tools

Part 1: Your tasks as a DAQ shifter

Context

- DAQ shifts take place at the CMS Control Room at Point 5 of the LHC in Cessy, France



- Three shifts: 7-15, 15-23, 23-7
- Five shifters
  - Shift leader: manage operations in line with daily plan, monitor data taking, communicate with LHC, safety
  - DCS shifter: slow control, access, safety
  - DAQ shifter: control, monitor & troubleshoot data taking
  - Trigger shifter: monitoring of the L1 trigger
  - DQM shifter: monitoring of data quality

  (the latter three shifts may get cancelled under certain circumstances)

Your responsibilities as a DAQ Shifter

- Main DAQ responsibilities:
  - Monitor the DAQ
    - Make sure the DAQ is running smoothly and that CMS is collecting high quality data!
    - This means:
      - Monitoring all stages of the DAQ from FEDs -> data sent to Tier 0.
        - Data rate, dead time, back-pressure, CPU usage, problems
      - Interfacing with the shift crew
  - Control data taking as the Shift Leader requests
    - Start / stop runs
    - Take in / out sub-detectors, FEDs
    - Manual re-syncs, hard resets, control random rate
  - Troubleshoot the DAQ in case of problems
    - But don’t hesitate to call the DAQ DOC (x76600) when you are stuck!
  - Document your shift: use the ELOG!

Your responsibilities as a DAQ Shifter

- **You are the main responsible for efficient data taking of CMS**
  - **The CMS efficiency will depend to a high degree on your abilities**
  - As a consequence you need to
    - Read and learn the necessary procedures
    - Keep yourself up-to-date
      - The run environment will evolve in time
      - Procedures will change
      - Monitoring systems will change
      -
    - This means you have to dedicate some time outside of your shift period to study the online DAQ system and how to operate it.

Your responsibilities as a DAQ Shifter

- **During your shift you have to communicate continuously with the shift leader and other shifters/DOCs in order to keep running efficiently**

  - You are needed since the job CANNOT be done by a computer program or a robot !

  - Since you are the key person which starts and stops the run, you also should be the key person to overcome or work around problems. This means:
    - You should be able to localize where the problem is
      - In the central DAQ, or in a subsystem, or in the computing infrastructure
    - You must communicate efficiently to the relevant experts and the shift leader in the control room
    - You are involved in suggesting workarounds where you can
    - You are co-responsible to solve problems in the DAQ and online system
      - This often means to efficiently communicate to the expert on call
      - You must be precise & concise when reporting problems on the phone

Your responsibilities as a DAQ Shifter

- **Be active!** **Do not just wait for instructions**
-
  - This will increase the data taking efficiency.
  - If you think time is being wasted, talk to the shift-leader.
  - If you think a specific sub-system is blocking for too long the data taking talk to the shift leader.
    - You as a cDAQ shifter probably have the best feeling which subsystem is blocking data-taking.

  - When you are active, you learn more about the other systems, too. This makes shifting much more fun, and is a good thing for CMS!

- **But always:** **Do nothing without informing the shift leader**
  - In particular when subsystem experts contact you directly: Always involve the shift leader that he/she knows what is going on!

  - Always make sure that either you or the shift leader have contacted the relevant sub-system expert (DOC) before deciding to remove subsystems or FEDs from the run

E-Log

- Document your shift in the e-log
  - Your entries are essential to make CMS run efficiently. The e-log is a primary source of information for improving the system. With your entries in the e-log you are part of the team which tries to improve the online system to achieve “smooth operation”.
- There should be an e-log window already open. Logout the old shifter and log in yourself as soon as you start your shift
  - Use the Subsystems>DAQ>DAQ area – there is a link to it in the shifters guide
- Please “submit” comments in a timely manner
  - That way people offsite can monitor what’s going on
    - Many short entries are preferable to one long log entry at the end of your shift!
- Document any issues that come up or observations you have about the DAQ, e.g.
  - If you have to constantly restart or resync a subdetector, If the DAQ goes into error at any time, anything that seems funny or you don’t understand
  - Make sure to copy / paste any error messages that are relevant
    - From hotspot / RCMS / handsaw / etc.
  - Add context information to the errors
- Give a meaningful & correct subject to your log message
    - E.g.: “Run blocked due to HCAL FED 1122 sending events out of sequence” (instead of “DAQ crashed”)

Shift Bulletin Board

- Before each shift check the Shift Bulletin Board
  - This Twiki page contains
    - Procedures
    - Settings (e.g. sub-system RUN Keys, FEDs that are out for a period of time)
    - Known problems (and workarounds)
    - (temporary) instructions
  - If you do not fully understand, check with the previous shifter or the DAQ on-call **at the beginning of your shift**

- You must keep the Shift Bulletin Board up-to-date
  - The next shifter will rely on it

General Rules and Policies

- **Security (computing):**
-
  - Never write down passwords in public places where other people have access to (files in your home account, paper on your desk, etc)
  - Do not give the passwords to other people
    - It is the on-call experts which give the relevant information to the shifters

General Rules and Policies

- **Take your work seriously:**
  - If you have time (during a long smooth run in the night):
    - Of course you may check your email on your portable
    - Of course you may write an email
    - You SHOULD eat and drink something during your shift...
  - BUT : You must continuously watch the screens to check that the data taking proceeds as it should.

- **It is unacceptable that the run stops due to some problem and you do not realize this for 5 minutes.**
-
- **Shifting is real work, and unfortunately not always exciting or breath-taking...**

- **If you manage to do efficiently other work during your shift, then you do not take the shifting seriously !**

General Rules and Policies

- **In case you are in trouble:**
  - FIRST INFORM THE SHIFT LEADER
    - Tell him/her what you suggest to do next. (Sometimes he/she does not know what to do...)
    -
- **You cannot get a run going and beam is there or imminent**
  - You have serious doubts that the data is taken correctly and/or there is a problem in the central DAQ
  - **CALL THE DAQ ON-CALL EXPERT AT ANY TIME**
  - The experts are there to help you at any time. (This is why they do not need to do other shifts)

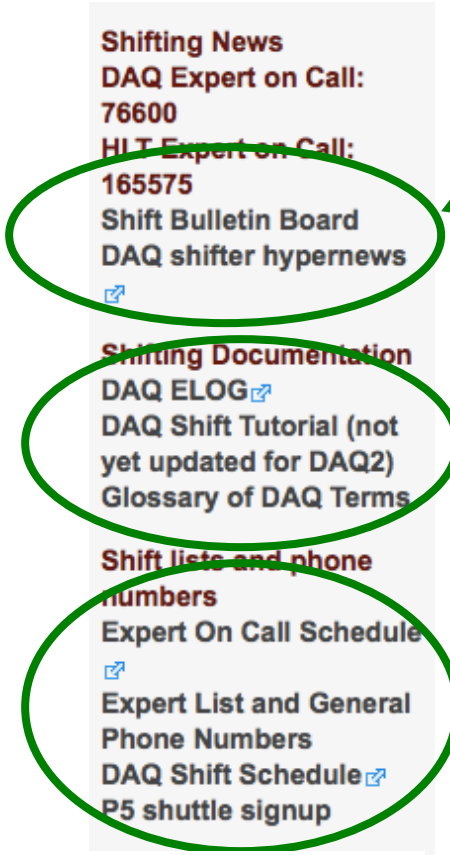
- **If you have a problem or question which you are sure is NOT critical to efficient data taking**
  - Document your problem in e-log
  - Call the expert at any time during the day/morning/evening.
  - The experts are also there in order to make you more expert!!

Your shift outline

- Arrive 15 minutes early!
  - Discuss with previous shifter any problems / issues / requests
  - Read through the shift bulletin board and make sure you understand it
  - Log into the Elog and begin documenting your shift
- If a run is ongoing, check all the monitoring screens
  - Is the data flowing ok?
  - Do you understand the trigger rate? Is the trigger correct?
- Follow any requests from the shift leader
  - ***Never*** include / remove subdetectors or FEDs without talking to the shift leader
- When you have time, take a tour of both the central control room and the subdetector room
  - Introduce yourself to your fellow shifters!
- If questions/problems do not hesitate to call the DAQ DOC (x76600)

Documentation and Resources

- DAQ2 shifters guide twiki page
  - https://twiki.cern.ch/twiki/bin/view/CMS/ShiftPourNuls2014
- The leftbar of the DAQ2 shifters guide has many valuable links:



DAQ shifter bulletin board: read before every shift.
DAQ shifter hypernews: subscribe to this! All DAQ shift related announcements are sent here

DAQ ELOG: Link to DAQ area of the ELOG
DAQ Shift Tutorial: link to slides from shift tutorial
Glossary of DAQ Terms: definition of all the DAQ acroynyms.

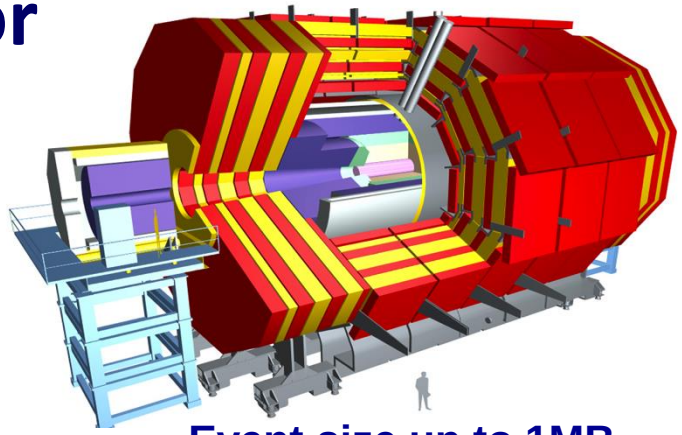
Expert on call: link to DAQ DOC area of shift tool
Expert List: link to list of DAQ and HLT experts
DAQ shift schedule: link to DAQ shifters area of shift tool
P5 shuttle: link to shuttle schedule

- Shifter bookmarks : http://cmsdaqweb/daqpro/ShifterBookmarks.html
- Questions about shifts: cms-daqshift-office@cern.ch

Part 2: The central DAQ system

The Central DAQ during Run-1
DAQ-1

CMS DAQ for LHC Run-1

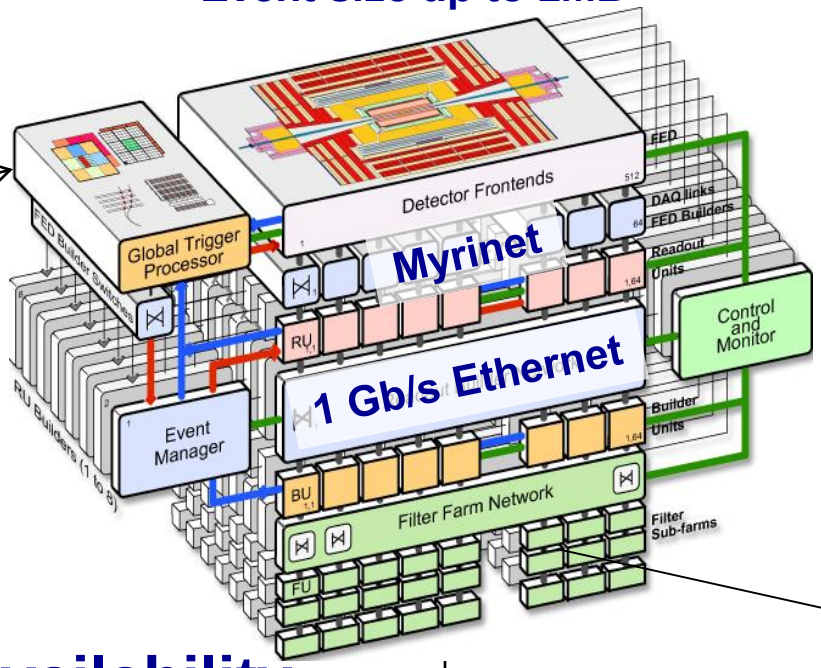


Bunch crossing rate
40 MHz nominal

Only 2 trigger levels in CMS

Event size up to 1MB

Level-1 Trigger accepting 100 kHz
Custom electronics



DAQ: 100 GB/s bandwidth

2-stage event builder
Myrinet
Gigabit Ethernet

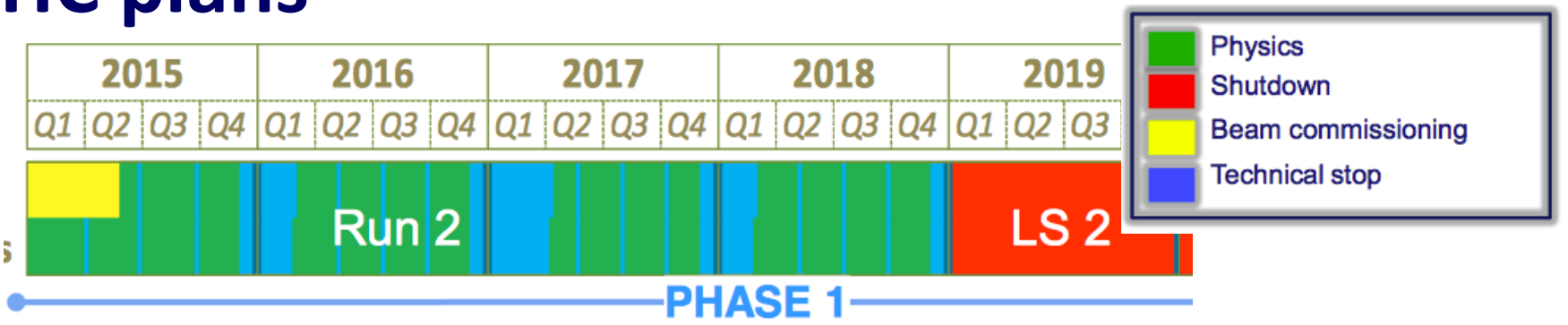
High-Level Trigger
working on full events
13000 cores
~500 Hz accept rate

99.6 % cDAQ availability
(2010-2013 physics runs)

up to 1.2 GB/s to storage

Why build a new DAQ?

LHC plans



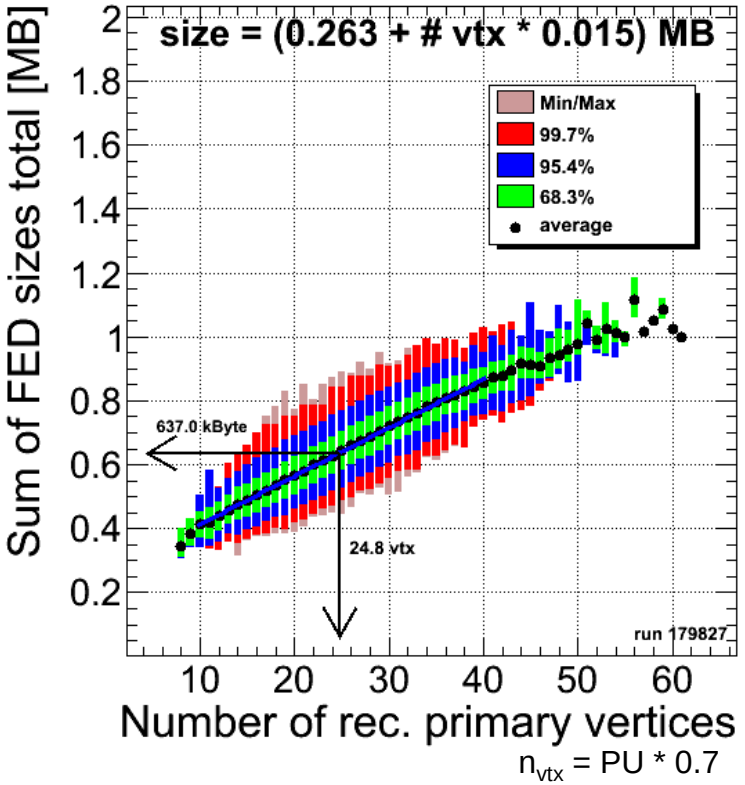
13 TeV center-of-mass energy
40 MHz (25 ns) operation
targeting 40 fb^{-1} / year
higher pile-up

Plan for startup after LS1 (2015)

BX spacing [ns]	Beam current [$x10^{11}$ e]	Emittance [μm]	Peak Lumi [$x\ 10^{34} cm^{-2} s^{-1}$]	Pileup
25	1.15	3.5	0.92	21
25	1.15	1.9	1.6	43
50	1.6	2.3	0.9-1.7	40-76 (*)
50	1.6	1.6	2.2	108 (*)

LHC run-2 pile-up scenarios

(*) LHC will perform luminosity leveling to limit pile-up to 50



n_{vtx} = PU * 0.7

CMS event size

(Run-1 subsystems)

New or upgraded detectors in CMS

- Several detectors / online-systems being upgraded to cope with higher luminosity
- Increase of event size
- Readout electronics of upgraded systems based on µTCA

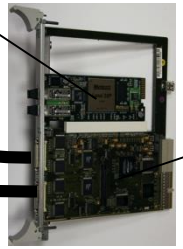
- 2014: New Trigger Control and Distribution System
- 2014: Stage-1 calorimeter trigger upgrade
- 2014/15: new HCAL readout electronics
- 2016: Full trigger upgrade
- 2017: New pixel detector and readout electronics

SLINK sender mezzanine plugged onto VME electronics

Fragment size 1..4 kB

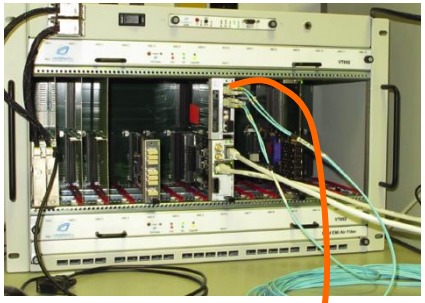
SLINK-64 copper cable 400 MB/s

Myrinet NIC



Frontend-Readout Link

640 Legacy Links: SLINK-64
(600 after pixel upgrade)

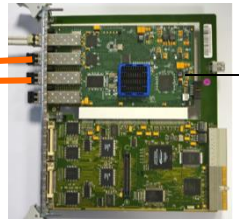


SLINK express sender = IP-core µTCA electronics

"AMC-13" card used by many subsystems

Fragment size 2..8 kB

Optical SLINK-express 4 Gb/s or 10 Gb/s - retransmit



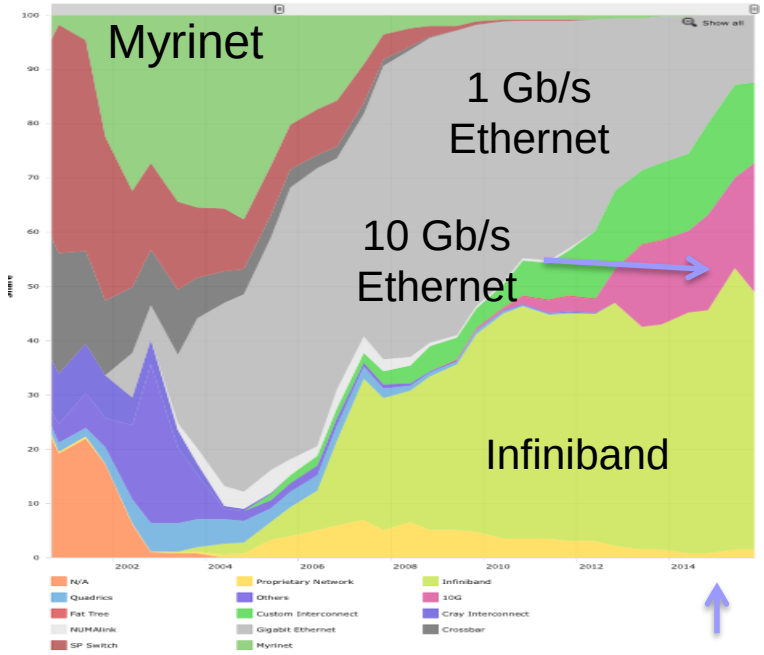
Frontend-Readout Optical Link

+ 50 new Links: SLINK-express
(170 after pixel upgrade)

Other reasons to upgrade

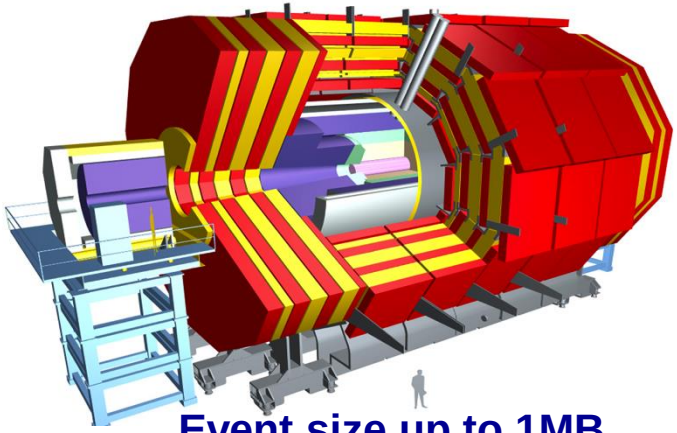
- Ageing hardware
  - Most PCs of Run-1 system at end of life cycle
  - NICs of Run-1 based on PCI-X
- New technologies
  - Myrinet widely used when DAQ-1 was designed
  - Today Ethernet and Infiniband dominate the Top-500 supercomputers

DAQ1 TDR (2002)

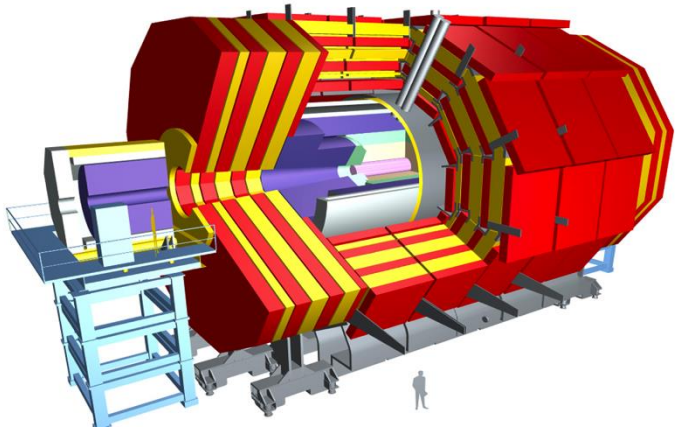


2014

Top500.org share by Interconnect family

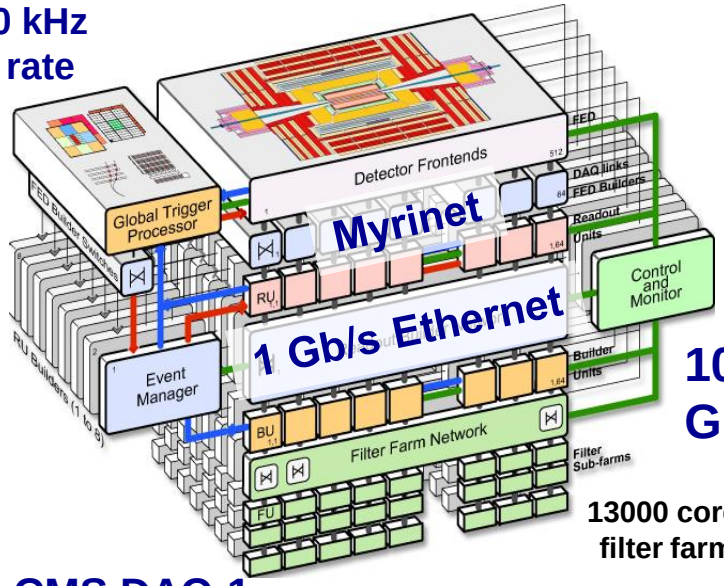


Event size up to 1MB



Event size up to 2MB
(large margin)

