ALICE, ATLAS, CMS & LHCb Second Joint Workshop on DAQ@LHC

Run 2 DAQ systems

ALICE

A bit of history

- The current ALICE DAQ is the evolution of the system used to support R&D sites, test beams, commissioning, and exploitation during Run 1 and 2015.
  - This explains some slightly exotic “ifdef”s in the code.
  - It also explains certain choices of architectures, libraries, and implementations.
- The system was gradually improved to satisfy the requirements until it reached its current capabilities.
  - And it’s not over yet!

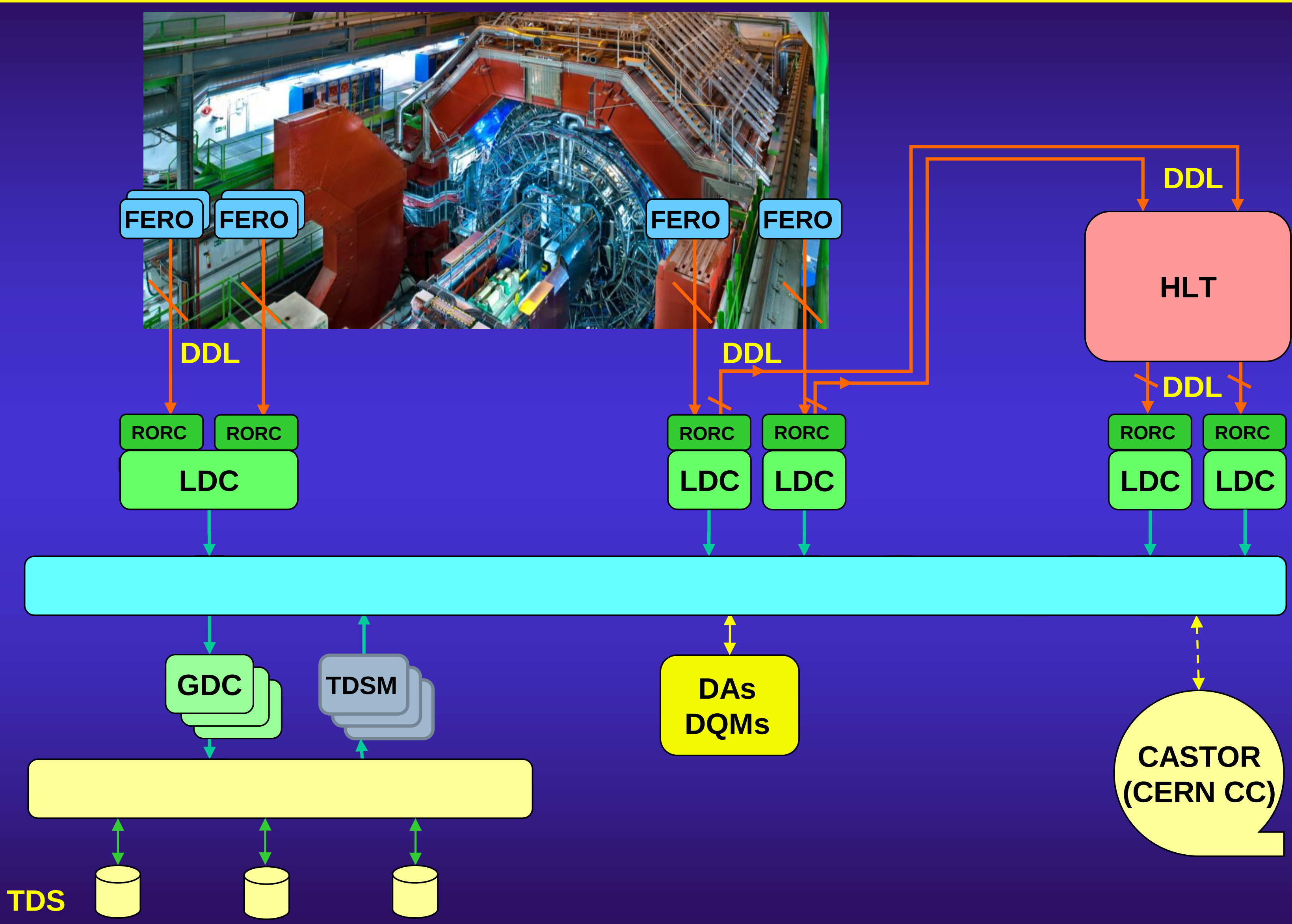
Major (recent) milestones:

1. February 2013: end of LHC Run 1.
2. 2013 – 2014: total renovation of hardware and infrastructure.
3. 2015: new data links and readout receiver cards (see talk from Heiko Engel).
4. 2016: front-end upgrade for the Time Projection Chamber (TPC).

Acronyms

- Point 2: LHC access point hosting ALICE.
- DAQ: Data AcQuisition system.
- HLT: High Level Trigger.
- FERO: Front End ReadOut.
  - Attached to the detector, gives the input to the DAQ.
- DDL: Detector Data Link.
  - Link between FERO/HLT and DAQ/HLT.
- LDC: Local Data Concentrator.
  - Entry point to the DAQ.
  - One or more LDCs are assigned to each Detector.
  - Fixed assignment Detector-FERO-DDL-LDC.
- GDC: Global Data Collector.
  - Event building and data recording.
  - Farm-like approach.
- DA: Detector Algorithms.
- DQM: Data Quality Monitoring.
- TDS: Transient Data Storage.
  - Disks local to ALICE used to store the data before sending it to the Grid.
  - TDSMs: TDS Movers.
    - Handle the TDS.
- CASTOR: CERN Advanced STOrage Manager.
  - DAQ's entry point to the Grid.

HW architecture



Detailed HW architecture

- DDL links:
  - The only custom components of the entire DAQ chain (all the rest is COTS).
  - D-RORC/DDL1:
    2 Gb/s serial optical
    2 links/card (1 in + 1 out).
  - C-RORC/DDL2:
    Max 6.25 Gb/s serial optical
    600 MB/s max per channel
    4000 MB/s max aggregate
    Up to 12 links/card.
    - 1.7 GB/s max aggregate
    - Link speeds:
      5.3 Gb/s (HLT)
      4 Gb/s (TRD)
      3 Gb/s (TPC, CTP)
    - Links:
      6 in + 6 out (36x TPC)
      2 in + 2 out (9x TRD)
      1 in + 1 out (1x CTP)
      2 in (14x HLT)
- LDCs:
  - 60x Dell R720, CPU E5-2640, 64GB.
  - 90x Supermicro X9SRE-F, CPU E5-1650, 16GB.
- GDCs:
  - 24x Dell R720, CPU E5-2690, 64GB.
- DQM/DAs:
  - 10x Dell R720, CPU E5-2665, 64GB.
- Network:
  - Force10: 3x S6000 (32x40G), 2x S4810 (48x10G), 12x S55 (44x1G).
- TDS:
  - 10x DELL MD3660f+MD3060 FC array plus extension: 60+60 disks@10kRPM, 108 TB each, total: 1080 TB (885 TB usable).
- TDS network:
  - 2x Brocade 6520.
- Uplink to LGC/CASTOR:
  - 8x 10G, standard CERN backbone.

Software architecture

- Cooperating multi-processes.
- IPC-based shared memory for synchronisation.
- Off-kernel pinned memory for data buffering.
- Out-of-the-box Linux (Scientific Linux CERN 6).
- Standard TCP/IP sockets and system calls (main streams and monitoring).
- Controls based on SMI and DIM.
- In-house messaging, alarms, configuration, and control.
- Shared file system for recording/migration with affinity (to avoid simultaneous write & read) and load distribution (by controller).

