


Data flow operations and analysis of data at LHC

T.Camporesi,C. Clement,C.Garabatos, Cuadrado,R. Jacobsson, L. Malgeri, T. Pauly

The operation of the LHC detector - 28/6/2010 - CERN

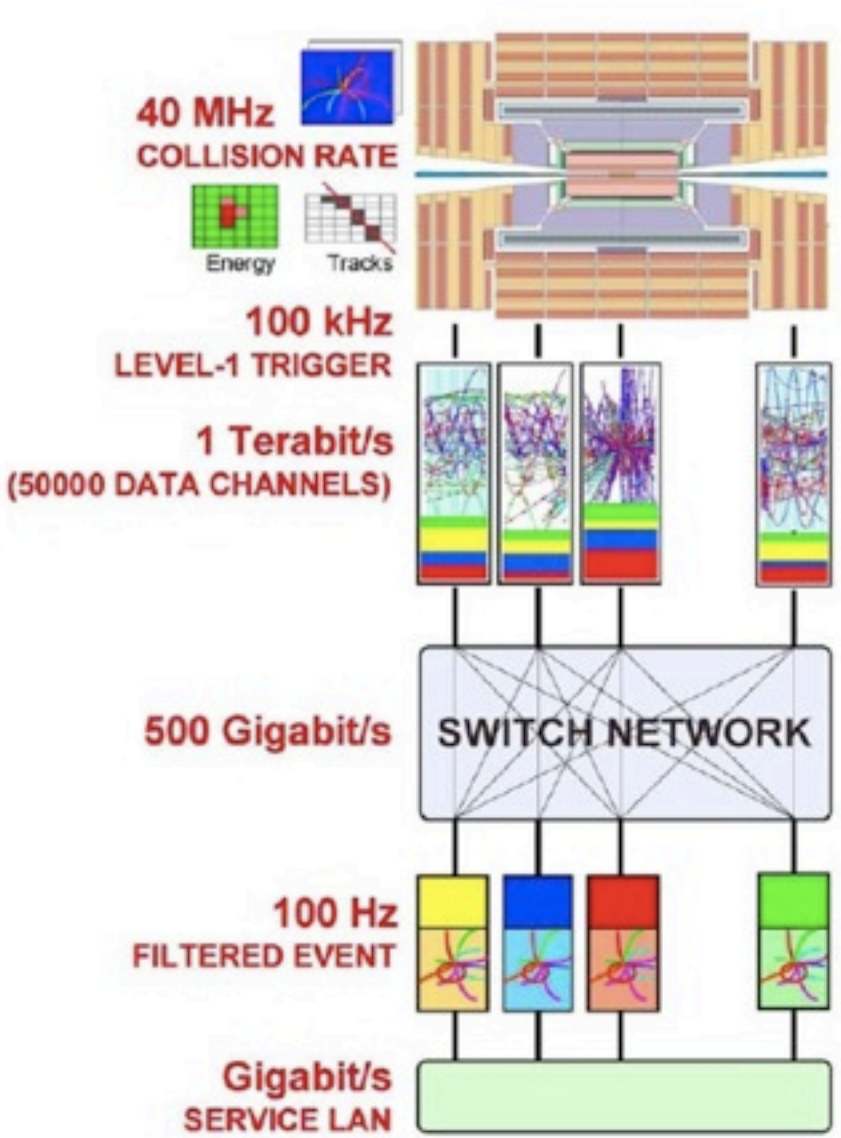
Outline

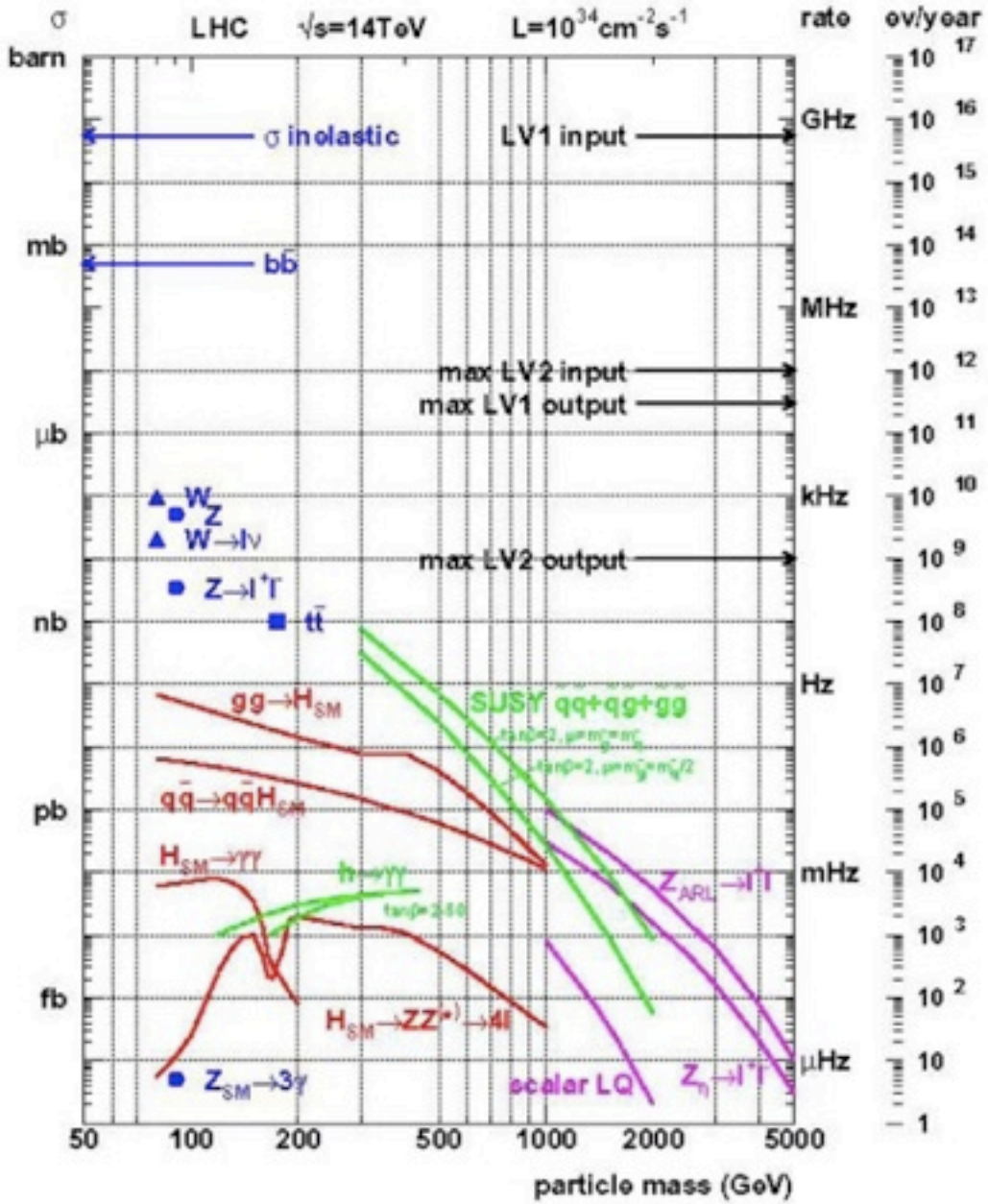
- From trigger bit to dataset for physics
- Data- and work-flow from Tier0 to Tier3
  - Calibration loops
- Data processing and usage (after the first 24h)
- Validation of data for physics analyses
  - data quality monitoring online
  - data quality monitoring offline

Some premises

- The LHC is a machine where most of the collisions must be thrown away and "physics" selection should be performed early on. It must be really clear which kind of events we want to retain and how to efficiently select them.
- In terms of dataflow, the LHC is a very complex pipe work. Data must flow continuously (no buffer is big enough to keep them too long). Any interruption of the flow has back propagation problems.
- Data must be available in a short delay: parallelize, split resources, split storage.

Nominal data rate





Real trigger rate these days

We are still far from nominal luminosities. The trigger rate at L1 is currently ~10-20 kHz (with mild pre-scaling of the Minimum Bias triggers).

The idea, for all experiments, is anyhow to keep the data flow at the edge of the bandwidth reducing all thresholds and adapting the trigger menu.

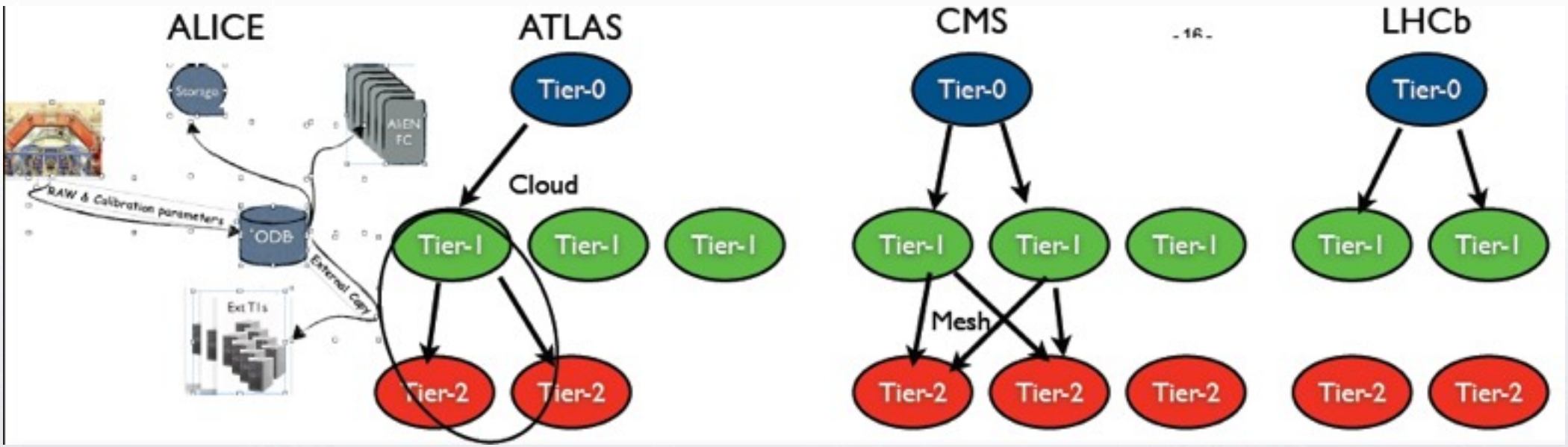
Few typical numbers:

- L1 rate: 10-20 kHz
- HLT rate (out of the pit): ~300-500 Hz
- Max bandwidth: ~300-500 MB/s

The name of the game is to optimize the constraints vs physics and commissioning needs

