ALICE Tier-1/Tier-2 Workshop

23-25 February 2015

Plans for Run2 and the ALICE upgrade in Run3

Predrag Buncic

CERN

ALICE

RUN 2 detector upgrades

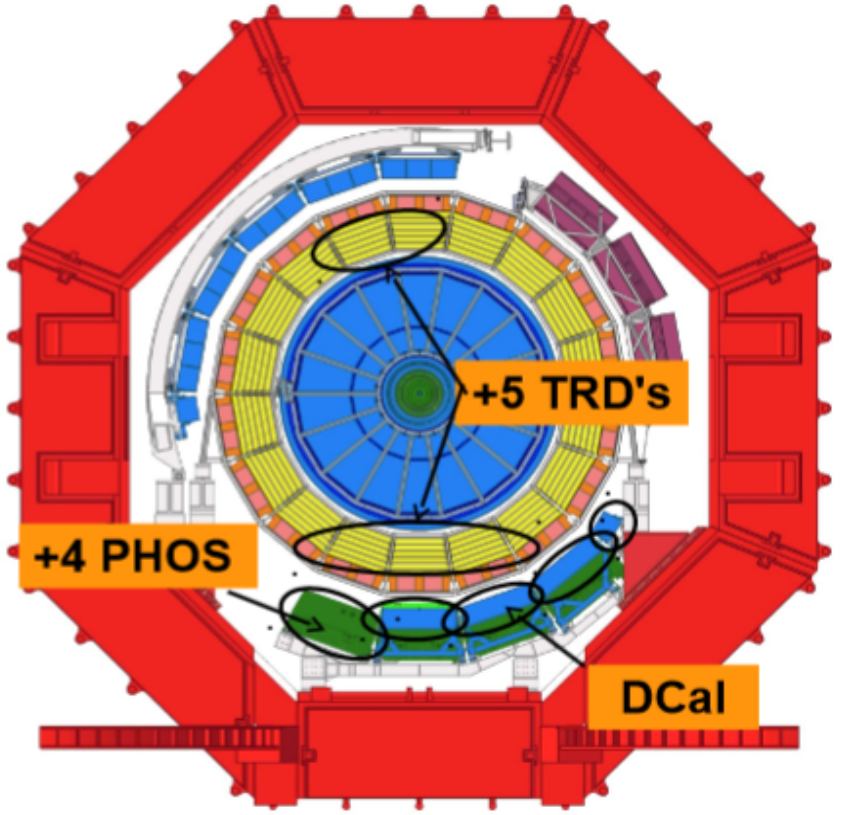
TPC, TRD readout electronics consolidation

+5 TRD modules

full azimuthal coverage

+1 PHOS calorimeter module

+ DCAL calorimeter



- Double event rate => increased capacity of HLT system and DAQ
  - Rate up to 8GB/sec to T0

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)

