LHCC Referee Meeting

March 3, 2015

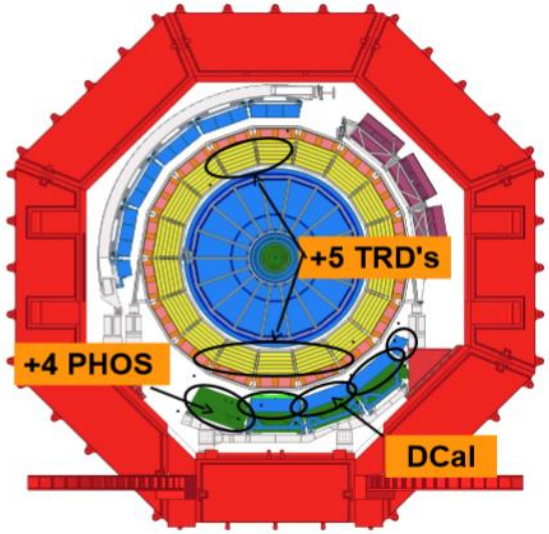
ALICE Status Report

Predrag Buncic

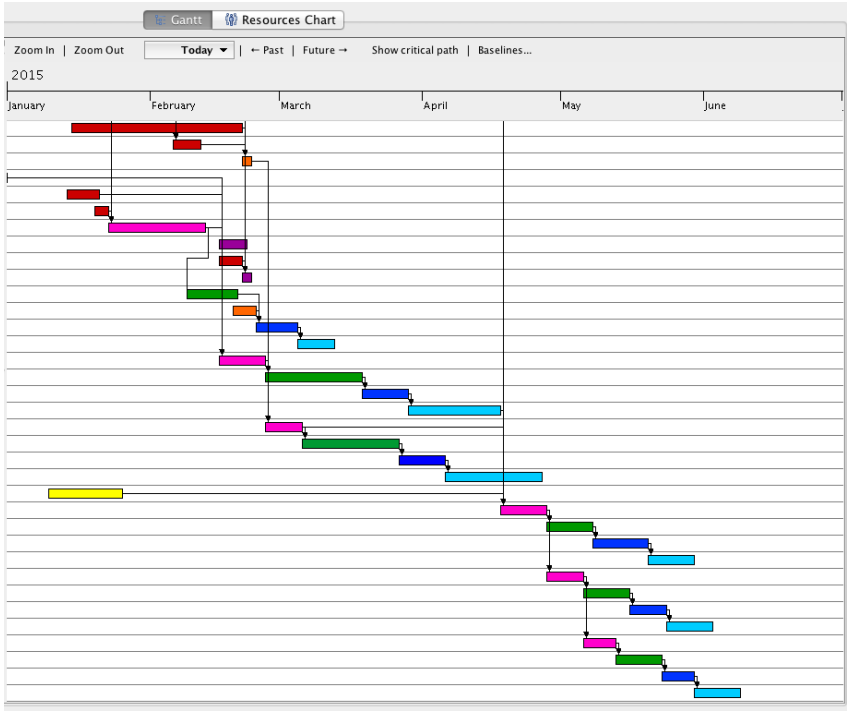
CERN

Preparations for Run2

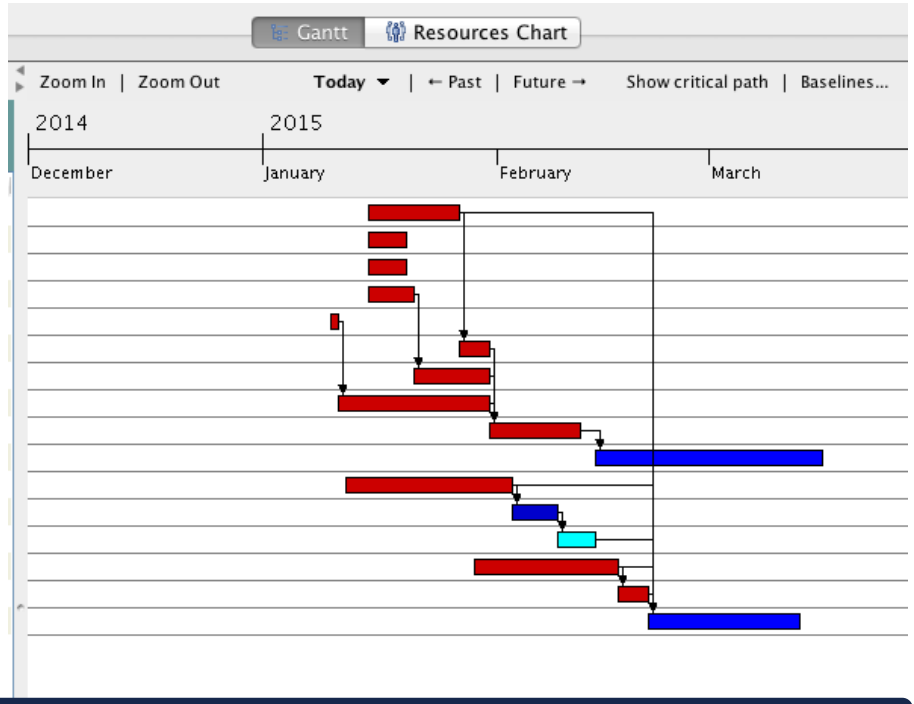
- Expecting increased event size
  - 25% larger raw event size due to the additional detectors
  - Higher track multiplicity with increased beam energy and event pileup
- Concentrated effort to improve performance of ALICE reconstruction software
  - Improved TPC-TRD alignment
  - TRD points used in track fit in order to improve momentum resolution for high p_T tracks
  - Streamlined calibration procedure
  - Reduced memory requirements during reconstruction and calibration (~500Mb, the resident memory is below 1.6GB and the virtual - below 2.4 GB)