Distributed computing

All experiments have adopted some variation of a "tiered" distribution approach

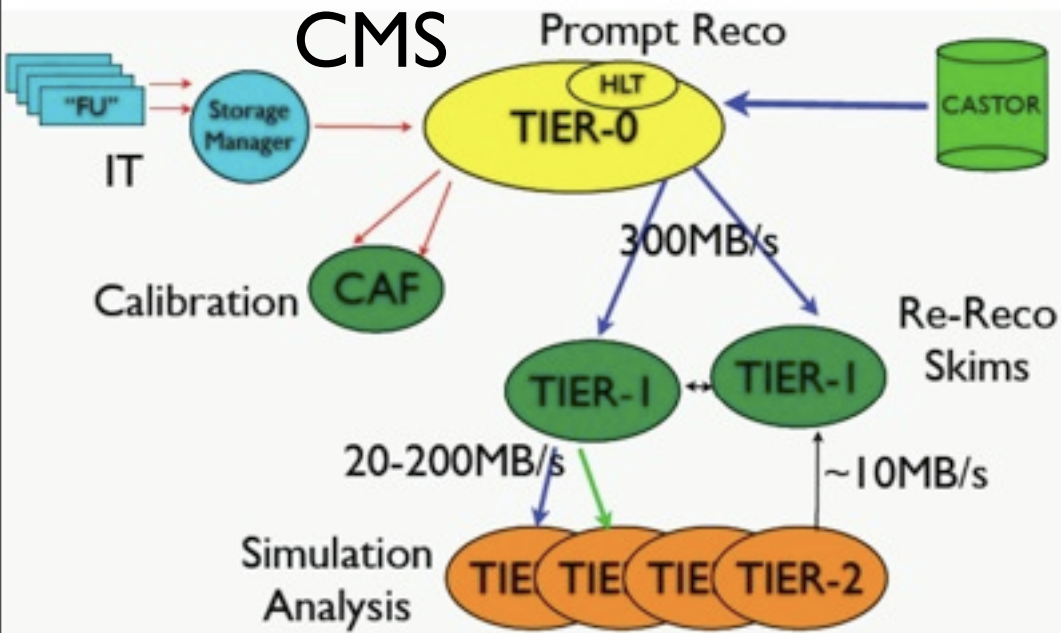


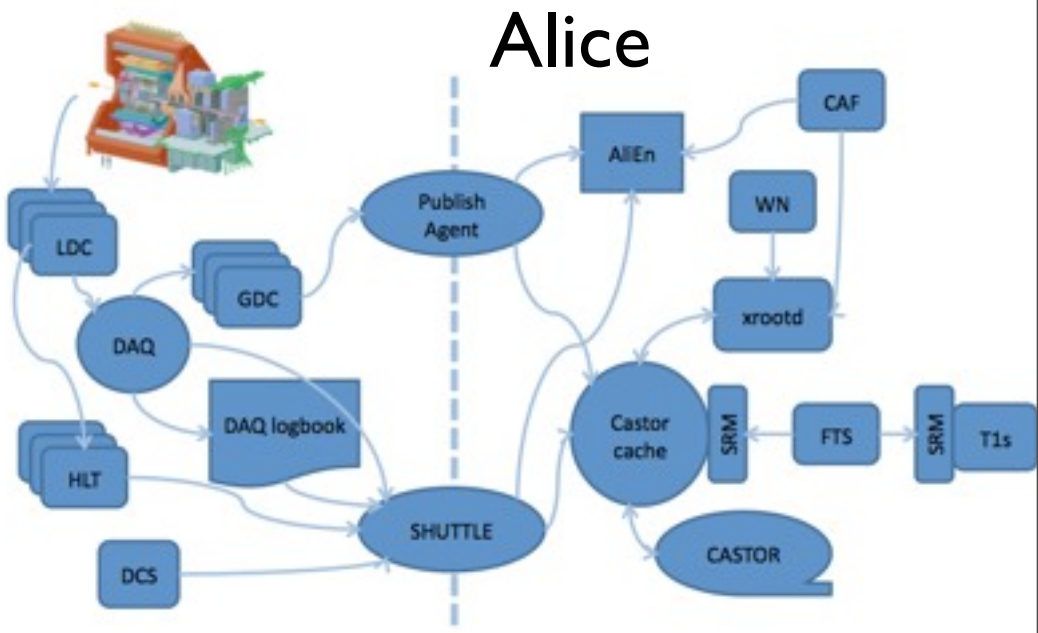
A bit of jargon

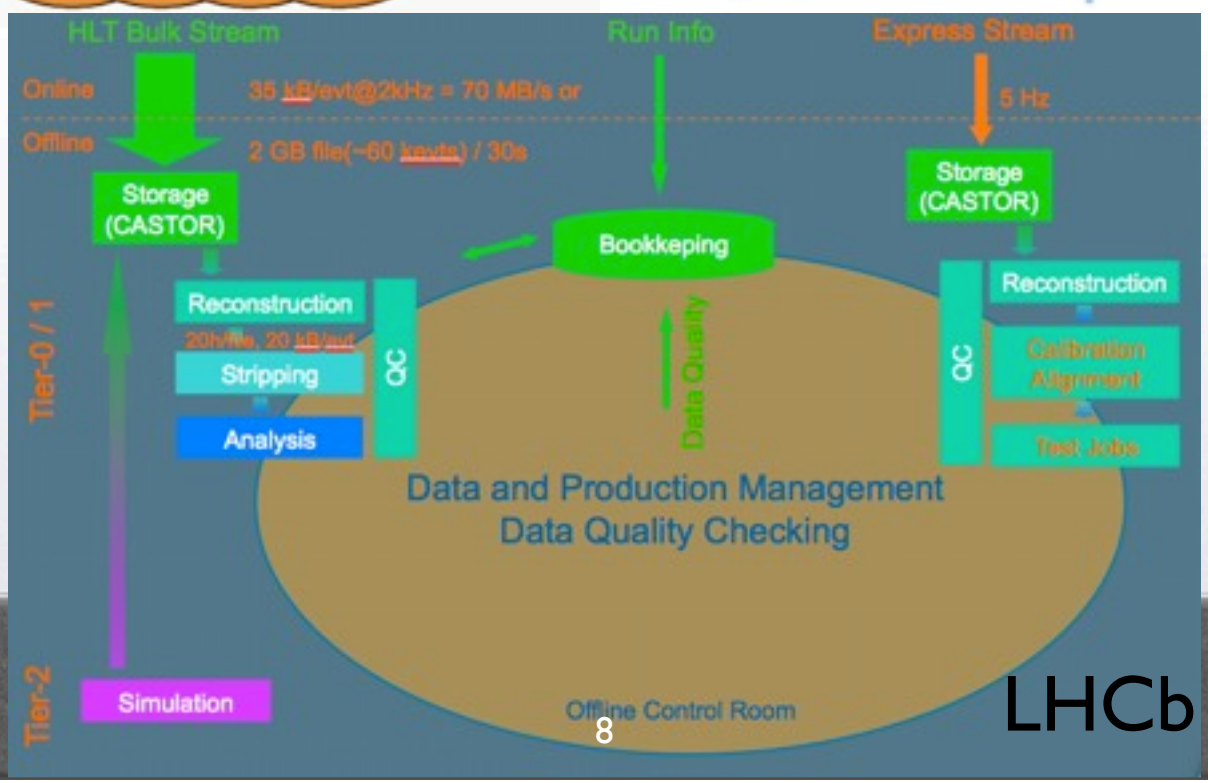
The tiers are a structured level of cpu/disk farms needed to parallelize data storage, processing and access:

- Tier0 → CERN
- Tier1 → big centres (<10) with large disk/tape capacity used for long term storage and central data crunching
- Tier2 → smaller center for user analysis (or MC productions)
- Tier3 → even smaller center for pure data analysis
- CAF → Central Analysis Facility: dedicated center for fast turn-around analyses or prompt feedback/ calibrations

Data flow sketch







Optimization

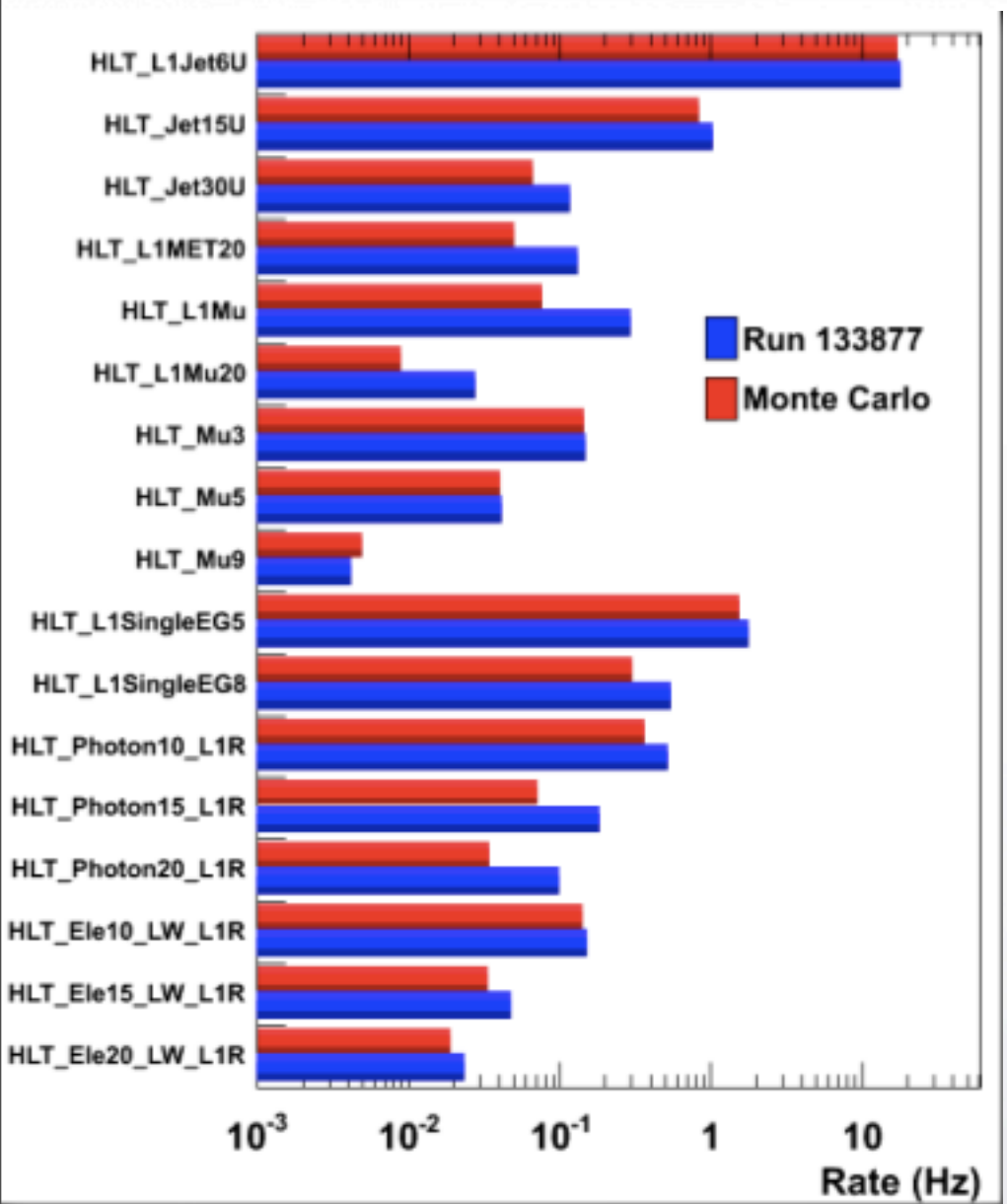
Optimization = parallelization

Events are classified according to triggers (ECAL, JET, MET, MUON, etc.) and bundled in combinations of triggers interesting for specific physics analyses.

The bundles (named streams/datasets/etc.) are sent in parallel to Tier0 and then processed. Their history continue in separate lives.

Optimization means also balancing small data size with large numbers of small files.

An example menu



Run2010A

MinimumBias
HcalCalIsoTrk ALCARECO
SiStripCalMinBias ALCARECO
SiStripCalZeroBias ALCARECO
TkAlMinBias ALCARECO
DQM

ZeroBias
SiStripCalZeroBias ALCARECO
DQM

EG
EcalCalElectron ALCARECO
DQM

EGMonitor
DQM

JetMETTau
HcalCalDijets ALCARECO
DQM

JetMETTauMonitor
DQM

Mu
DtCalib ALCARECO
MuAlCalIsolatedMu ALCARECO
MuAlOverlaps ALCARECO
TkAlMuonIsolated ALCARECO
DQM

MuMonitor
DQM

Commissioning
DQM

Rereco datasets

From **Commissioning**: MuonDPG_skim

From **MinimumBias**:
CSCSkim_BeamHalo_MinBias
MuonDPG_skim
ECALRECHIT
Skim_StoppedHSCP
BEAMBKGV3
Skim_logerror
valskim
TPGSkim

From **ZeroBias**: GOODVERTEX

From **JetMETTau**:
CS_DiJetAve
CS_Tau

From **Mu**: CS_Onia

Skim datasets

Datasets for physics

What are the criteria for bundling trigger bits in datasets?

- the single dataset should be manageable both by the workload management and by the end users: how would you read/ analyse a single dataset of >1PB?
- the list of triggers in a dataset should be logically grouped such to minimize the number of dataset needed by a typical analysis.

The drawback of such approach is that inevitably some of the events are duplicated: triggering two or more paths heading to different dataset.

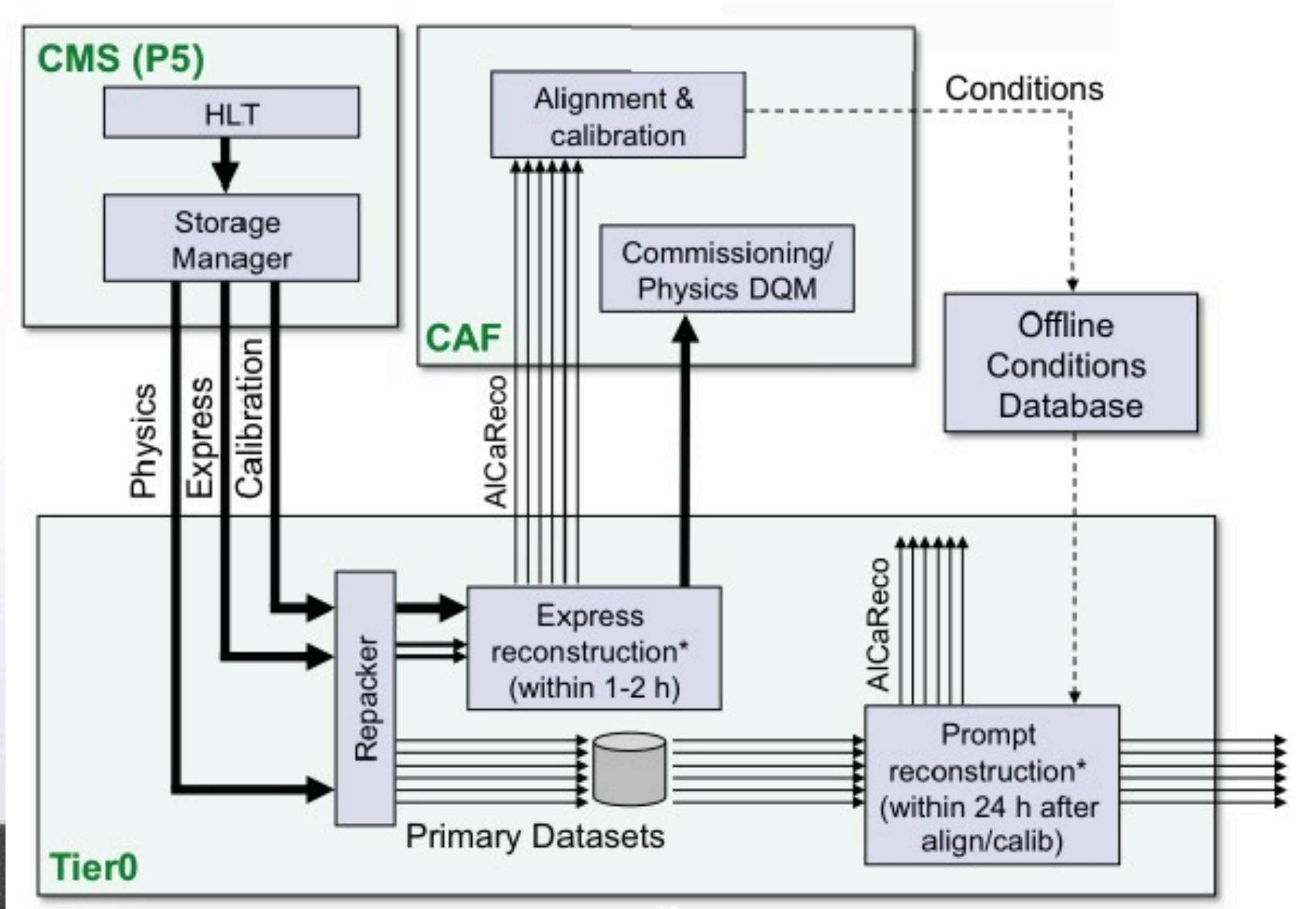
- The overhead in data transfer might be as high as 40% especially at the beginning (not many handles to differentiate MinBias events).
- The overlap need to be taken into account in the analyses.