100 kHz
L1 rate

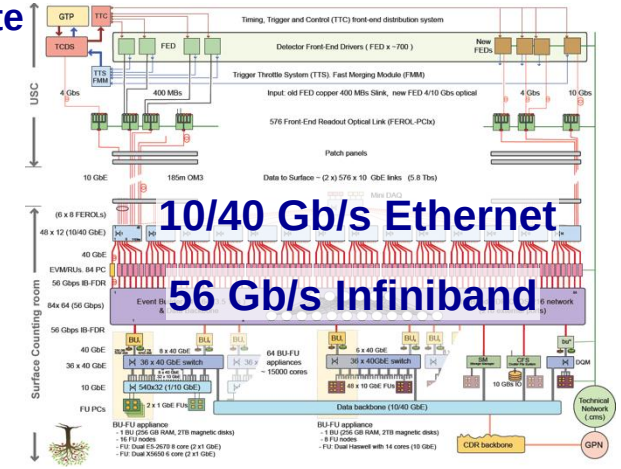


100
GB/s

13000 core
filter farm

CMS DAQ 1

100 kHz
L1 rate



200
GB/s

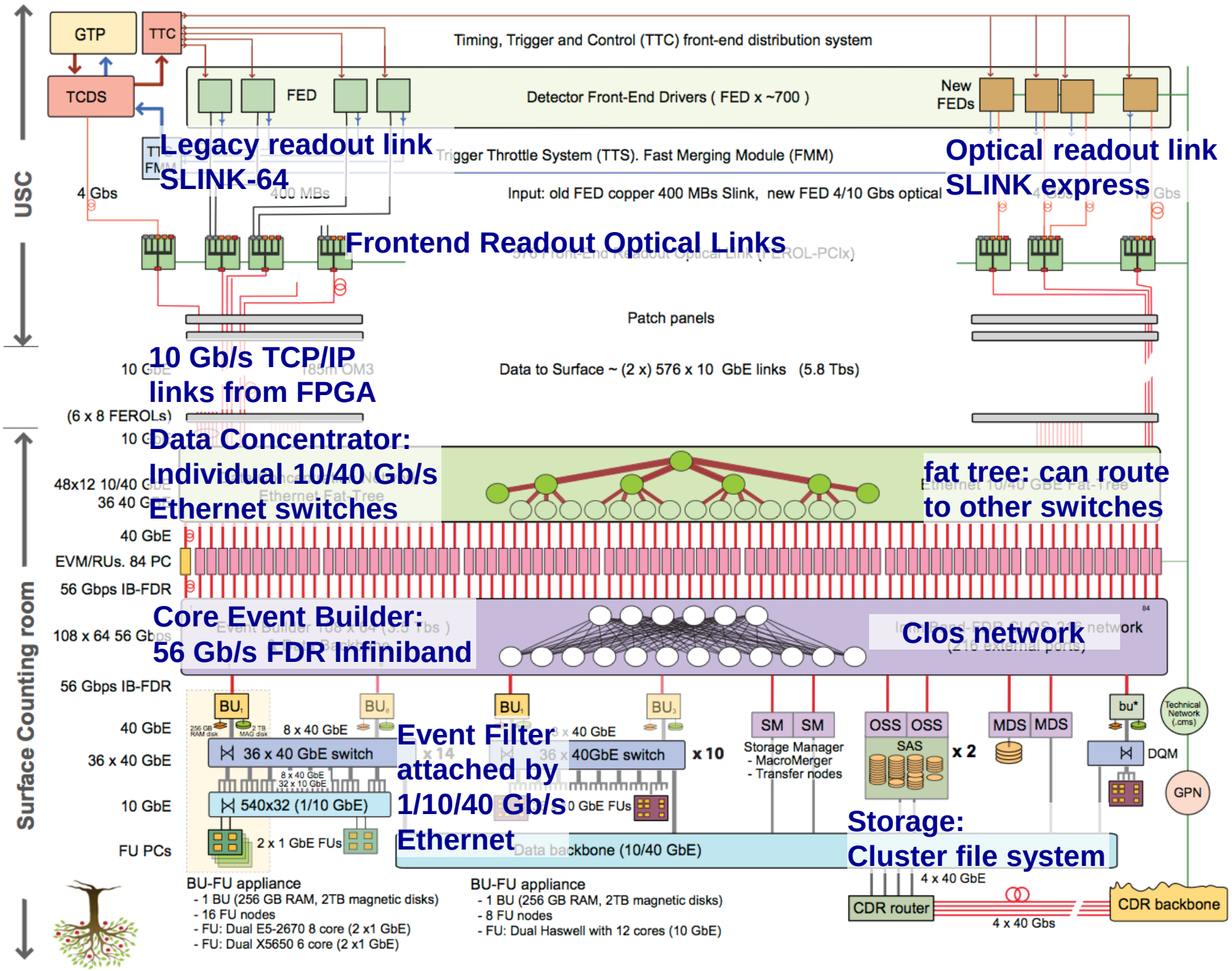
16000+ core
filter farm

CMS DAQ 2

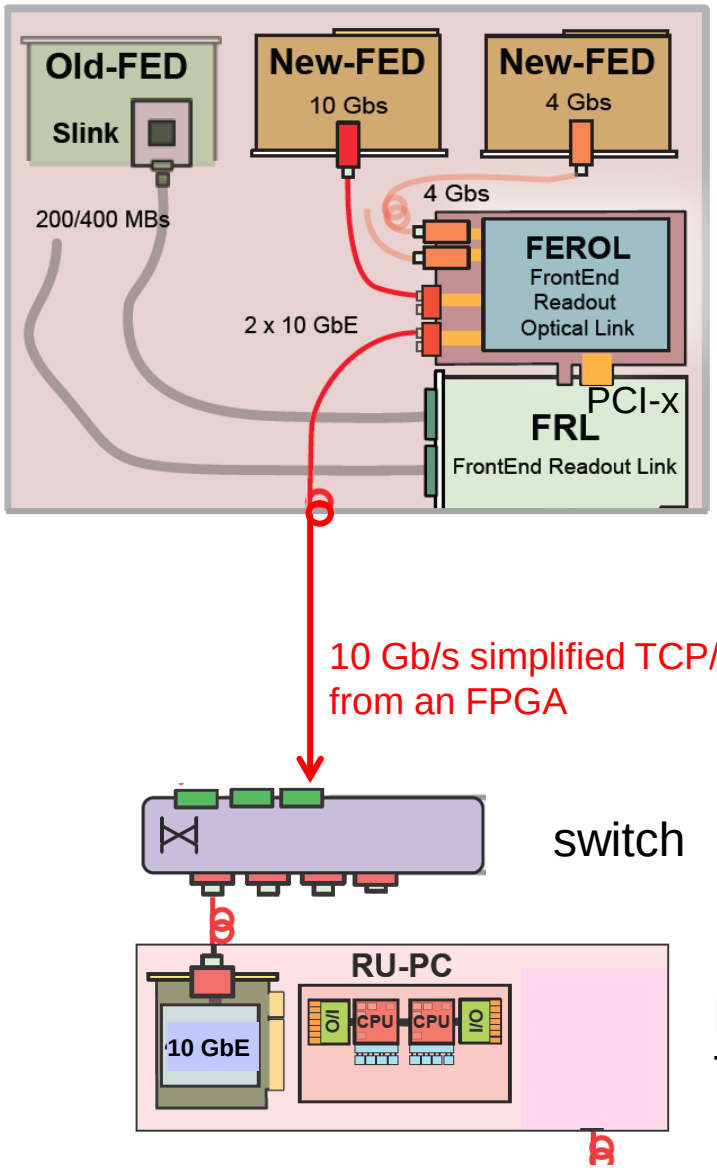
max. 1.2 GB/s to storage

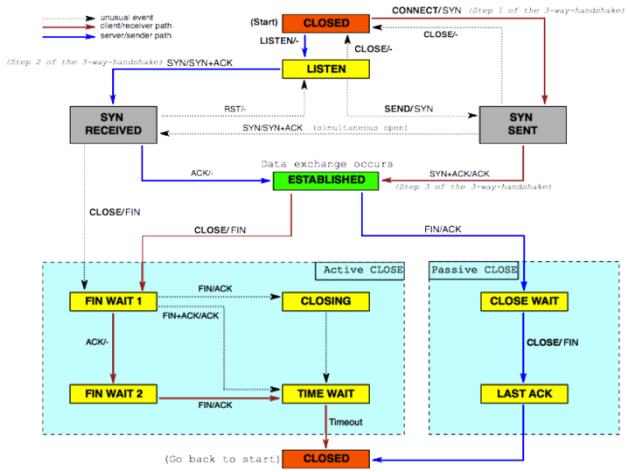
~ 3 GB/s to storage

The Central DAQ during Run-2
DAQ-2


USC
Surface Counting room
GTP
TTC
TCDS
Timing, Trigger and Control (TTC) front-end distribution system
FED
Detector Front-End Drivers (FED x ~700)
New FEDs
Legacy readout link
SLINK-64
Optical readout link
SLINK express
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
Frontend Readout Optical Links
Patch panels
10 Gb/s TCP/IP
links from FPGA
Data to Surface ~ (2 x) 576 x 10 GbE links (5.8 Tbs)
(6 x 8 FEROLs)
Data Concentrator:
Individual 10/40 Gb/s
Ethernet switches
fat tree: can route
to other switches
48x12 10/40 GbE
36 40 GbE
40 GbE
EVM/RUs. 84 PC
56 Gbps IB-FDR
Core Event Builder:
56 Gb/s FDR Infiniband
108 x 64 56 Gbps
Clos network
56 Gbps IB-FDR
BU
40 GbE
8 x 40 GbE
36 x 40 GbE switch
36 x 40 GbE
10 GbE
540x32 (1/10 GbE)
FU PCs
2 x 1 GbE FUs
Event Filter
attached by
1/10/40 Gb/s
Ethernet
x 10
SM SM
Storage Manager
- MacroMerger
- Transfer nodes
OSS OSS
SAS
x 2
MDS MDS
bu*
DQM
Technical Network (.cms)
GPN
Storage:
Cluster file system
4 x 40 GbE
CDR router
CDR backbone
4 x 40 Gbs
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)

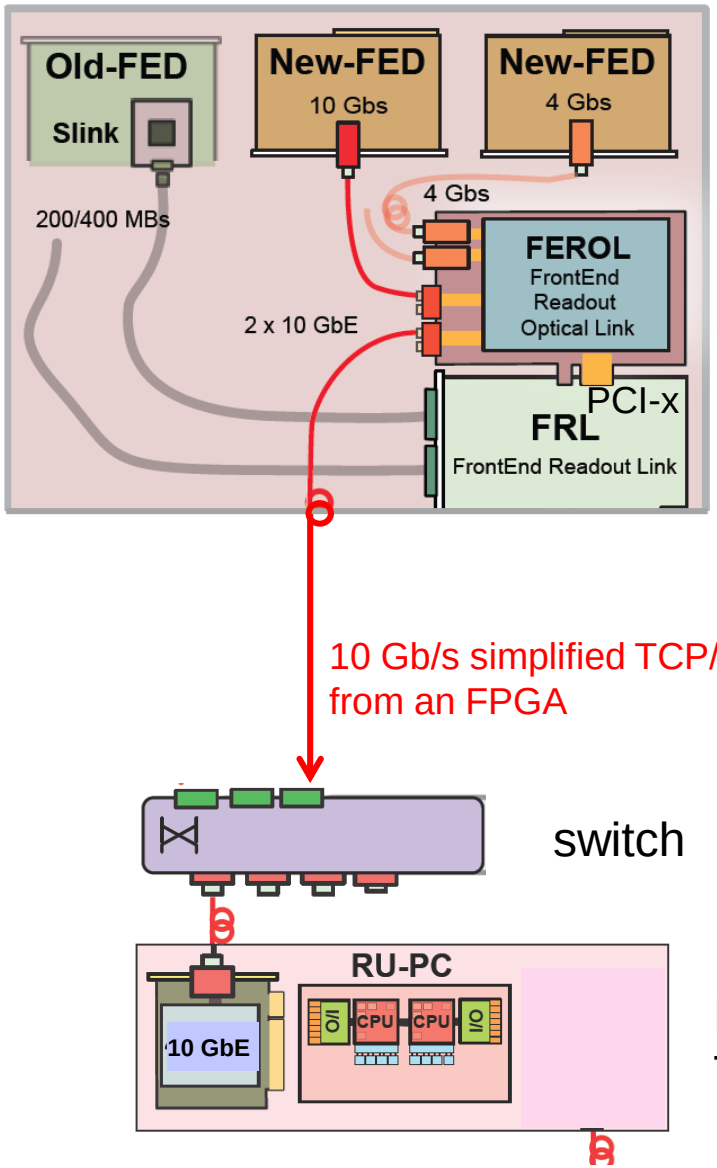
Frontend-Readout Optical Link

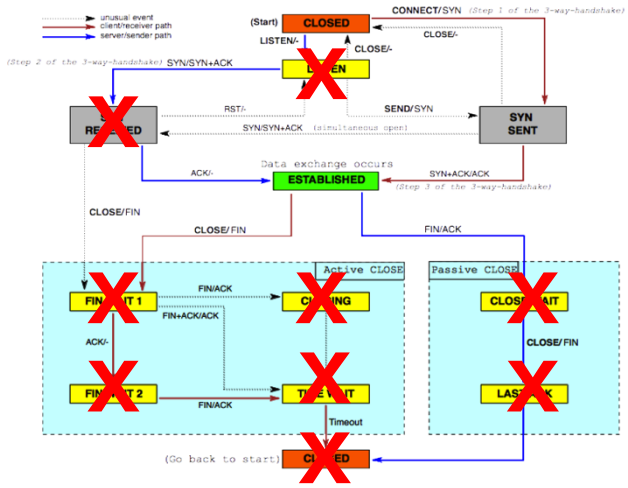




TCP/IP in principle difficult to implement in an FPGA …

Frontend-Readout Optical Link





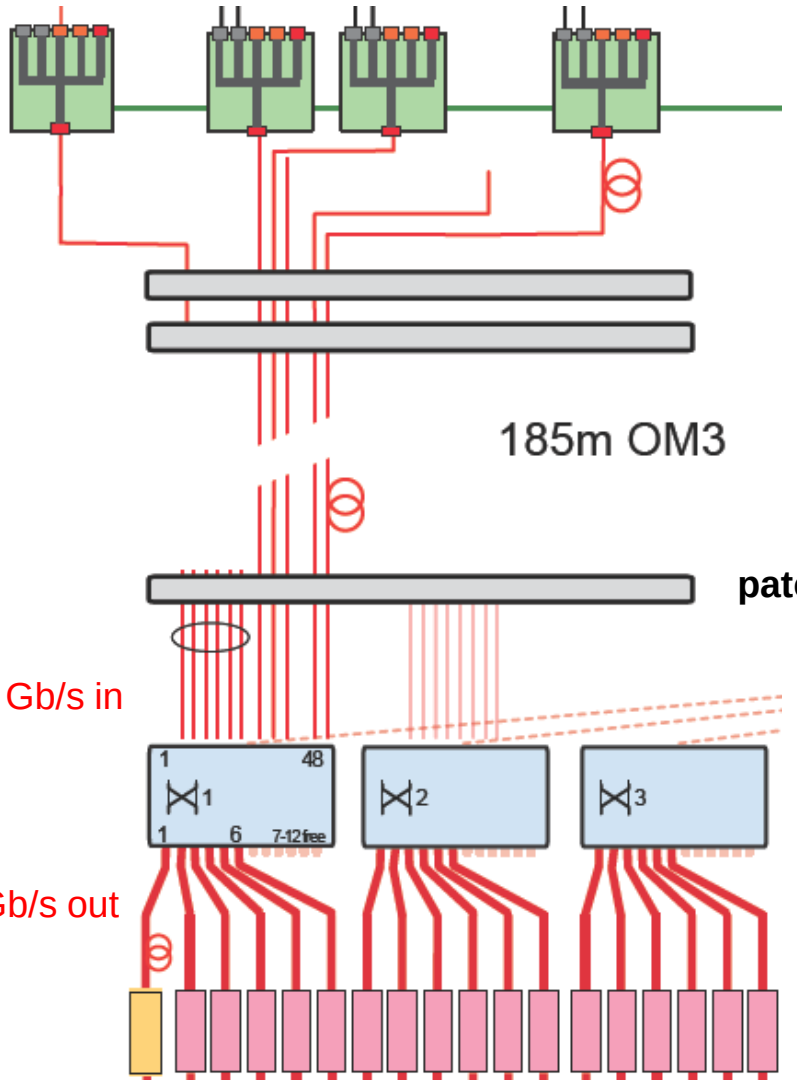
Simplified unidirectional TCP/IP only needs 3 states

PC running standard TCP/IP Linux stack

New in 2017: FEROL 40 board

- 4x 10 Gb/s SLINK Express in
- 40 Gb/s (4x 10 Gb/s) TCP/IP to DAQ
- uTCA standard
- Used to read out Pixel Upgrade

Data concentrator



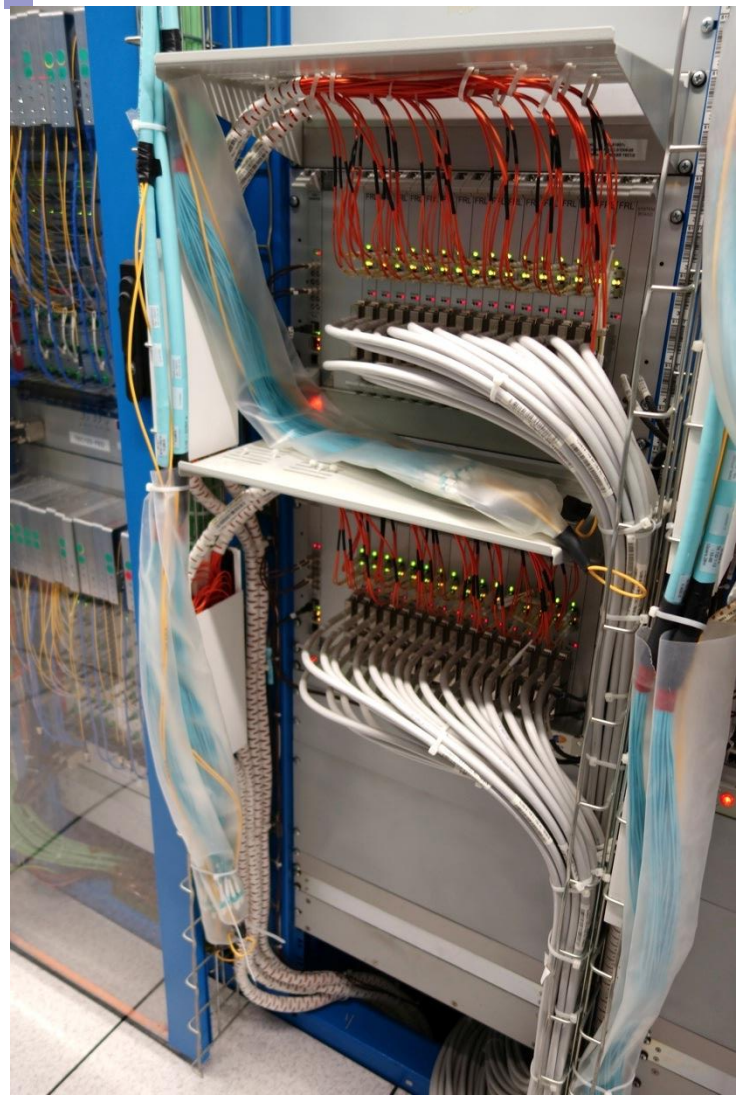
48 Frontend Readout Optical Links (FEROLs) per data concentrator switch

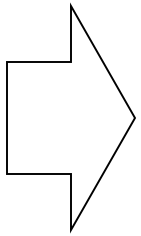
patch panels

links to fat tree

Individual 10/40 Gb/s Ethernet switches
Mellanox MSX1024

Readout Unit (RU) PCs

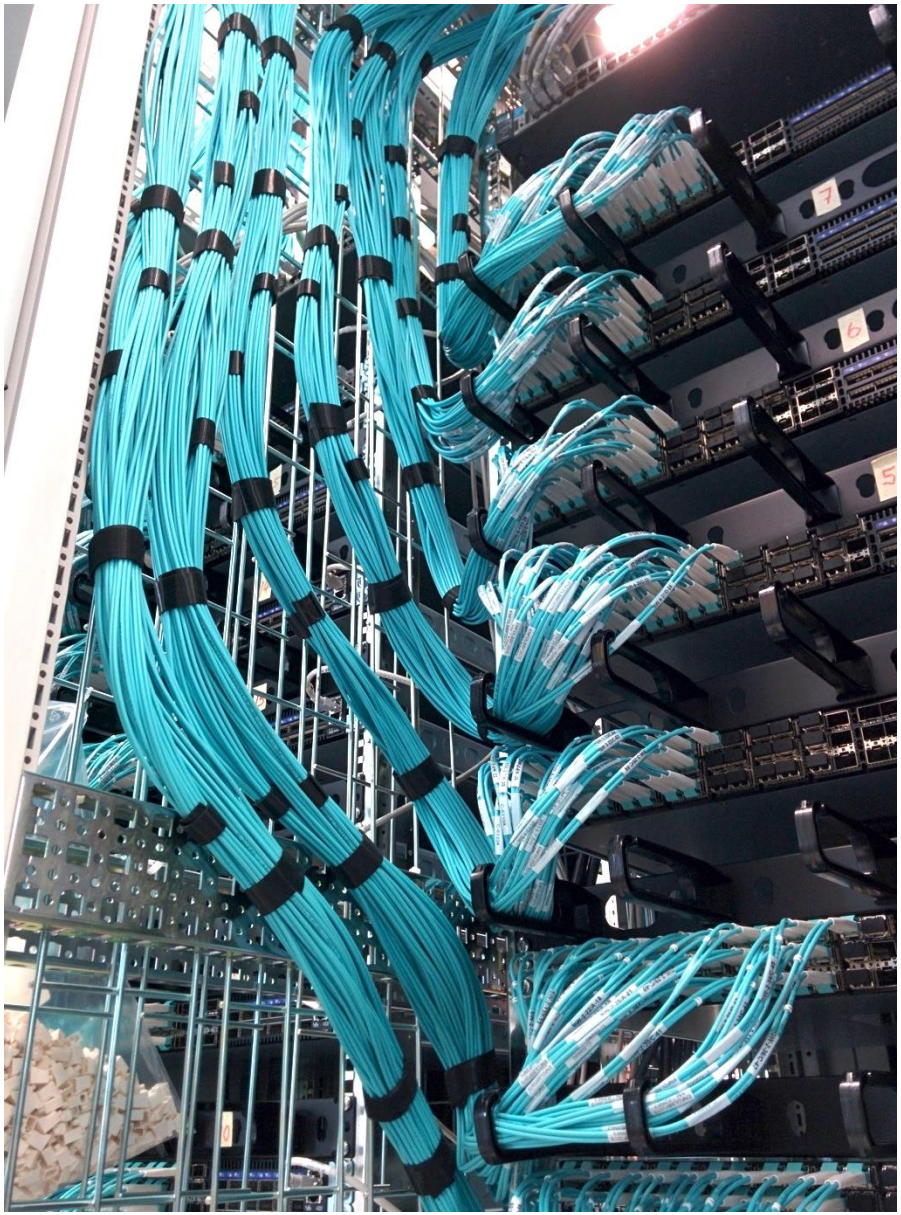




DAQ-1: FRL/Myrinet

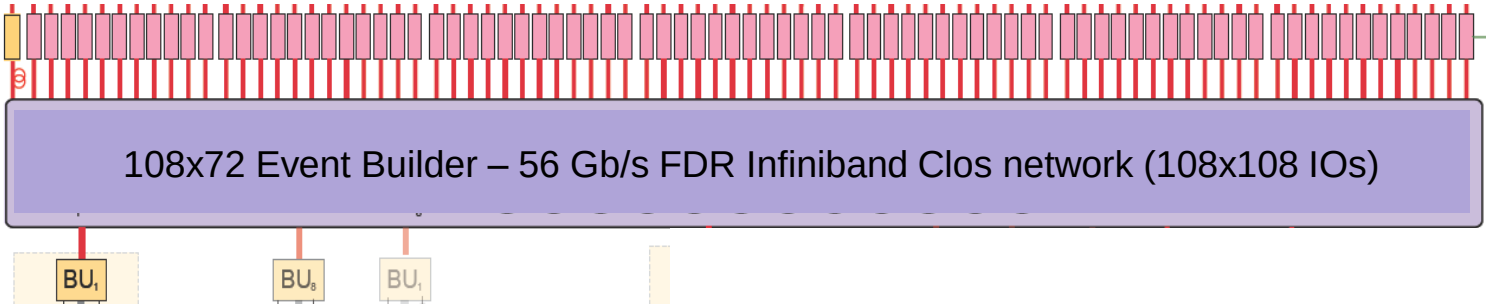
DAQ-2: FRL/FEROL 10 Gb/s Ethernet

Switchover completed



Data concentrator patch panels **and switches**

Core Event Builder



3.5 Tb/s

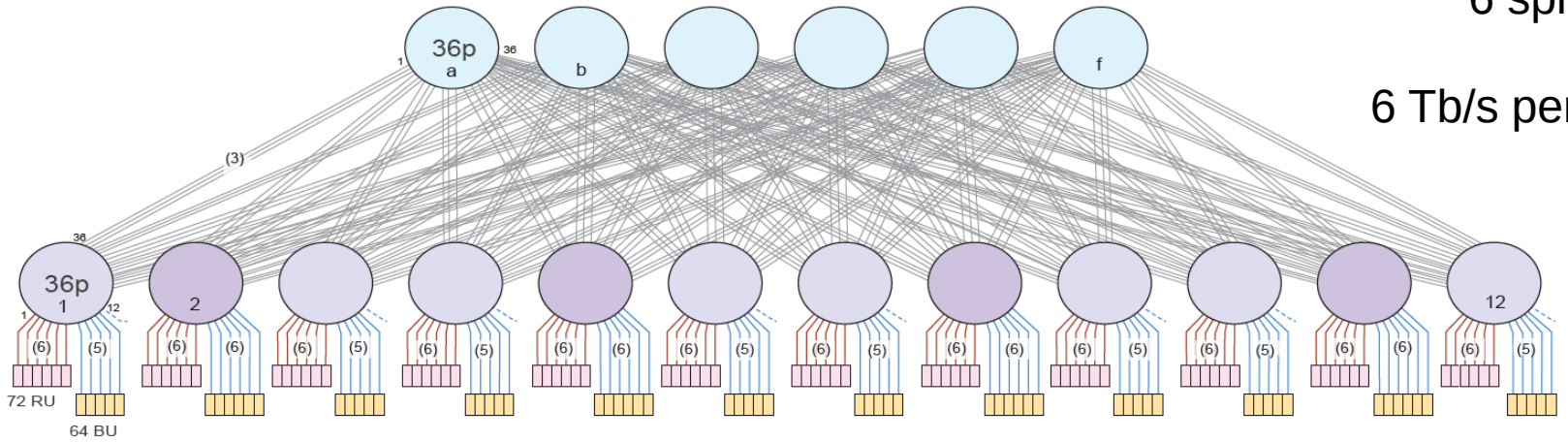
Infiniband

- reliable in hardware at link level (no heavy software stack needed)
- supports credit-based flow control
- switches do not need to buffer
- can construct large network from smaller switches
- cost effective