Re-processing & re-commissioning

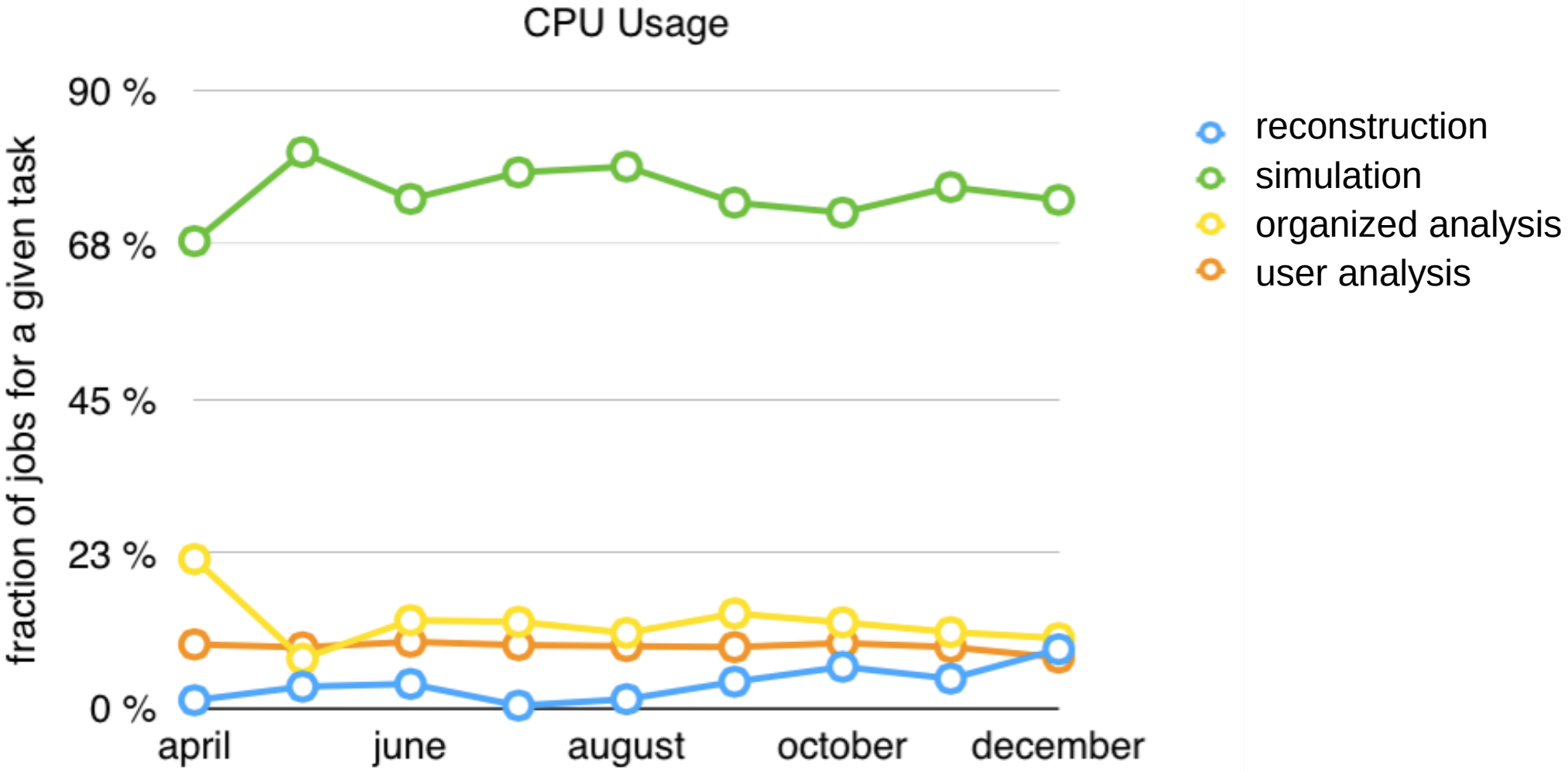


- Re-processing (Run 1)
  - Steady RAW and MC activities
  - Full detector re-calibration and 2 years worth of software updates
  - All Run 1 RAW data processing with the ***same*** software

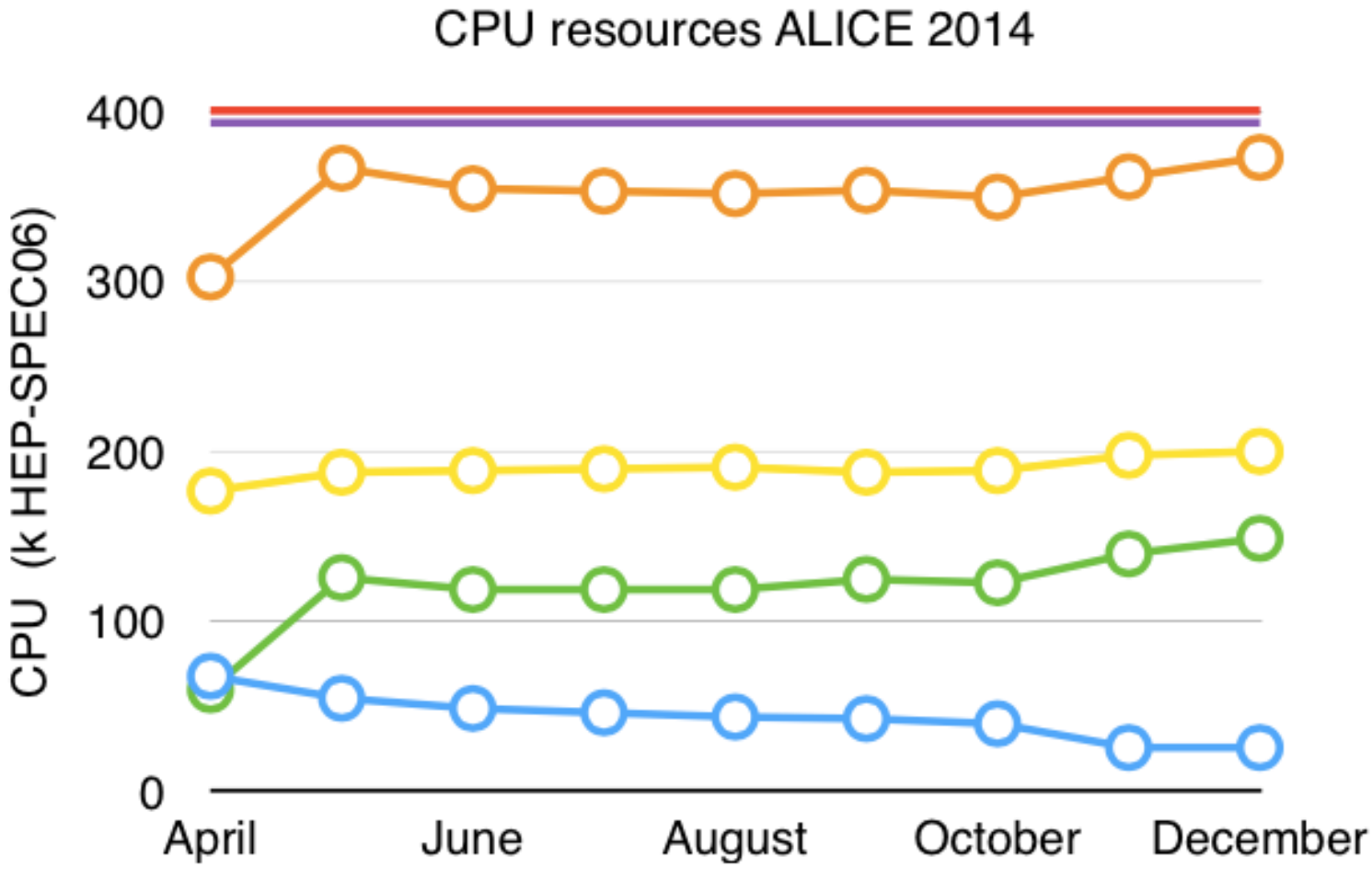


- Re-commissioning (Run 2)
  - Test of upgraded detectors readout, Trigger, DAQ, new HLT farm
  - Full data recording chain, with conditions data gathering
  - Cosmics trigger data taking with Offline processing