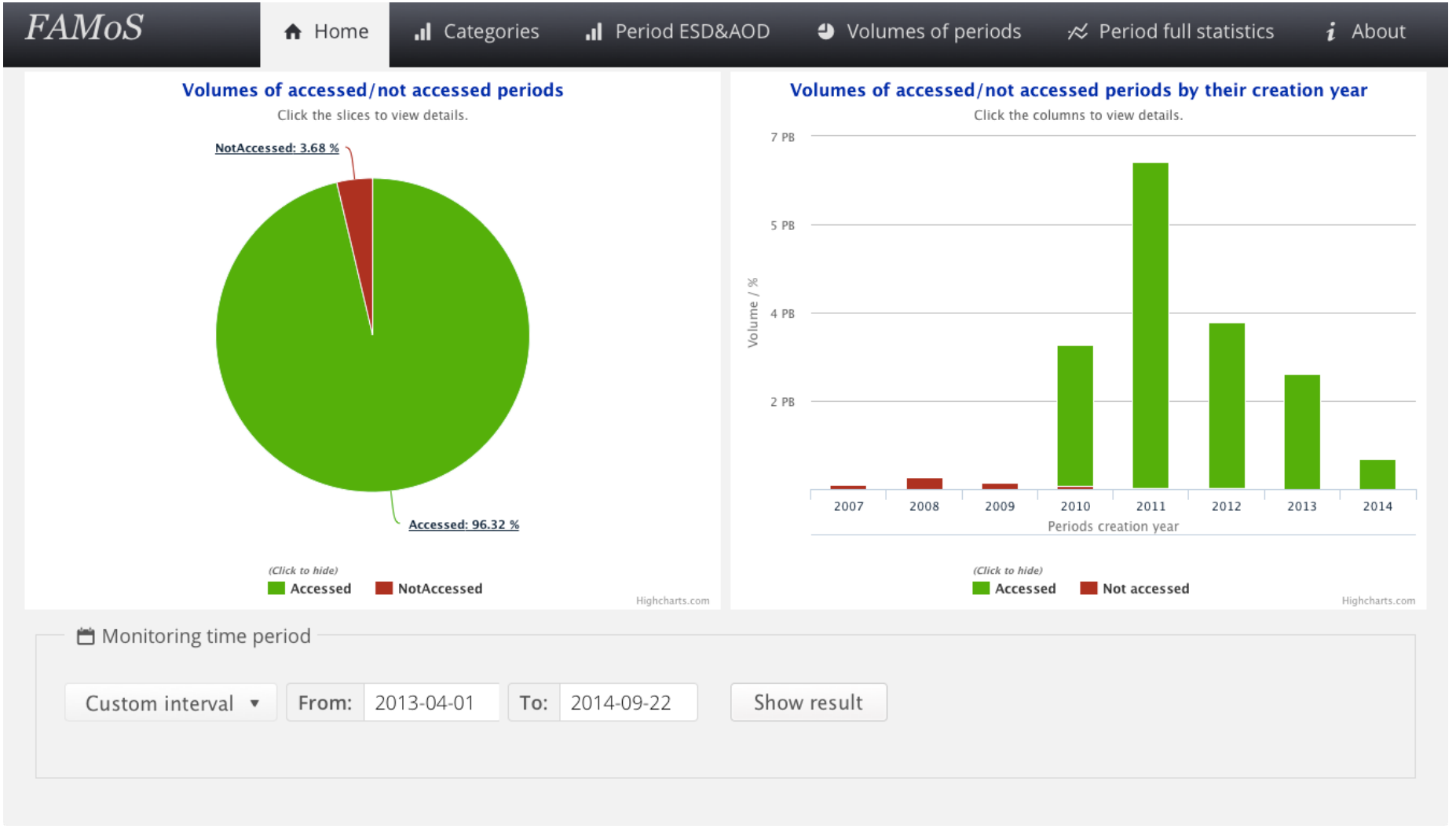
Simulation

- Geant4 v10 Physics Validation has started
  - First test production (done) Pythia6, pp, 7 TeV
  - QA in progress
- CPU performance still 2x worse compared to simulation with G3
  - Some gains that we made with G4 v9.6 are gone with v10
  - But, we can use G4 multithreaded capabilities to put our hands on resources that would otherwise be out of reach
- Next Step
  - Comprehensive comparison of detector response with data