LS1 upgrade

During LHC Long Shutdown 1 (Feb 2013 – end 2014) we have substantially renewed our HW:

- C-RORCs replaced several D-RORCs (TPC, TRD, HLT).
- Newer hosts to take over overloaded/obsolete LDCs, GDCs, DQMs, DAs, and servers.
- Network completely renewed (Ethernet and FCS).
- All racks (the original ones were from LEP/L3) replaced.
- New cooling.
- Infrastructure almost totally rebuilt.

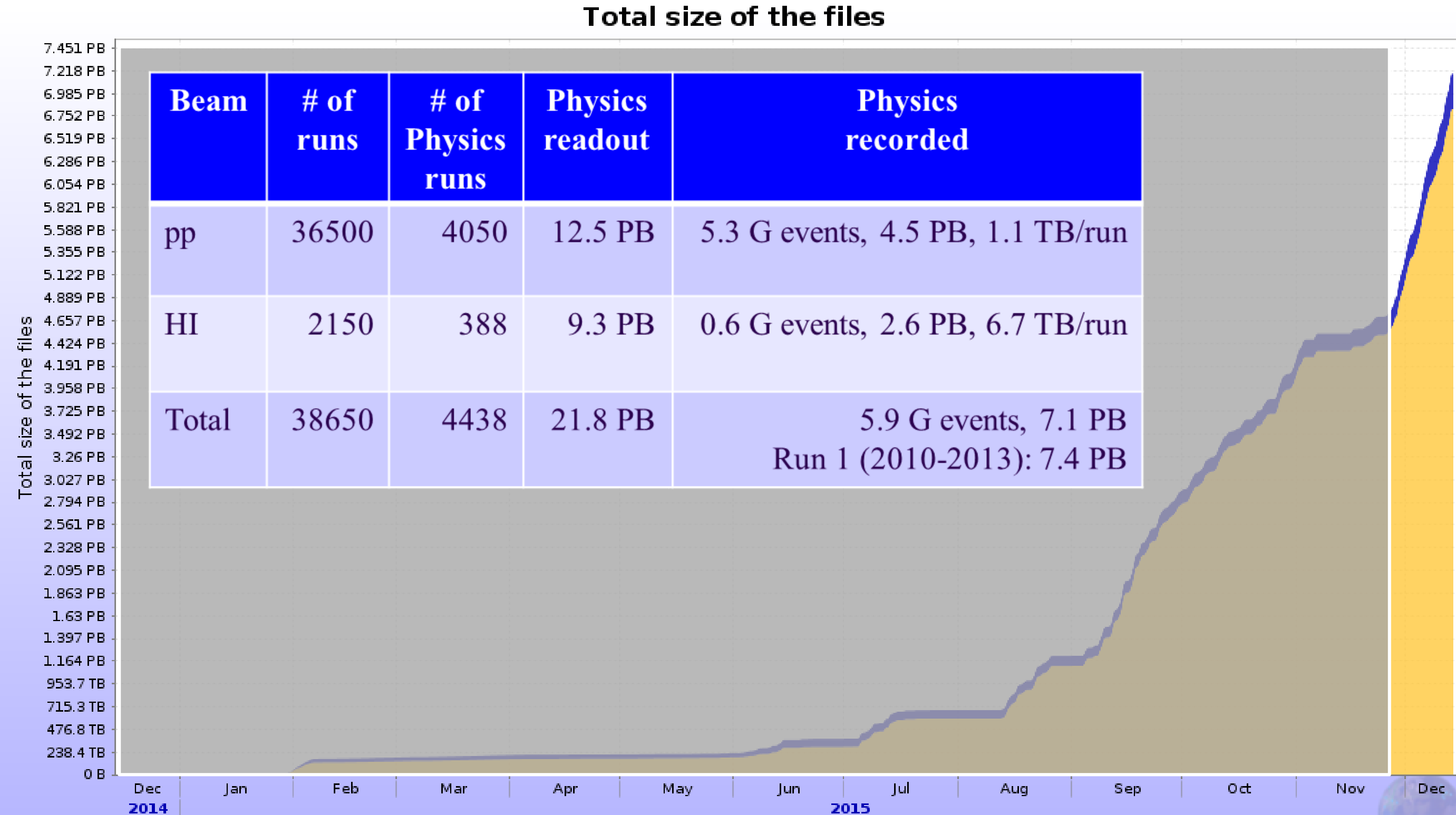
Achievements

- End of LHC Run 1 (2012 – Feb 2013):
  1. 13 GB/s from the detectors.
  2. 7 GB/s RAW to disk.
  3. 4 GB/s HLT-COMPRESSED sustained to disk/Grid.
- First year of LHC Run 2 (2015):
  1. 17 GB/s from the detectors @ [1.4 .. 3.5] KHz.
  2. 6 GB/s HLT-COMPRESSED sustained to disk.
  3. 7 GB/s to the Grid.

Beam	# of runs	# of Physics runs	Physics readout	Physics readout TPC	Physics readout HLT	Physics readout TRD	Physics recorded
pp	36500	4050	12.5 PB	10.0 PB	1.7 PB	0.4 PB	5.3 G events, 4.5 PB, 1.1 TB/run
HI	2150	388	9.3 PB	6.8 PB	1.8 PB	0.6 PB	0.6 G events, 2.6 PB, 6.7 TB/run
Total	38650	4438	21.8 PB	16.8 PB	3.5 PB	1.0 PB	5.9 G events, 7.1 PB Run 1 (2010-2013): 7.4 PB