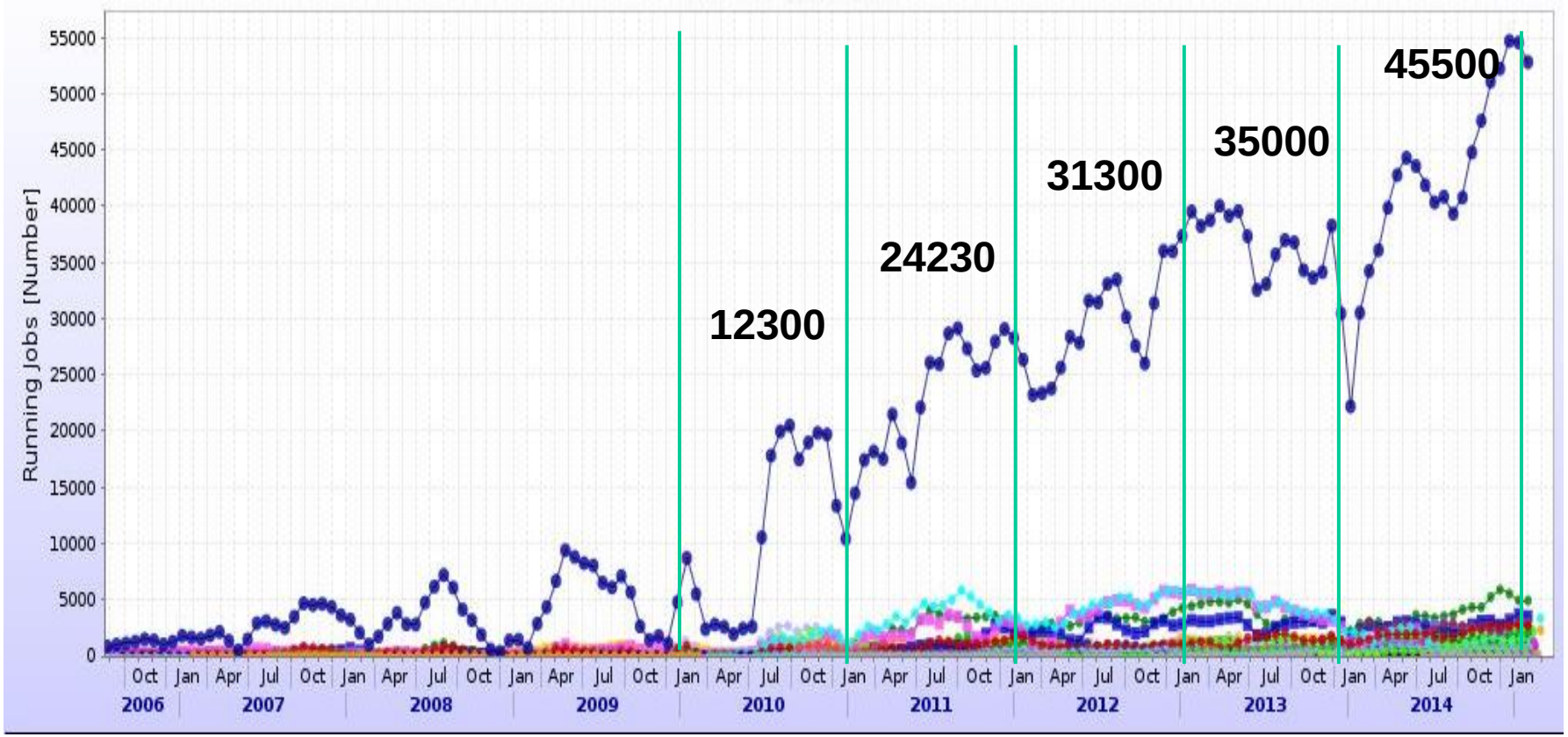
CPU Shares


CPU Usage
fraction of jobs for a given task
90 %
68 %
45 %
23 %
0 %
april
june
august
october
december
reconstruction
simulation
organized analysis
user analysis