Other Datasets

In addition there are dedicated streams of data which contain selected type of events for:

- calibration (reduced format events to collect high statistics of pi0, isolated track, etc.)
- alignment (reduced format of high pT tracks events)
- data quality monitoring (selected sample of "physics"-trigger events)
- trigger monitoring (dedicated stream with all details of the trigger decision)
- conditions/detector status
- visualization
- etc.

A Dataflow Example



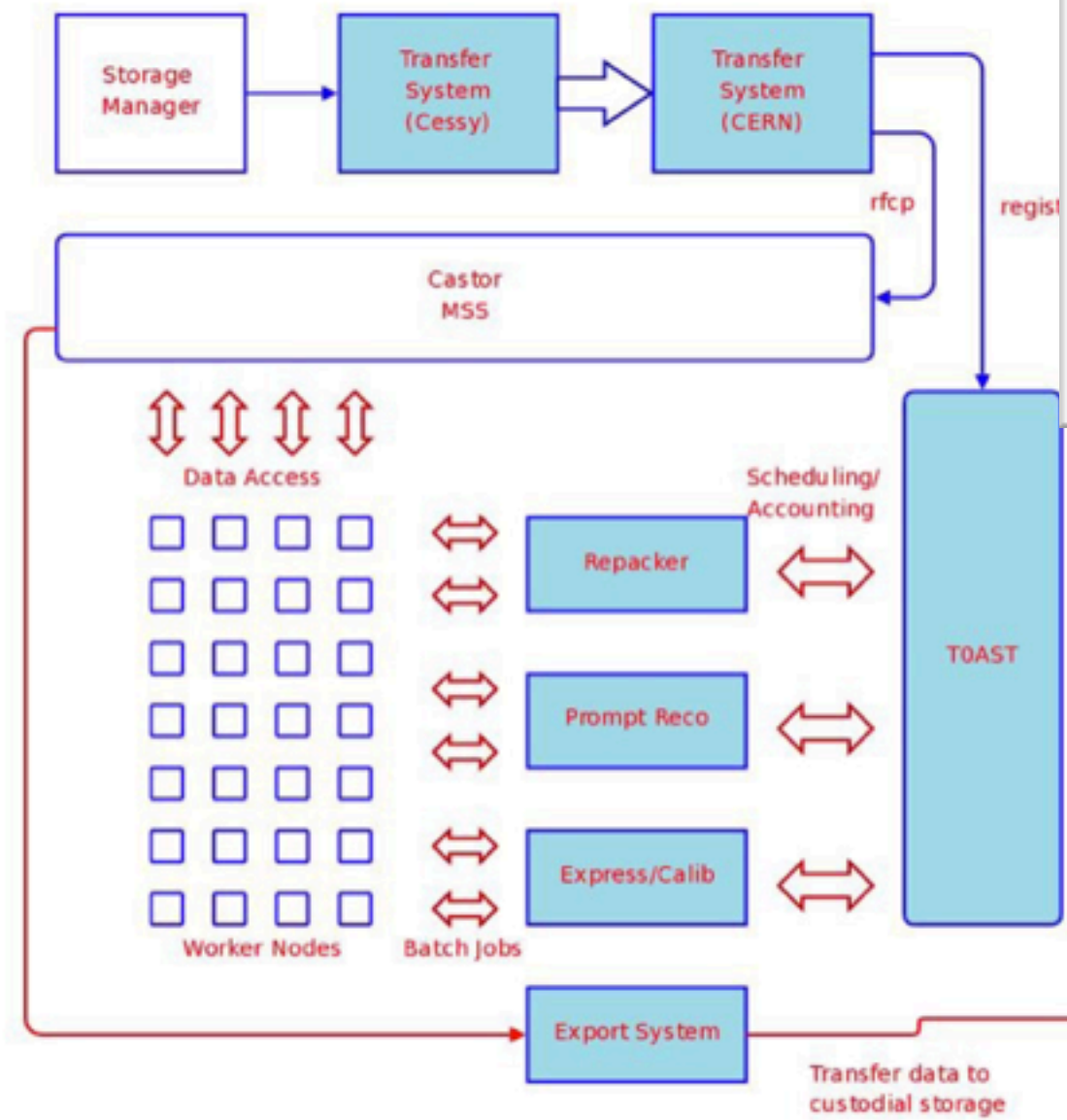
A Dataflow Example

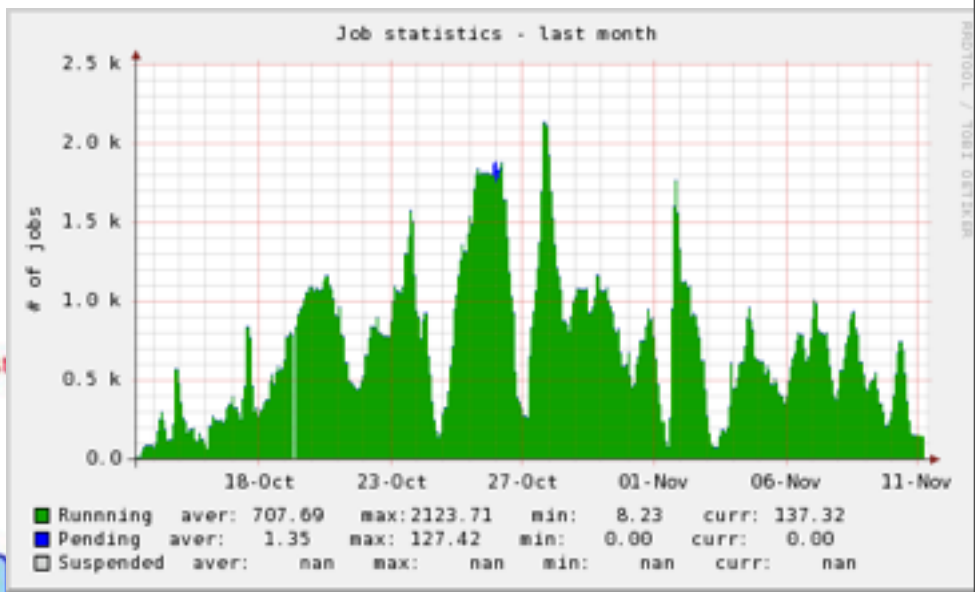
Data is transferred in three principal streams:

- Physics stream: The main data stream intended for physics analysis, comprising the datasets
- Express stream: Specially selected subset of events (few percent, initially more, of the total data bandwidth), chosen according to trigger bits, for prompt physics analysis, calibration and alignment, detector and trigger commissioning, fast physics validation
- Calibration Stream: Dedicated streams for calibration and alignment. Includes:
  - Hardware calibration streams: Events taken during the LHC orbit gap, important e.g. for pedestal and noise calibration, when no signal is present
  - "AlCaRaw" streams: dedicated and highly compact data format with only needed information for Alignment and Calibration studies (allows higher rate in pre-allocated bandwidth)

A Tier0 workflow

The Tier0 is a data crunching system





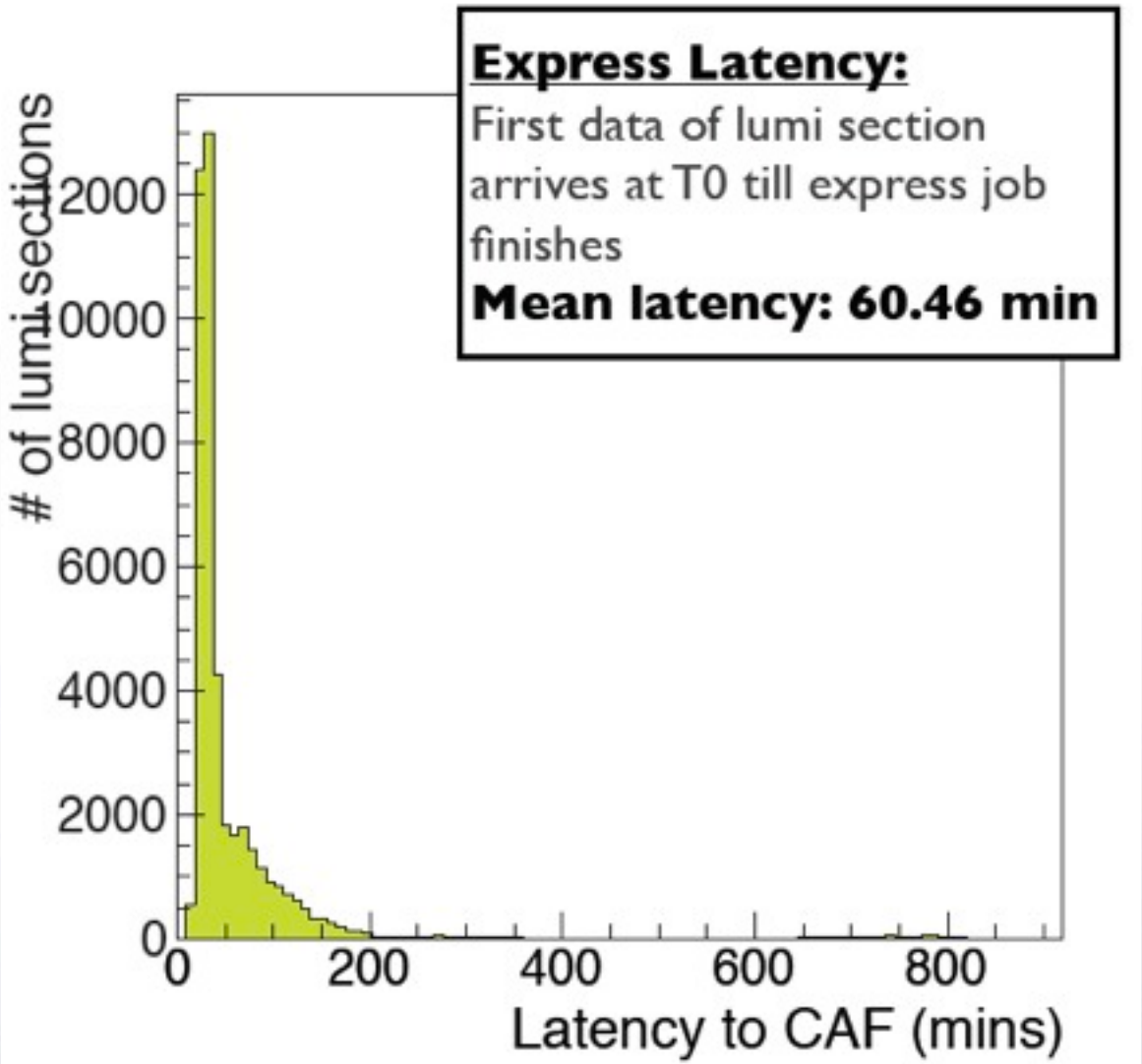
It is strictly coupled to online and it should be fail safe.
It is not designed to stop: relatively small caches.
If something fails will be recovered in a reprocessing step.

Two words on Express

The express stream and its processing is particularly important because it is at the front-line of the validation of the collected data.

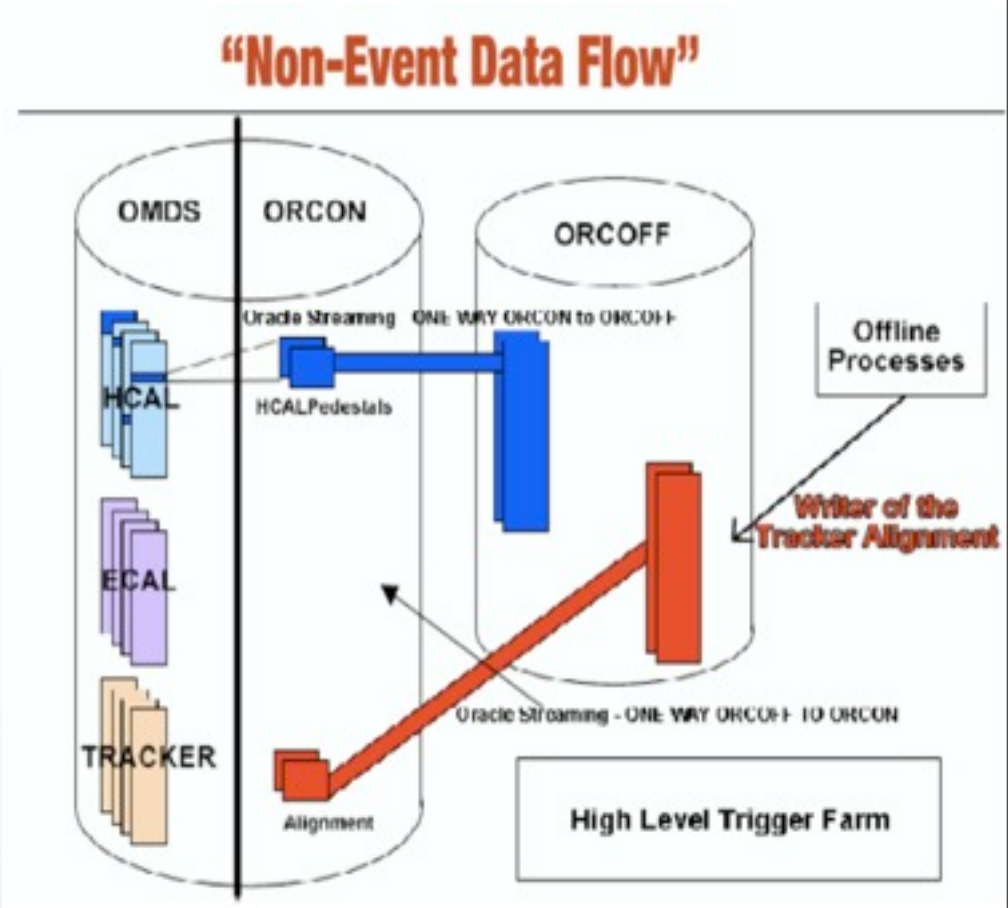
A prompt evaluation of the quality of the data allows:

1) quick reaction in case of problems
2) feedback on the full reconstruction step
3) an hotline for fast physics analyses



Conditions

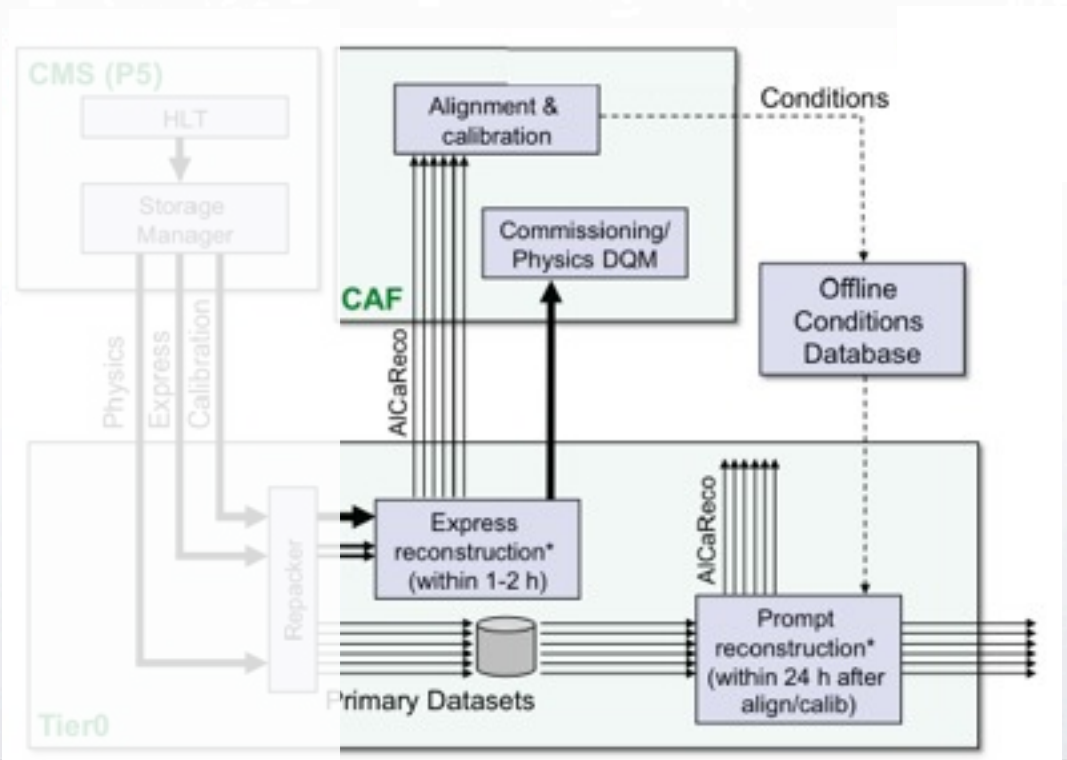
- A fundamental aspect in the dataflow of data for physics analysis are the "conditions" used to reconstructed the data.
- Trivially we need to convert ADC/TDC counts to a physical observable, the conversion parameters are the calibration constants.
- Calibrations might vary in time (think about pedestals/temperature dependent calibrations/timing drift, etc.) and need to be derived and stored in dedicated DB.
- The amount of "conditions" data might be big: from few GB/day to hundreds of GB/day (few numbers per readout channel every few seconds).
- It must be kept fully in sync with data taking and reconstruction.



Prompt Calibration

Prompt alignment & calibration workflows produce constants from the express & calibration stream data. These constants are validated and uploaded to the database.

When prompt alignment & calibration are completed (target latency: ~24h), the full data sample is passed through prompt reconstruction with the updated constants, and generally made available for analysis



Prompt Calibration

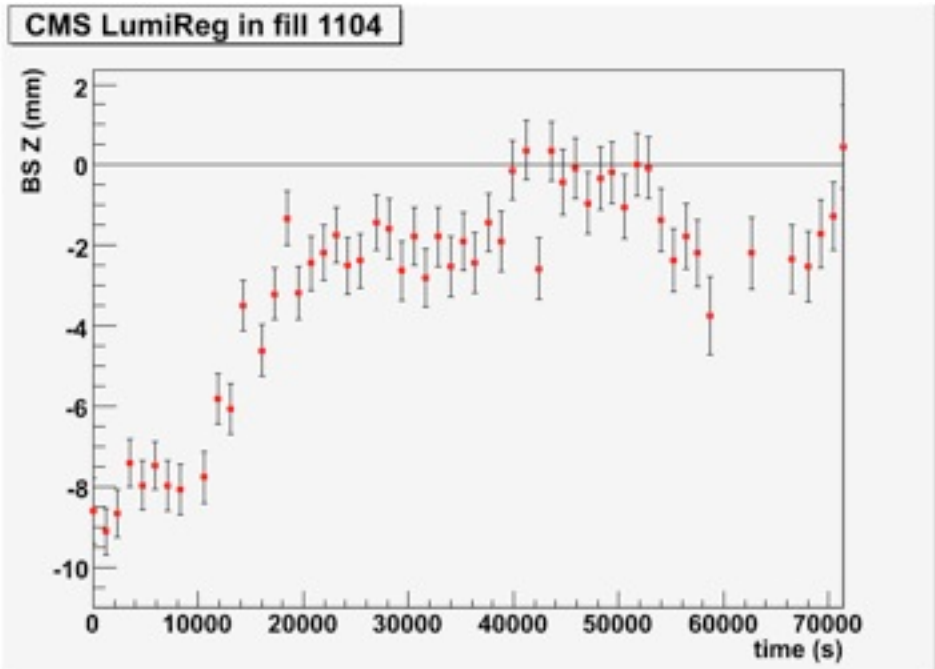
From first running experience things which are part of the prompt calibration loop:

- pedestals
- masked/noisy channels

online

- beam spot: needed to "clean-up" tracking with initial vertex assumption
- time calibrations

offline

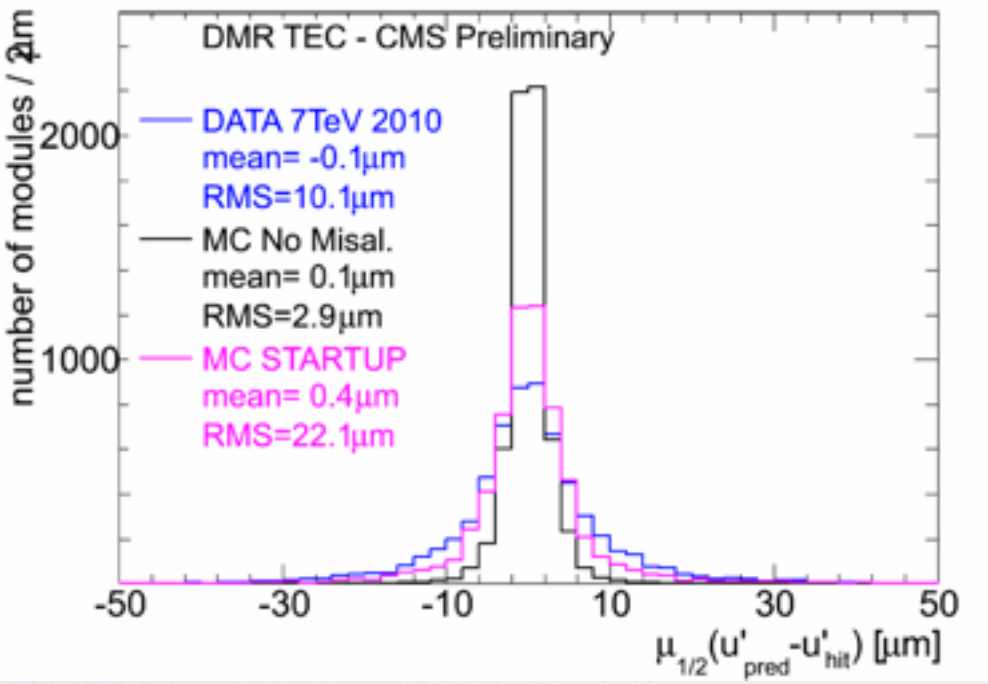


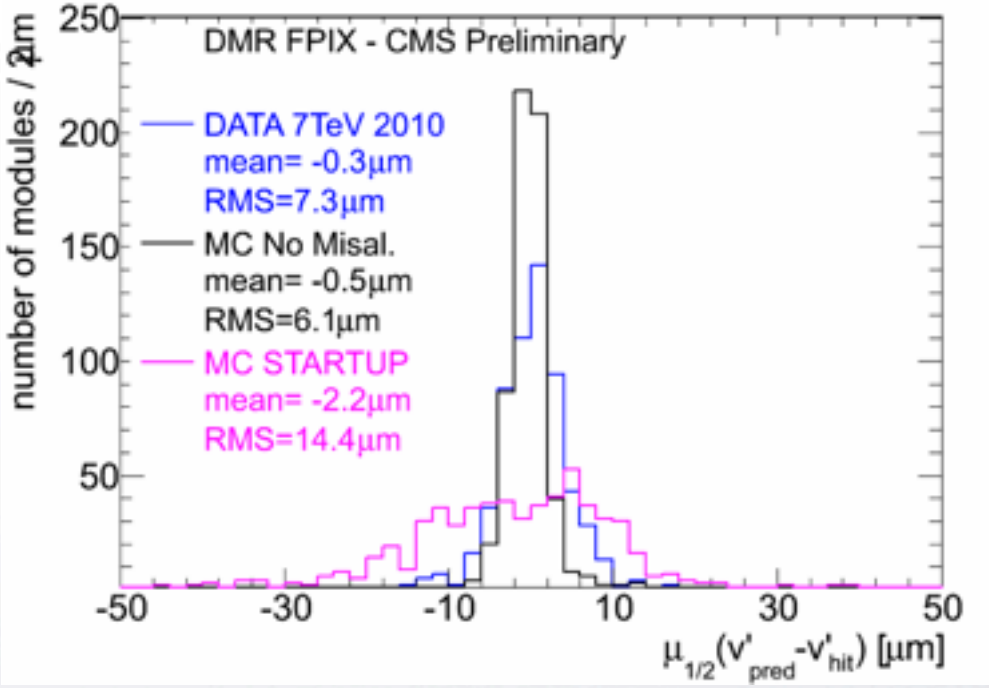
Things that will go in prompt calibrations with high luminosity:

- alignments (if sizeable changes are detected)
- calorimeters intercalibrations
- .....

Long term calibrations

Long term calibrations are the ones needing statistics larger than what available in one or two fills (24h). They enter in the data processing only at subsequent re-processing steps.





Processing lines summary

Data arriving from the experiment are in order:

1) converted (if needed) from RAW data to root readable format
2) stored in RAW format at T0/1
3) separated in "physics" interesting streams
4) reconstructed using the latest available conditions
5) (prompt reco) at the T0
6) shipped to remote sites (T1/T2)
7) re-reconstructed several times when:
   - new/improved reconstruction algorithms are deployed
   - new/improved calibrations are available

Examples:

- the data collected so far were fully reprocessed between 4 and 7 times
- this rate will not be possible anymore when luminosity (and data size) will increase

What after the first 24h

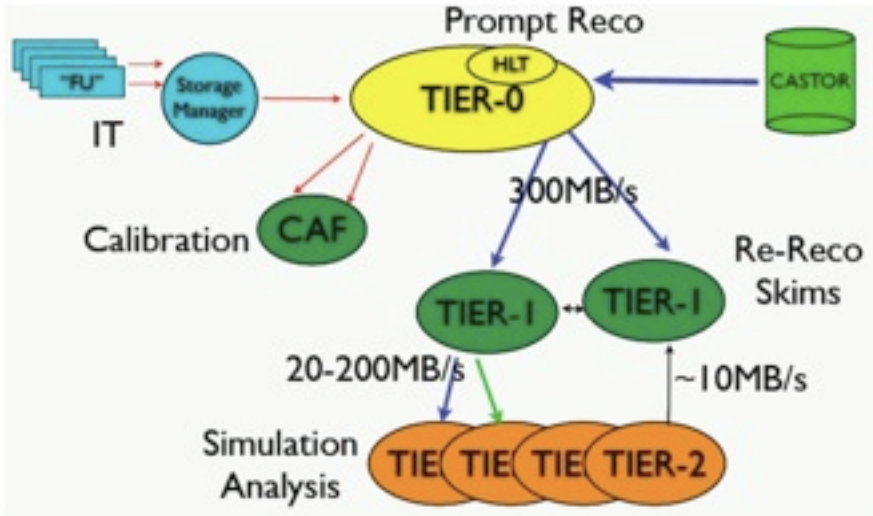
We have promptly reco-ed events. They are separated in bundles according to trigger bits. They are available at CERN (and possibly outside).