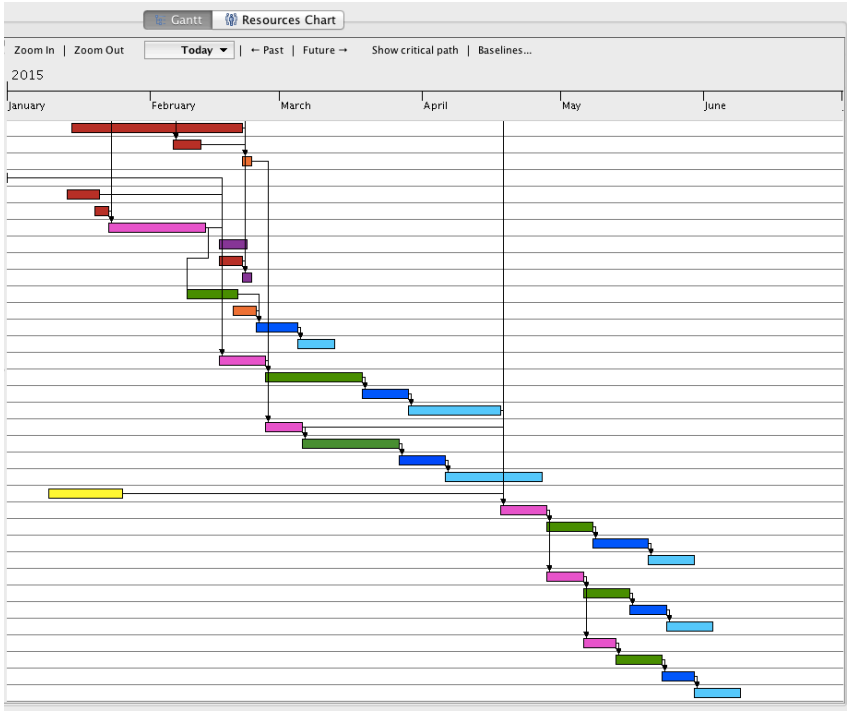
Distributed Computing & Analisys

- No news is a good news
  - With occasional hiccups, things work and continue to grow
- Switch to CVMFS is fully completed
  - Including OCDB repository that is mirrored from AliEn
- Consolidation of AliEn development branches is ongoing behind the scene
  - Overall update of dependencies
  - Becomes increasingly important in order to address security issues that seem to be more and more frequent
- New AliEn/ARC interface
- AliEn on HLT farm tested on development cluster
- Work in progress
  - CAF on demand
  - AliEn in the box – virtualized site on OpenStack
  - AliEn + PanDA on HPC