CPU Resources in 2014



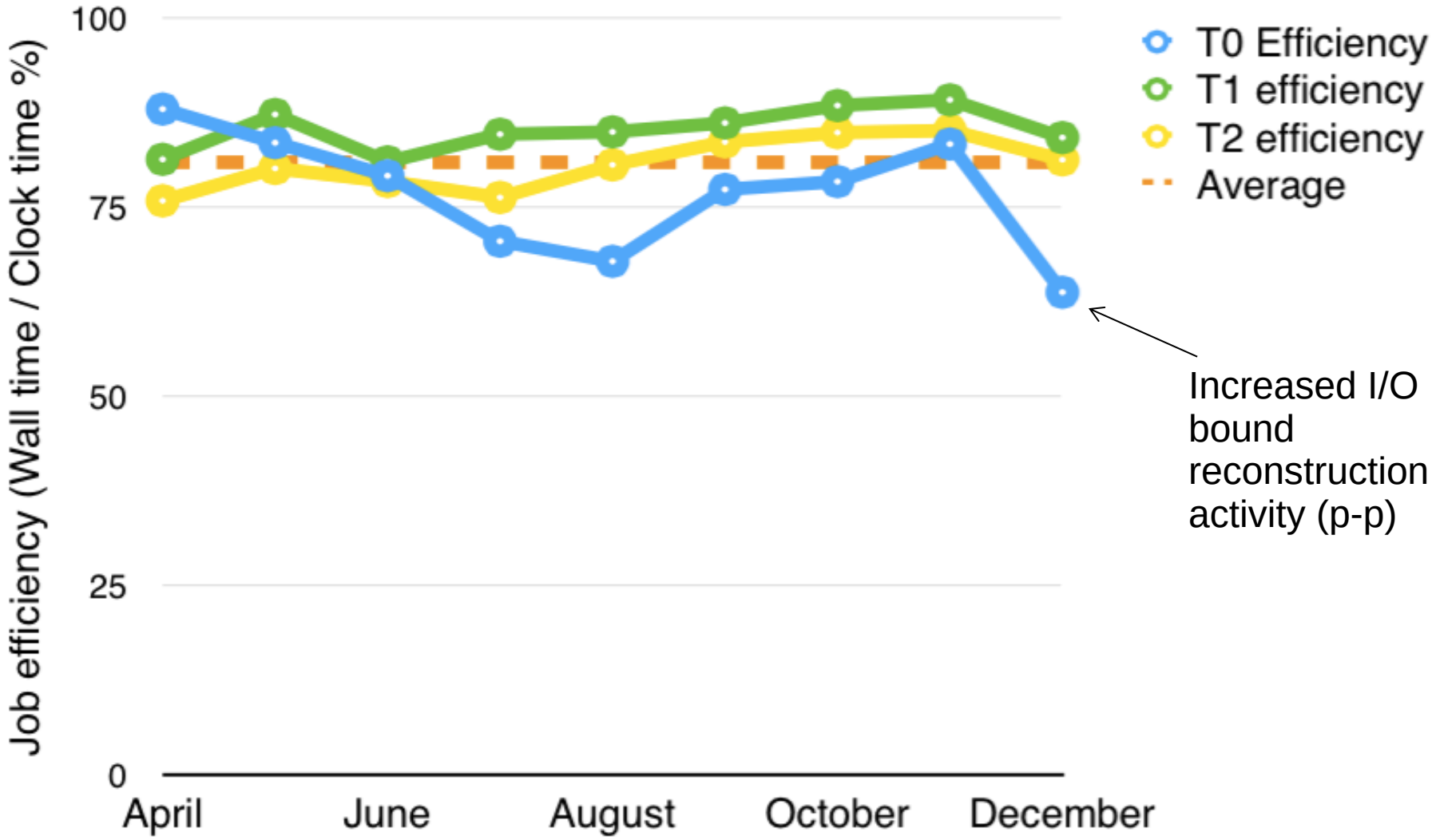
Available CPU evolution



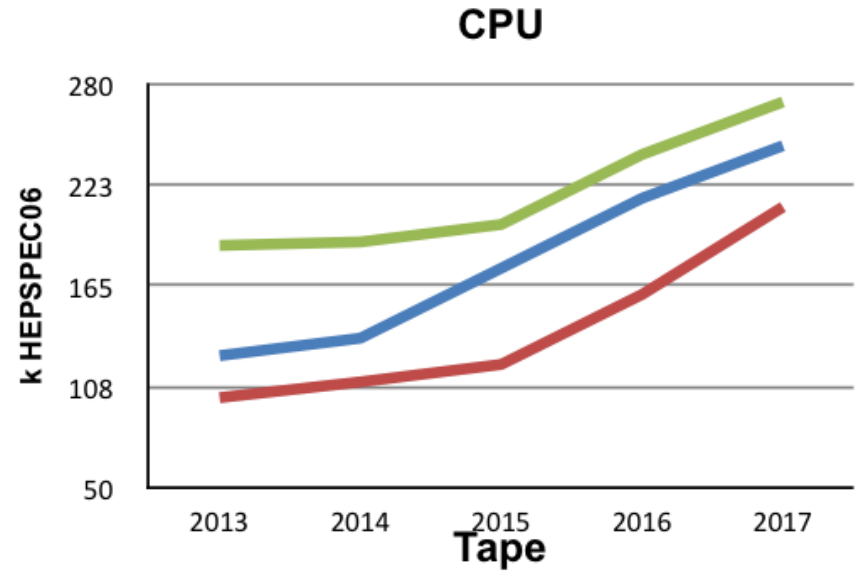
Year on year increase	+97 %	+30 %	+12 %	+30 %

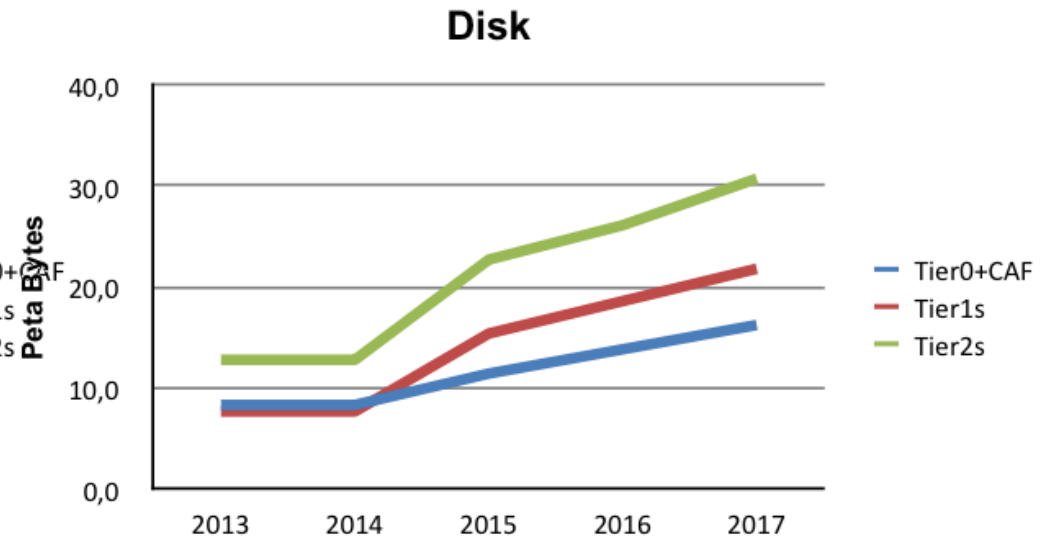
- 22% average per year – slightly above the WLCG projection

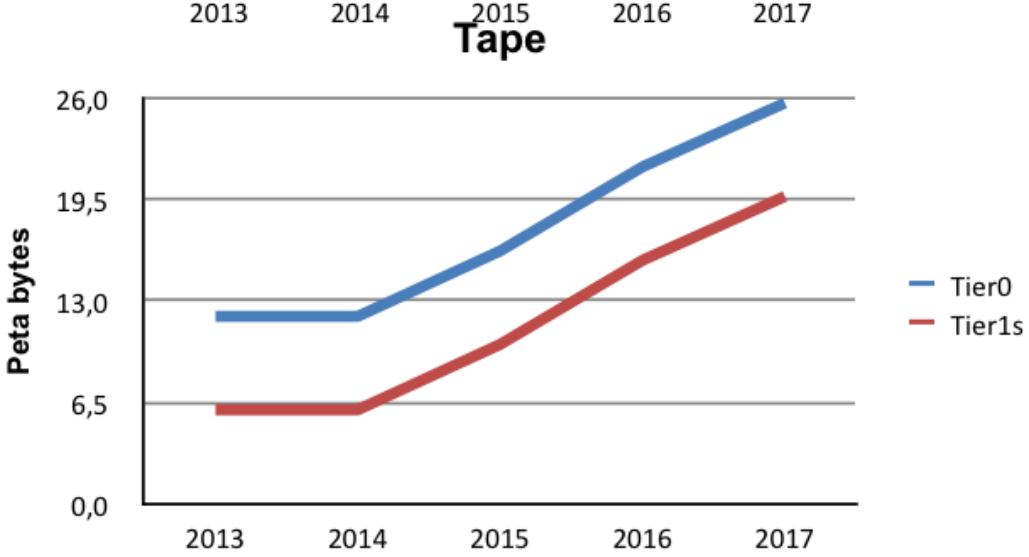
CPU Shares


100
75
50
25
0
Job efficiency (Wall time / Clock time %)
April
June
August
October
December
T0 Efficiency
T1 efficiency
T2 efficiency
Average
Increased I/O bound reconstruction activity (p-p)