Some experiments foresee an additional processing workflow at T0/T1 (a.k.a. skimming) refining the selection of events in dedicated streams/datasets for:

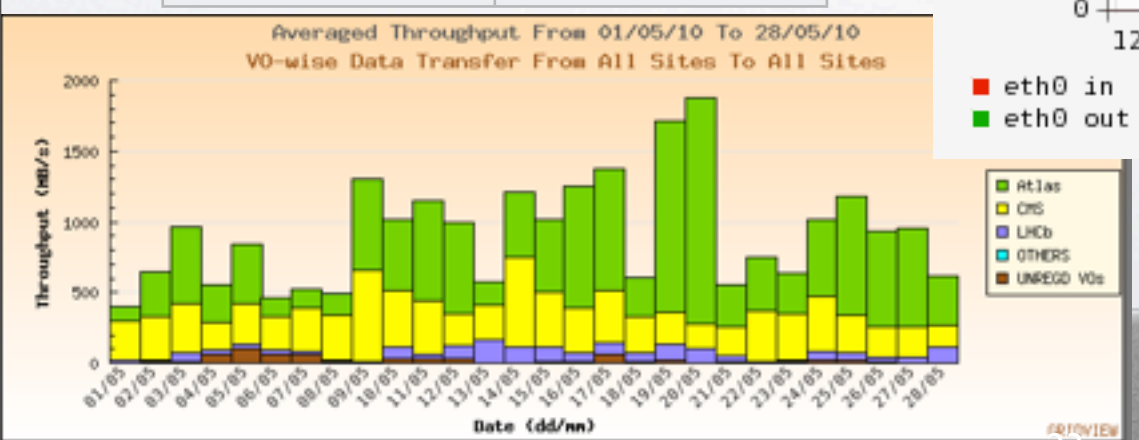
- physics studies (J/Psi, Y, W, etc.)
- detector commissioning studies (beam halo, noise, machine backgrounds)
- trigger studies (Zero Bias, Min Bias, single object)
- validation studies
- etc.

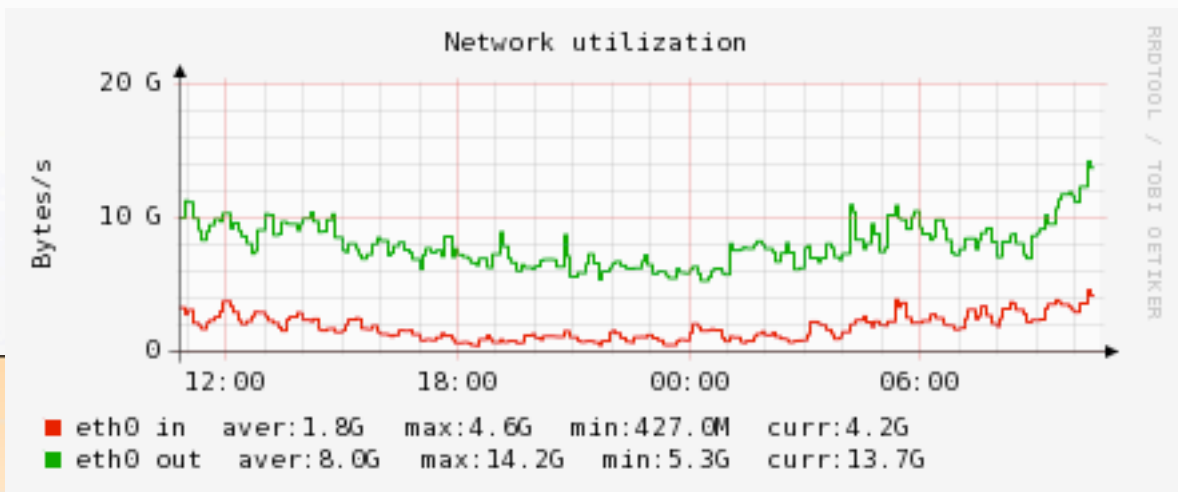
Data leaving CERN

We need to keep the flux going in the pipes.
Currently not testing the system at its maximum (still low luminosity).



Experiment	Size
ATLAS	120TB
CMS	27TB
LHCb	18TB





Need to access data for analysis

The user's analyses run where data are located (paradigm of the Tiered system)

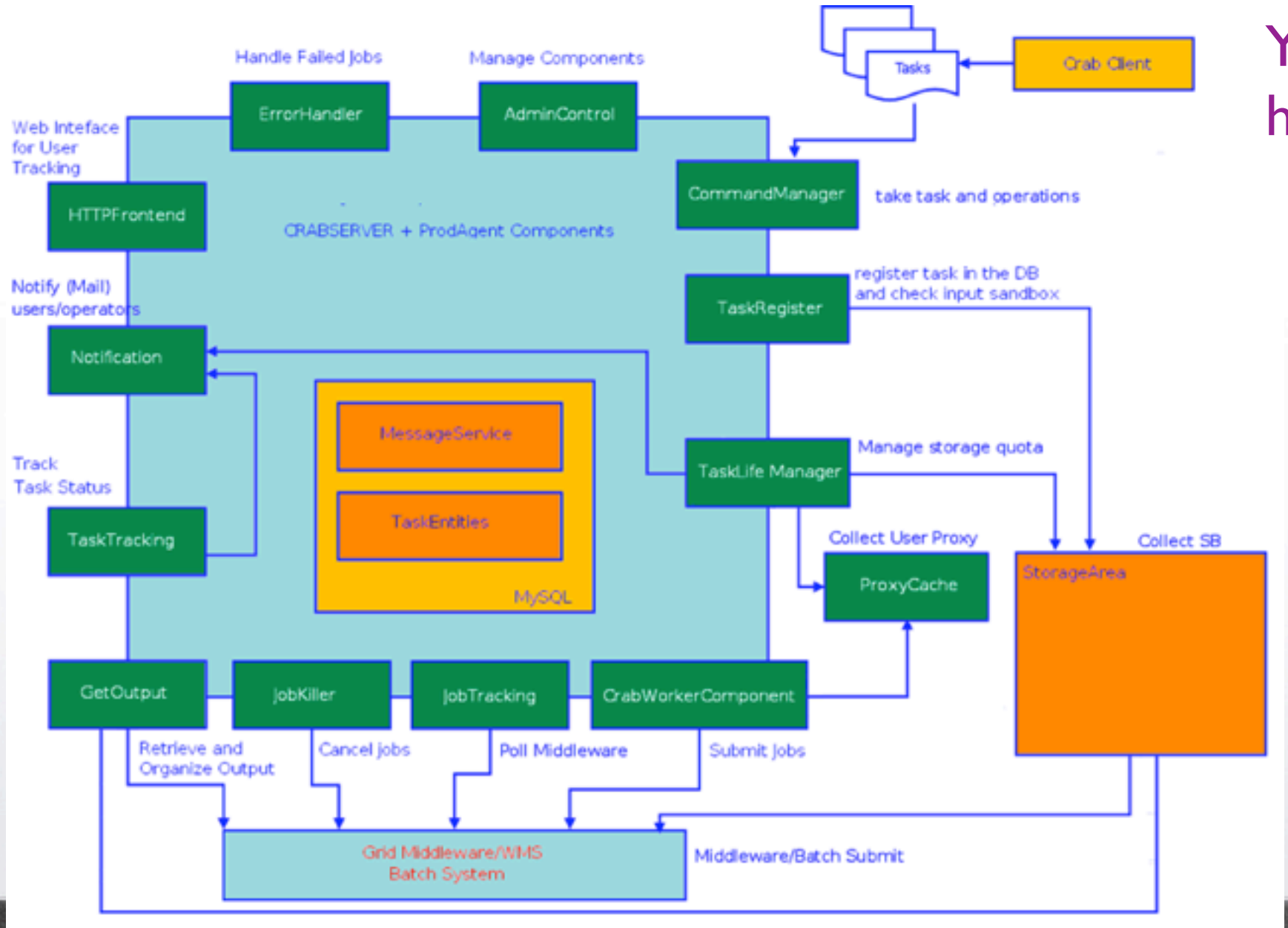
- User runs interactively on small data sample developing the analysis code.
- User selects large data sample to run the very same code.
- User's analysis code is shipped to the site where sample is located.
- Results are made available to the user for the final plot production.

Need to access data for analysis

There are three essential components in the analysis path:

1. Discover where the events are (they might be even "shared" in multiple sites). To give an idea of the scale of the problem:currently we have several hundreds M events, several hundreds datasets, several hundreds K files (only from data)
2. Submit jobs using Grid Middleware technology on all sites where data are sitting. Tracking of all jobs and transient local storage must be guaranteed.
3. Collect output from different places, possibly merge them, and bring them back to the submitters
4. Possibly make this output available to other users re-injecting them in the loop.

Need to access data for analysis

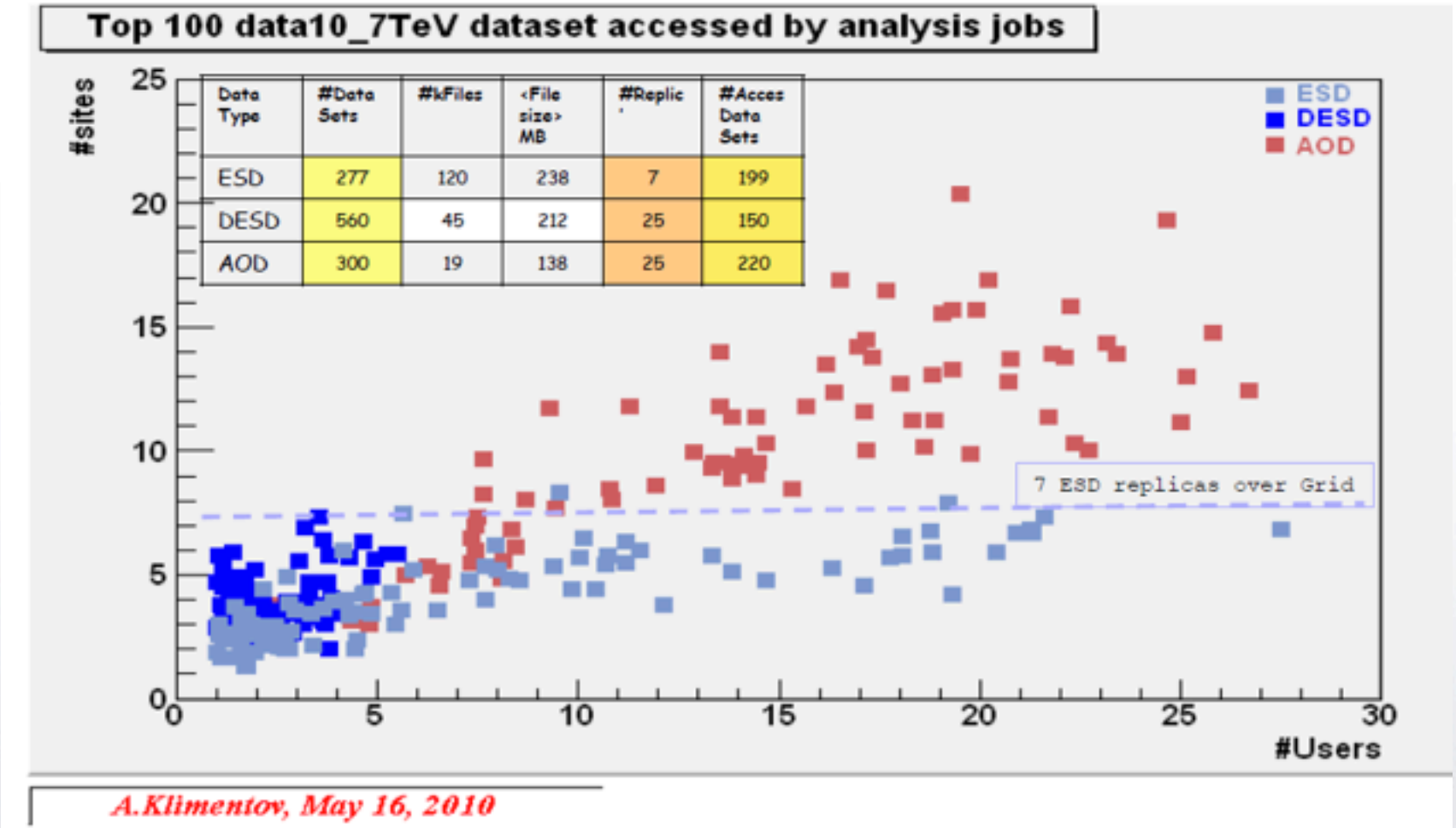


You are here

and you get your root tree there...