pp vs HI

2 ½ weeks of HI (with several breaks)…

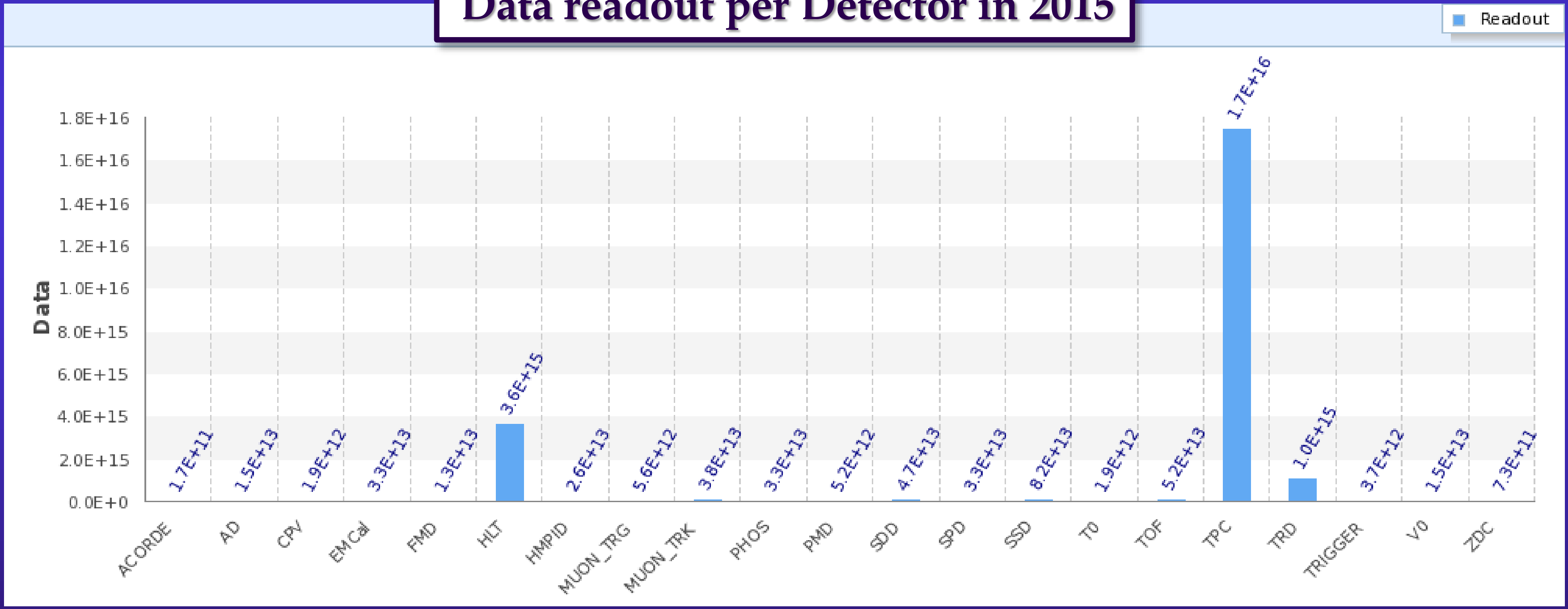


Beam	# of runs	# of Physics runs	Physics readout	Physics recorded
pp	36500	4050	12.5 PB	5.3 G events, 4.5 PB, 1.1 TB/run
HI	2150	388	9.3 PB	0.6 G events, 2.6 PB, 6.7 TB/run
Total	38650	4438	21.8 PB	5.9 G events, 7.1 PB Run 1 (2010-2013): 7.4 PB