Resource Request 2013-2017

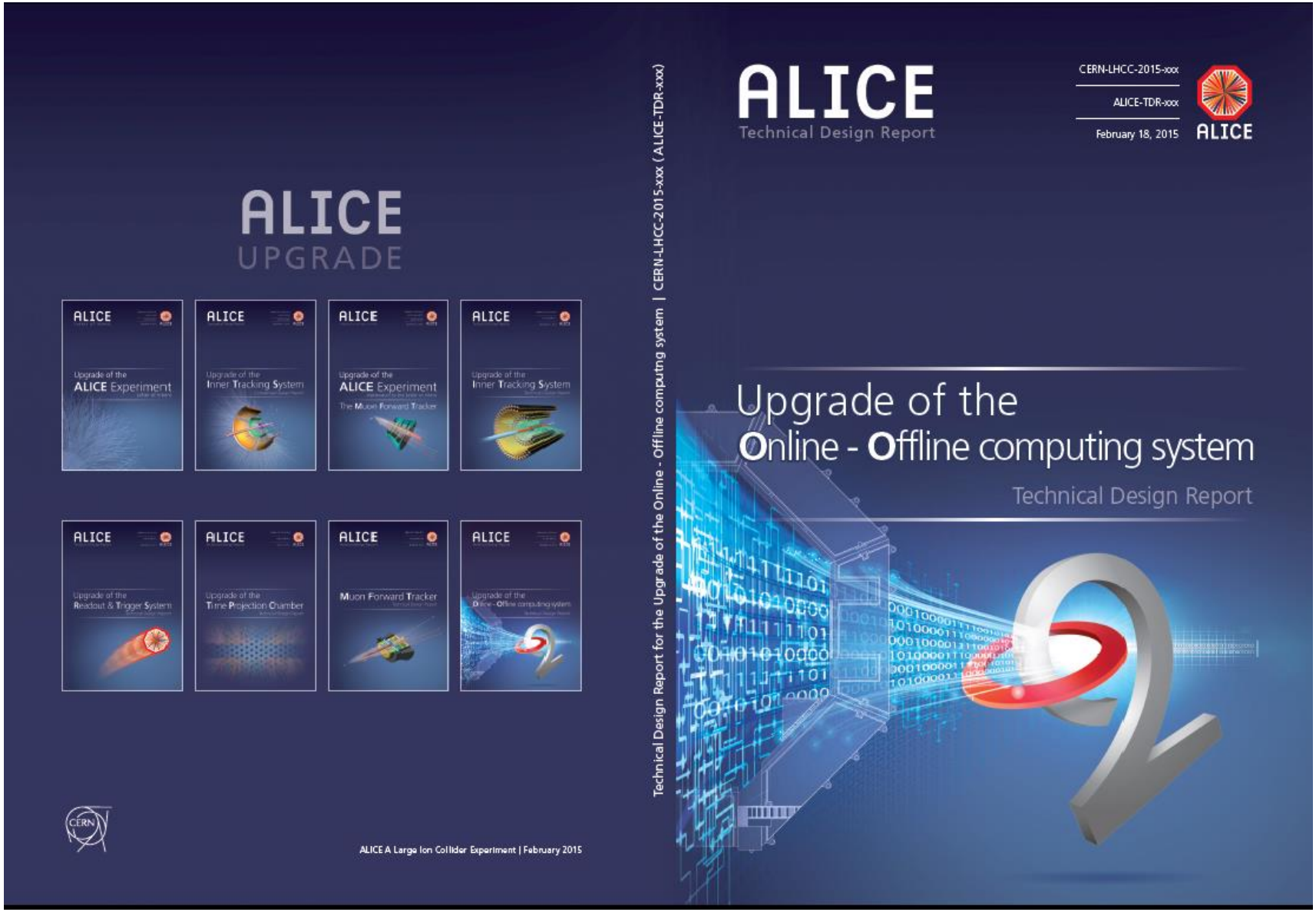






No significant changes compared to what we showed last time

O2 TDR


ALICE
UPGRADE
ALICE
Upgrade of the ALICE Experiment
ALICE
Upgrade of the Inner Tracking System
ALICE
Upgrade of the ALICE Experiment
The Muon Forward Tracker
ALICE
Upgrade of the Inner Tracking System
ALICE
Upgrade of the Readout & Trigger System
ALICE
Upgrade of the Time Projection Chamber
ALICE
Muon Forward Tracker
ALICE
Upgrade of the Online - Offline computing system
CERN
ALICE A Large Ion Collider Experiment | February 2015
Technical Design Report for the Upgrade of the Online - Offline computing system | CERN-LHCC-2015-xxx (ALICE-TDR-xxx)
ALICE
Technical Design Report
CERN-LHCC-2015-xxx
ALICE-TDR-xxx
February 18, 2015
ALICE
Upgrade of the
Online - Offline computing system
Technical Design Report

Basic premises

- Reduce data volume early and by a large factor
  - TPC x20, overall x14
  - Requires considerable computing capacity @ P2
- Use h/w accelerators to speed up the computation and reduce the cost
  - Requires new and flexible s/w framework that can adapt itself to different environments
- Assume that Grid capacity will continue to grow within constraints of a flat budget
  - Available resources must be used with maximum efficiency
  - 2/3 of raw data processing will happen @ P2
  - All intermediate data formats will be transient or temporary
  - Only the compressed raw data and analysis level data will be persistent

The ALICE O2 Project: Architecture

Detector

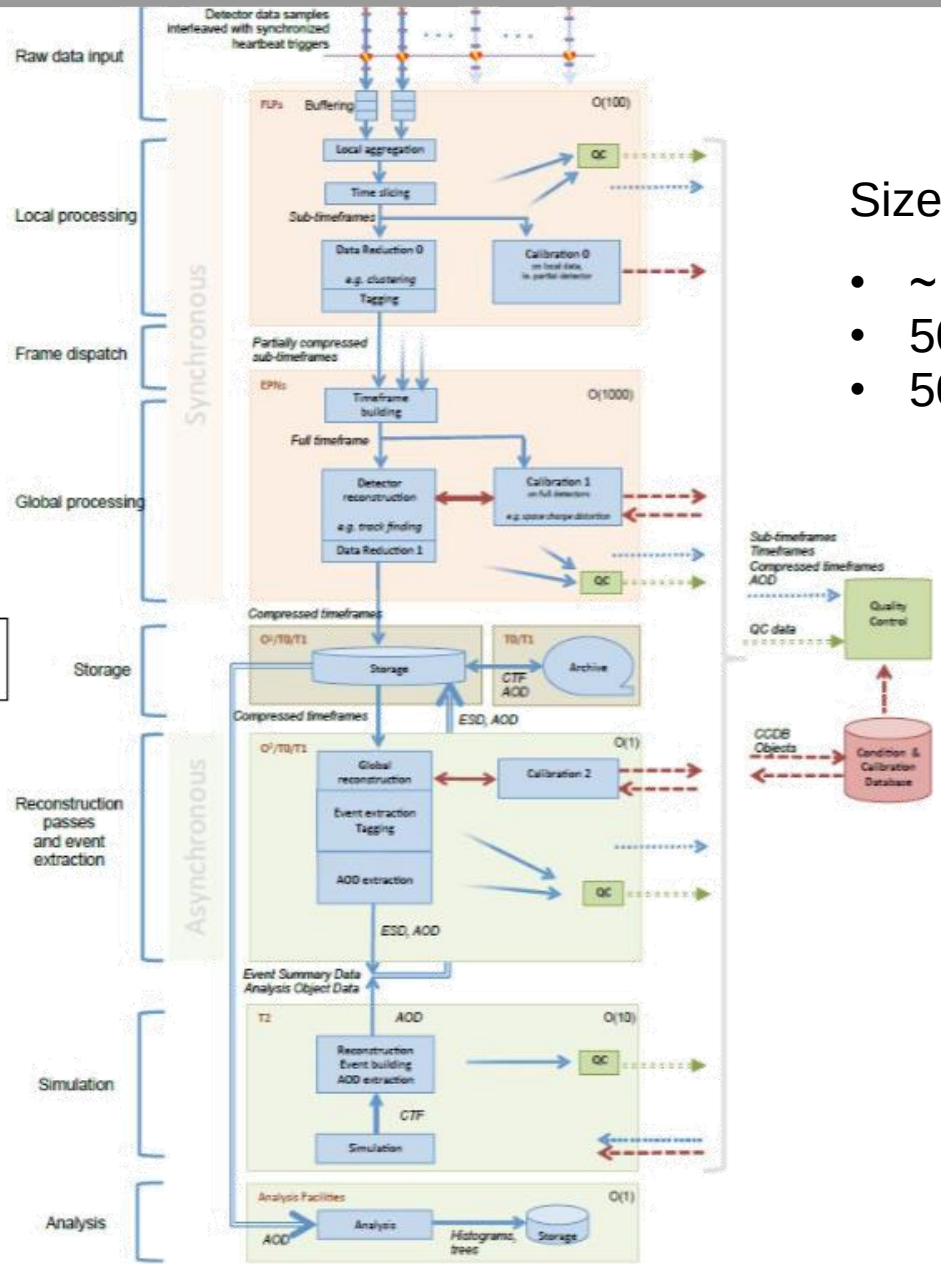
Synchronous readout, aggregation, compression
- FLP 2.5x
- EPN 8x