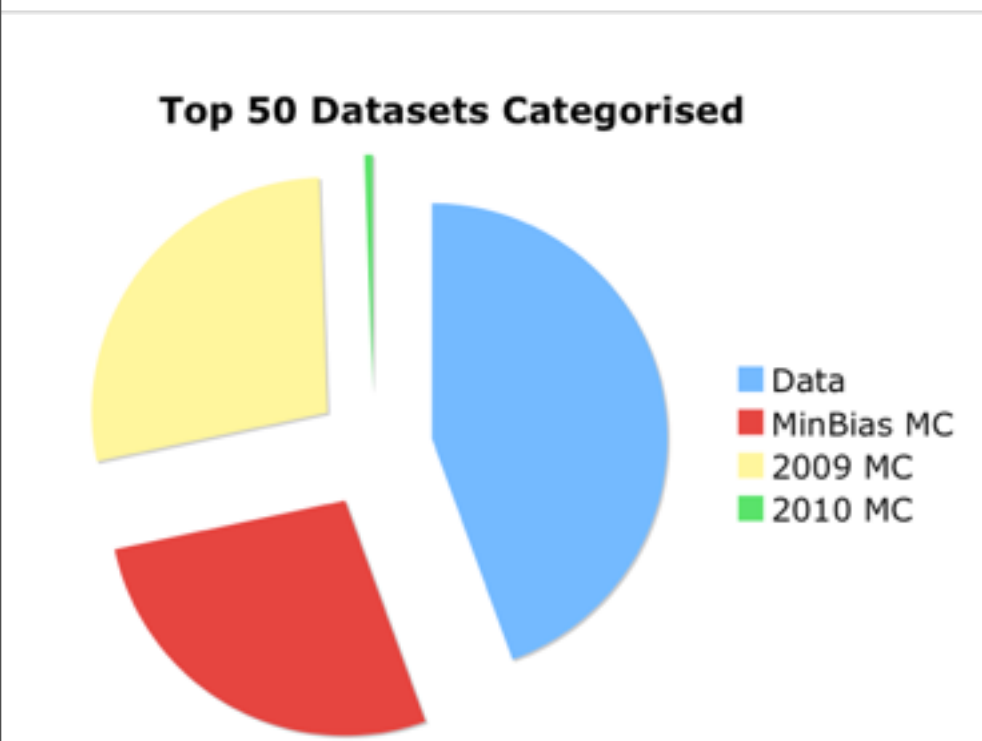
Need to access data for analysis

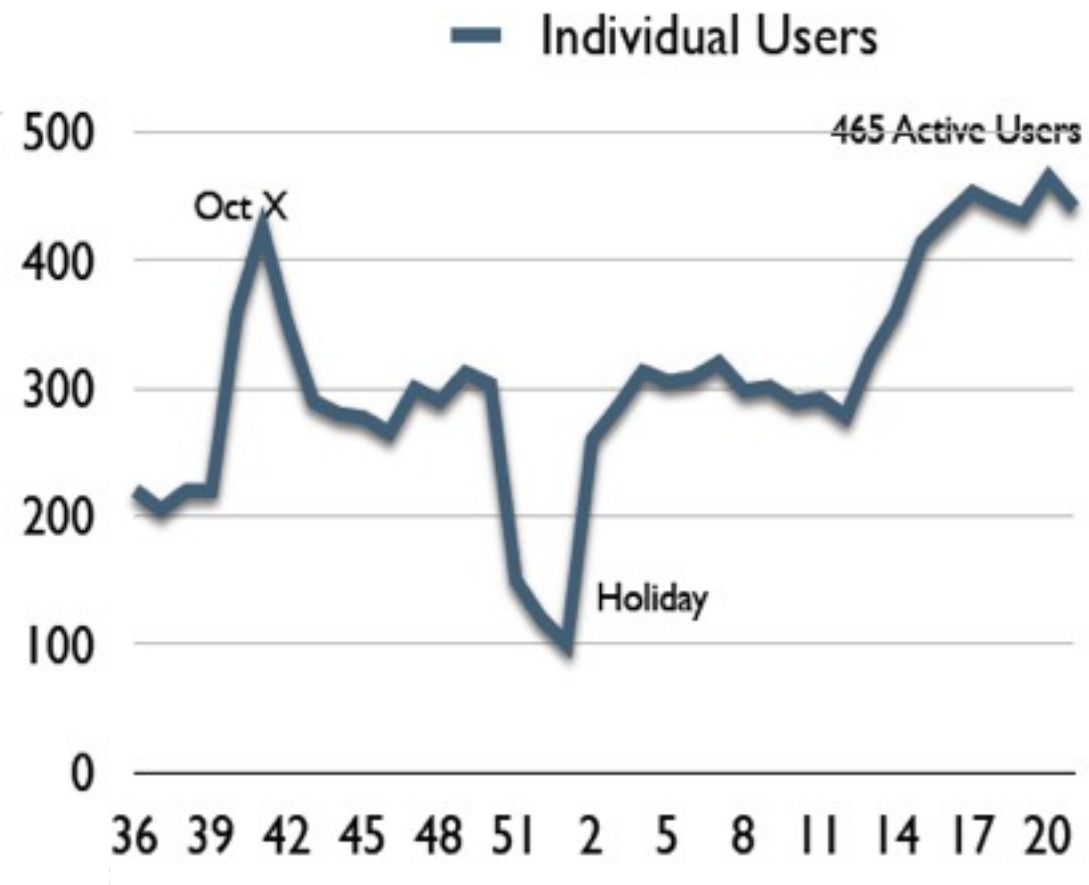
Pattern of recent data access in the ATLAS experiment



Need to access data for analysis

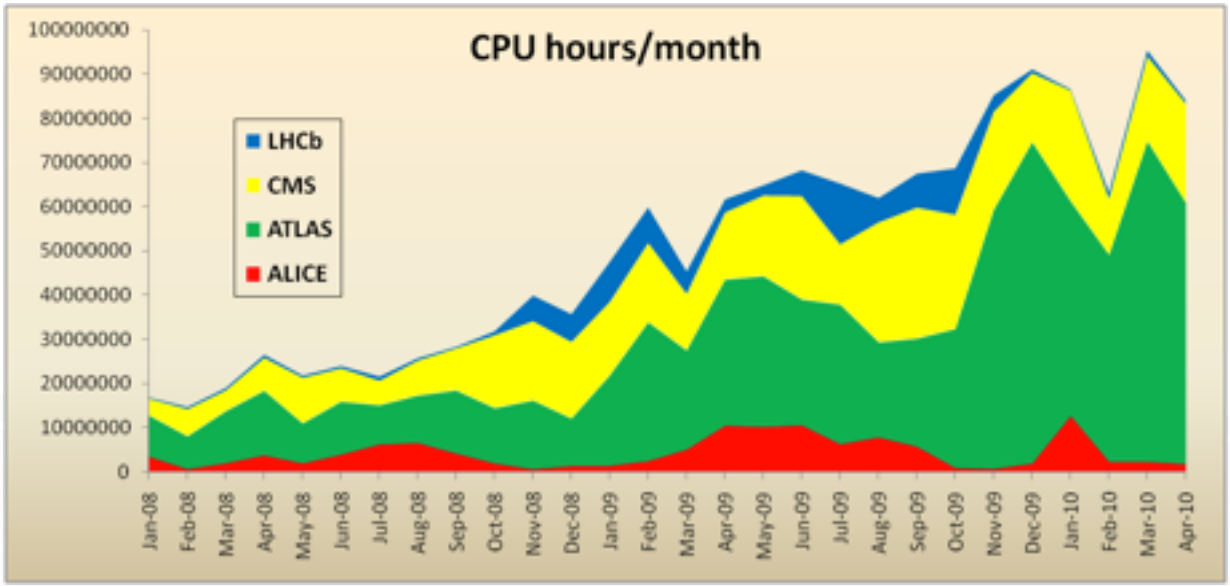
.....similarly in CMS

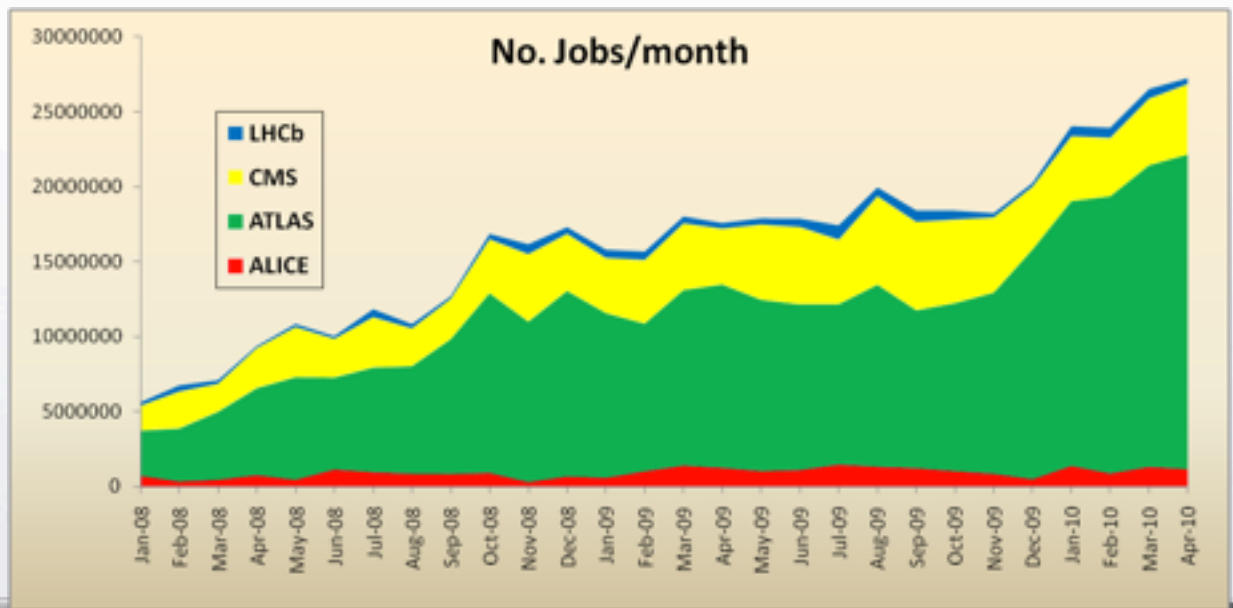




The grid is huge

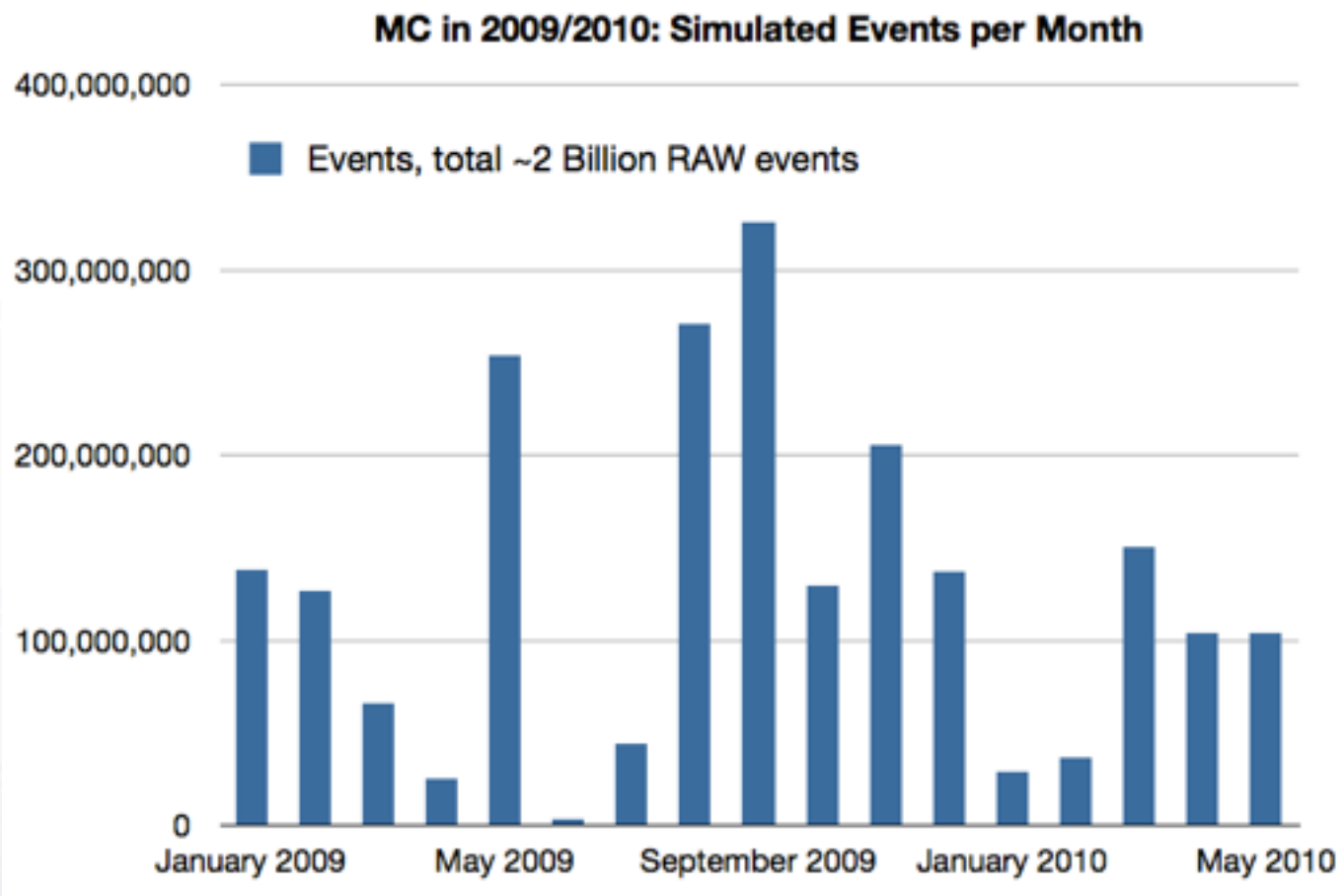
The amount of computing and storage needs are such that there are not many options. Central running (everybody running at CERN on data stored locally) is impossible.





MC production

The same structure/ workflow used also for MC production.
Jobs running in parallel with data reprocessing cycles, usually following the data reprocessing (if you change code, you'd like a MC following it)
Need to optimize and prioritize.



The day-by-day life
from the offline side

- Revisit the trigger menu and adapt to physics/commissioning needs
- Revisit reconstruction code:
  - improve detector noise cleaning
  - apply bug fixes/patches
- Discuss and test new calibrations before applying to prompt reco (see DQM section)
- Check automatic and manual reports from quality monitoring processes (see DQM section)
- Trigger a new full reprocessing when new code/calibrations have reached a critical mass
- Monitor the reprocessing and "certify" it for physics analyses
- Monitor the MC production to see if it fits the physics analysis needs.

The day-by-day life
from the offline side

- Revisit the trigger menu and adapt to physics/commissioning needs
- Revisit reconstruction code:
  - improve detector noise cleaning
  - apply bug fixes/patches
- Discuss and test new calibrations before applying to prompt reco (see DQM section)
- Check automatic and manual reports from quality monitoring processes (see DQM section)
- Trigger a new full reprocessing when new code/calibrations have reached a critical mass
- Monitor the reprocessing and "certify" it for physics analyses
- Monitor the MC production to see if it fits the physics analysis needs.