Challenges for 2016

- The Time Projection Chamber is upgrading its Readout Control Units, effectively doubling its readout capacity by sensibly reducing its deadtime.

Data readout per Detector in 2015



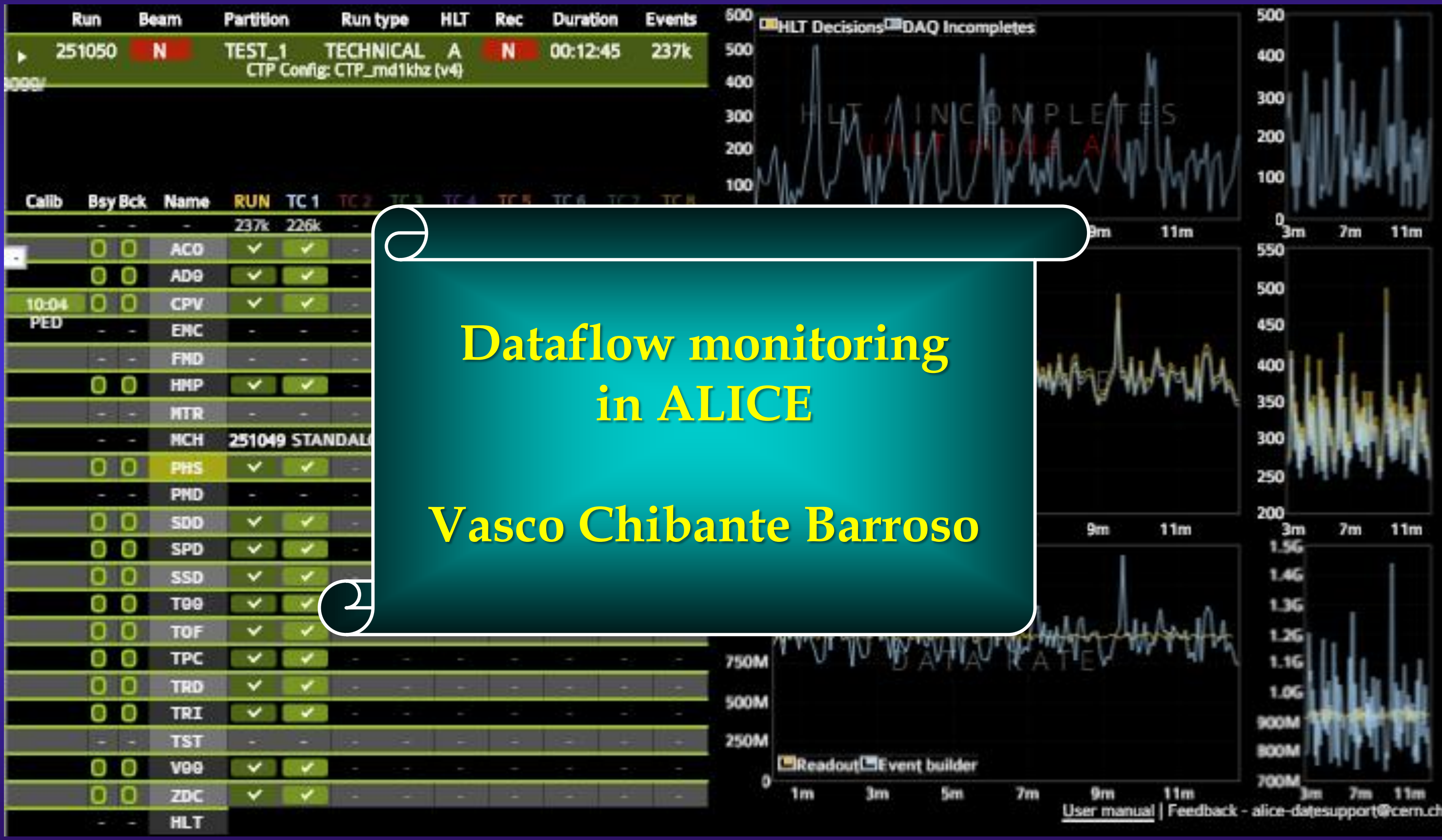
- HLT and DAQ will have to follow the trend.