Storage

Asynchronous reconstruction
75 % EPN / 25 % T0-T1

Simulation

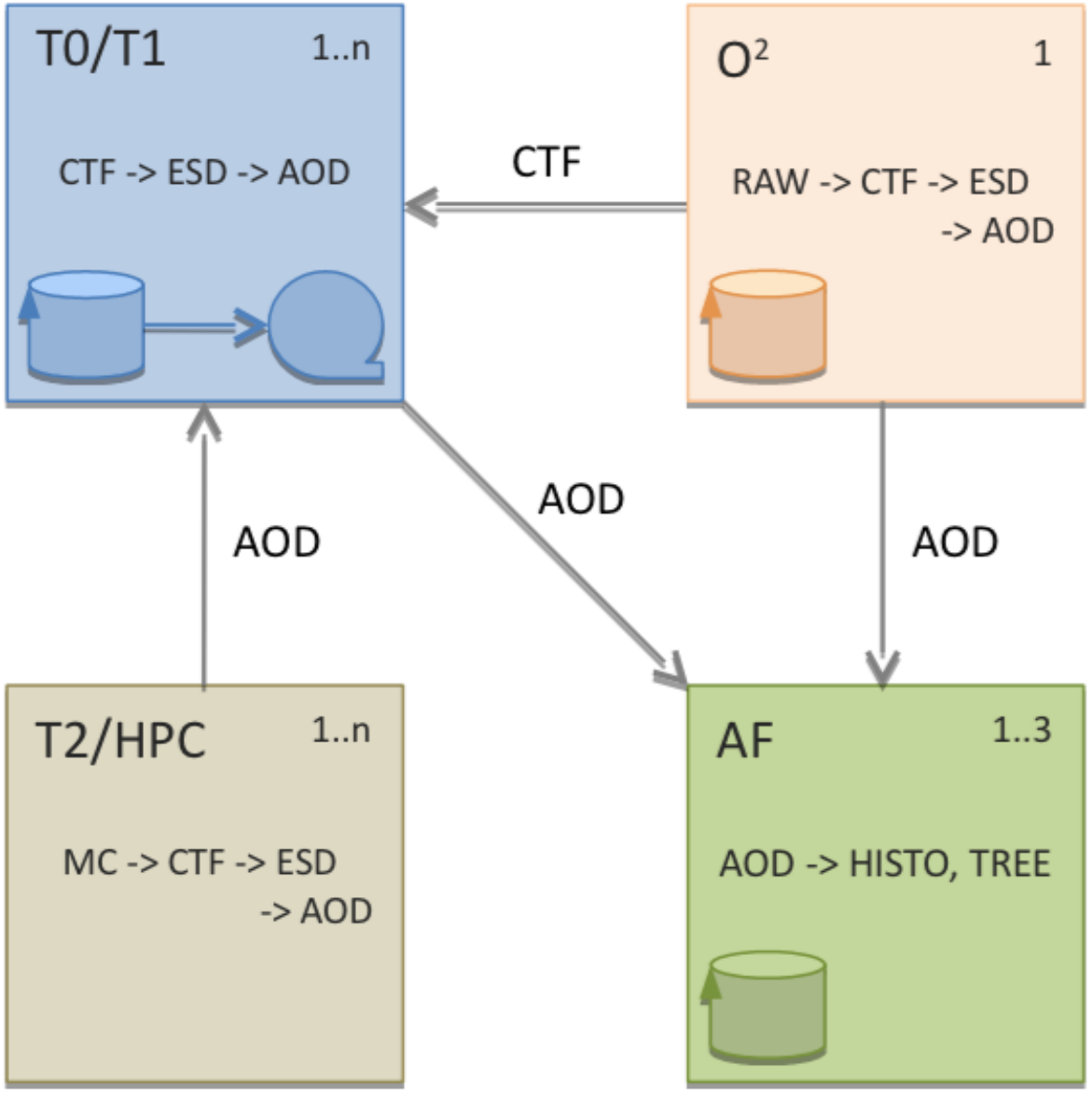
Analysis



Size of O2 facility:

- ~100’000 CPU cores
- 5000 GPUs
- 50 PB of disk

Roles of Tiers



Analysis Facilities

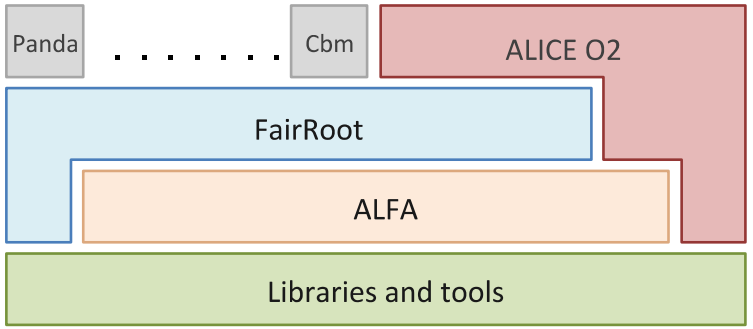
- Motivation
  - Analysis is the least efficient of all workloads that we run on the Grid
  - I/O bound in spite of attempts to make it more efficient by using the analysis trains
  - Increased data volume will only magnify the problem
- Solution
  - Collect AODs on a few dedicated sites that are capable of locally processing quickly large data volume
  - Typically (a fraction of) HPC facility (20-30’000 cores) and 5-10 PB of disk on very performant file system
  - Run organized analysis on local data like we do today on the Grid

Impact on Grid T1/T2 sites

- ALICE expects grid resources to evolve and grow at 20% per year rate which is consistent with a flat funding
- We expect 20Gb/s share of network connectivity between CERN and T1s in order to be able to export 1/3 of raw data to T1s
- On T1s data will need to be archived on tape and subsequently processed (calibration & reconstruction)
- Since T2s will be used almost exclusively for simulation jobs (no input) and resulting AODs will be exported to T1s/AFs, we expect to significantly lower the future needs for storage on T2s and would like to use available funding to buy more CPUs
- While in this model the sites will be mostly specialized for a given task, we still want to retain ability to run any kind of job on any available resource
- Data management is going to be the biggest issue, we need a uniform solution