Validation flow

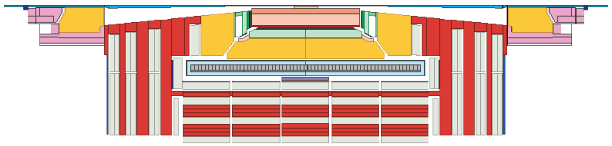
Validation of data is an important component of the "road to physics"

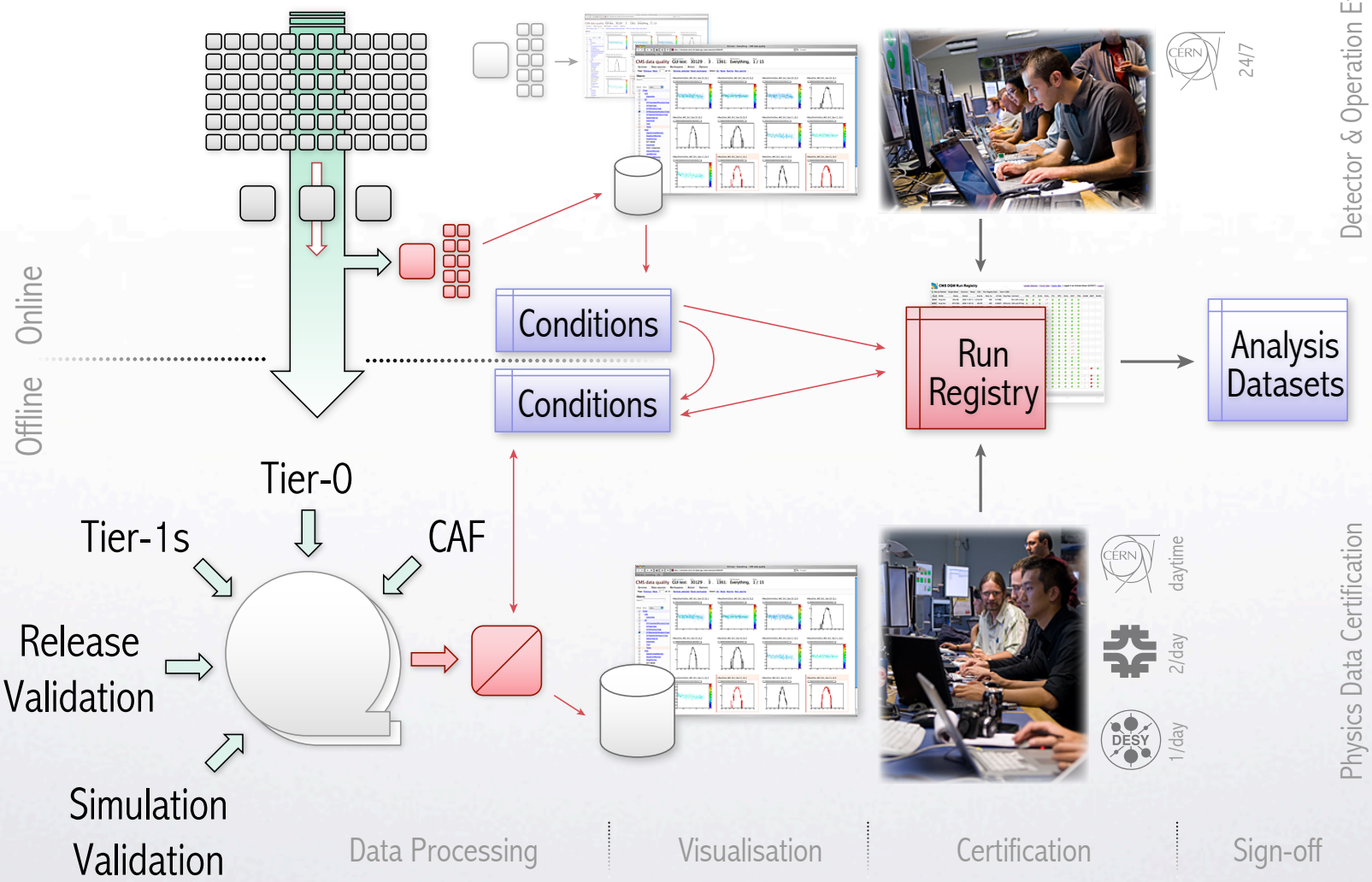
Data might be "invalid" for several reasons:

- detector faults
- wrong configuration/calibrations
- induced noise (physics induced noise or electronic ones)
- data transfer corruption/losses
- mistake or unexpected "features" in the reprocessing steps
- bug in the analysis code

all but the last should be (are) centrally monitored.

Validation flow





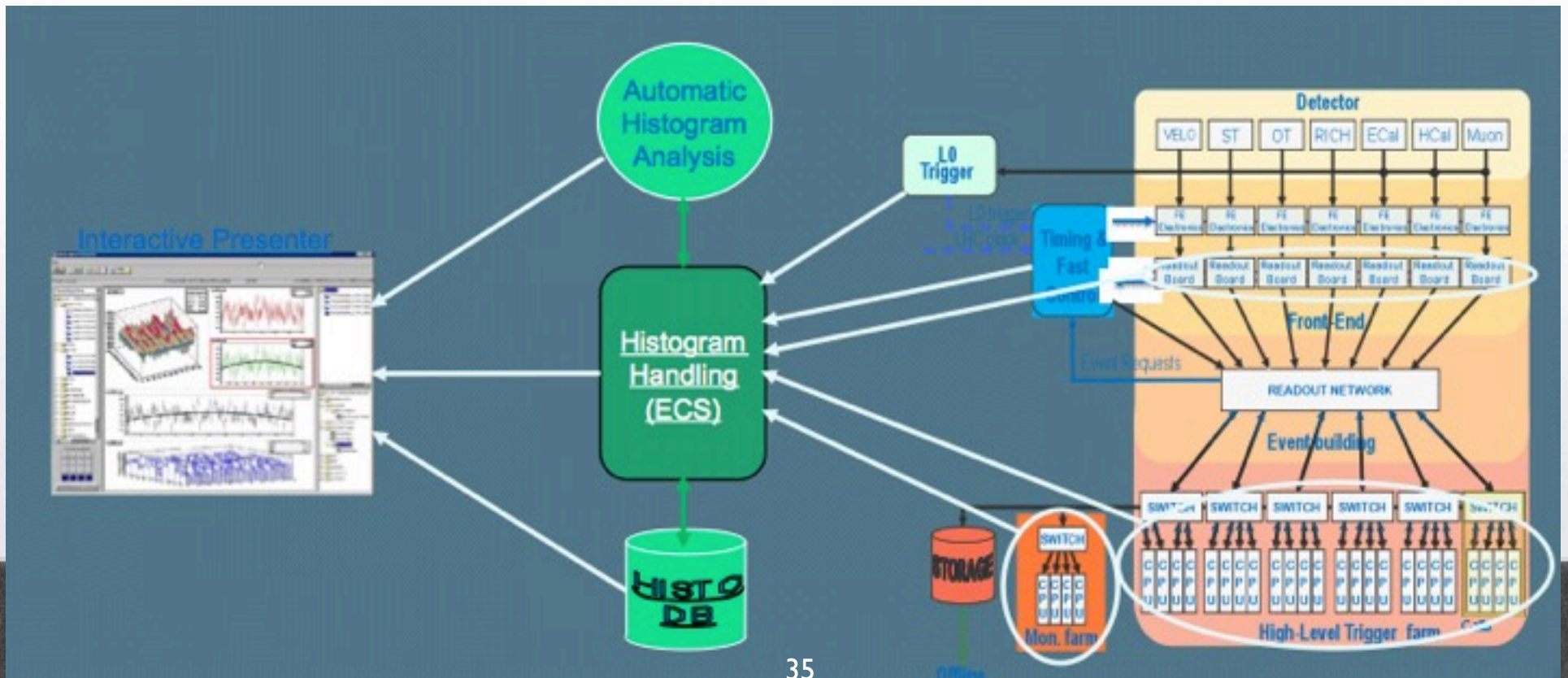
DQM online

Aim is *efficient detector and operation* by giving detector and trigger status feedback to experts and shifters.

- Live display at $\Delta t \sim$ seconds
- Several thousand of histograms produced on DQM cluster,
- Need a proper GUI allowing multiple connection to be attached to it.

Online results, initial run summary made available to offline analysis and processing.

- Includes online detector quality summary and other key values in conditions database.



DQM online

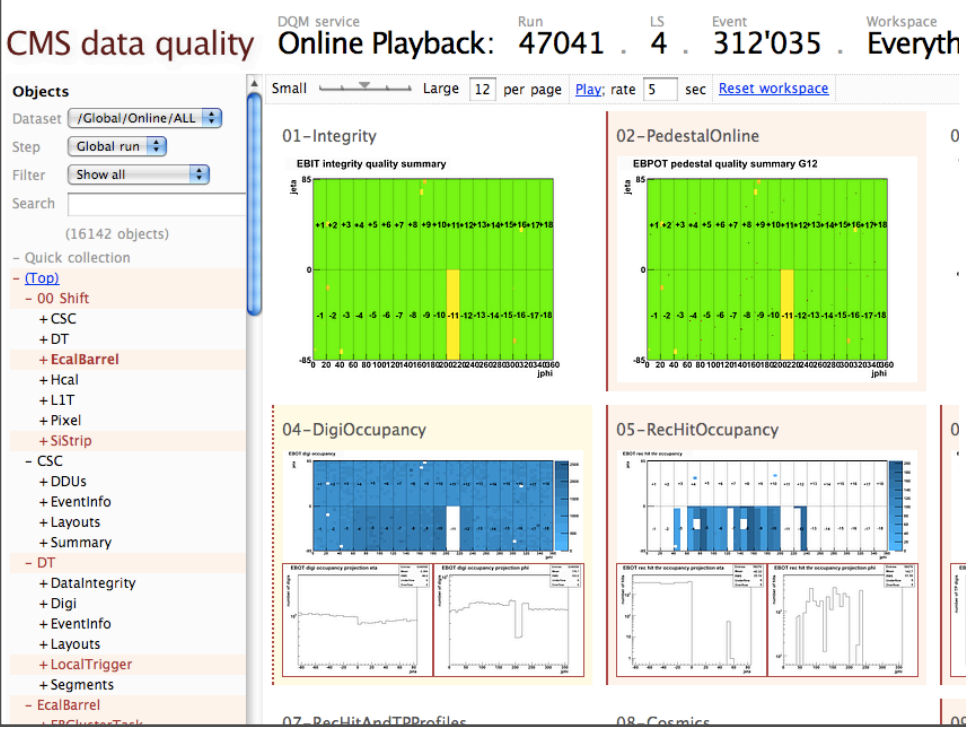
Operationally:

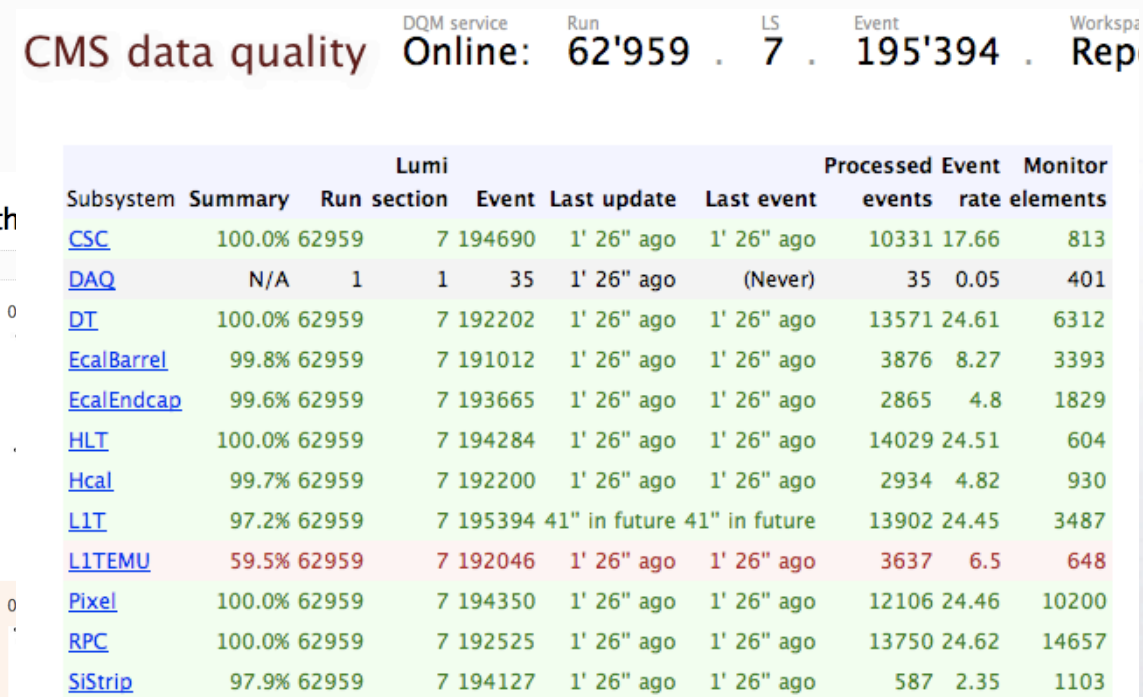
histograms of sensitive quantities are filled in real time

(detector status, energy distributions, track distributions, correlations, trigger rates, etc. etc.)