Making a better DAQ

- During LHC Run 1, it became obvious that Detectors do fail and that configurations must be changed.
  - Particularly true when you have 19 Detectors and 100 Trigger Classes combinable in 8 independent Trigger Clusters.
- We worked a lot on improved error handling and quicker ways to change the configurations:

1. Efficient monitoring.
2. Faster Start and End run.
3. Mid-Run error recovery.
4. Lightweight End/Start run.

Efficient monitoring

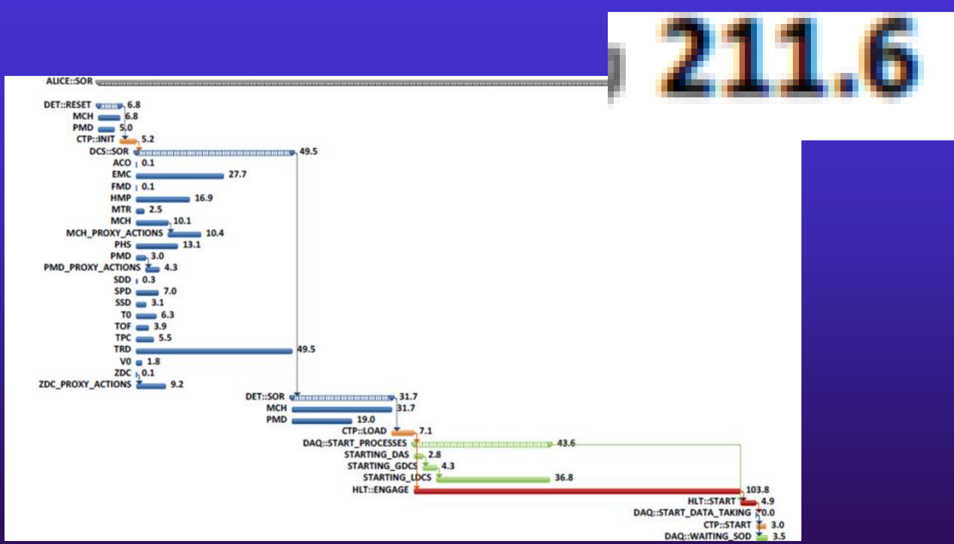


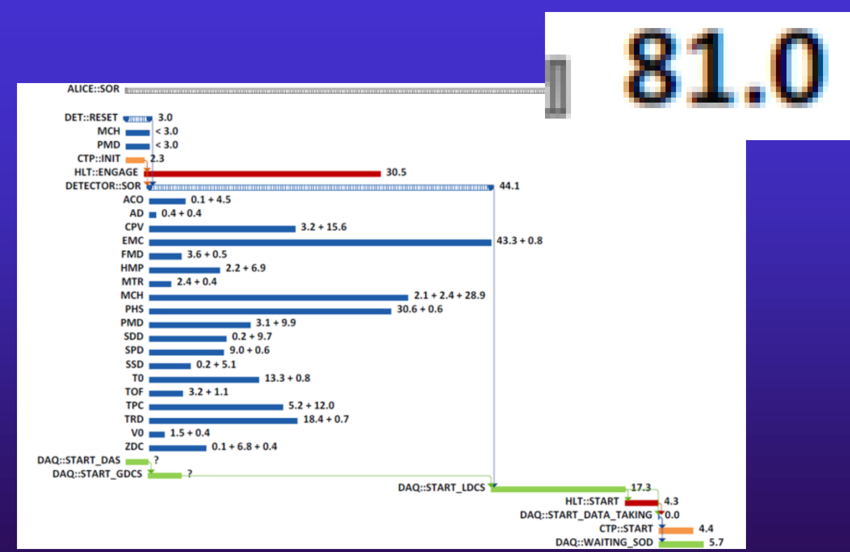
Dataflow monitoring in ALICE

Vasco Chibante Barroso

Faster Start and End run

- A detailed profiling of the procedures to start and end the runs was done on the live systems.
- We proceed to parallelize what could be parallelized and optimize what could be optimized for all the Systems (without effecting their functionality).
  - Example: time to start a run, before and after…





Mid-Run error recovery

We have established a common Pause-and-Reconfigure (PAR) procedure to:

1. Recover individual Detectors triggered by messages in the data, state changes in the DCS or commands from the Shift Crew.
2. "Ping" the Detectors to check their health and eventually recover them.
3. Keep the healthy detectors r[illegible]ning whenev[illegible]possible[illegible]

Beam	# of PARs (Physics)	# of su[illegible]
pp	438	[illegible]
HI	184	[illegible]

During the last 4 days of HI:
46 recoveries/49 activations
94% success rate

PARs example

The top plot (the "EVENT RATE") shows the Hz total (in yellow) and the individual Trigger Clusters (TCs) in various colours:

- Purple TC: MUONs.
- Green TC: TPC.

Note that Purple TC and Green TC drop independently:

- Dip in Purple:
  MUON being PAR'ed.
- Dip in Green:
  TPC being PAR'ed.

These are not incomplete events but rather disabled TCs.

- Best achievable efficiency under the circumstances!

Lightweight End/Start run

- For some types of configuration changes, we can now execute a much faster end/start of run procedure:
  - Triggers are paused.
  - All the operations needed to close the ongoing run are done.
  - The configuration is changed (e.g. downscaling of trigger inputs, recovery of a HW failure etc…).
  - The steps required to start the new run are taken.
    - Detectors are <u>not</u> reset/restarted and the links are kept open.
  - Triggers are resumed.
- Some changes can also be preloaded <u>before</u> the EOR…
- All Systems played the game and implemented the above.
- Will be re-validated and used in production this year.

Conclusions

- The ALICE Data Acquisition kept increasing its throughput throughout the years (thanks COTS!).
  - 2015's sustained average (6 GB/s) is much higher than 2010's (first HI) peak (2.5 GB/s).
  - Comparable data has been written in 2015 alone (7.1 PB) as during the whole of LHC Run 1 (7.4 PB in 4 years).
- Tomorrow's challenges will be even more challenging thanks to improved detectors front-ends and faster data links.
- Running efficiency more a more a hot issue…
  - Effective error detection/recovery and lightweight reconfiguration.

We will have many hints for the DAQ of LHC Run 3 and HL-LHC.

But this is another story…