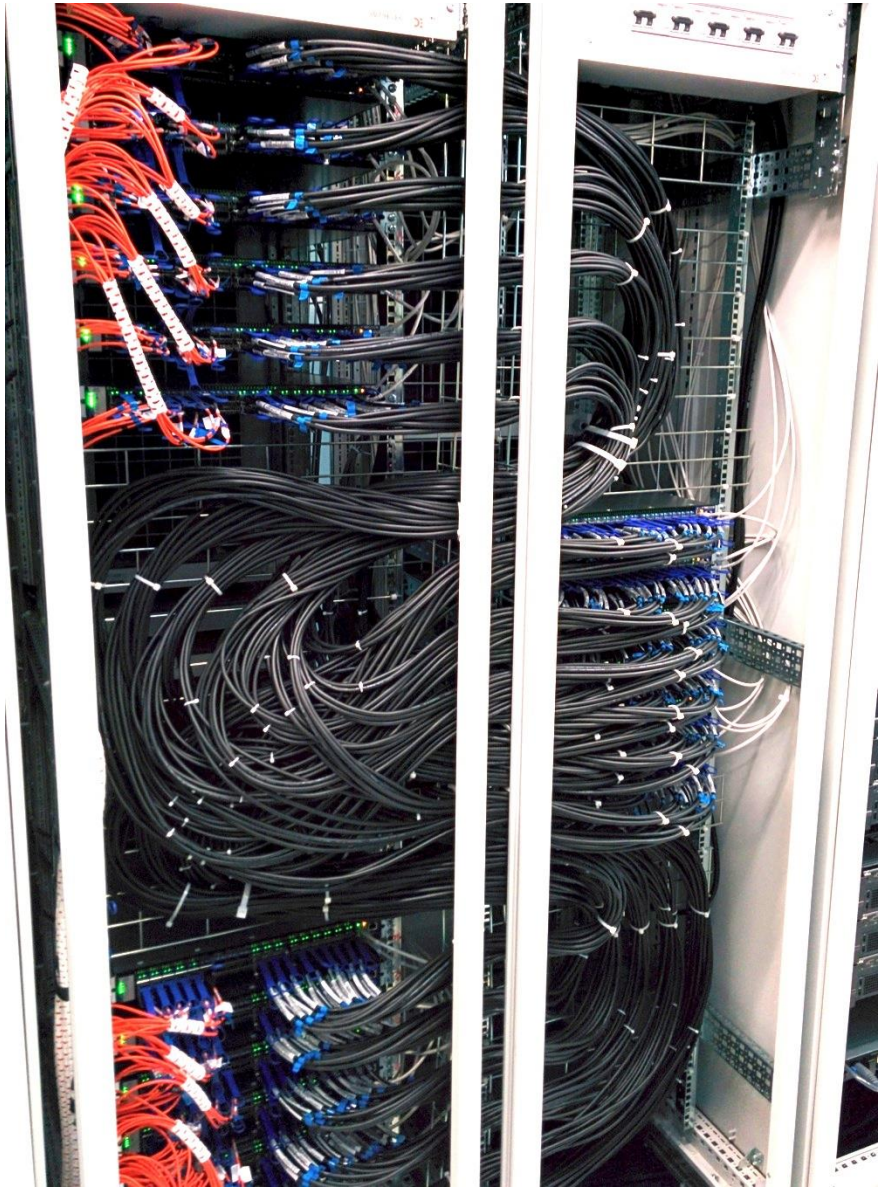
6 spine switches

6 Tb/s per direction

12 leaf switches



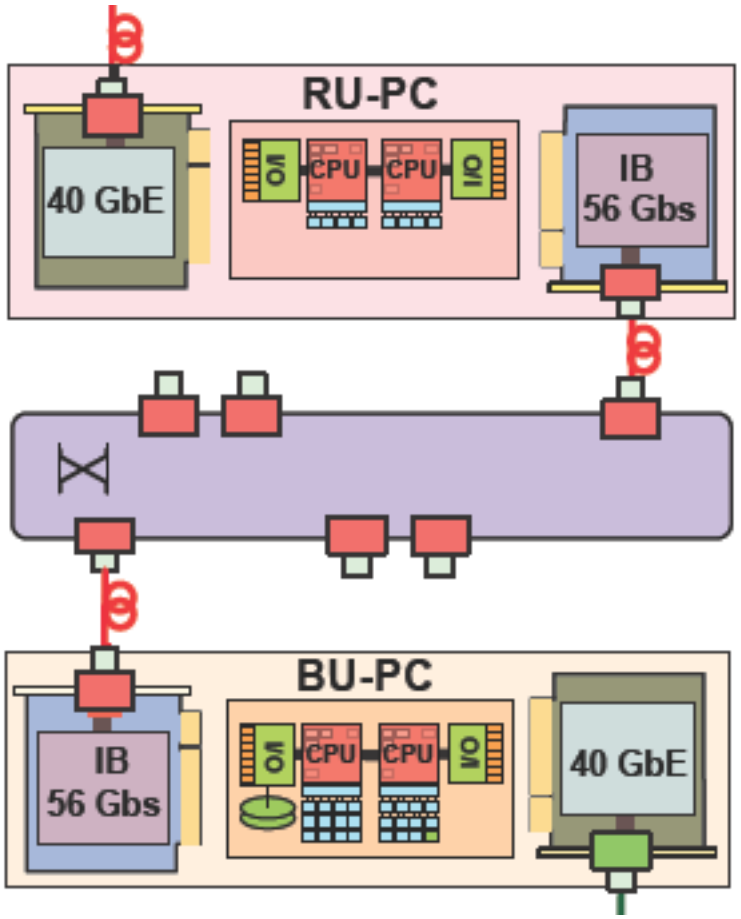
Inputs and outputs mixed on leafs to better utilize leaf-to-spine connections



Infiniband Clos network

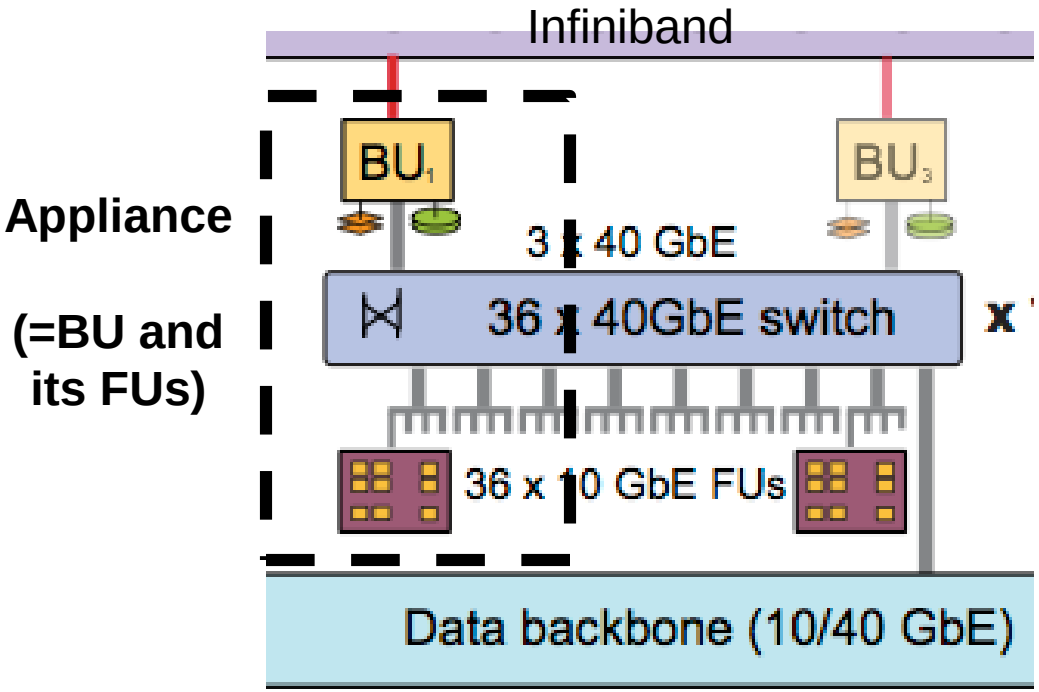
Core event builder performance tuning

- 40 Gb/s Ethernet
  - Linux stack with performance tuning
- 56 Gb/s Infiniband
  - Software based on Infiniband Verbs API
  - All data transport by RDMA
- In both cases:
  - Multiple threads for data reception and writing
  - CPU affinities tuned
    - threads
    - memory pools
    - interrupts



Non-Uniform Memory Access (NUMA)

Filter Farm


Infiniband
Appliance
(=BU and its FUs)
BU1
BU3
3 x 40 GbE
36 x 40GbE switch
x
36 x 10 GbE FUs
Data backbone (10/40 GbE)

HLT farm, DAQ2

2012
64x

2015
90x

2016
81x

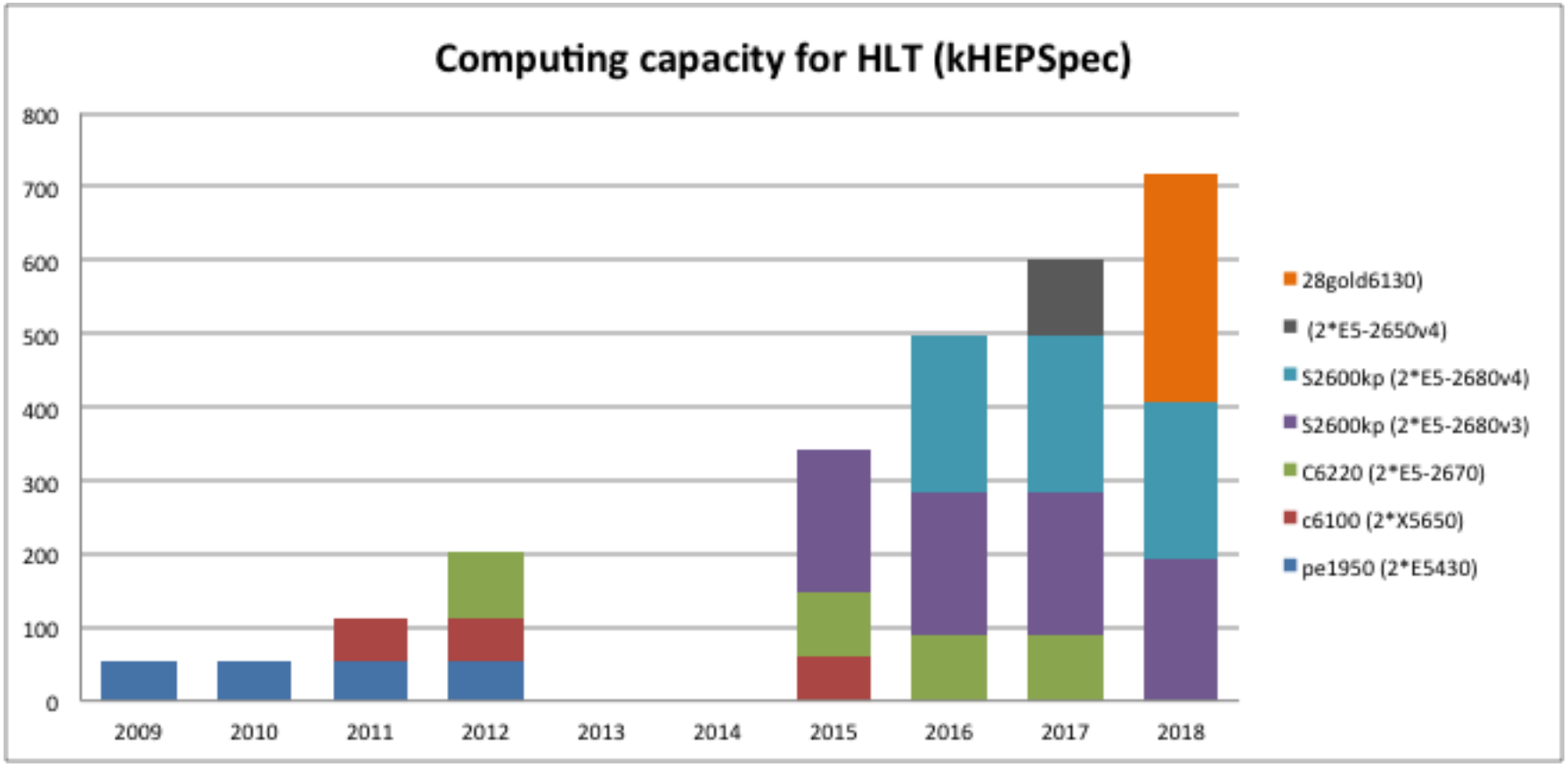
2018
100 x

	2012 extension of DAQ-1 Dell Power Edge c6220	HLT PC 2015 Megware S2600KP	HLT PC 2016 Action S2600KP	HLT PC 2018 Maguay
Form factor	4 motherboards in 2U box	4 motherboards in 2U box	4 motherboards in 2U box	4 motherboards in 2U box
CPUs per mother-board	2x 8-core Intel Xeon E5-2670 **Sandy Bridge**, 2.6 GHz, hyper threading, 32 GB RAM	2x 12-core Intel Xeon E5-2680v3 **Haswell**, 2.6 GHz, hyper threading, 64 GB RAM	2x 14-core Intel Xeon E5-2680v4 **Broadwell**, 2.5 GHz, hyperthreading, 64 GB RAM	2x 16-core Intel Xeon Gold 6130 **Skylake**, 2.1 GHz, hyperthreading, 92 GB RAM
#boxes	64 (=256 motherboards)	90 (=360 motherboards)	81 (=324 montherboards)	100 (=400 motherboards)
#cores	4096	8640	9072	12800
Data link	2x 1Gb/s	1x 10 Gb/s, 1x 1Gb/s	1x 10 Gb/s, 1x 1Gb/s	1x 10 Gb/s, 1x 1Gb/s

Cloud / Spare

Total 2018: 31k cores on 1100 motherboards

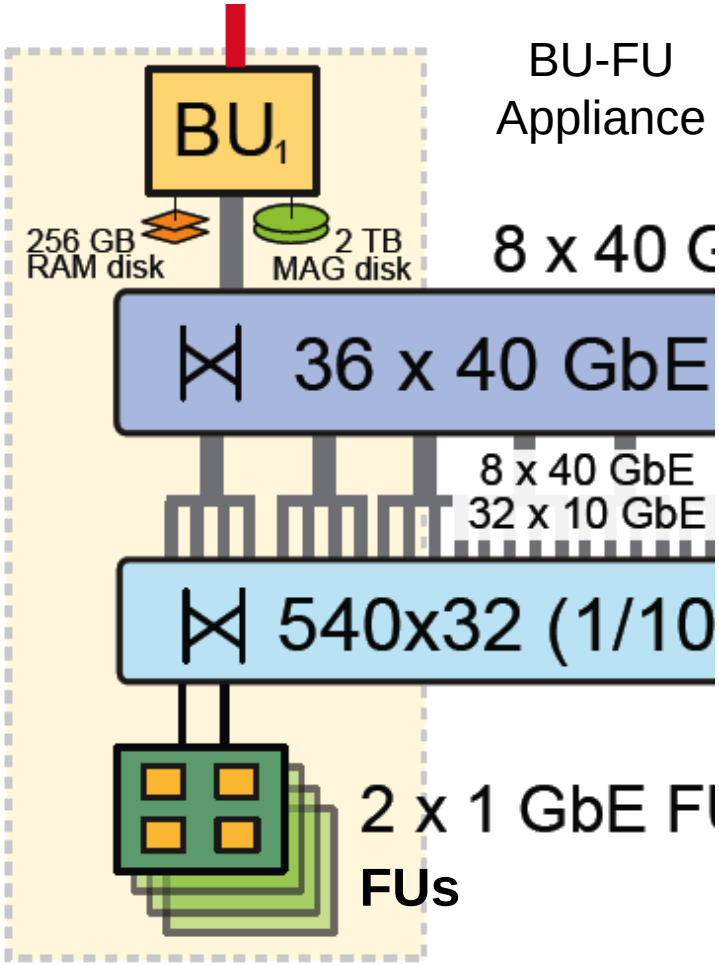
HLT farm, DAQ2



Total 2018: 31k cores on 1100 motherboards

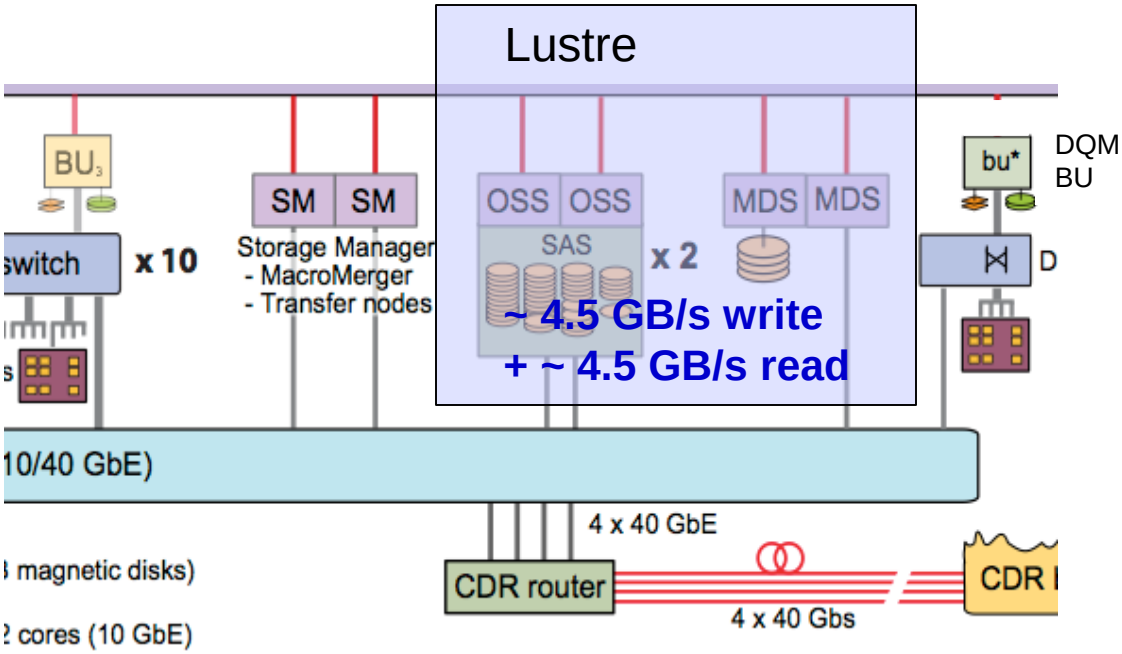
File based Filter Farm Data Flow

- BUs build full events and write them to RAM disk
- Several FU machines per BU run CMSSW processes to reconstruct / filter the events
  - CMSSW input/output is file based
  - BU-FU data transfer uses file systems as a protocol
  - 8-16 FUs mount ram disk via NFSv4 over the BU-FU network :
    - 40 to 10 Gbit Ethernet (1 Gbit on legacy FU)
- The output files of the CMSSW processes are merged by the **Micro-Merger** on the FU and written back to a hard disk on the BU over NFS.



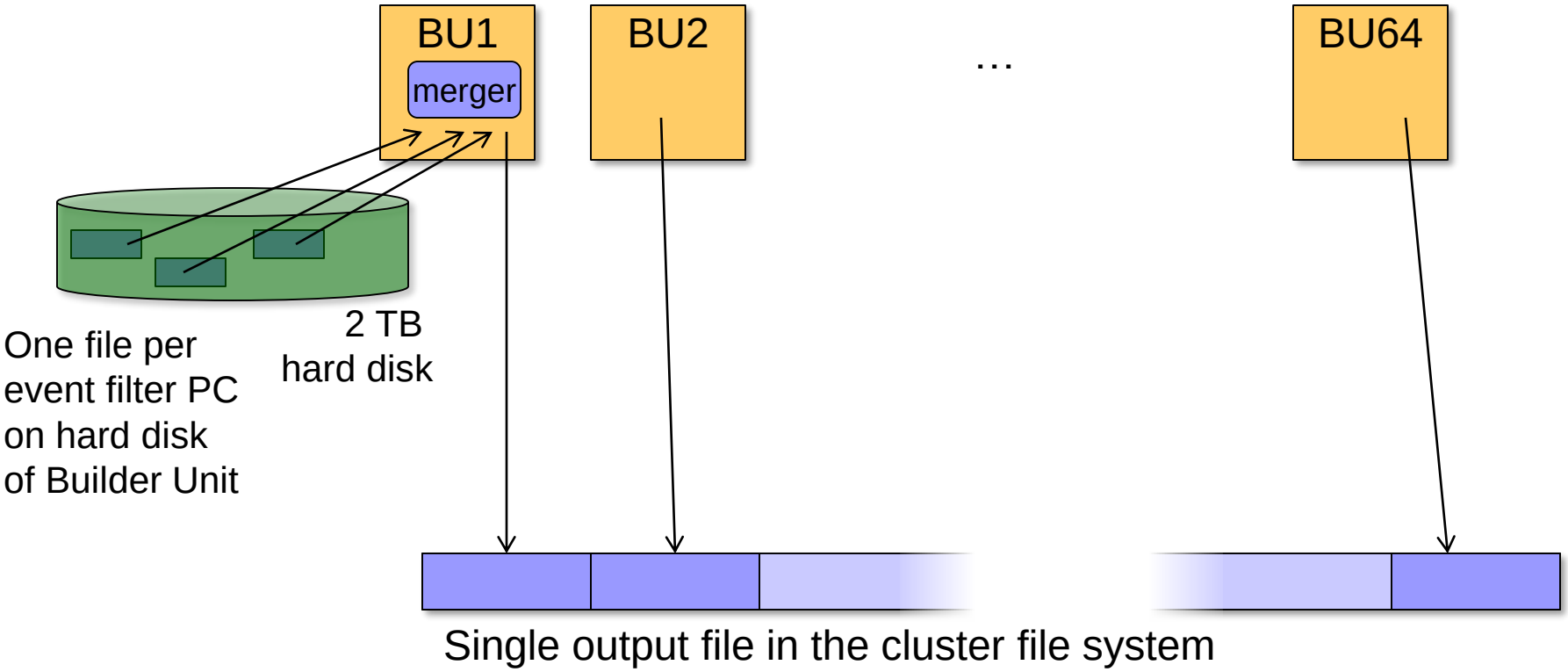
Merging

- File-Based Filter Farm produces output files
  - After micro-merging on FU: 1100 files x 10 streams per lumi section (23s) scattered over hard disks on 72 BUs
  - To be merged to 1 file per stream and lumi section in a central place



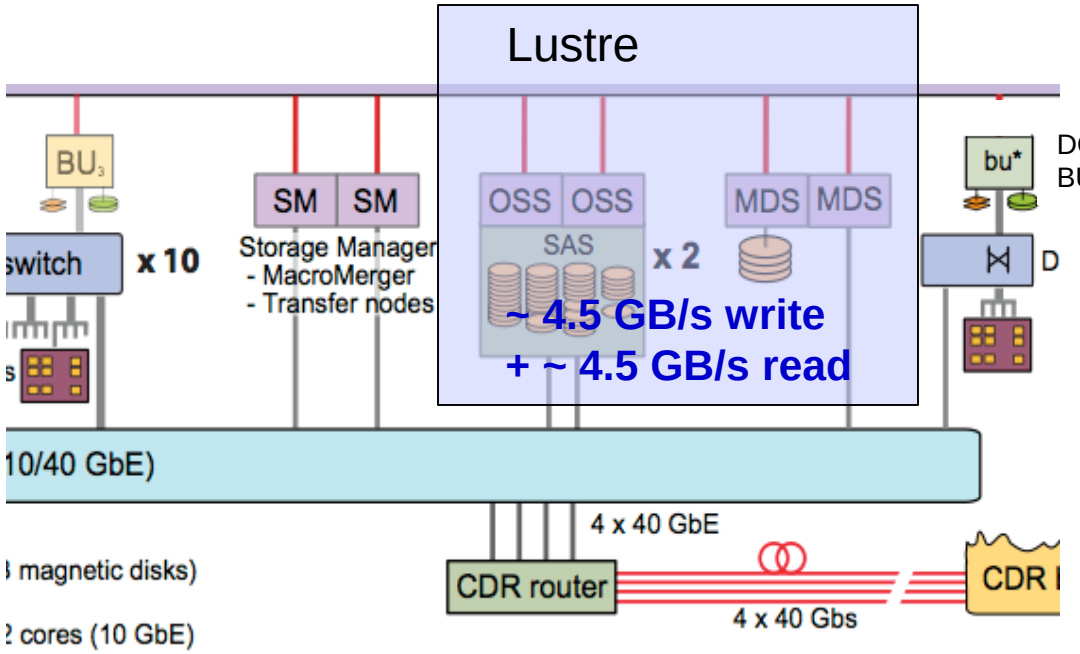
- Merging is done by a Global File System (Lustre)
  - Micro-merging on FU
  - Mini-Merging on BU
  - Macro-merging on dedicated merger nodes

Mini-Merge step

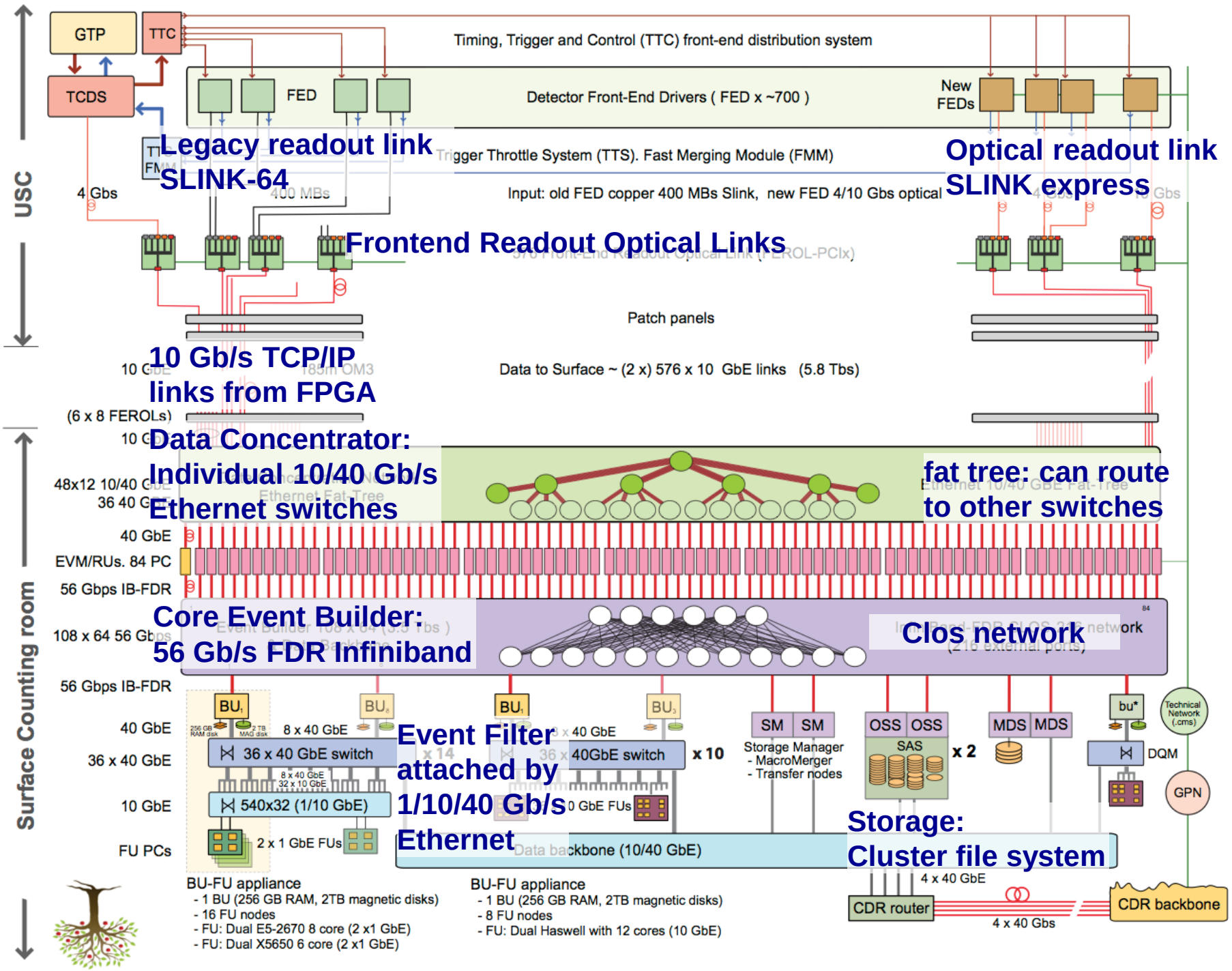


- **Mini-Merge** step
  - Merger process on BU reads data from all FUs in the appliance
  - Data are written directly from the BUs to a single output file per stream in the global file system
    - Exception: DQM streams: One file per BU in the global file system