Checks are done w.r.t. reference histograms or quality criteria are applied

- output are reduced to flags (red/green)
- shifters are requested to look at flags and react:
  - inspect histograms
  - call experts
  - ask to stop the run and fix





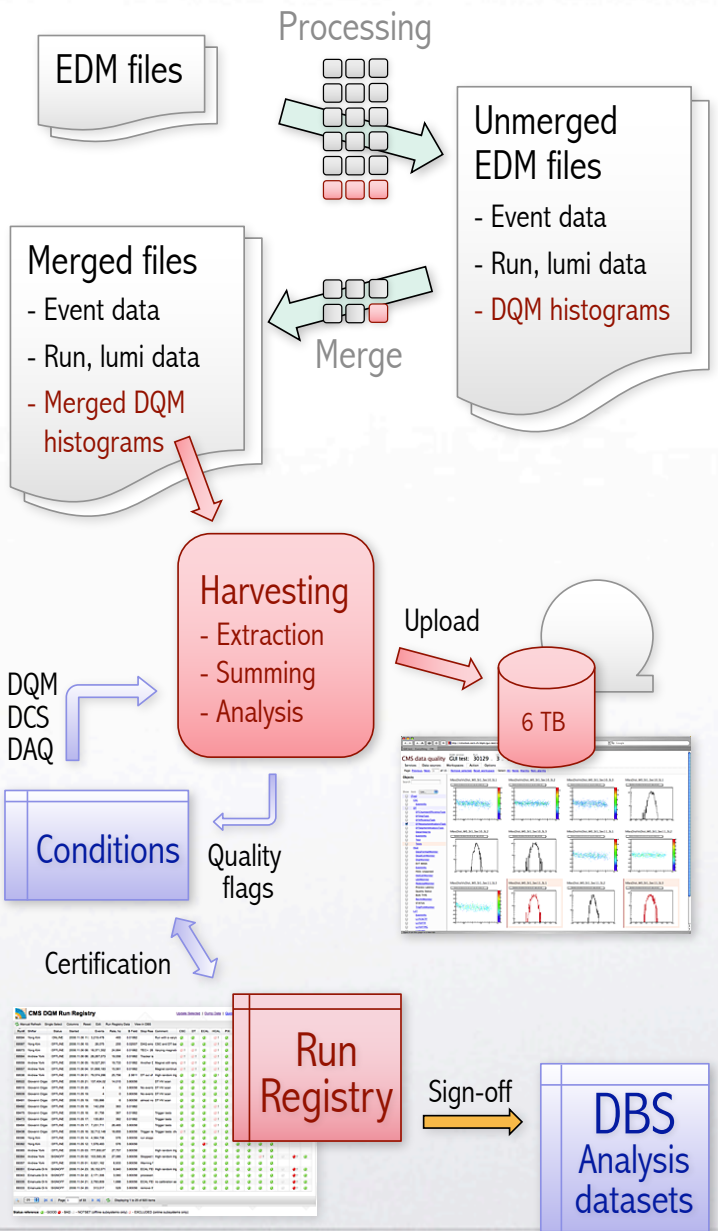
Subsystem	Summary	Lumi Run	Lumi section	Event	Last update	Last event	Processed events	Event rate	Monitor elements
CSC	100.0%	62959	7	194690	1' 26" ago	1' 26" ago	10331	17.66	813
DAQ	N/A	1	1	35	1' 26" ago	(Never)	35	0.05	401
DT	100.0%	62959	7	192202	1' 26" ago	1' 26" ago	13571	24.61	6312
EcalBarrel	99.8%	62959	7	191012	1' 26" ago	1' 26" ago	3876	8.27	3393
EcalEndcap	99.6%	62959	7	193665	1' 26" ago	1' 26" ago	2865	4.8	1829
HLT	100.0%	62959	7	194284	1' 26" ago	1' 26" ago	14029	24.51	604
Hcal	99.7%	62959	7	192200	1' 26" ago	1' 26" ago	2934	4.82	930
L1T	97.2%	62959	7	195394	41" in future	41" in future	13902	24.45	3487
L1TEMU	59.5%	62959	7	192046	1' 26" ago	1' 26" ago	3637	6.5	648
Pixel	100.0%	62959	7	194350	1' 26" ago	1' 26" ago	12106	24.46	10200
RPC	100.0%	62959	7	192525	1' 26" ago	1' 26" ago	13750	24.62	14657
SiStrip	97.9%	62959	7	194127	1' 26" ago	1' 26" ago	587	2.35	1103

DQM offline

Usually the same workflow as in online is used, but with different timing and scale.

All "data quality" histograms are re-produced (new conditions/different codes) and are checked automatically or manually to existing references.

The results of the certification is made persistent in a dedicated database and used in the analysis to select "good data" periods. The atomic unit might be an entire run or a subset of it (fixed time interval)

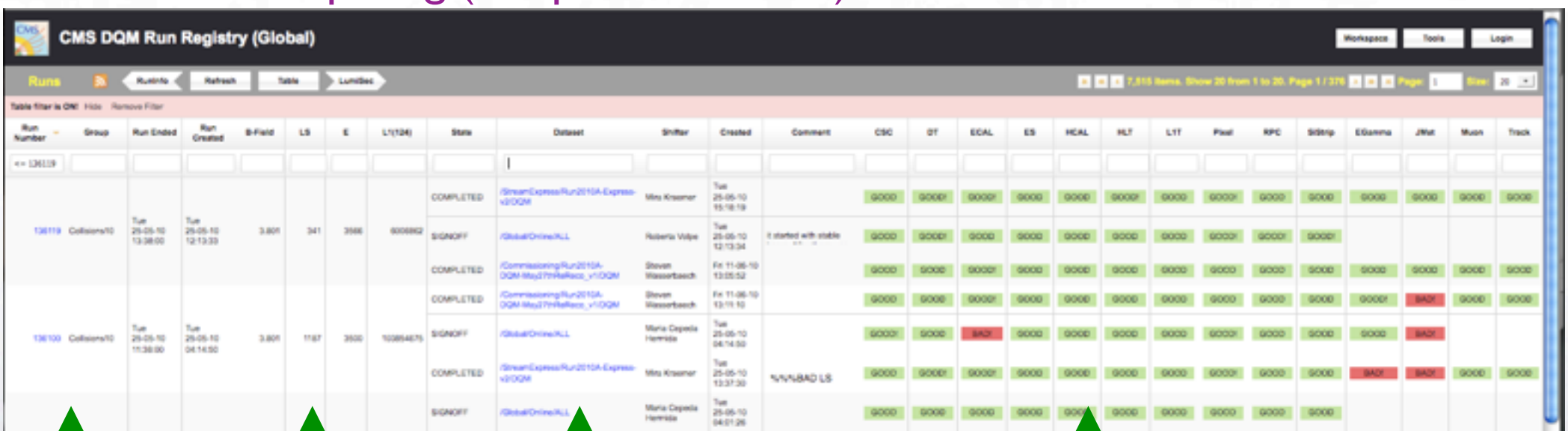


DQM offline

The certification must be redone at all reprocessing steps and the analysis chain should be in sync.

The bookkeeping of the whole process is essential and must as automatic as possible.

Technically it must be coupled with all tools used for physics analysis and distributed computing (see previous slides).



Run

Time

Dataset

Reduced det status

DQM operations

A typical quality monitoring workflow before data are "blessed" for physics analysis:

- automatic online DQM flags runs/lumi-block
- online DQM shifters monitor (and react in case)
- daily meeting to discuss about quality of data taken the day before
- offline shifter (can be remote) have a detailed look at histograms and classify runs
- weekly the runs (and the lumi-block) are "certified" for physics analyses
- the information is stored for future reference

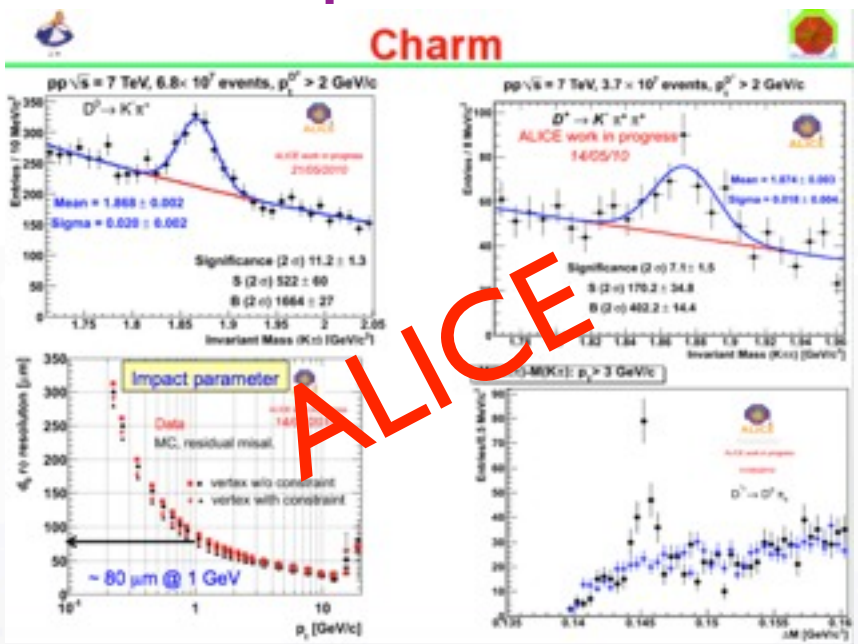
* any reprocessing step re-triggers the validation workflow

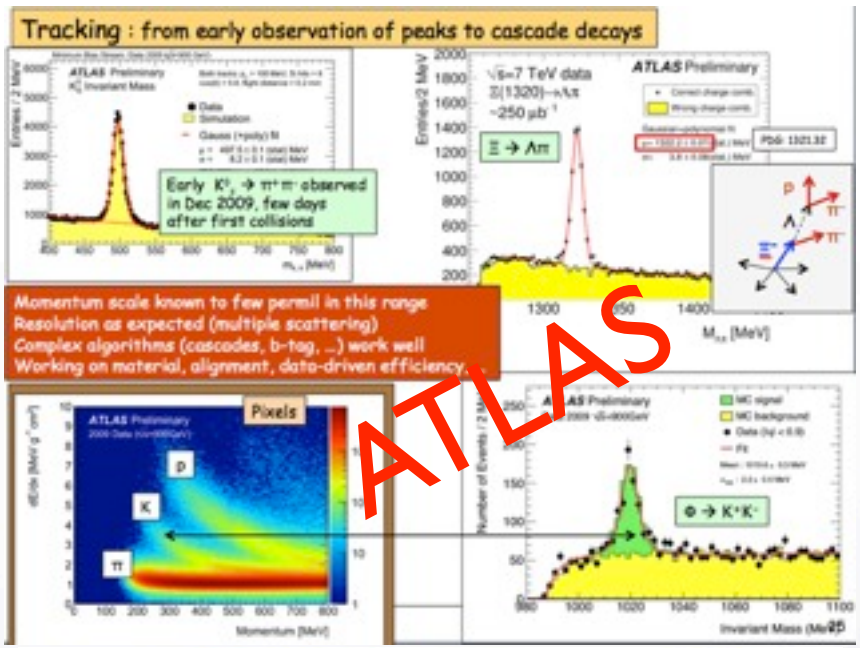
DQM operations

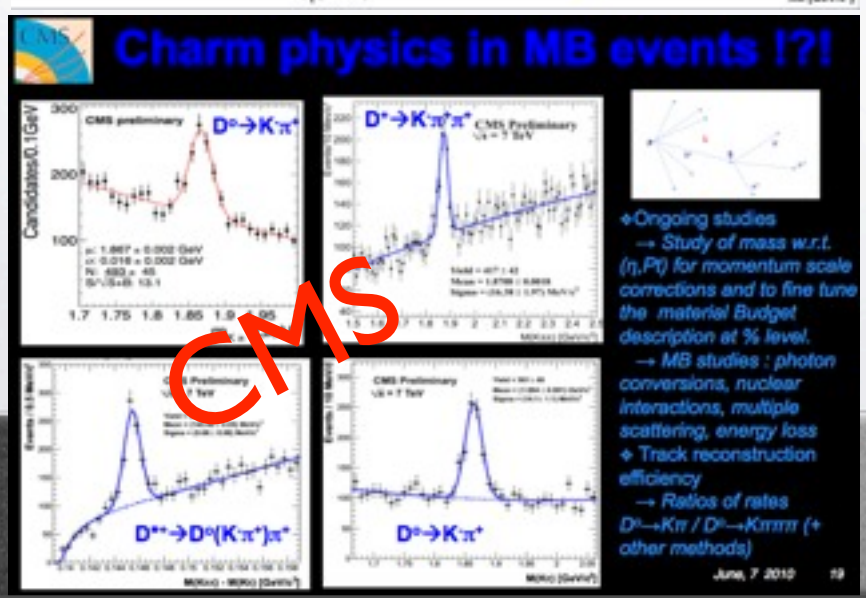
A control room and its quality monitoring shifter place

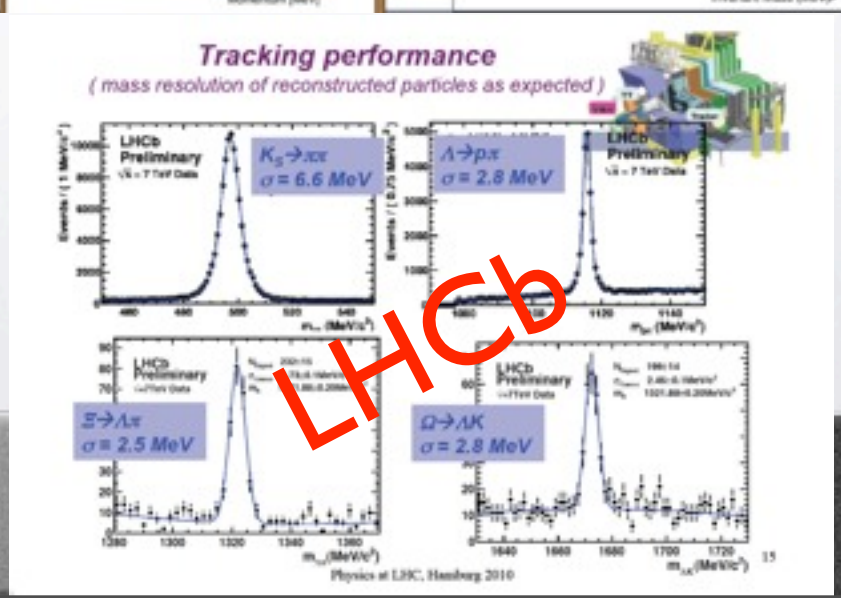
All this is essential

None of the (first) physics results would have been possible without all steps seen so far









Summary

The only way to cope with the amount of data of the LHC experiment is to parallelize:

- data flow
- data storage
- data access

Reprocessing the data will be an expensive task: need to have as much as possible "prompt/express" processing:

- calibration
- validation
- selection (skimming)

The data validation (data quality monitoring) is essential:

- for quick online feedback
- for blessing data for physics analysis

Never forget the scaling and complexity of the LHC experiment.
Investing in the above steps will pay off.