Software Developments

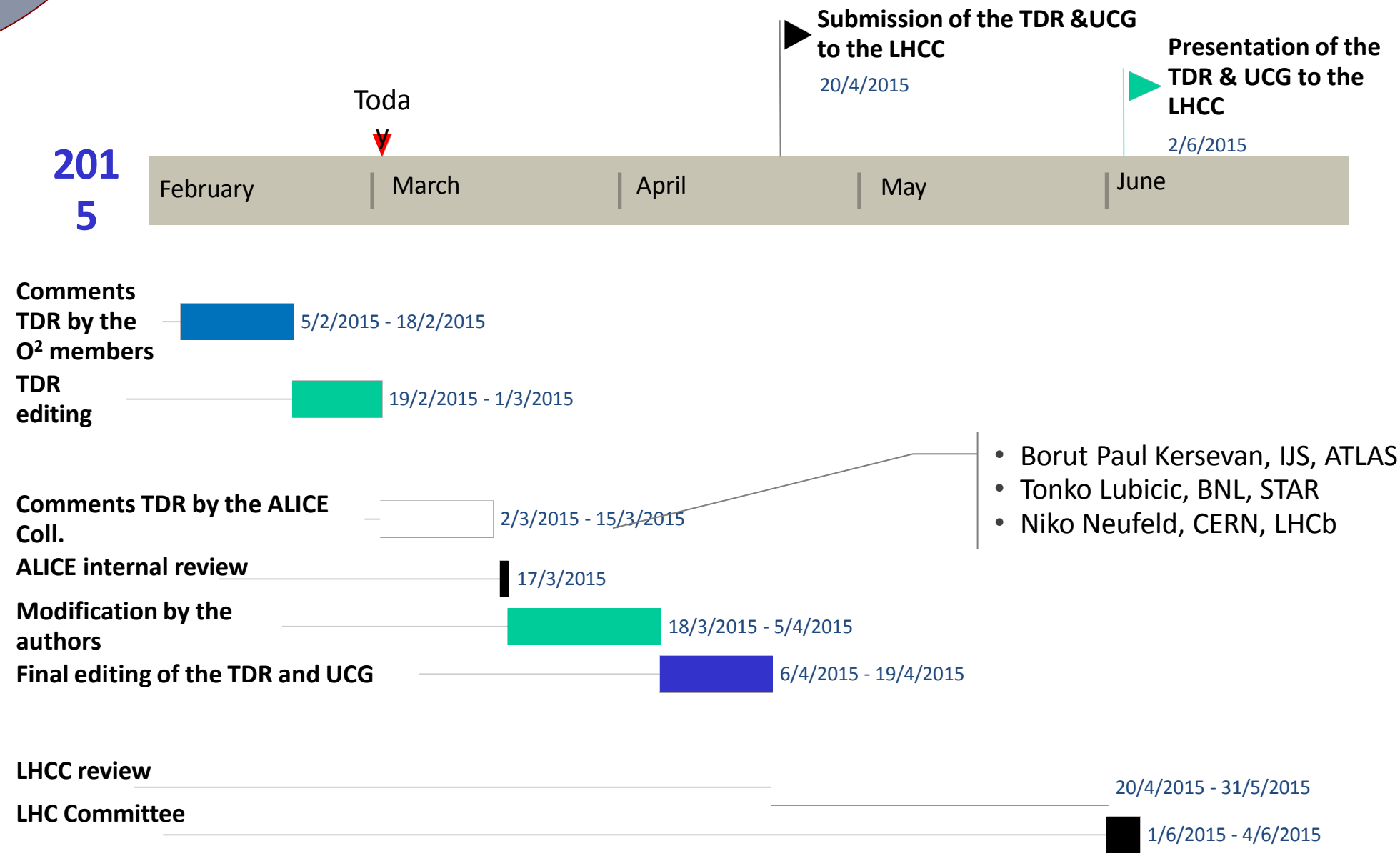
- ALFA & AliceO2
  - Common software layer between FairRoot and AliceO2
  - Git repository, Cdash/Ctest infrastructure in place
  - Template for implementation of a detector code in a new frameworks
  - DDS – Dynamic Deployment System v0.8 released
  - Interface to existing HLT modules in place allowing to test existing algorithms (TPC reconstruction on GPUs)



- AliEn & PanDA
  - Collaboration between ATLAS and ALICE to allow Grid like interface to HPC resources
  - Architecture agreed and manpower finally identified

- ALICE & OpenLab
  - Collaboration on productizing EOS storage (COMTRADE)
  - Rack-scale computing (Intel, Cisco..)

O2 TDR schedule



Conclusions

- Resources for Run 2 are sufficient and we are showing that we can used the efficiently
- The primary goal if O2 facility is data compression
- In a new computing model we try to minimize amount of data moving between Tiers and carry out most of processing on local datasets
- We expect the Grid to grow by 20% per year
- We expect 1/3 of raw data processing to be done on T1s
- Since T2s will be used mostly for simulation and we can rebalance CPU/disk ratio and buy more CPUs
- Dedicated Analysis Facilities are new element that should improve analysis efficiency
- TDR is on schedule