Macro-Merge step and Transfer System



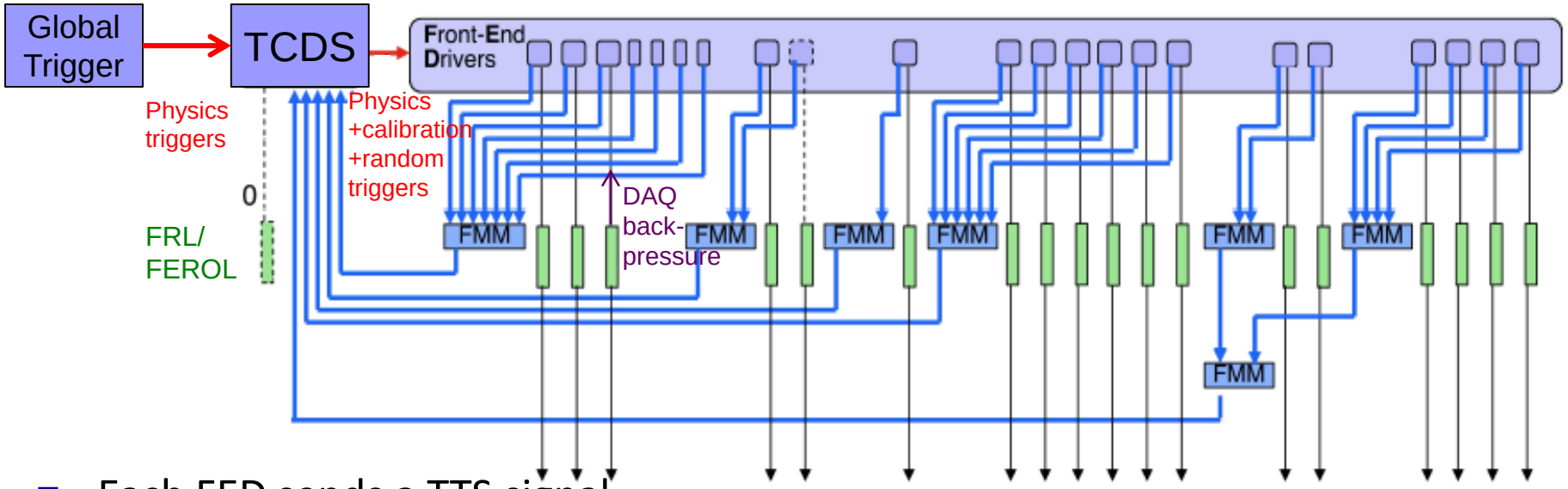
- **Macro-Merge** step
  - Single output file in Lustre is checked and resized
  - Exception: DQM streams
    - Output files per BU are read from Lustre and written to single output
      - Cut on size (2 GB) and timeout of 15s to ensure DQM data gets to DQM in time
    - Histograms are added
- Transfer
  - Merged data are then transferred from Lustre to Tier 0 or to consumers (e.g. DQM/Event Display) at pt.5


USC
Surface Counting room
GTP
TTC
TCDS
Timing, Trigger and Control (TTC) front-end distribution system
FED
Detector Front-End Drivers (FED x ~700)
New FEDs
Legacy readout link
SLINK-64
Optical readout link
SLINK express
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
Frontend Readout Optical Links
Patch panels
10 Gb/s TCP/IP
links from FPGA
Data to Surface ~ (2 x) 576 x 10 GbE links (5.8 Tbs)
(6 x 8 FEROLs)
Data Concentrator:
Individual 10/40 Gb/s
Ethernet switches
fat tree: can route
to other switches
48x12 10/40 GbE
36 40 GbE
40 GbE
EVM/RUs. 84 PC
56 Gbps IB-FDR
Core Event Builder:
56 Gb/s FDR Infiniband
108 x 64 56 Gbps
Clos network
56 Gbps IB-FDR
BU
40 GbE
8 x 40 GbE
36 x 40 GbE switch
36 x 40 GbE
10 GbE
540x32 (1/10 GbE)
FU PCs
2 x 1 GbE FUs
Event Filter
attached by
1/10/40 Gb/s
Ethernet
x 10
SM SM
Storage Manager
- MacroMerger
- Transfer nodes
OSS OSS
SAS
x 2
MDS MDS
bu*
DQM
Technical Network (.cms)
GPN
Storage:
Cluster file system
4 x 40 GbE
CDR router
CDR backbone
4 x 40 Gbs
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)

Flow control

- The entire DAQ from detector to storage is loss-less
- If cDAQ cannot handle the data throughput, back-pressure is propagated all the way back to the FED
  - Bandwidth limitation in any part of CDAQ
  - CPU limitation in the filter farm
  - A failure / crash
- Buffers in the FED may fill up
  - If too much data is coming from the detector
    - Backgrounds, noise, high trigger rate, wrong settings
  - Or if the FED is back-pressured by CDAQ
- In this case the FED throttles the trigger
  - Through the Trigger Throttling System (TTS)
    A tree of Fast Merging Modules (FMMs)

Trigger, TCDS + flow control



- Each FED sends a TTS signal
  - Possible sTTS signals: Busy, Warning, OutOfSync, Error, Disconnected, Ready
- FEDs are grouped into partitions
- FMMs (Fast Merging Modules) merge sTTS signals from FEDs in each partition
- Merged signals sent to TCDS (Trigger Control and Distribution System) which reacts according to the signal
  - i.e. blocks triggers from Global Trigger for all states except Ready
- Special cases of TTS signals that are only seen by TCDS (not by the FMMs):
  - For tracker partitions also the emulated APV state is an input to the partition controller
  - New uTCA FEDs send their TTS status through the TTC Partition Interface to TCDS

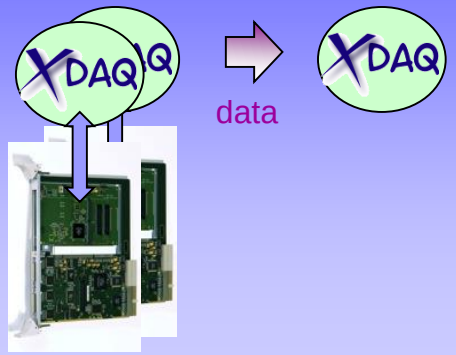
Software

The online software

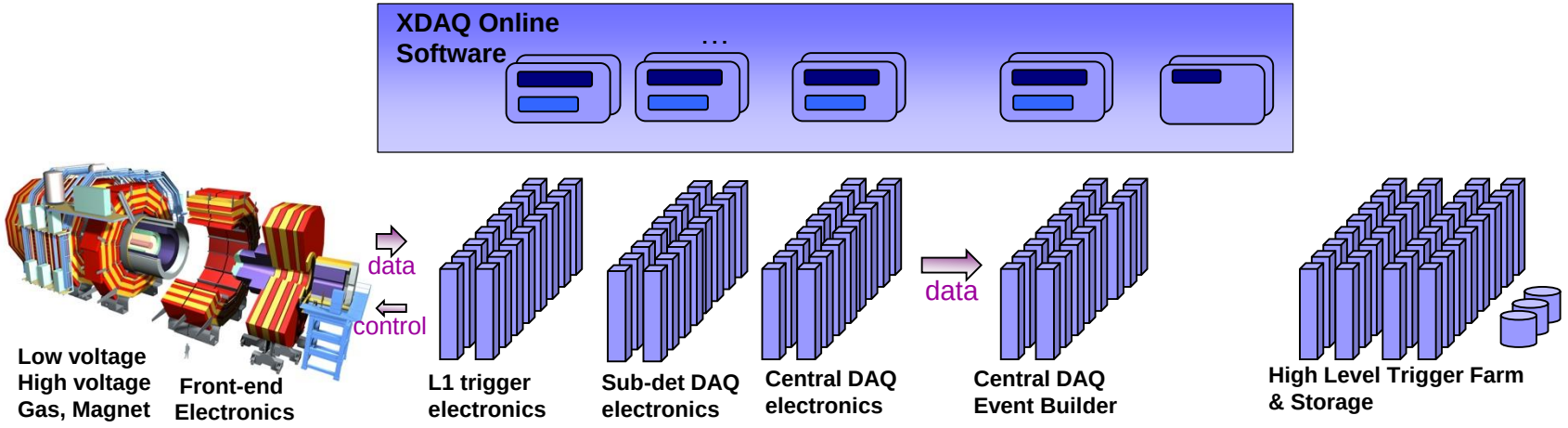
XDAQ Framework – C++, XML, SOAP

XDAQ applications control hardware and data flow

XDAQ is the framework of CMS online software
It provides Hardware Access, Transport Protocols, Services etc.



XDAQ Application



The online software

Run Control System – Java, Web Technologies

Defines the control structure

GUI in a web browser

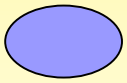
HTML, CSS, JavaScript, AJAX

Run Cont

Apache To

Java Serve

Web Servi

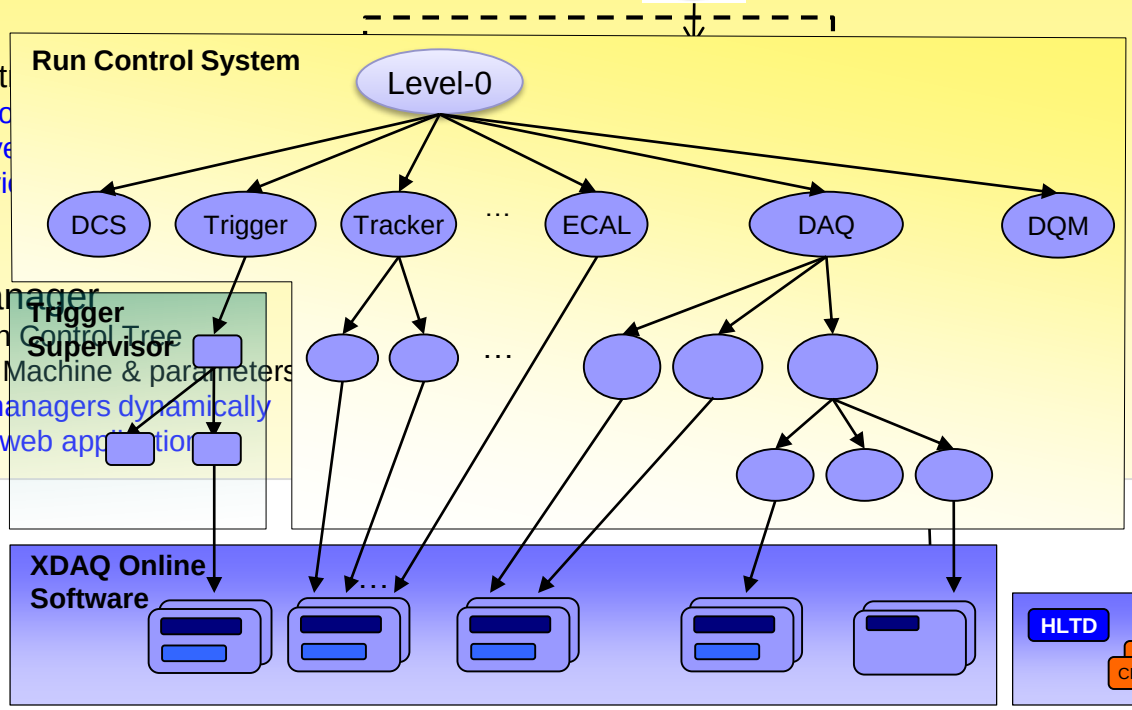


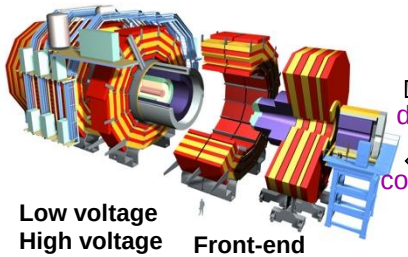
Function Manager

Node in the Run Control Tree

defines a State Machine & parameters

User function managers dynamically loaded into the web appl tion





Low voltage
High voltage
Gas, Magnet

Front-end Electronics

data

control

L1 trigger electronics

Sub-det DAQ electronics

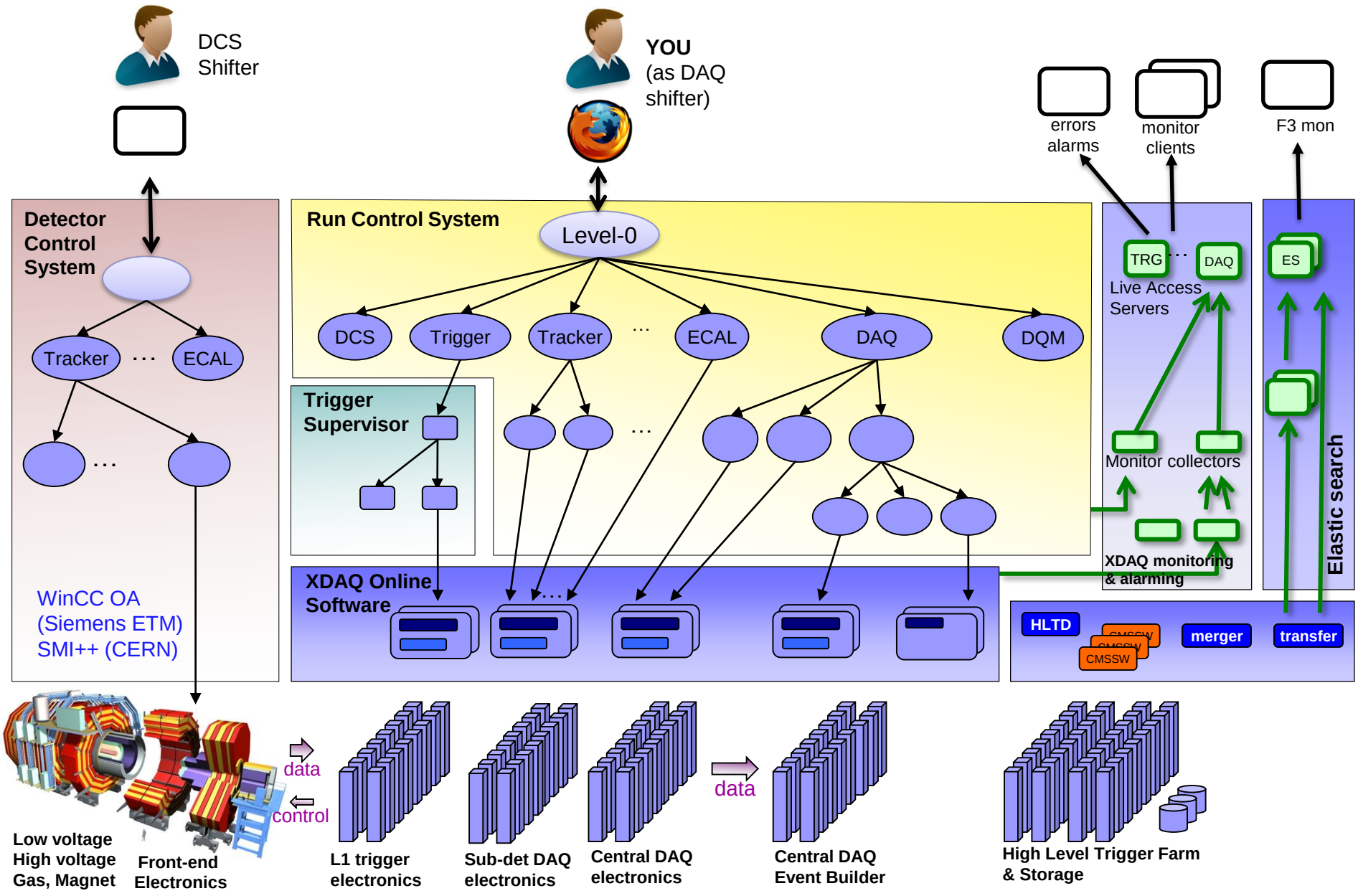
Central DAQ electronics

data

Central DAQ Event Builder

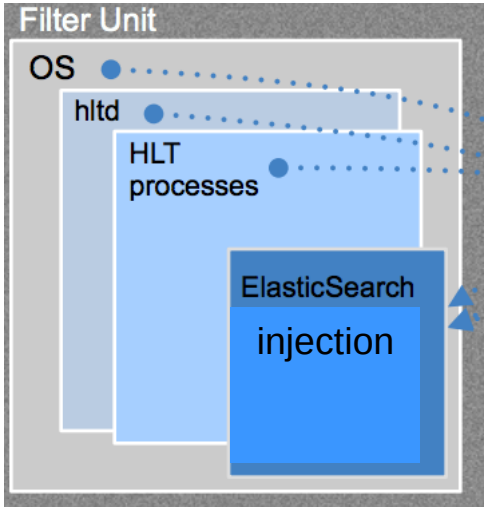
High Level Trigger Farm & Storage

The online software

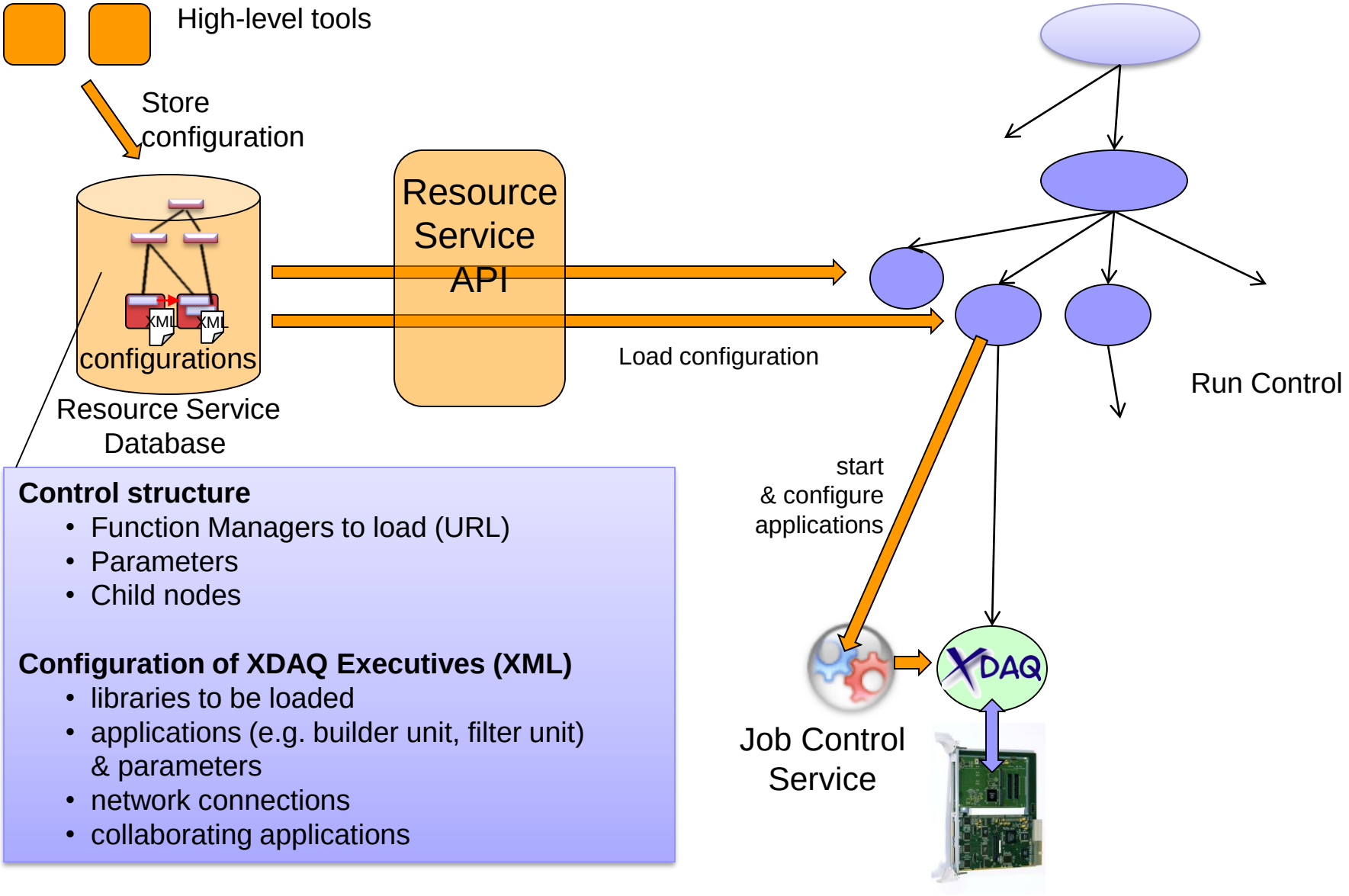


File-based Filter Farm Software

- FFF on appliances is controlled by a service (hltd), asynchronous to Run Control
- hltd is running on BUs and FUs and responsible for:
  - Detecting a new run (run directory in ramdisk appears)
  - CMSSW runs as standard cmsRun jobs, process input files
  - Output bookkeeping and copying merged data files to BU
- Monitoring
  - Using elasticsearch (a search engine)
  - Data is indexed, searching for specific information is available in near-realtime
  - Running on central ES server clusters
  - Insertion of information by hltd or merger services
    - CPU usage, event processing statistics, merging completion, logs (more details later in F3Mon description)



How Run Control starts up a sub-system:
Run Control + XDAQ system structure configurable



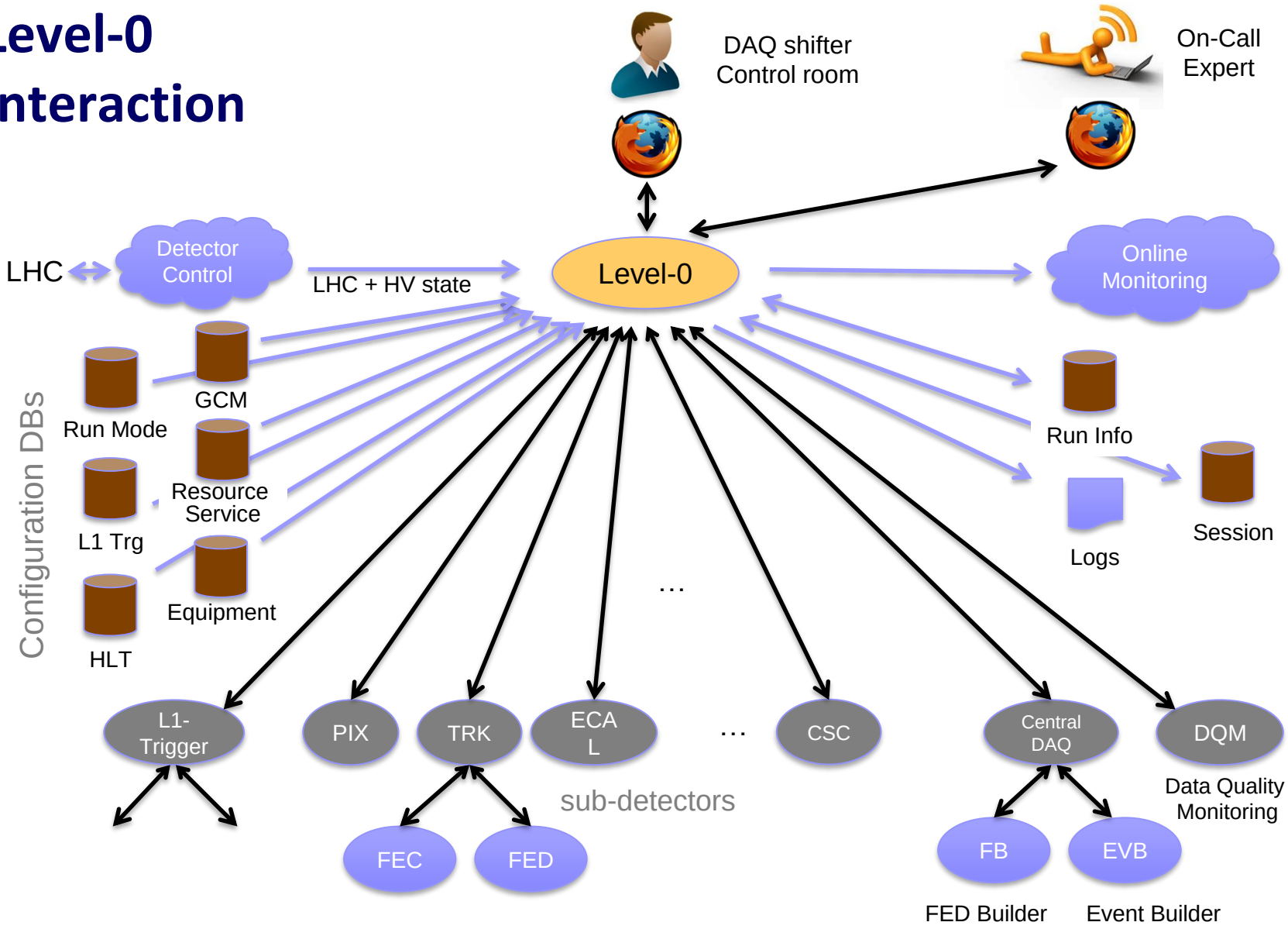
Control structure

- Function Managers to load (URL)
- Parameters
- Child nodes

Configuration of XDAQ Executives (XML)

- libraries to be loaded
- applications (e.g. builder unit, filter unit) & parameters
- network connections
- collaborating applications

Level-0 interaction



RCMS: Registration and Keys

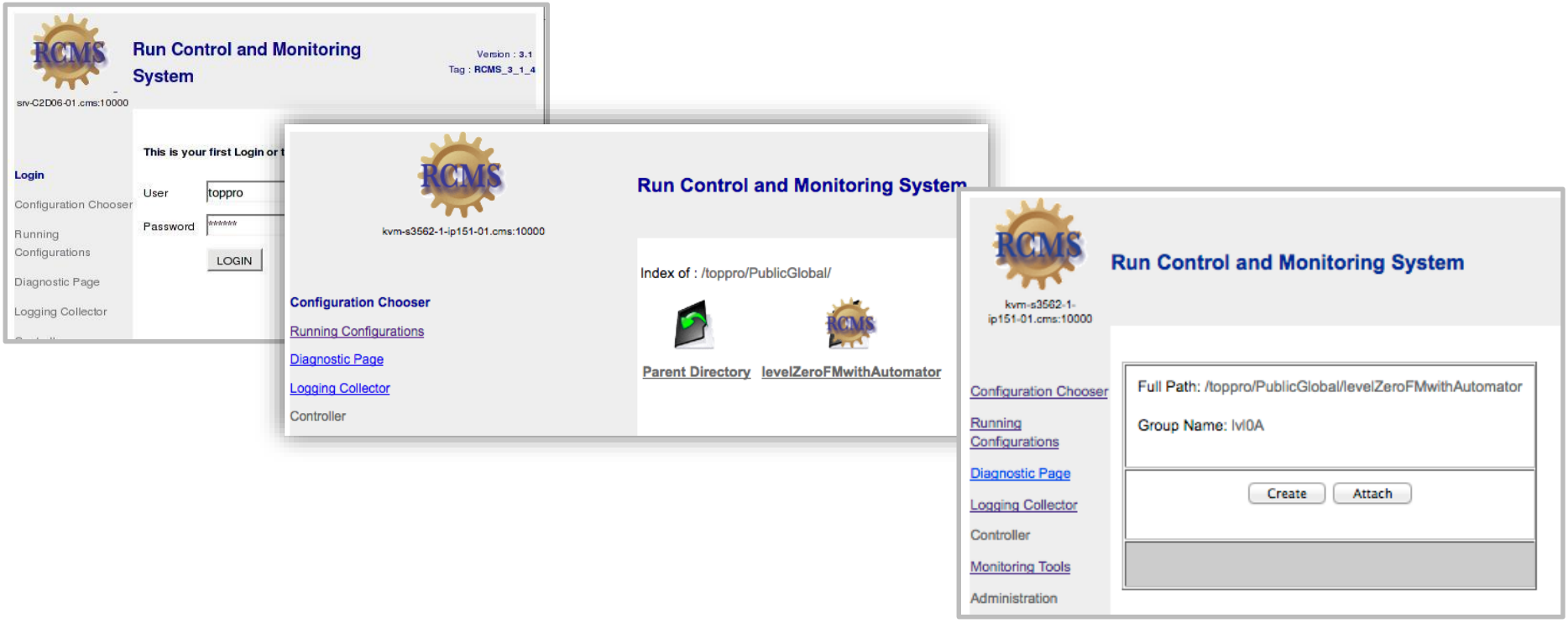
- Sub-system configurations need to be "registered" with the Global Configuration Map Database. This is done by the DAQ on-call expert or sub-system experts

- The Level-0 Function Manager queries the DB
  - to know what configuration to start for a subsystem
    - **Important: when a new configuration is registered you need to destroy (red recycle) the corresponding FM**
  - to know what RUN KEYS are available for a subsystem
    - You can parameterize a subsystem's configuration by selecting a run key (unless the key is are selected by a CMS Run Mode)

Part 3: Controlling data taking through Run Control

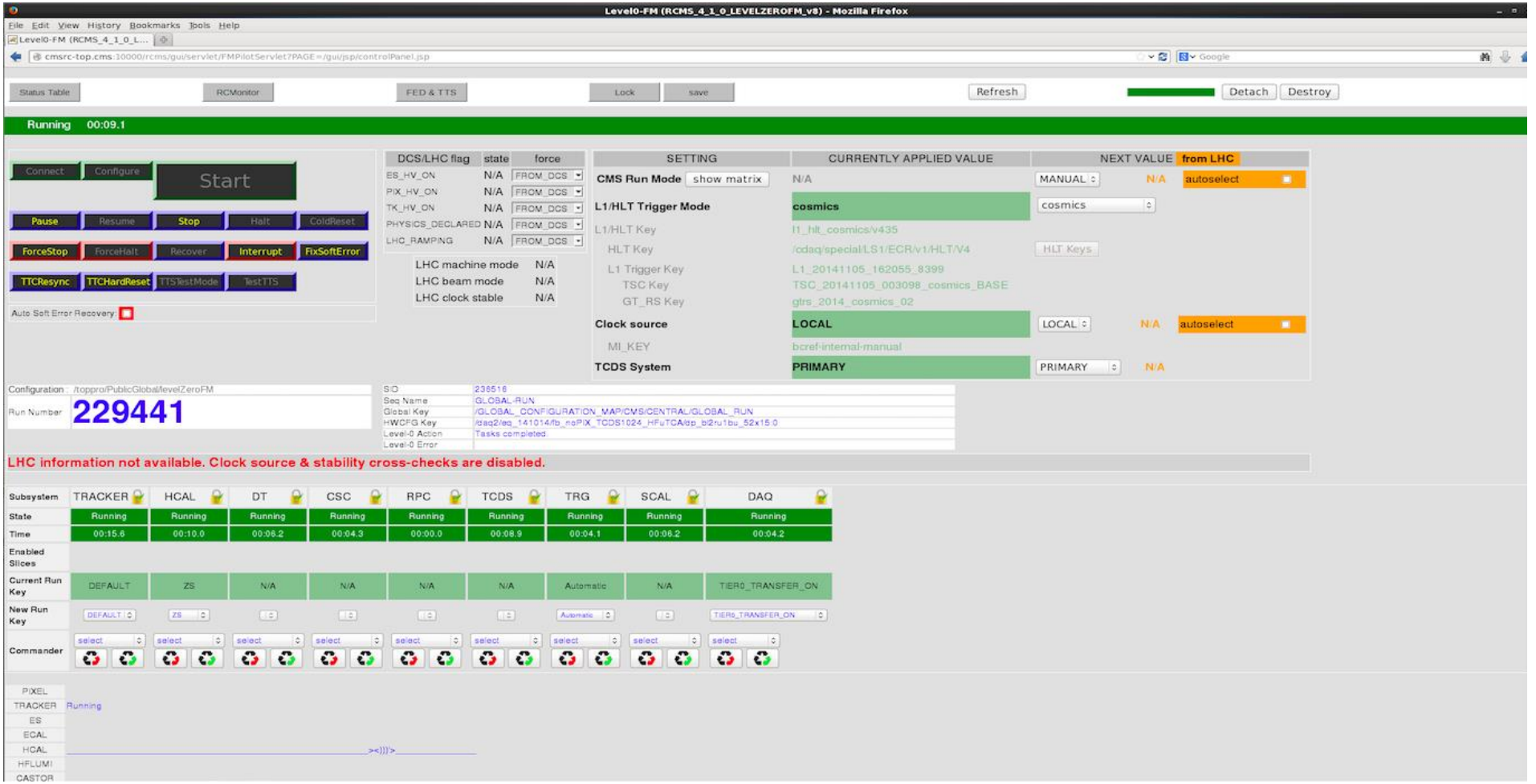
Create FM Level Zero

- RCMS workflow to create the FM Level Zero
  - Log into RCMS as toppro (http://cmsrc-top.cms:10000/rcms)
  - Configuration Chooser: *Path: PublicGlobal/LevelZeroFMwithAutomator*
  - *Press on "Create"*
    - *At this point the Level 0 Function Manager is created in the tomcat.*



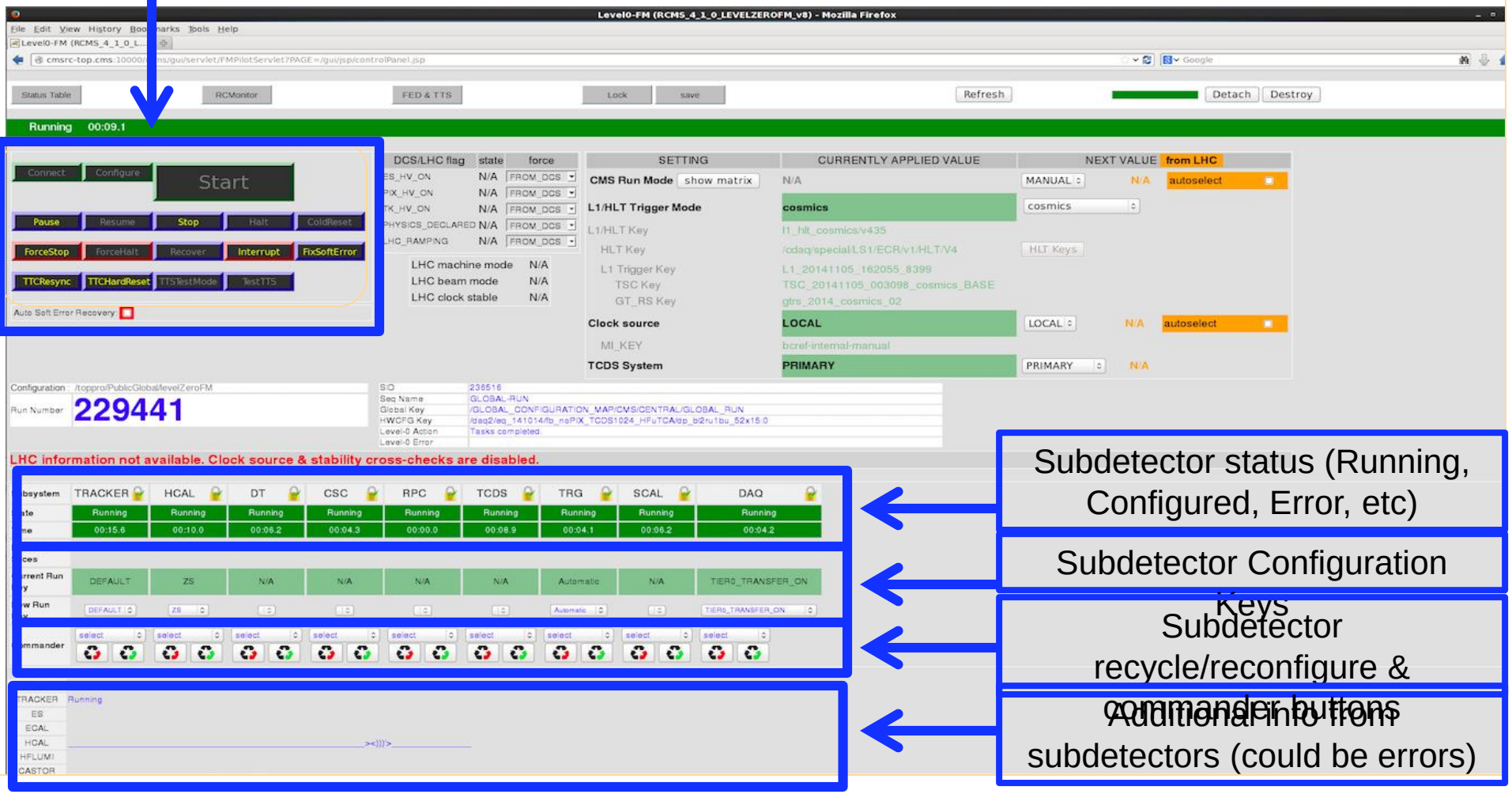
1. RCMS interface

- This is your main interface to DAQ management. This is the window where you will configure the CMS running mode, start / stop runs, remove / add subdetectors and FEDs, and look for errors.



1. RCMS interface

Main activity (start / stop / TTCResync etc)



Subdetector status (Running, Configured, Error, etc)

Subdetector Configuration Keys

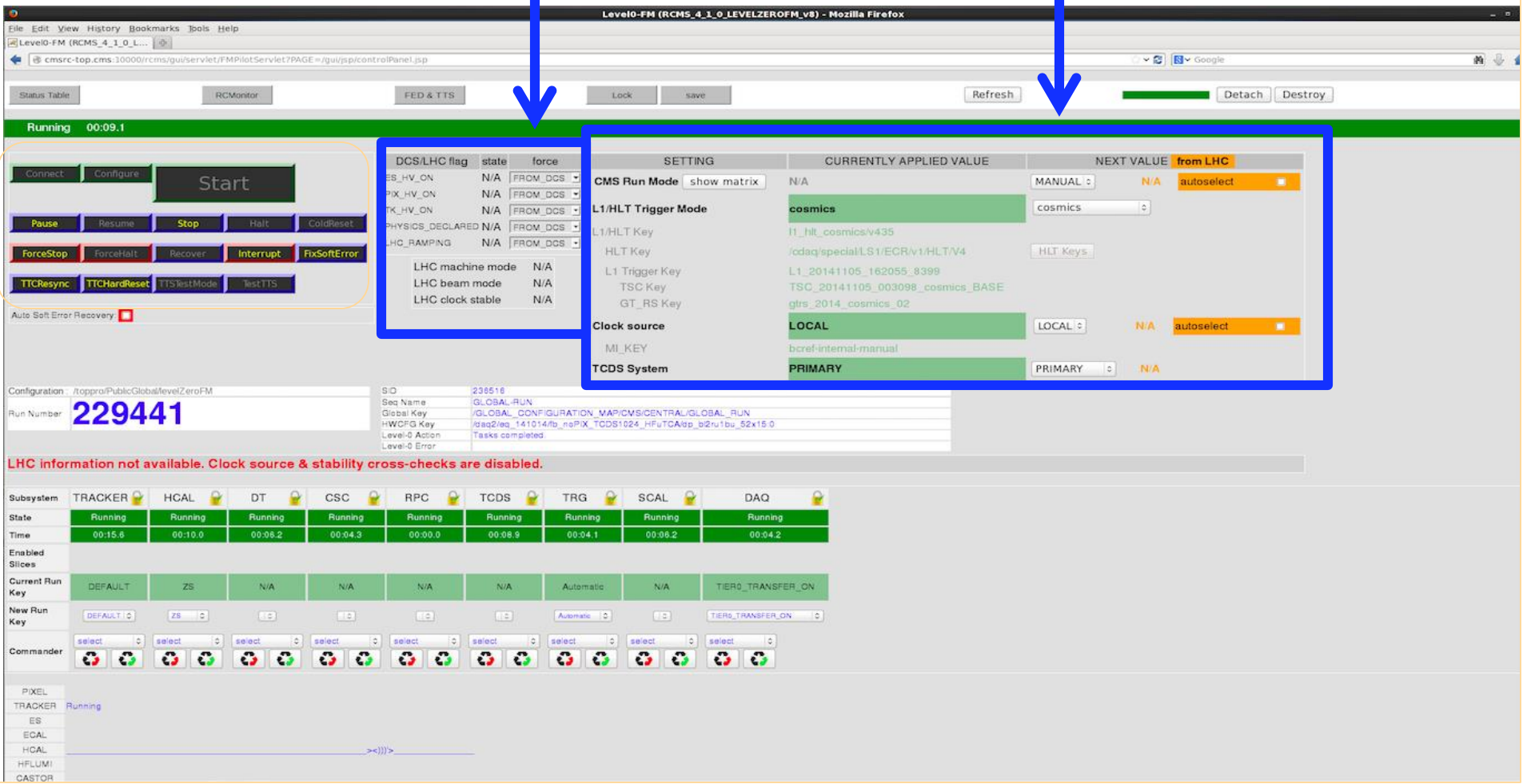
Subdetector recycle/reconfigure & commander buttons

Additional info from subdetectors (could be errors)

1. RCMS interface

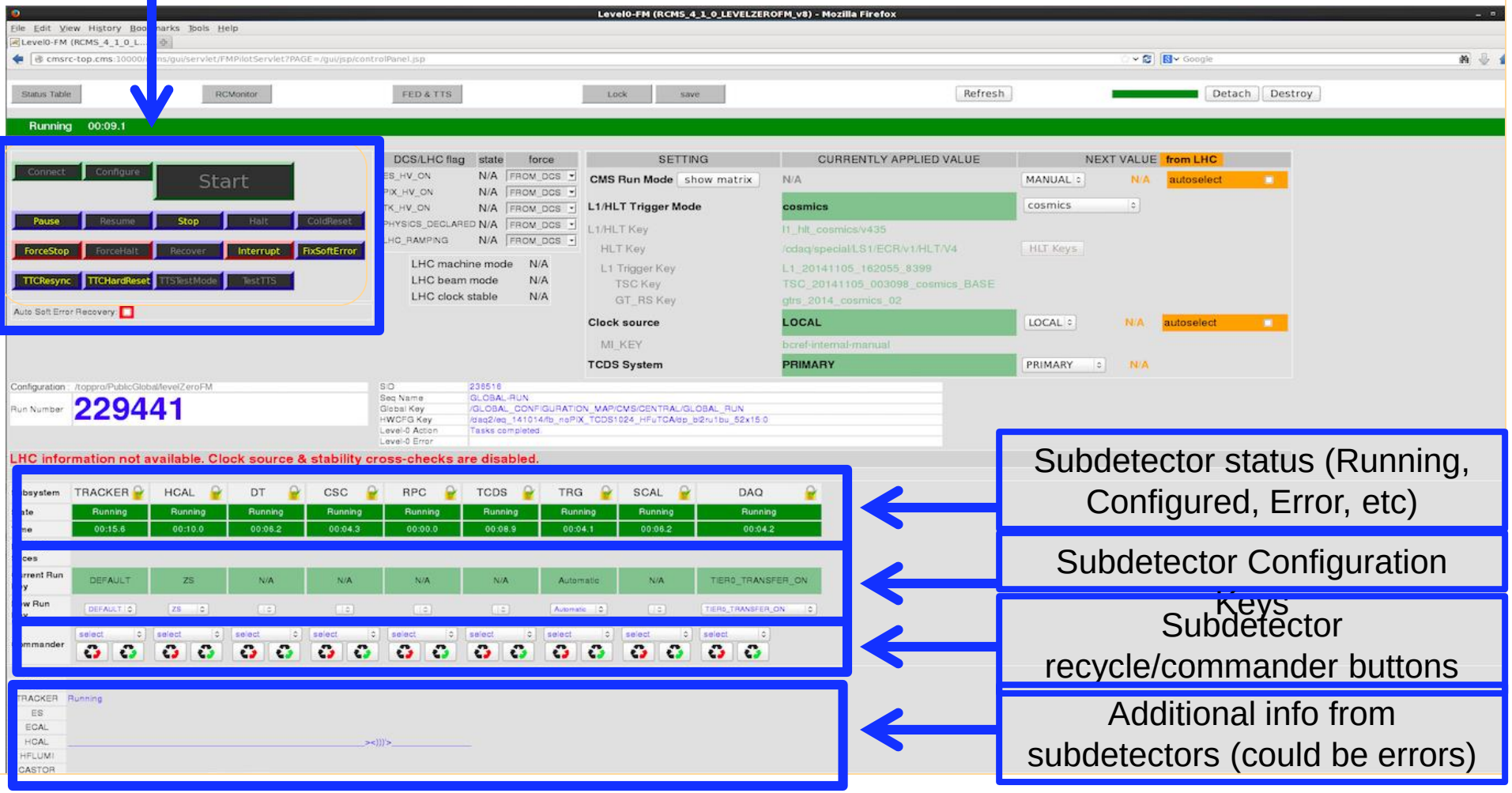
Slow Control (DCS / LHC) information (when available)

Information (and settings) for Trigger keys, Clock source, TCDS



Managing the run: start / stop runs

Main activity (start / stop / TTCResync etc)



Subdetector status (Running, Configured, Error, etc)

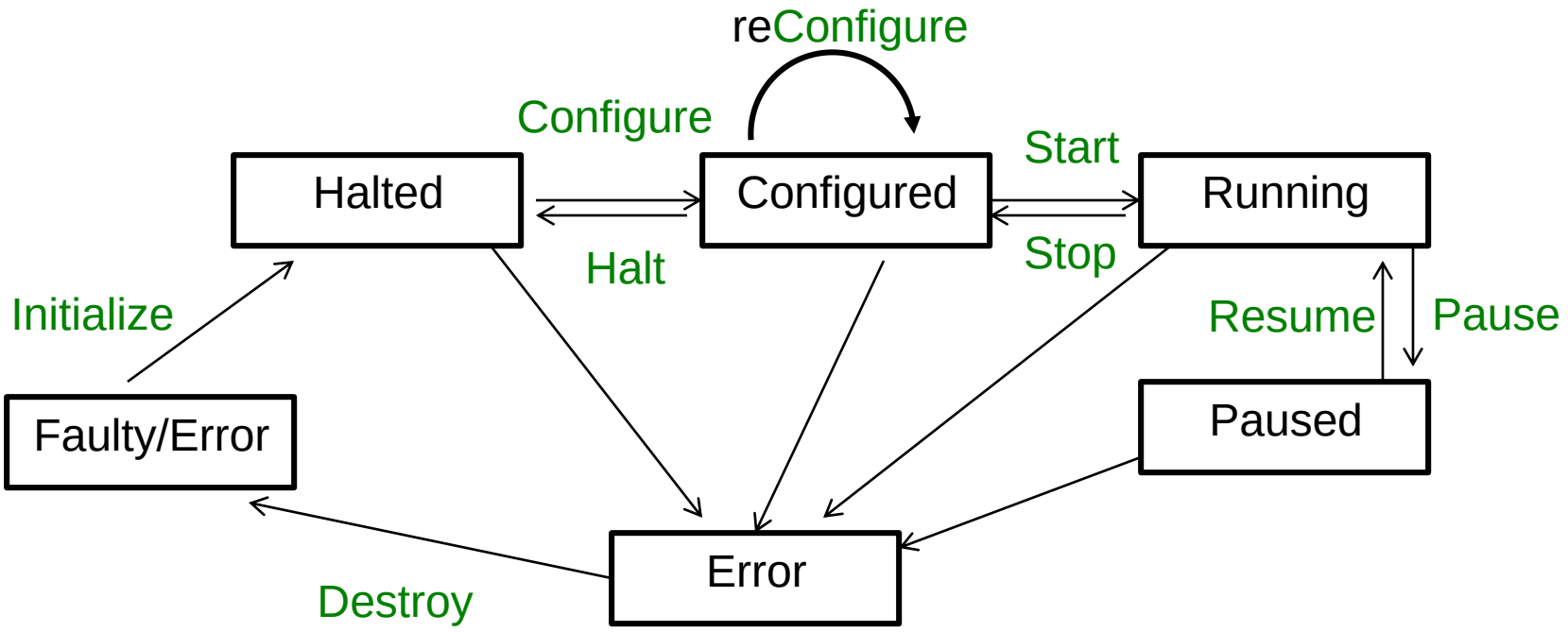
Subdetector Configuration Keys

Subdetector recycle/commander buttons

Additional info from subdetectors (could be errors)

Simplified state diagram for run control

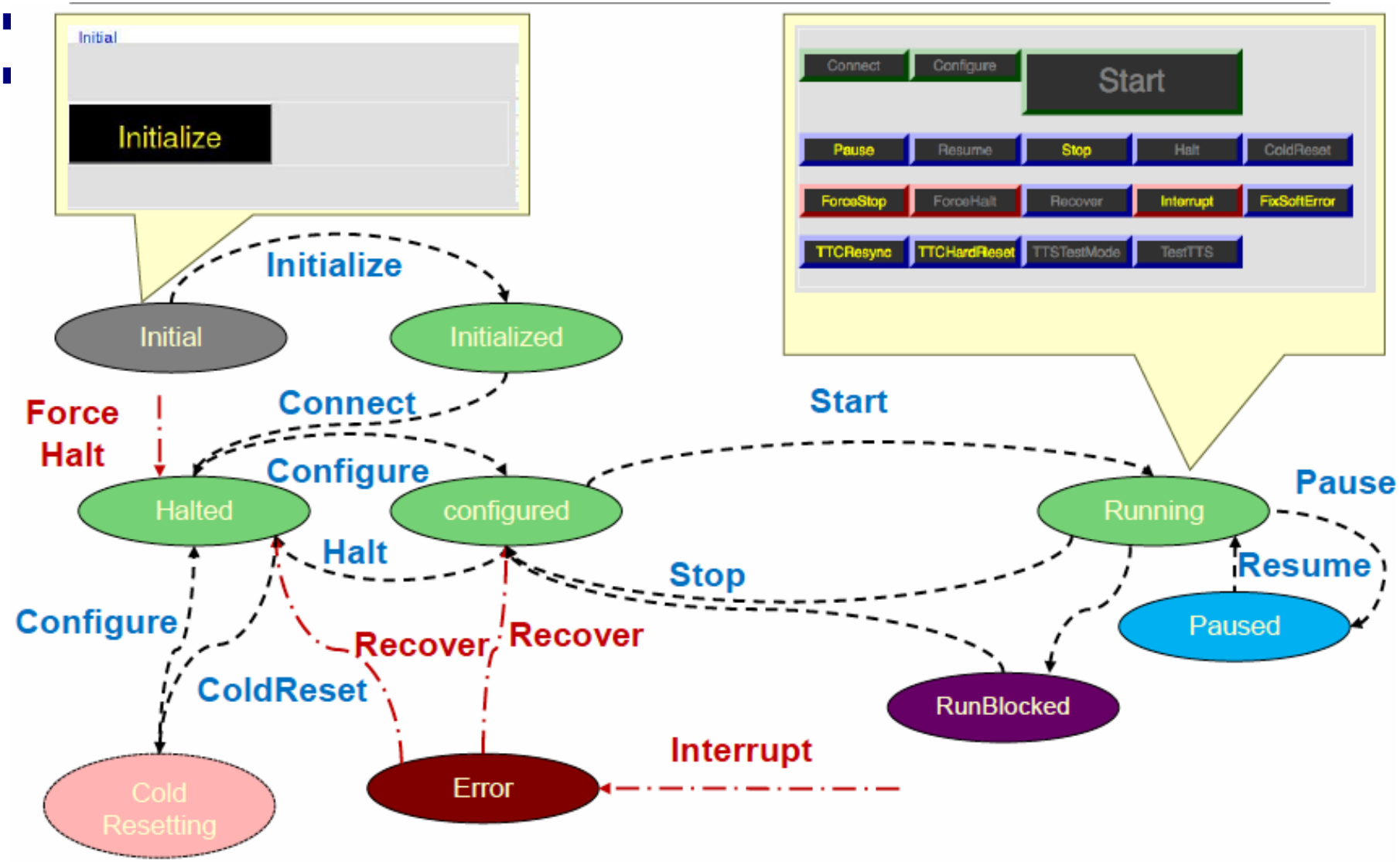
- In green are commands, in black are states reached after command is executed
- Note that “Error” state can happen during any step
- Only valid commands in each state are enabled in the Run Control screen