Conclusions

Backup slides

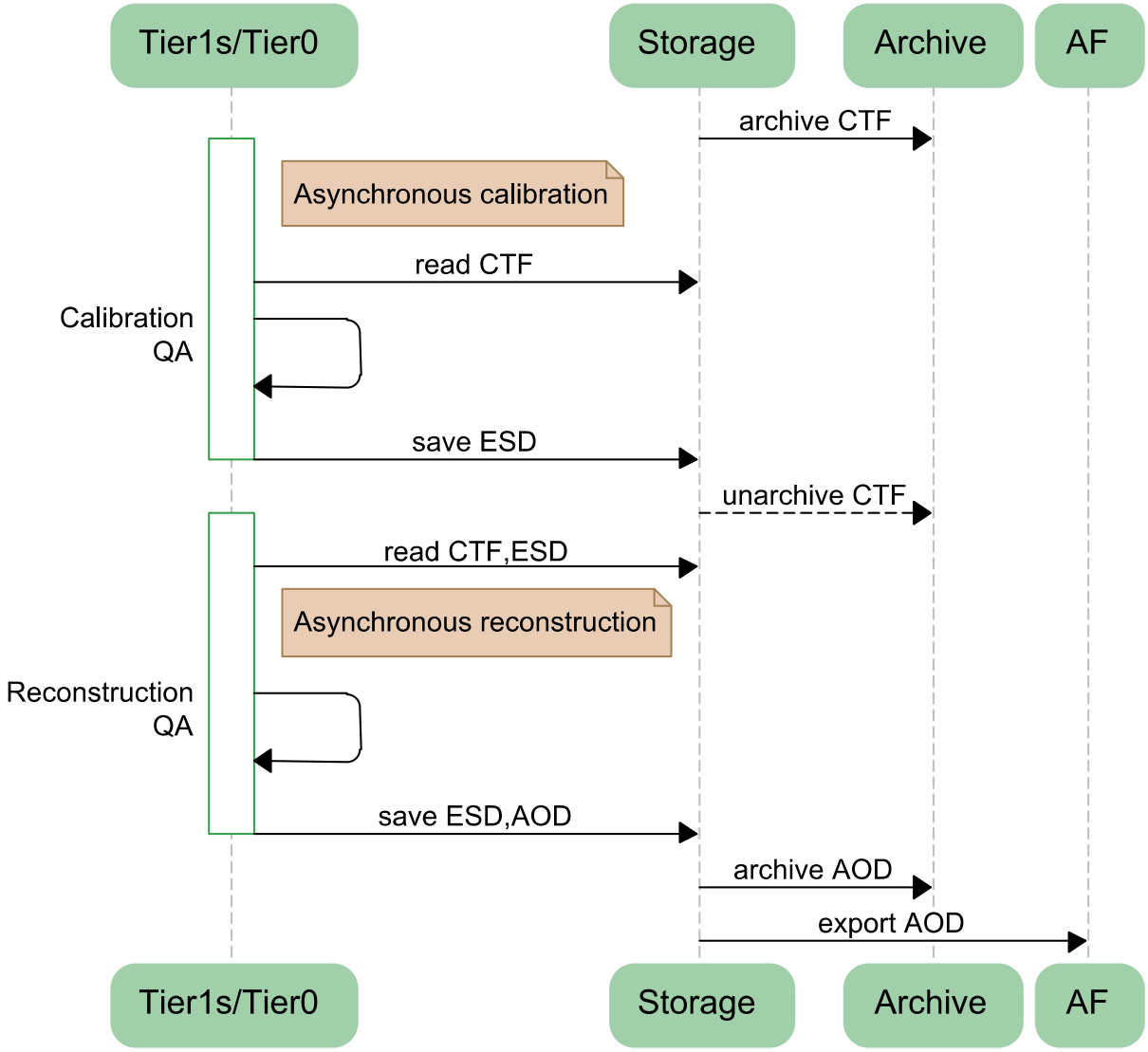
Running scenario

Year	System	$\sqrt{s_{\mathrm{NN}}}$	L_{int}	$N_{\mathrm{collisions}}$
2020	pp	14 TeV	6 pb^{-1}	$4 \cdot 10^{11}$
	Pb–Pb	5.5 TeV	2.85 nb^{-1}	$2.3 \cdot 10^{10}$
2021	pp	14 TeV	4 pb^{-1}	$2.7 \cdot 10^{11}$
	Pb–Pb	5.5 TeV	2.85 nb^{-1}	$2.3 \cdot 10^{10}$
2022	pp	14 TeV	4 pb^{-1}	$2.7 \cdot 10^{11}$
	pp	5.5 TeV	6 pb^{-1}	$4 \cdot 10^{11}$
2025	pp	14 TeV	4 pb^{-1}	$2.7 \cdot 10^{11}$
	Pb–Pb	5.5 TeV	2.85 nb^{-1}	$2.3 \cdot 10^{10}$
2026	pp	14 TeV	4 pb^{-1}	$2.7 \cdot 10^{11}$
	Pb–Pb	5.5 TeV	1.4 nb^{-1}	$1.1 \cdot 10^{10}$
	p–Pb	8.8 TeV	50 nb^{-1}	10^{11}
2027	pp	14 TeV	4 pb^{-1}	$2.7 \cdot 10^{11}$
	Pb–Pb	5.5 TeV	2.85 nb^{-1}	$2.3 \cdot 10^{10}$

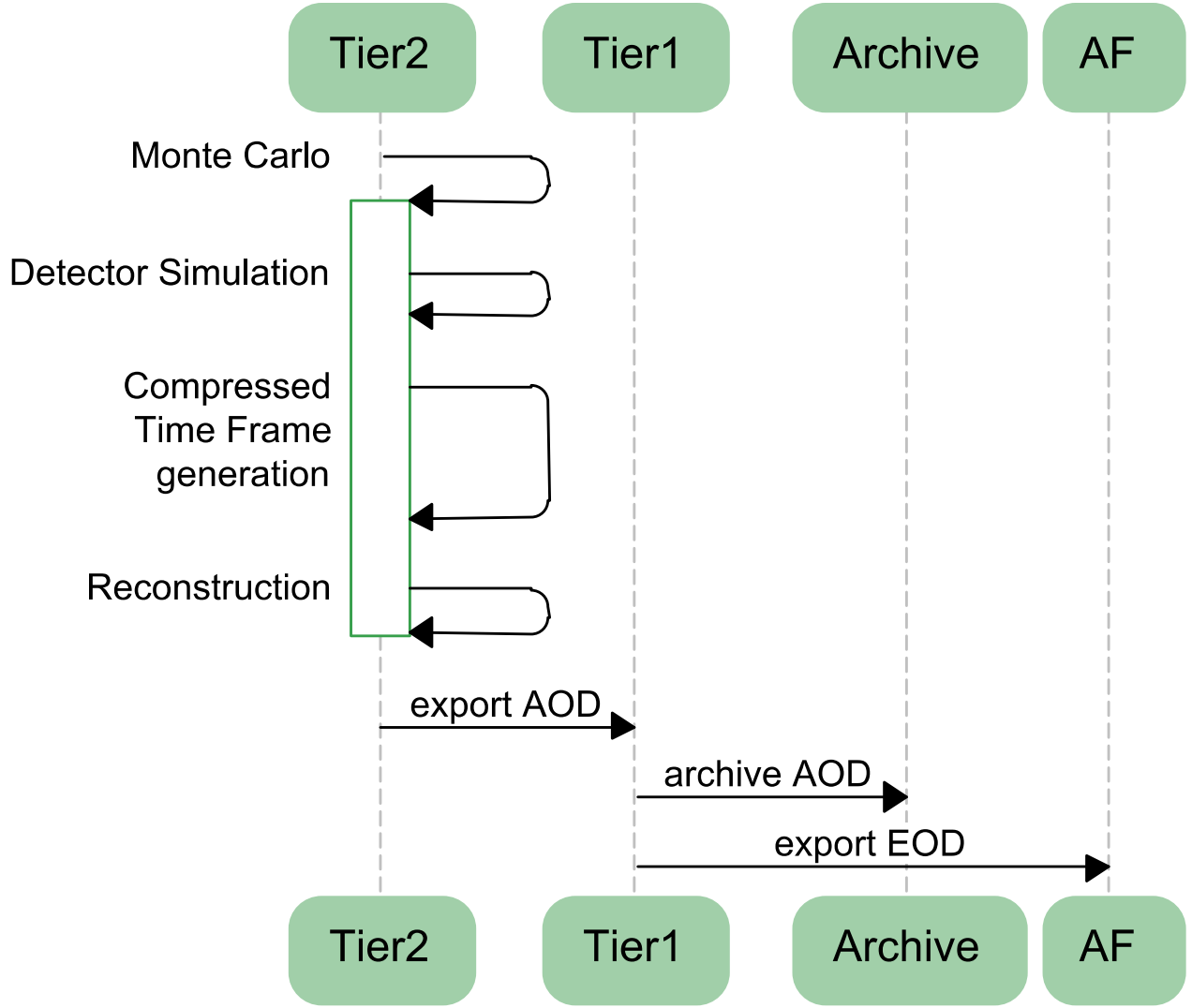
Data types

Acronym	Description	Persistency
RAW	Raw data as it comes from the detector	Transient
CTF	Compressed Time Frame containing processed raw data of for a period of time ≈ 100 ms. In the case of TPC clusters not belonging to tracks are rejected and the remaining information is compressed to the maximum. Once written, CTF becomes read only data.	Persistent
ESD	Event Summary Data. Auxiliary data to CTF containing the output of the reconstruction process that assigns tracks to vertices and identifies the individual collisions.	Temporary
MC	Simulated energy deposits in sensitive detectors. Removed once the reconstruction of MC data is completed.	Transient
AOD	Analysis Object Data containing the final track parameters in a given vertex and for a given physics event. AODs are collected on dedicated facilities for subsequent analysis.	Persistent
MCAOD	Analysis Object Data for a given simulated physics event. Same as AOD with addition of kinematic information that allows comparison to MC. MCAODs are collected on dedicated facilities for subsequent analysis.	Persistent
HISTO	The subset of AOD information specific for a given analysis. Can be generated during analysis but needs to be offloaded from the Grid.	Temporary

Tier 0/1


Tier1s/Tier0
Storage
Archive
AF
archive CTF
Asynchronous calibration
read CTF
Calibration
QA
save ESD
unarchive CTF
read CTF,ESD
Asynchronous reconstruction
Reconstruction
QA
save ESD,AOD
archive AOD
export AOD
Tier1s/Tier0
Storage
Archive
AF

Tier 2


Tier2
Tier1
Archive
AF
Monte Carlo
Detector Simulation
Compressed Time Frame generation
Reconstruction
export AOD
archive AOD
export EOD
Tier2
Tier1
Archive
AF