Re-configure = (Halt) + Configure

Re-cycle = Destroy + Initialize

Full state diagram for run control

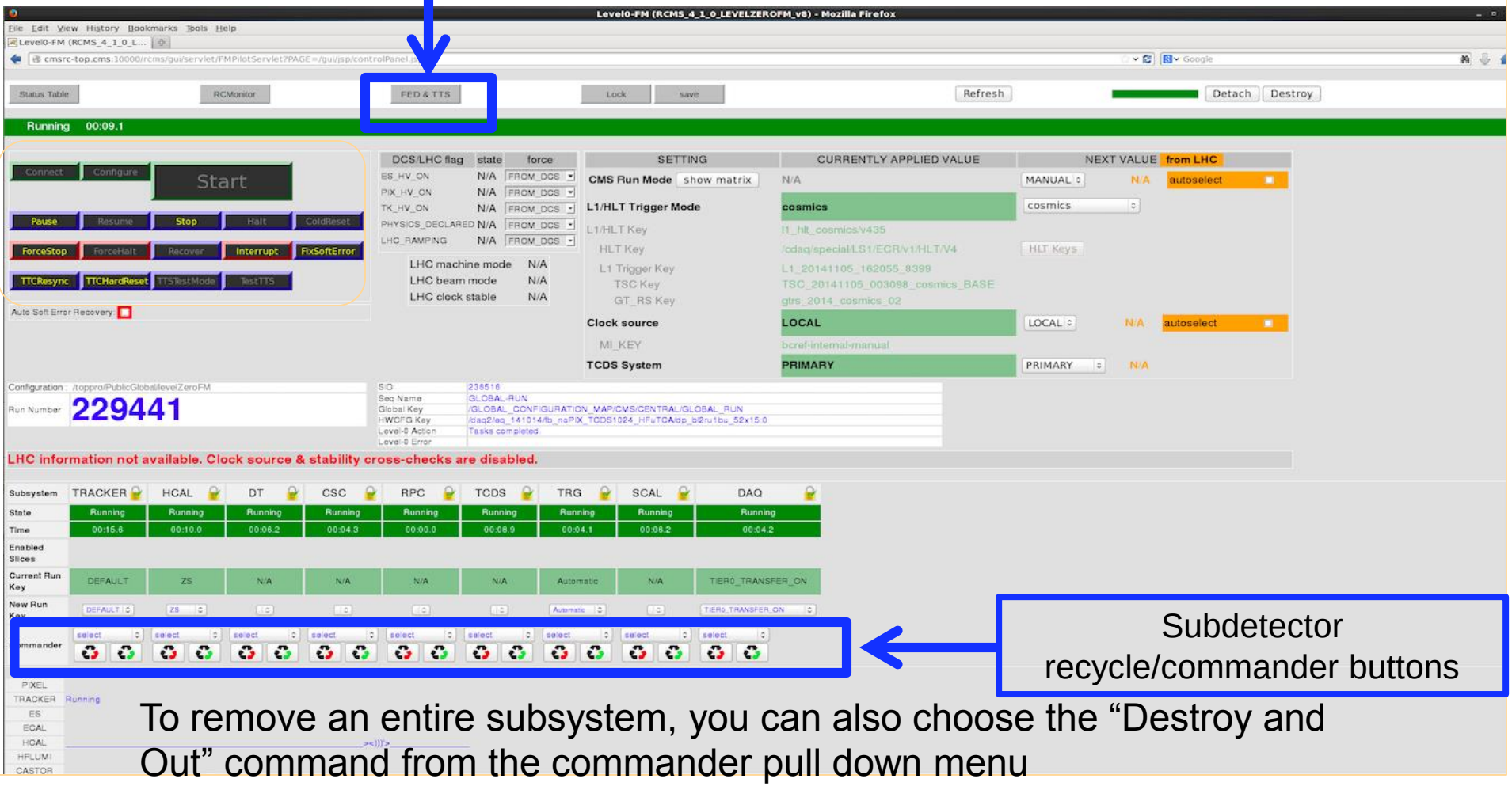


Starting / Stopping runs

- In general, you will use the buttons in the main activity area
  - Here the initialize / configure / start / stop buttons will address all subdetectors in the run
    - There is an order in which the commands must be sent to subdetectors, and the order is built into the L0 function manager
- It is possible also to use the “commander” and / or “recycle” buttons below the subdetector
  - Sends commands only to one subdetector
  - If this command requires subsequent action on another subdetector, then there will be a flashing message next to the subdetector that requires action
    - E.g. if you add a subdetector into the run when the TCDS or DAQ (or both) is already configured, a flashing message next to the TCDS / DAQ will tell you to reconfigure it
  - Recycle/Reconfigure buttons
    - There are two: “red” recycle and “green” reconfigure
    - Red recycle destroys the subdetector function manager and restarts it, ending in the “halted” state. You can do this from any state, you ***must*** do this from the “error” state
    - Green reconfigure reconfigures the subdetector only. You can do this either from the “halted” or “configured” state

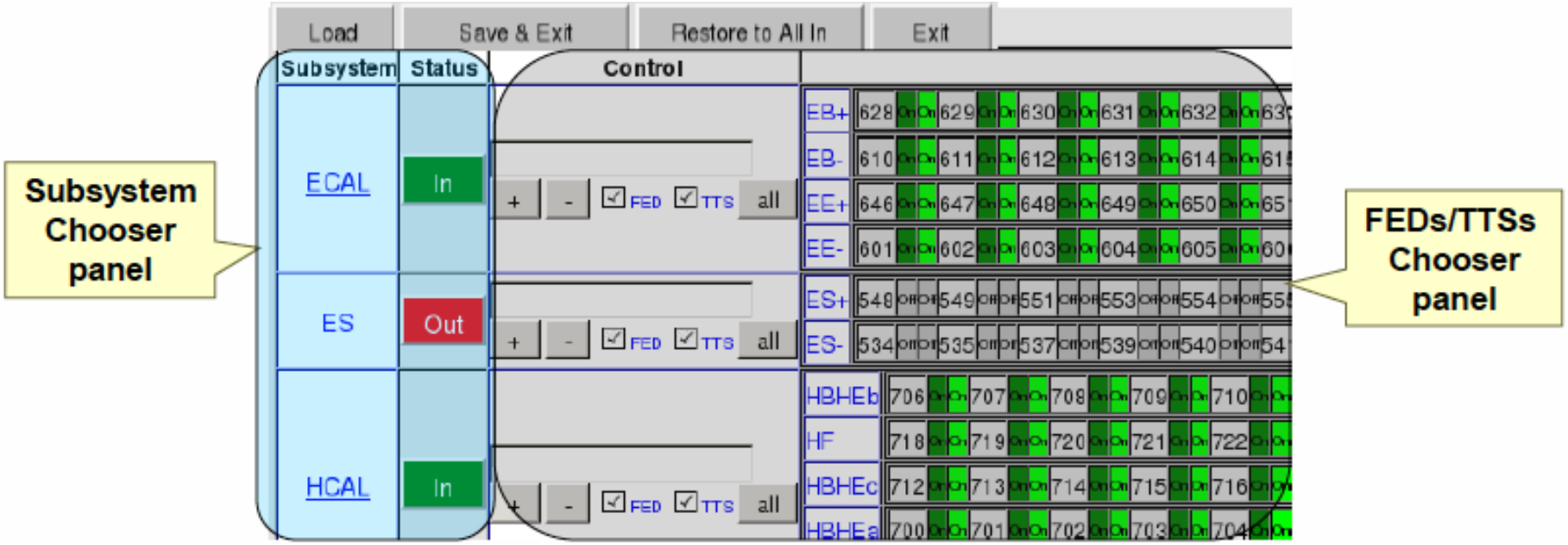
Adding / removing subsystems and/or FEDs

To add / remove FEDs or subsystems, click on the FED & TTS button
The window on the next slide will pop up



To remove an entire subsystem, you can also choose the “Destroy and Out” command from the commander pull down menu

Adding / removing subsystems and/or FEDs



Subsystem chooser panel:

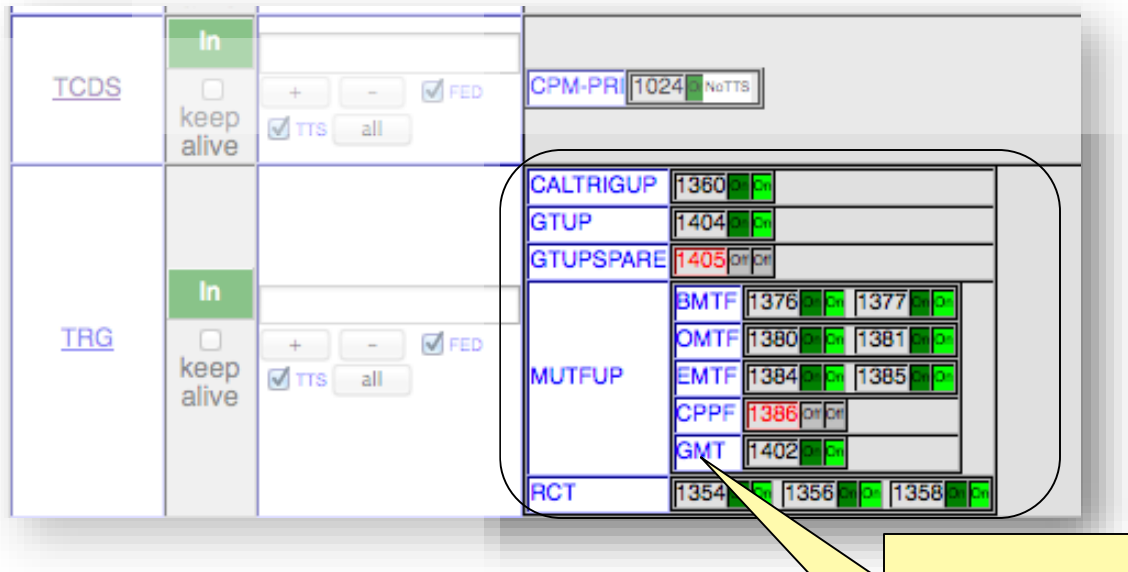
- you select the subsystems which participate in the run

FEDs/TTSs chooser panel:

- you choose the TTC partitions and with finer granularity the FEDs which participate in the run
- the Entry for selecting feds supports partition names, FEDIds and FEDID-intervals

NB: press the "Save & Exit" button if you want to save the new setting.

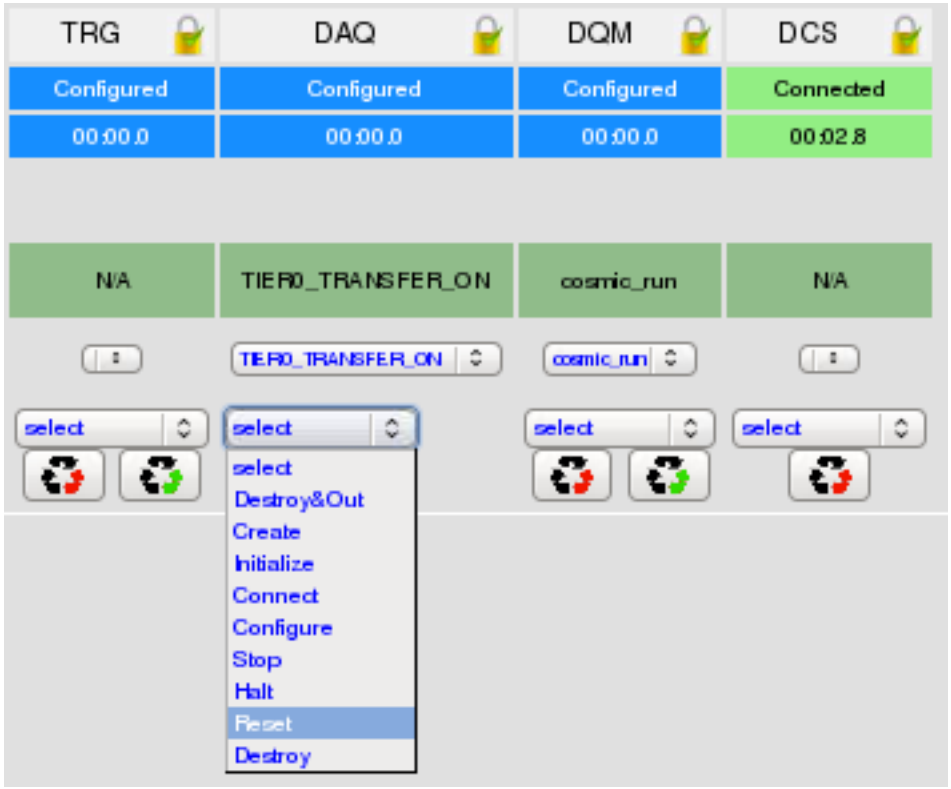
Adding / removing subsystems and/or FEDs



FED groups help to differentiate between subsystems within a partition

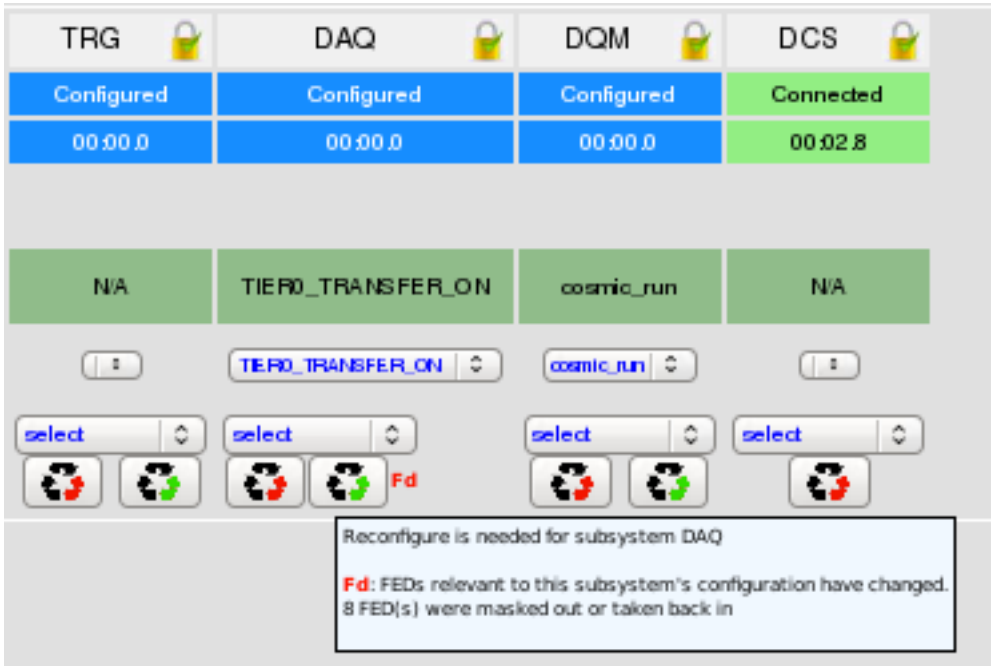
Sub-Systems Control Panel

- The Sub-Systems panel contains:
  - *All the subsystems included in the Global Run*
  - *State of each subsystem*
  - *Applied Run Key for each subsystem*
  - *Run Key selector*
  - *Commander for each subsystem:*
    - *The pull down menu allows to send command directly to the subsystem*
    - *Red re-cycle button allows to destroy the subsystem software and bring it to halted state.*
    - *Green re-cycle button allows to (re-)configure the subsystem software.*



FM L0 built-in cross-checks

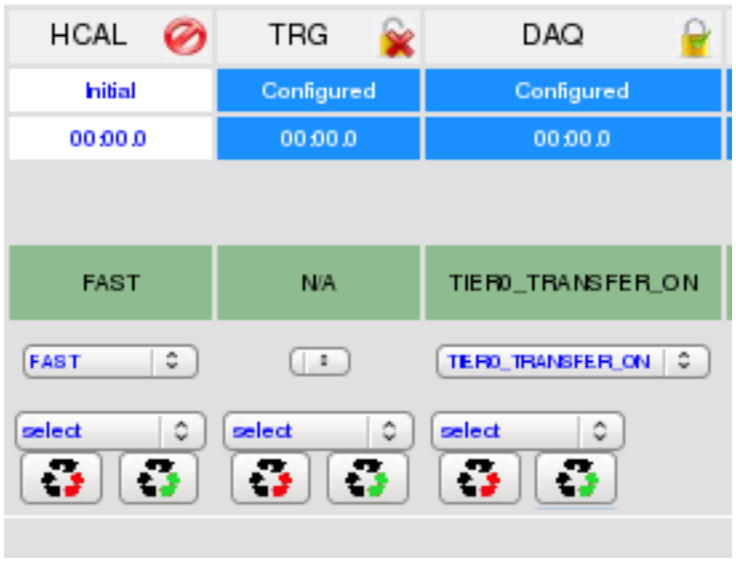
- Indicate sub-systems to re-configure if :
  - A parameter is changed in the GUI
  - A sub-system / FED is added/removed
  - External parameters change
- Enforce correct order of re-configuration
- Enforce procedure to follow if LHC clock stability changes



Access Control

Subsystems are created by the Level-0 in a locked state and the subsystem RCMS GUIs may be attached for read access but may not command the subsystem or set parameters.

If a subsystem-expert needs to access the sub-system through their RCMS GUI, you may unlock the subsystem by clicking on the lock icon. You should lock the subsystem again after the intervention is finished.

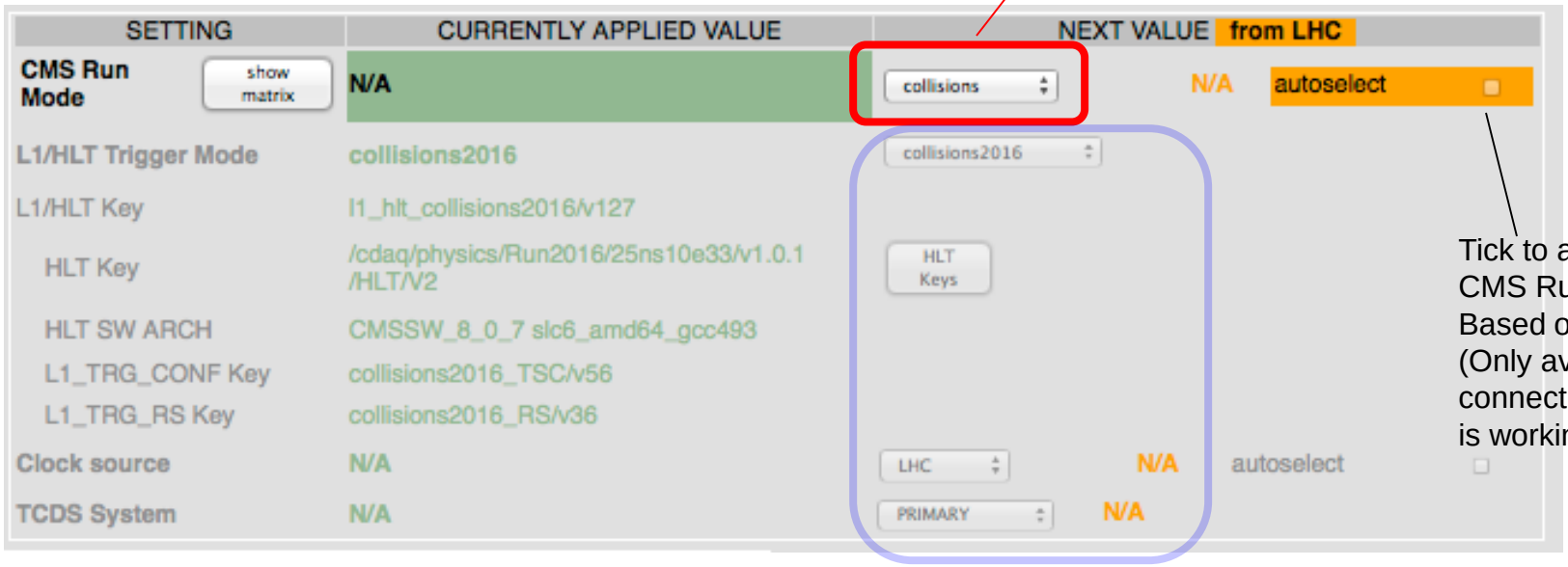


If a subsystem was created by a GUI or by a different Level-0, the central Level-0 may not be able to command this subsystem. In order to control the subsystem from the central Level-0, the subsystem must be destroyed from where it was created.

Destroy Backdoor (this applies for any Function Manager, whether it belongs to a subsystem or it is the Level-0 itself): If things went wrong and you cannot destroy a Function Manager in any regular way.

Run and Trigger mode selection

Select CMS Run Mode other than MANUAL



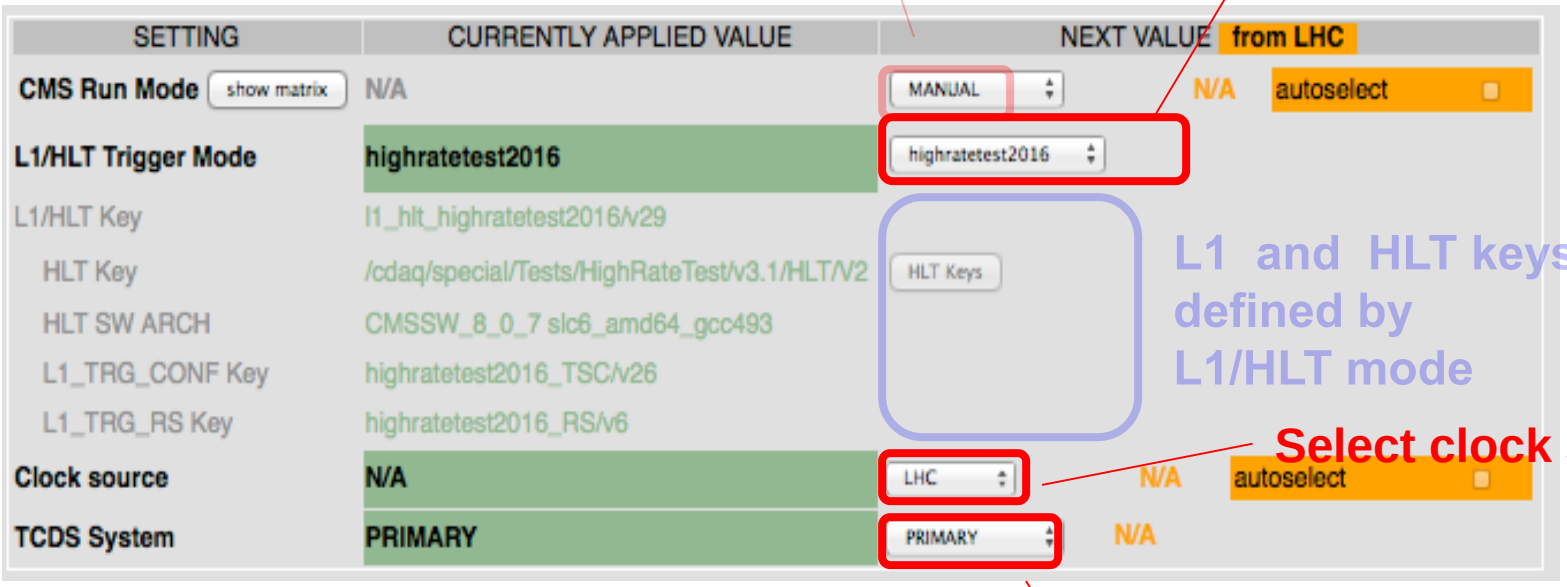
Tick to auto-select CMS Run Mode Based on LHC mode (Only available if connection to DCS is working)

All keys defined by CMS Run Mode including (some) sub-system RUN KEYs

You may sometimes need: Manual RUN MODE

Select CMS Run Mode is MANUAL

Select L1/HLT Mode



L1 and HLT keys defined by L1/HLT mode

Select clock source

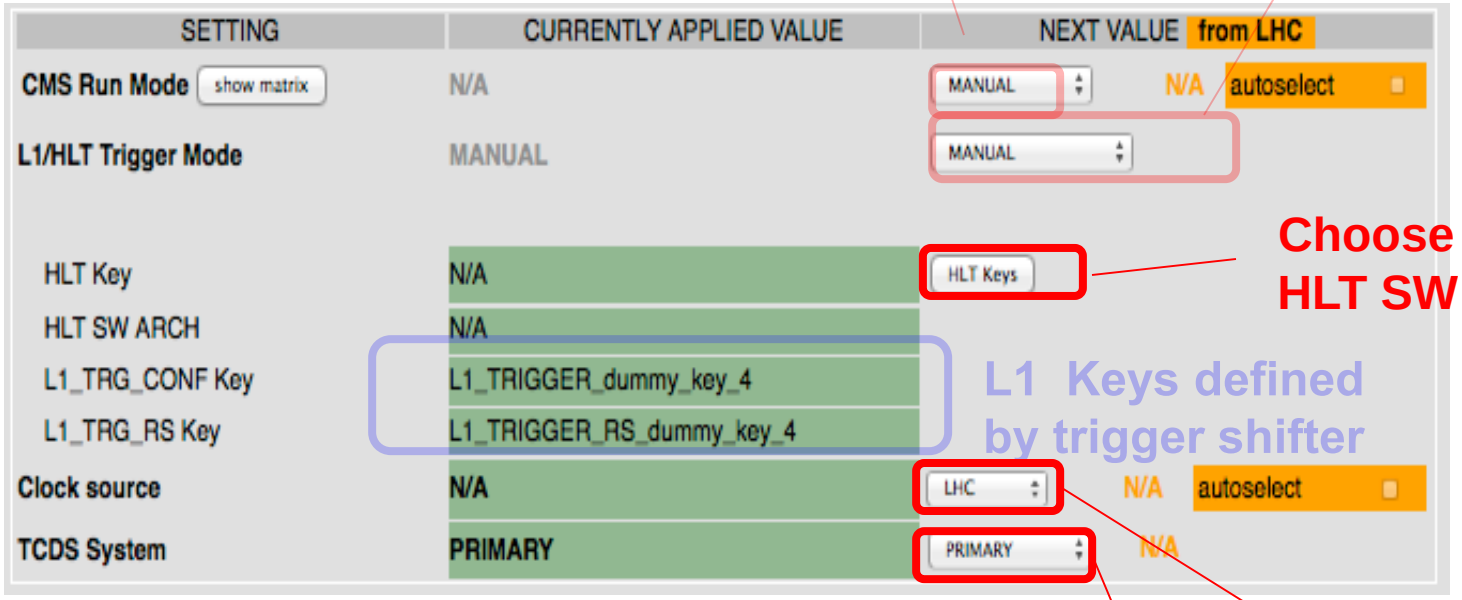
(Select primary/secondary TCDS system – not yet available)

Attention: You also need to select sub-system run keys manually

You may rarely need: Manual L1/HLT MODE

Select CMS Run Mode is MANUAL

L1/HLT Mode is MANUAL



Choose HLT key and HLT SW Architecture

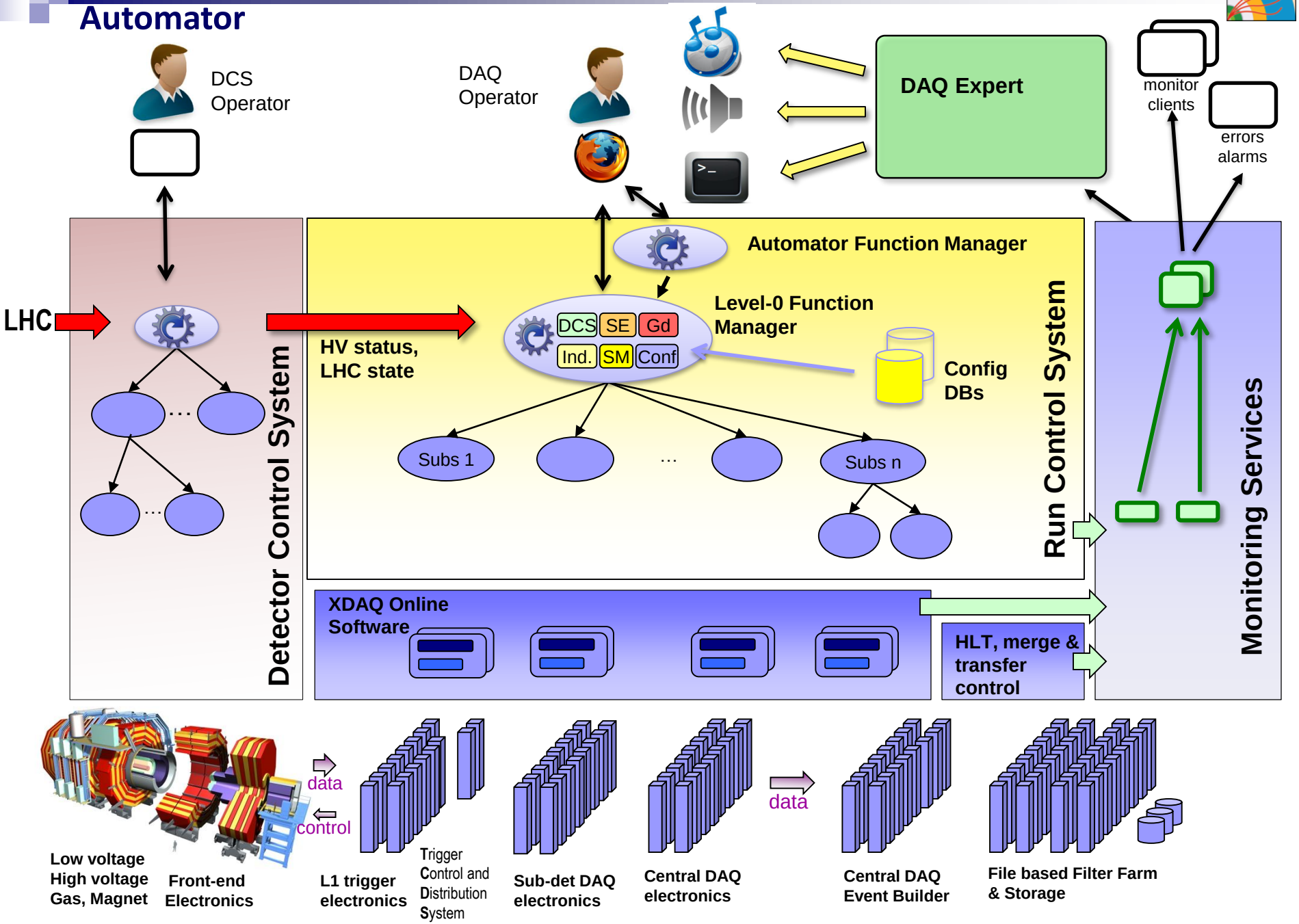
L1 Keys defined by trigger shifter

Select clock source

(Select primary/secondary TCDS system – not yet available)

Attention: You also need to select sub-system run keys manually

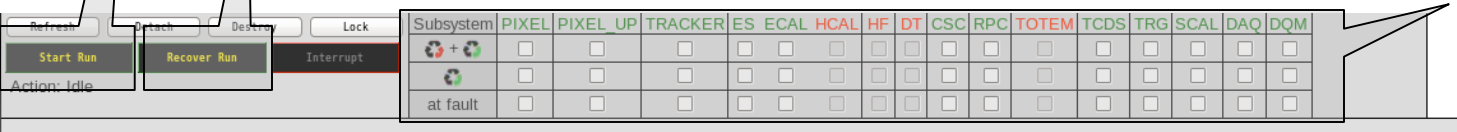
Automator


DCS Operator
DAQ Operator
DAQ Expert
monitor clients
errors alarms
LHC
Automator Function Manager
Level-0 Function Manager
DCS
SE
Gd
Ind.
SM
Conf
Config DBs
HV status, LHC state
Subs 1
Subs n
Detector Control System
Run Control System
Monitoring Services
XDAQ Online Software
HLT, merge & transfer control
data
control
data
Low voltage
High voltage
Gas, Magnet
Front-end Electronics
L1 trigger electronics
Trigger Control and Distribution System
Sub-det DAQ electronics
Central DAQ electronics
Central DAQ Event Builder
File based Filter Farm & Storage

Level-0 Automator

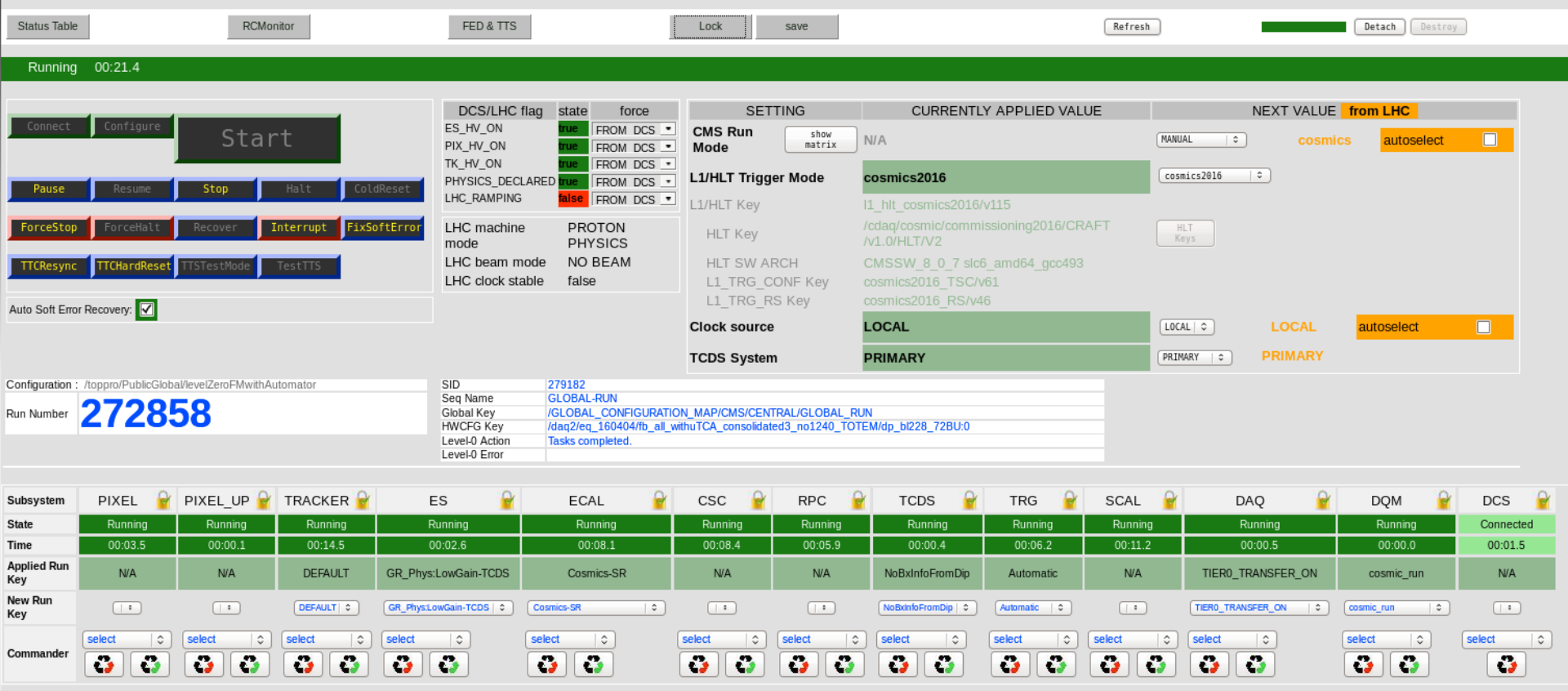
Start a run from any state
Stop a run then re-start

- Taking into account all cross-checks (except sanity checks)
- Taking into account scheduled actions
- Attempting to recover from failures (2 retries, currently)



Schedule

"at-fault" is for future automatic down-time splitting



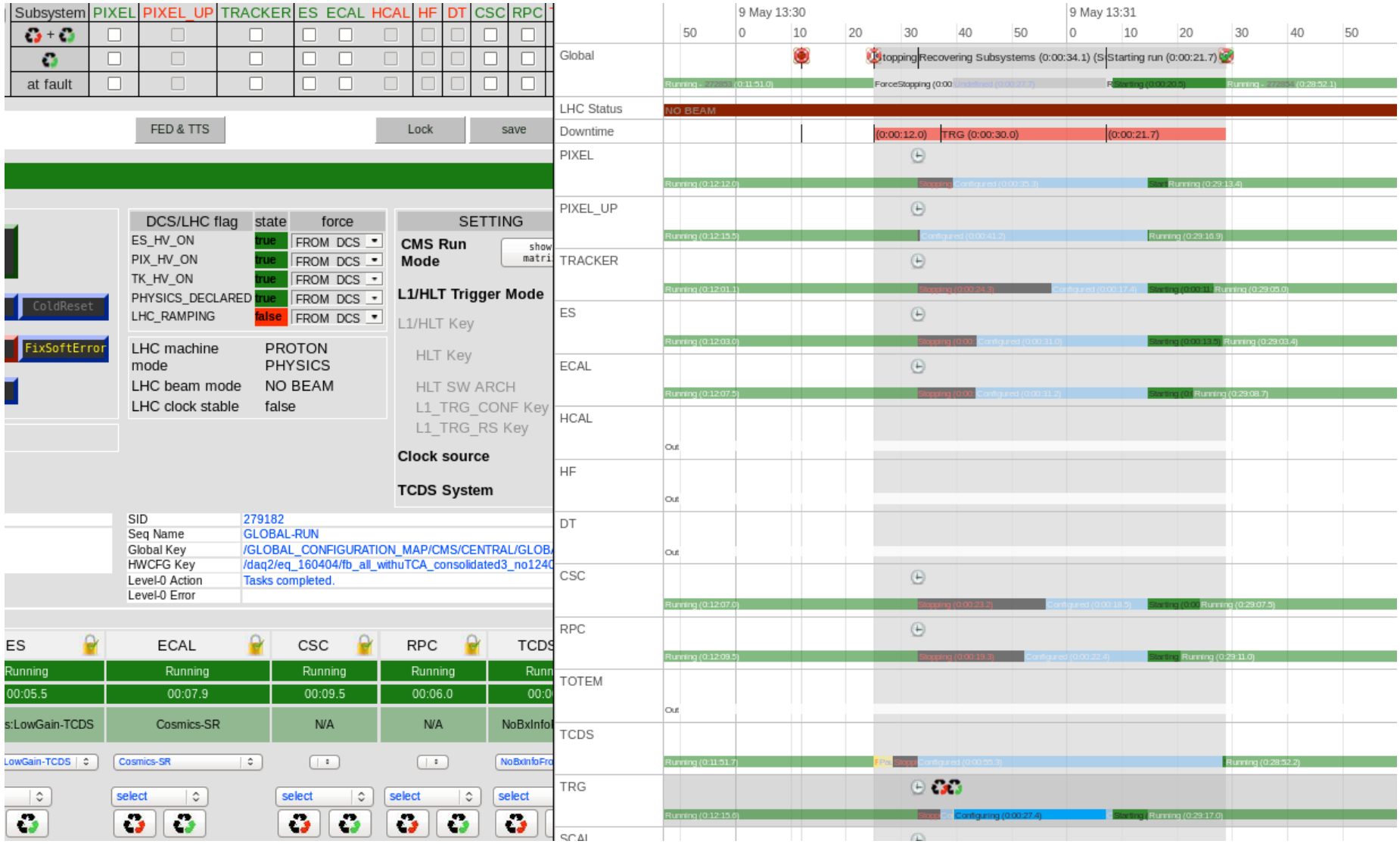
Level-0 Automator

- When to use what button
  - If we are not “Running”
    - “Start Run” will start a new run (“Recover Run” will do the same thing)
  - If we are “Running”
    - “Recover Run” will stop, then start a new run (“Start Run” has no effect)

- When to use the automator
  - To start a run
    - First set all settings (Subsystems & FEDs in/out, run mode etc.)
    - Then let the automator re-configure / re-cycle subsystems as necessary and start a new run
  - To recover from a problem (if you know the appropriate recovery action)
    - Select the recovery action in the Schedule
    - Let the automator stop the run, re-configure / re-cycle subsystems as necessary and start a new run
  - To apply changed settings such as a new run mode or trigger key
    - Just click “Recover Run” while a run is going
    - The automator stops the run, recycles/ reconfigures subsystems as requested by indicators and starts again

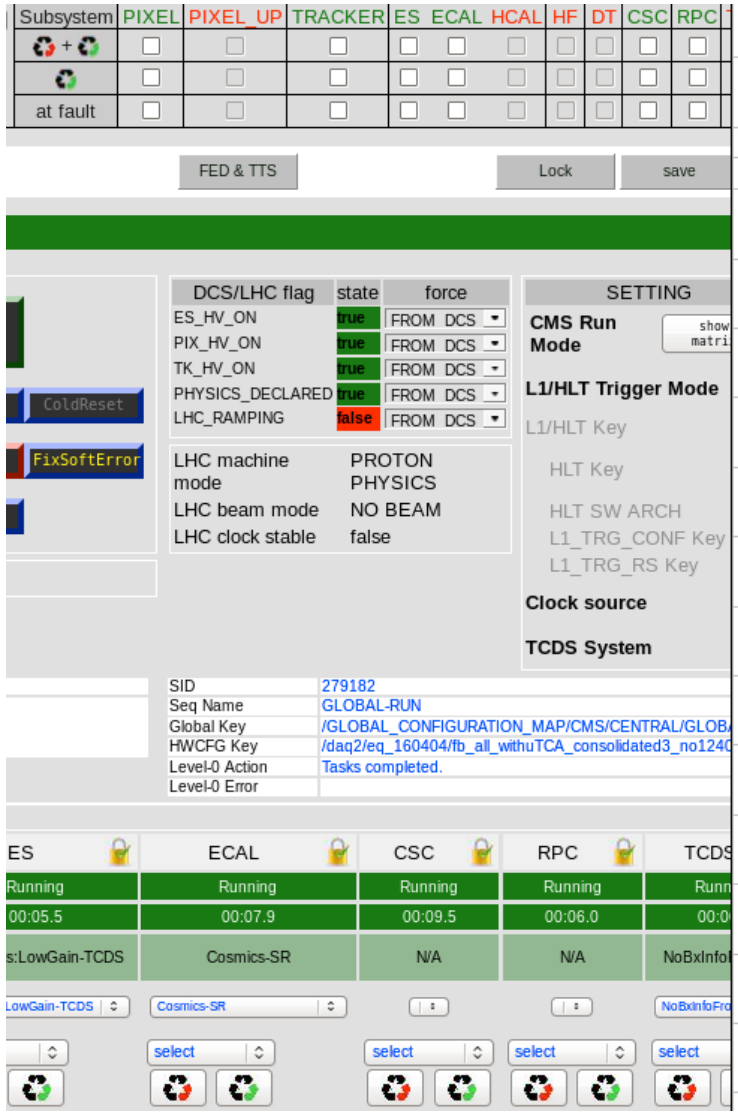
Level-0 timeline

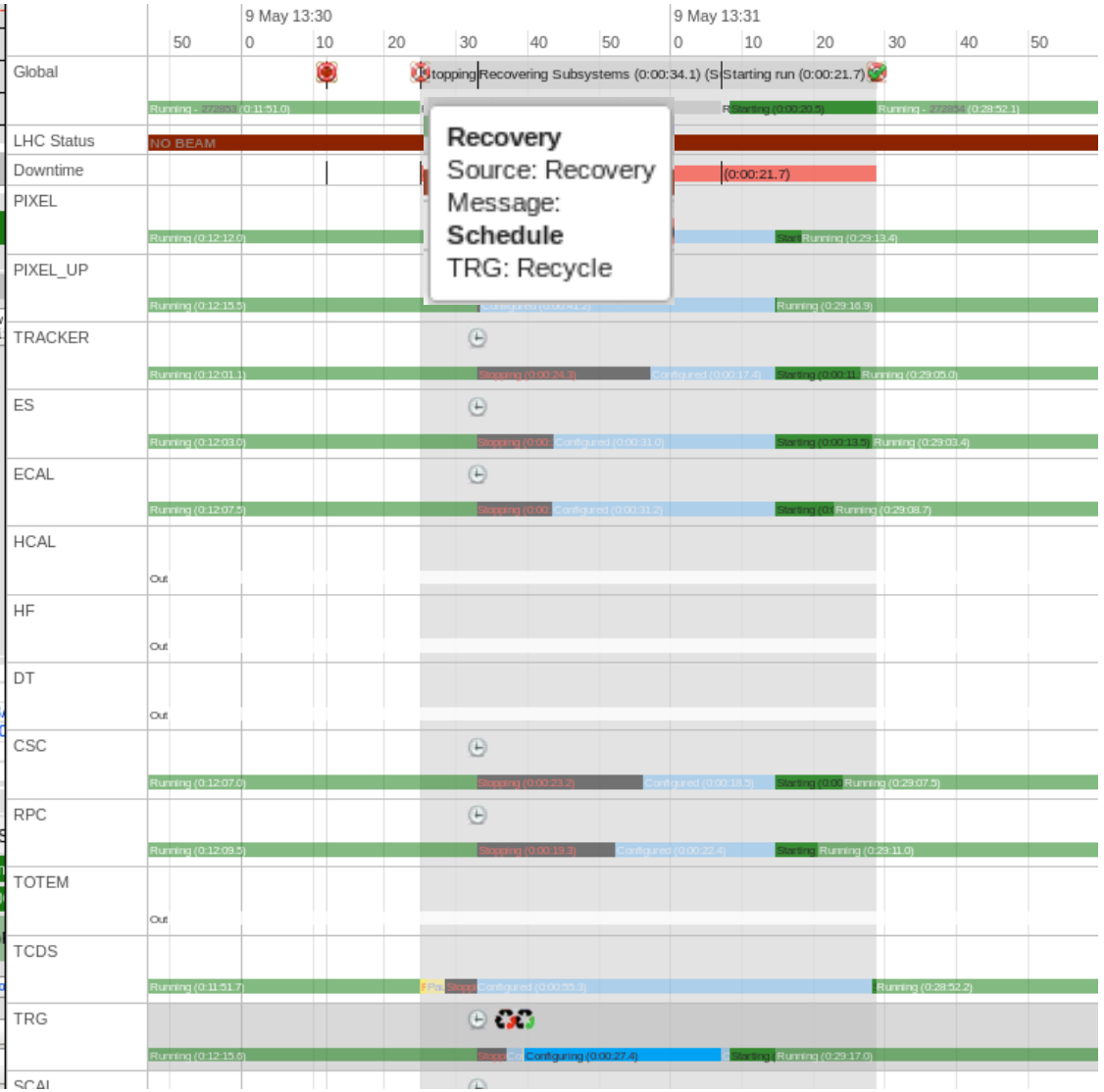
Shows history of subsystem states and all manual and automatic actions taken



Level-0 timeline

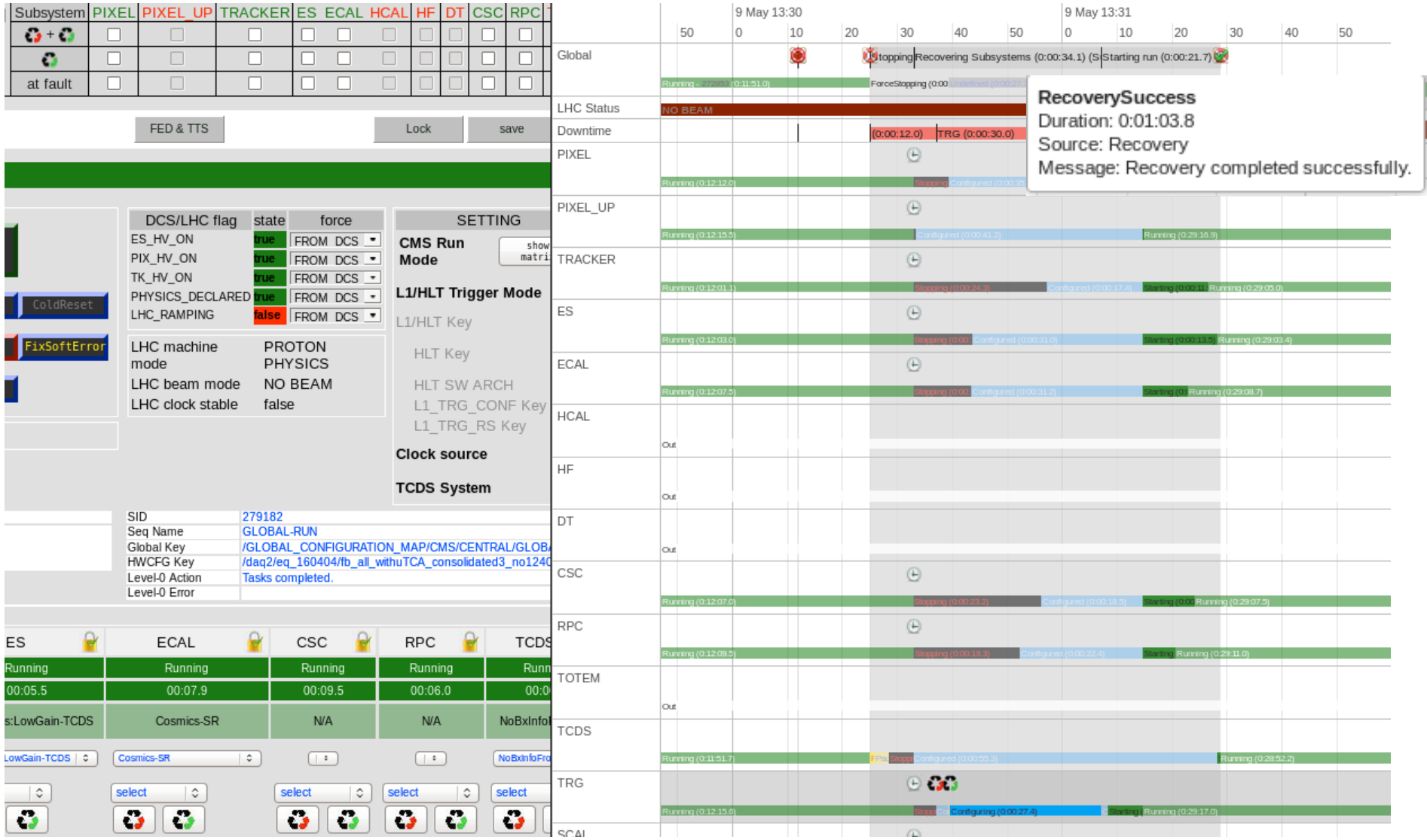
Tool tips give information about the reasons for action



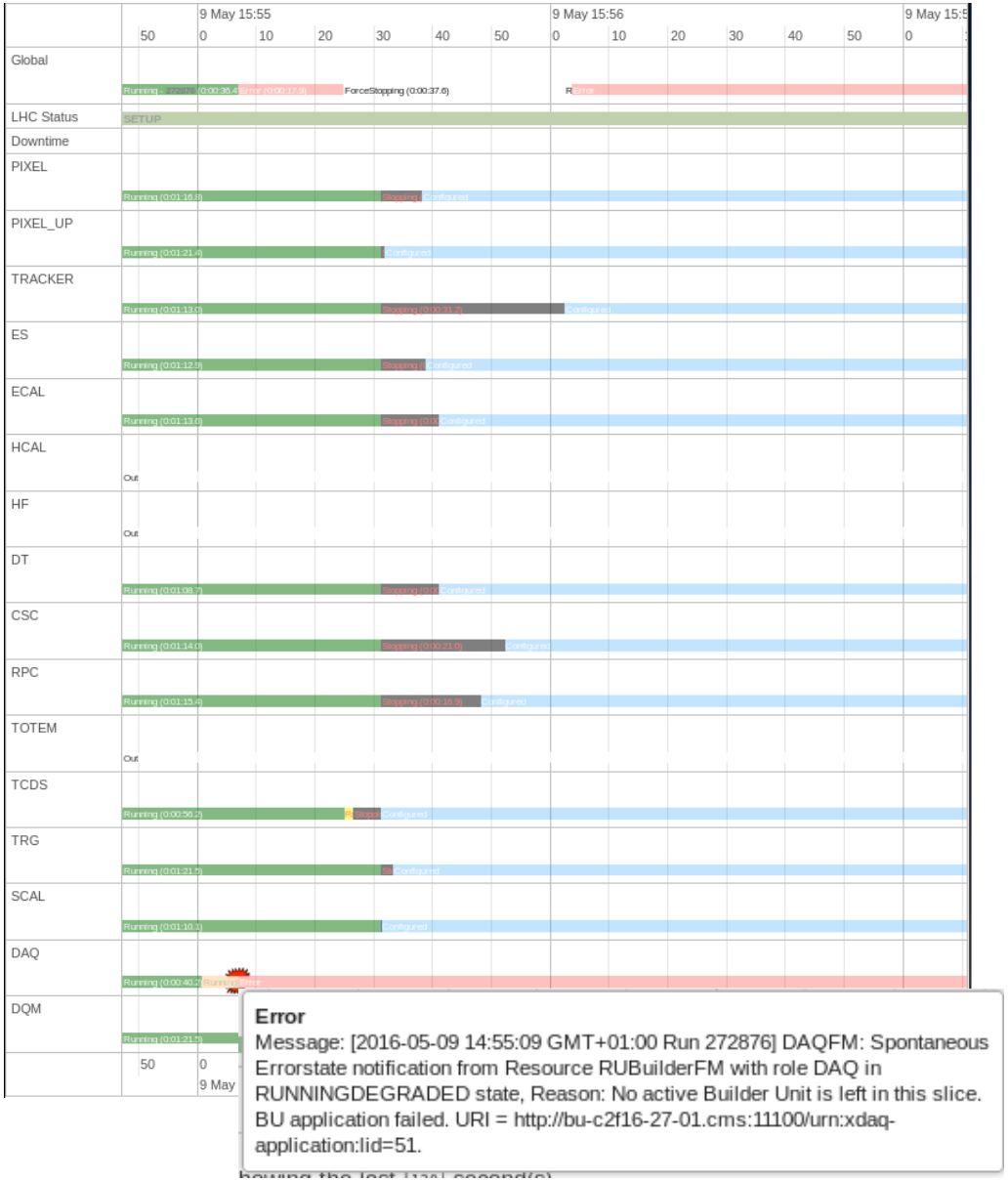


Level-0 timeline

Tool tips give information about the reasons for action … and their outcome

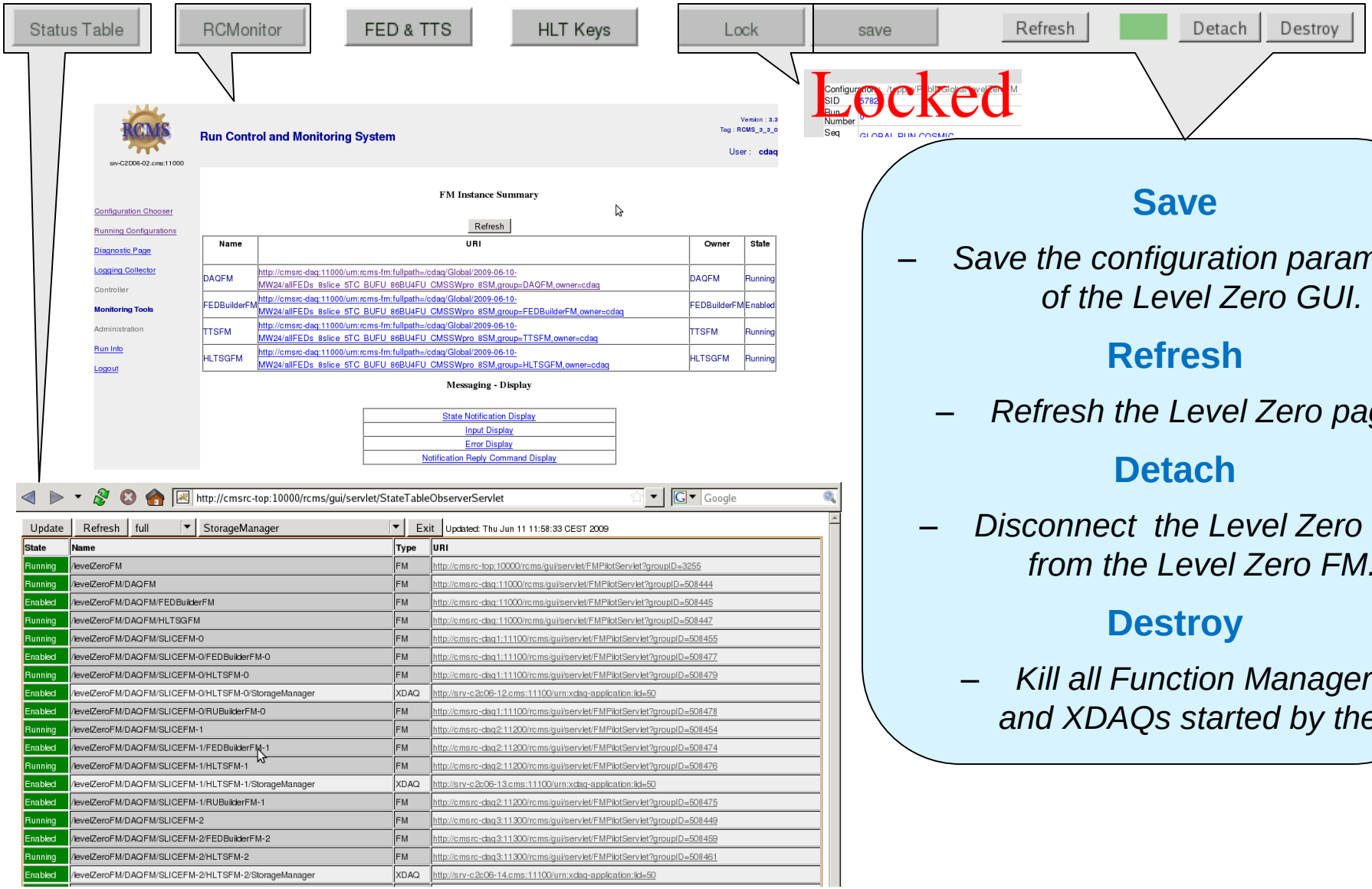


Tooltips for errors



Going back to analyze problems

FM L0 Links Panel



Name	URI	Owner	State
DAQFM	http://cmsrc-daq:11000/urn:rcms-fm:fullpath=/cdaq/Global/2009-06-10-MW24/allFEDs_8slice_5TC_BUFU_86BU4FU_CMSSWpro_8SM,group=DAQFM,owner=cdaq	DAQFM	Running
FEDBuilderFM	http://cmsrc-daq:11000/urn:rcms-fm:fullpath=/cdaq/Global/2009-06-10-MW24/allFEDs_8slice_5TC_BUFU_86BU4FU_CMSSWpro_8SM,group=FEDBuilderFM,owner=cdaq	FEDBuilderFM	Enabled
TTSFM	http://cmsrc-daq:11000/urn:rcms-fm:fullpath=/cdaq/Global/2009-06-10-MW24/allFEDs_8slice_5TC_BUFU_86BU4FU_CMSSWpro_8SM,group=TTSFM,owner=cdaq	TTSFM	Running
HLTSGFM	http://cmsrc-daq:11000/urn:rcms-fm:fullpath=/cdaq/Global/2009-06-10-MW24/allFEDs_8slice_5TC_BUFU_86BU4FU_CMSSWpro_8SM,group=HLTSGFM,owner=cdaq	HLTSGFM	Running

State	Name	Type	URI
Running	/levelZeroFM	FM	http://cmsrc-top:10000/rcms/gui/servlet/FMPilotServlet?groupID=3255
Running	/levelZeroFM/DAQFM	FM	http://cmsrc-daq:11000/rcms/gui/servlet/FMPilotServlet?groupID=508444
Enabled	/levelZeroFM/DAQFM/FEDBuilderFM	FM	http://cmsrc-daq:11000/rcms/gui/servlet/FMPilotServlet?groupID=508445
Running	/levelZeroFM/DAQFM/HLTSGFM	FM	http://cmsrc-daq:11000/rcms/gui/servlet/FMPilotServlet?groupID=508447
Running	/levelZeroFM/DAQFM/SLICEFM-0	FM	http://cmsrc-daq1:11100/rcms/gui/servlet/FMPilotServlet?groupID=508455
Enabled	/levelZeroFM/DAQFM/SLICEFM-0/FEDBuilderFM-0	FM	http://cmsrc-daq1:11100/rcms/gui/servlet/FMPilotServlet?groupID=508477
Running	/levelZeroFM/DAQFM/SLICEFM-0/HLTSFM-0	FM	http://cmsrc-daq1:11100/rcms/gui/servlet/FMPilotServlet?groupID=508479
Enabled	/levelZeroFM/DAQFM/SLICEFM-0/HLTSFM-0/StorageManager	XDAQ	http://srv-c2c06-12.cms:11100/urn:xdaq-application:lid=50
Enabled	/levelZeroFM/DAQFM/SLICEFM-0/RUBuilderFM-0	FM	http://cmsrc-daq1:11100/rcms/gui/servlet/FMPilotServlet?groupID=508478
Running	/levelZeroFM/DAQFM/SLICEFM-1	FM	http://cmsrc-daq2:11200/rcms/gui/servlet/FMPilotServlet?groupID=508454
Enabled	/levelZeroFM/DAQFM/SLICEFM-1/FEDBuilderFM-1	FM	http://cmsrc-daq2:11200/rcms/gui/servlet/FMPilotServlet?groupID=508474
Running	/levelZeroFM/DAQFM/SLICEFM-1/HLTSFM-1	FM	http://cmsrc-daq2:11200/rcms/gui/servlet/FMPilotServlet?groupID=508476
Enabled	/levelZeroFM/DAQFM/SLICEFM-1/HLTSFM-1/StorageManager	XDAQ	http://srv-c2c06-13.cms:11100/urn:xdaq-application:lid=50
Enabled	/levelZeroFM/DAQFM/SLICEFM-1/RUBuilderFM-1	FM	http://cmsrc-daq2:11200/rcms/gui/servlet/FMPilotServlet?groupID=508475
Running	/levelZeroFM/DAQFM/SLICEFM-2	FM	http://cmsrc-daq3:11300/rcms/gui/servlet/FMPilotServlet?groupID=508449
Enabled	/levelZeroFM/DAQFM/SLICEFM-2/FEDBuilderFM-2	FM	http://cmsrc-daq3:11300/rcms/gui/servlet/FMPilotServlet?groupID=508459
Running	/levelZeroFM/DAQFM/SLICEFM-2/HLTSFM-2	FM	http://cmsrc-daq3:11300/rcms/gui/servlet/FMPilotServlet?groupID=508461
Enabled	/levelZeroFM/DAQFM/SLICEFM-2/HLTSFM-2/StorageManager	XDAQ	http://srv-c2c06-14.cms:11100/urn:xdaq-application:lid=50

Save

– *Save the configuration parameters of the Level Zero GUI.*

Refresh

– *Refresh the Level Zero page.*

Detach

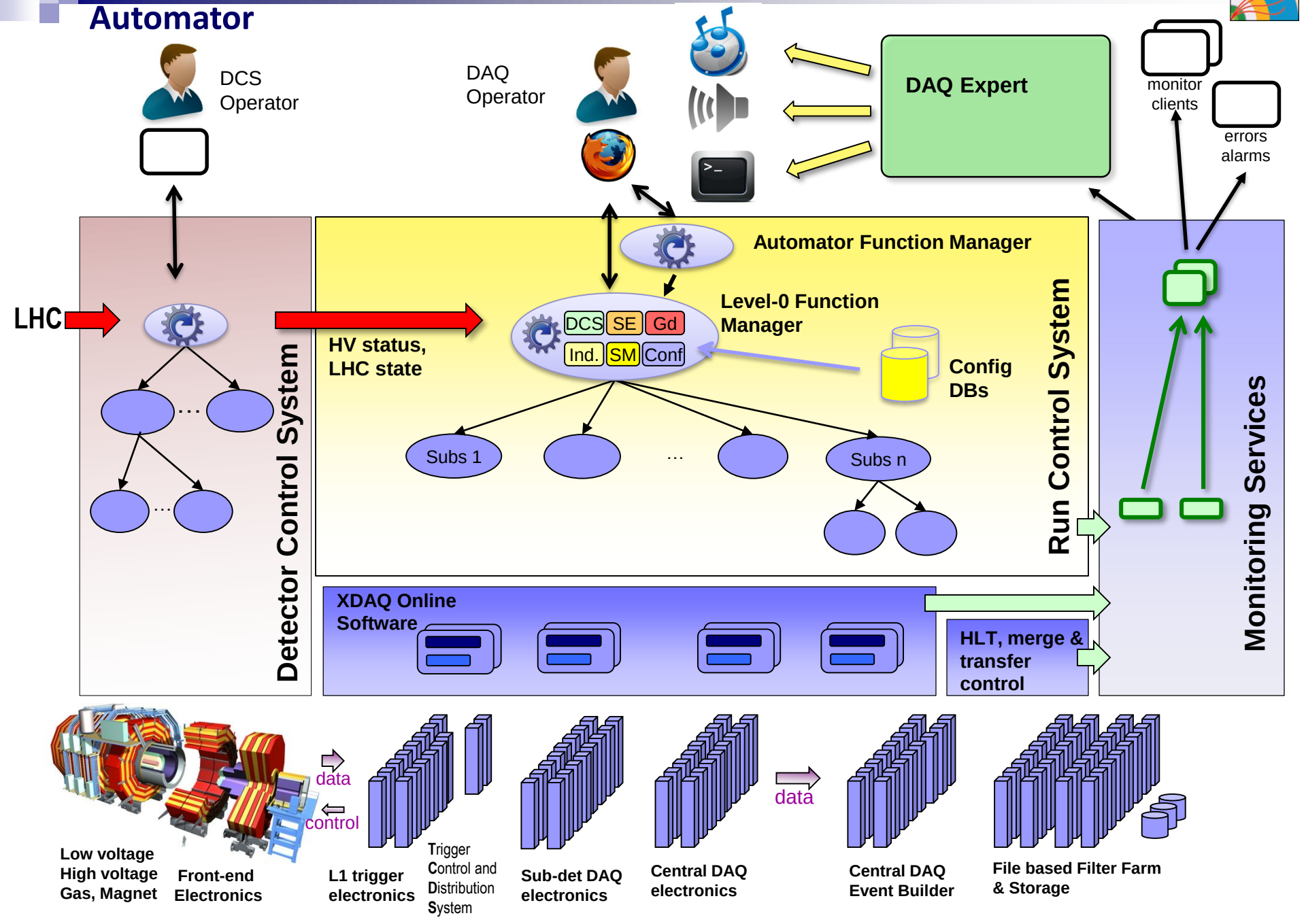
– *Disconnect the Level Zero GUI from the Level Zero FM.*

Destroy

– *Kill all Function Managers and XDAQs started by them*

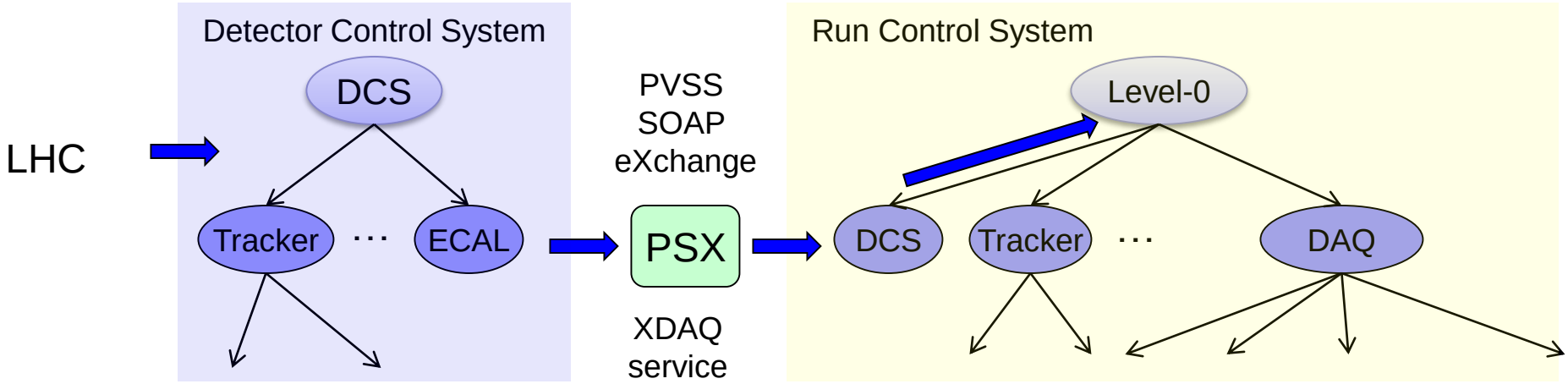
Automation: Automatic reaction to LHC beam/machine mode and DCS high voltage state

Automator

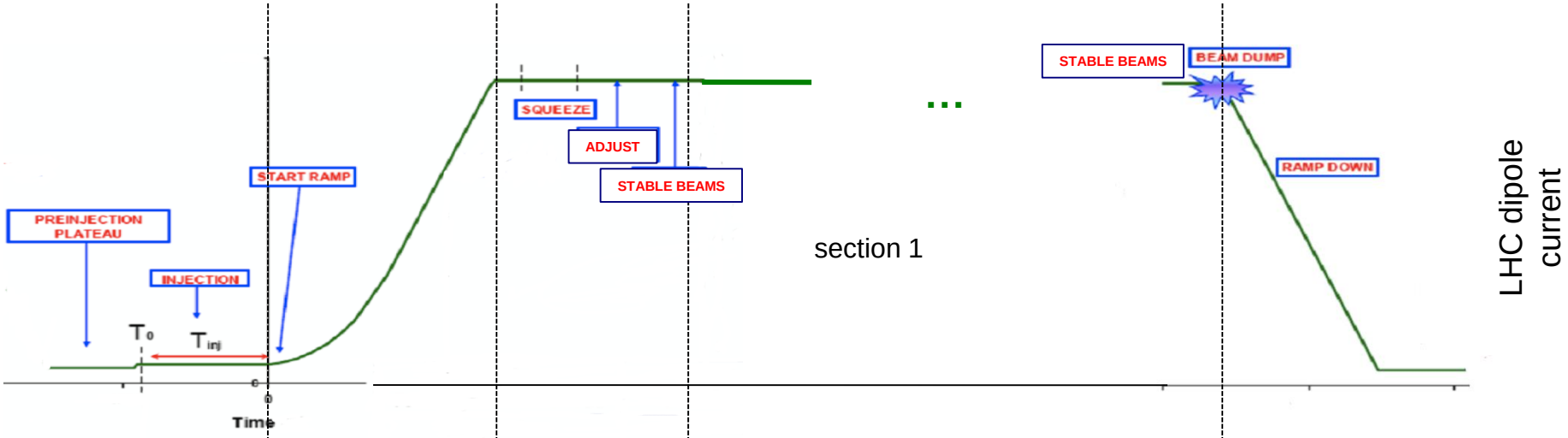

DCS Operator
DAQ Operator
DAQ Expert
monitor clients
errors alarms
LHC
Automator Function Manager
Level-0 Function Manager
DCS
SE
Gd
Ind.
SM
Conf
Config DBs
HV status, LHC state
Subs 1
Subs n
Detector Control System
Run Control System
Monitoring Services
XDAQ Online Software
HLT, merge & transfer control
data
control
data
Low voltage
High voltage
Gas, Magnet
Front-end Electronics
L1 trigger electronics
Trigger Control and Distribution System
Sub-det DAQ electronics
Central DAQ electronics
Central DAQ Event Builder
File based Filter Farm & Storage

DAQ actions on LHC and DCS state changes

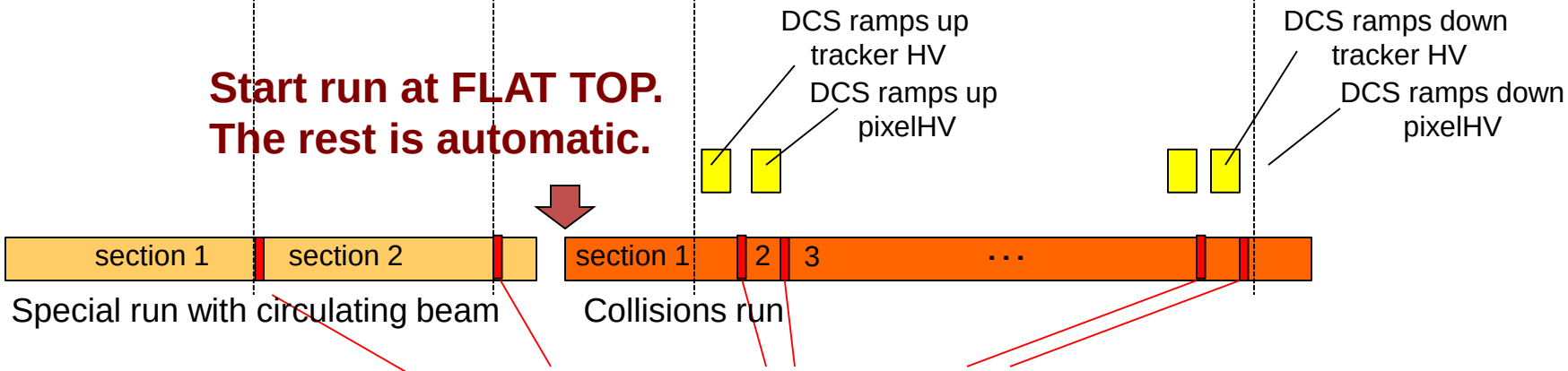
- Extensive automation in the Detector Control System (DCS)
  - Automatic handshake with the LHC
  - Automatic ramping of high voltages (HV) driven by LHC machine and beam mode
- Some DAQ settings depend on the LHC and DCS states
  - Suppress tracker payload while HV is off (noise)
  - Reduce pixel gain while HV is off
  - Mask sensitive channels while LHC ramps …
- Automatic new **run sections** driven by asynchronous state notifications from DCS/LHC



Automatic actions driven by the LHC …



Start run at FLAT TOP.
The rest is automatic.



Automatic actions in DAQ ("PerformingDCSPauseResume" state)

ramp start	**ramp done**	**Tracker HV on**	**Tracker HV off**
Mask sensitive trigger channels	Unmask sensitive trigger channels	Enable payload (Tk) raise gains (Pixel)	Disable payload (Tk) reduce gains (Pixel)

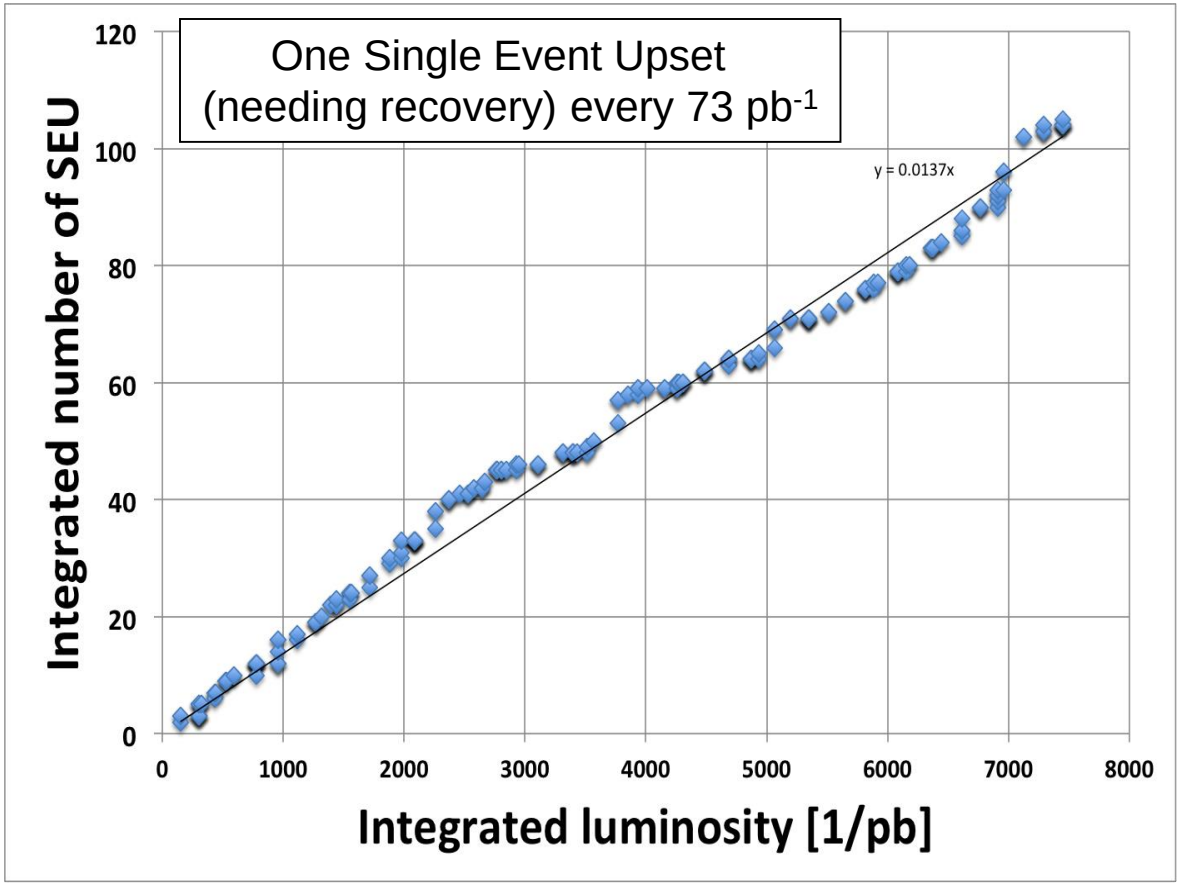
Automatic recovery from Soft Errors

Automatic soft error recovery

- With higher instantaneous luminosity in 2011 more and more frequent "soft errors" causing the run to get stuck
  - Proportional to integrated luminosity
  - Believed to be due to single event upsets

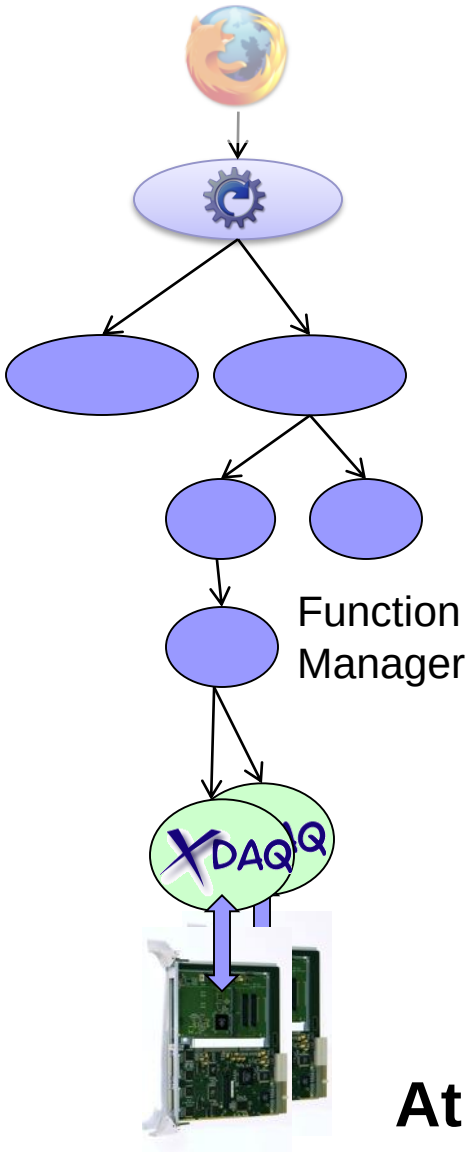
- Recovery procedure
  - Stop run (30 sec)
  - Re-configure a sub-detector (2-3 min)
  - Start new run (20 sec)

3-10 min down-time



Single-event upsets in the electronics of the Si-Pixel detector. Proportional to integrated luminosity.

Automatic soft error recovery



- From 2012, new automatic recovery procedure in top-level control node
  1. Sub-system detects soft error and signals by changing its state to **RunningSoftErrorDetected**
  2. Top-level control node invokes recovery procedure
     a) Pause Triggers (TCDS)
     b) Invoke newly defined selective recovery transition on requesting detector (**FixSoftError**)
     c) In parallel perform preventive recovery of other detectors
     d) Resynchronize
     e) Resume

12 seconds down-time

At least 46 hours of down-time avoided in 2012

Other special states

- RunningDegraded
  - Subsystems may change into **RunningDegraded** state if data taking is still continuing, but there is a problem requiring the attention of the shift crew
  - The subsystem message panel should contain a message describing the problem
  - Discuss with the shift leader how to proceed and check with the corresponding DOC if the message is not 100% clear.
- RunBlocked
  - The DAQ and Level-0 may change into the RunBlocked state if the DAQ received corrupted data from a subdetector
    - It is possible to (Force-)Stop the run from RunBlocked state.

Part 4: Monitoring tools