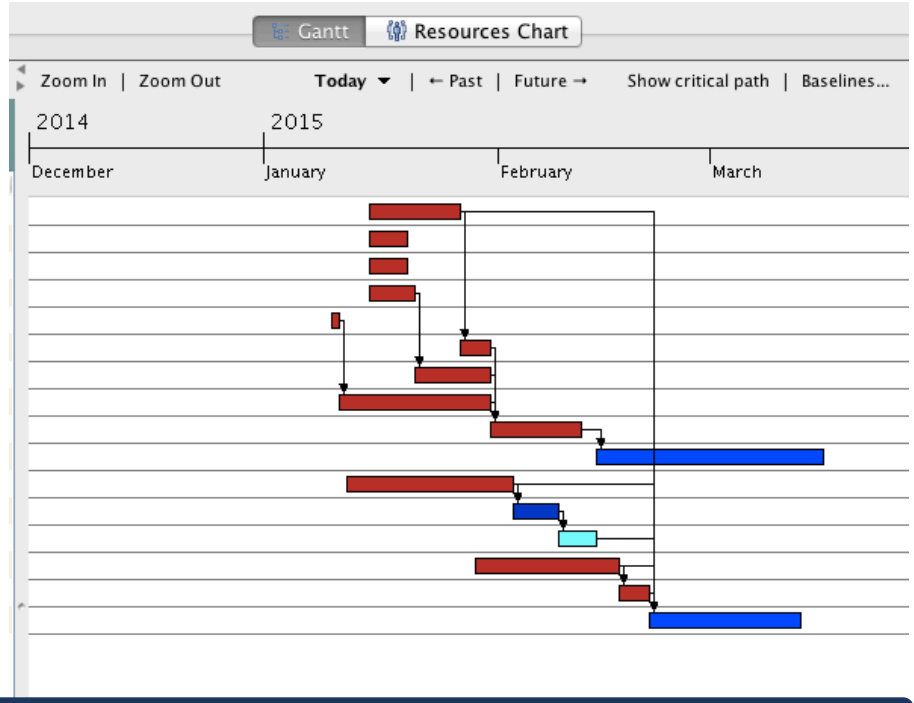
Data popularity



Re-processing & re-commissioning

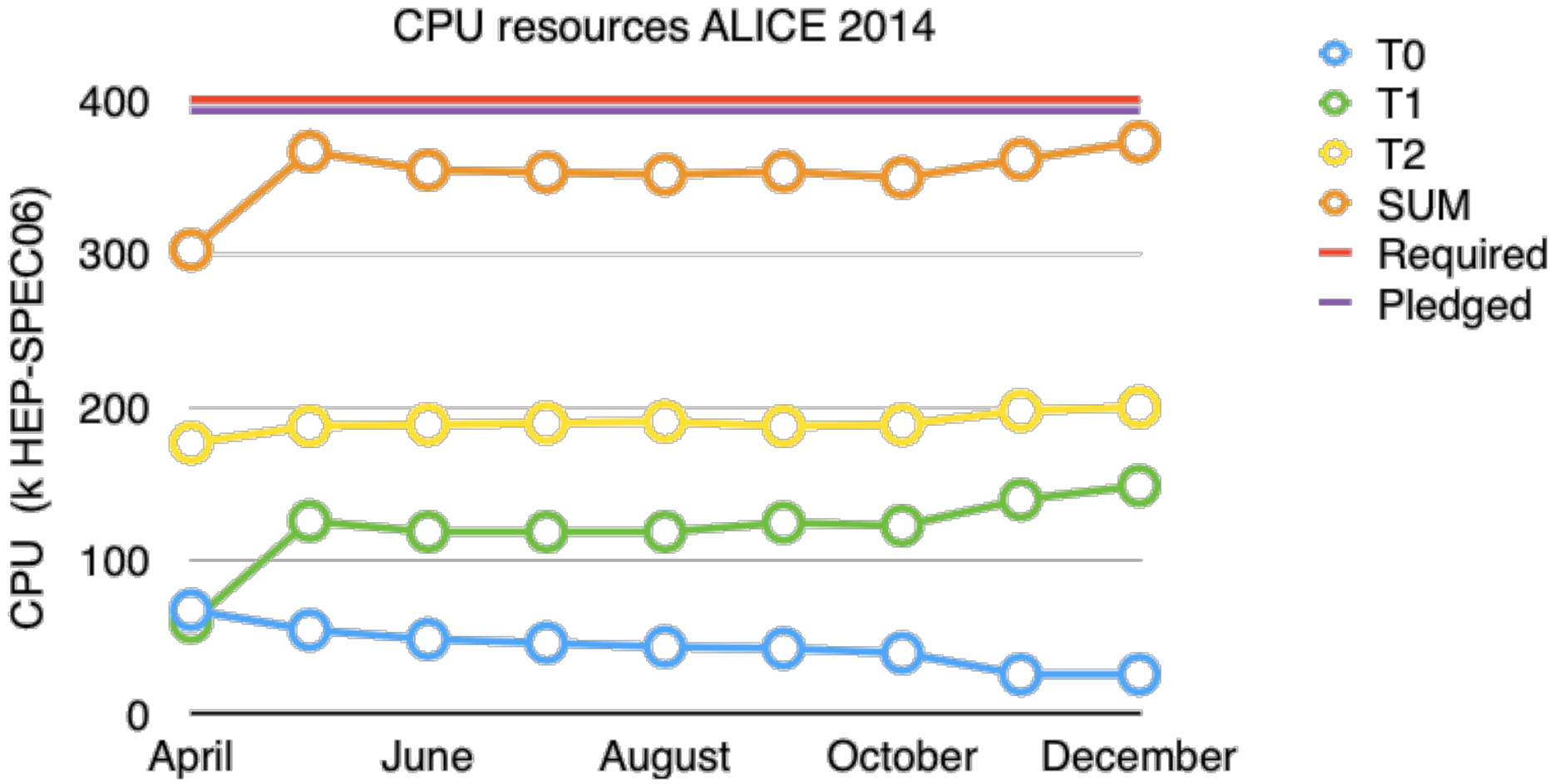


- Re-processing
  - Steady RAW and MC activities
  - Full detector re-calibration and 2 years worth of software updates
  - All Run 1 RAW data processing with the ***same*** software

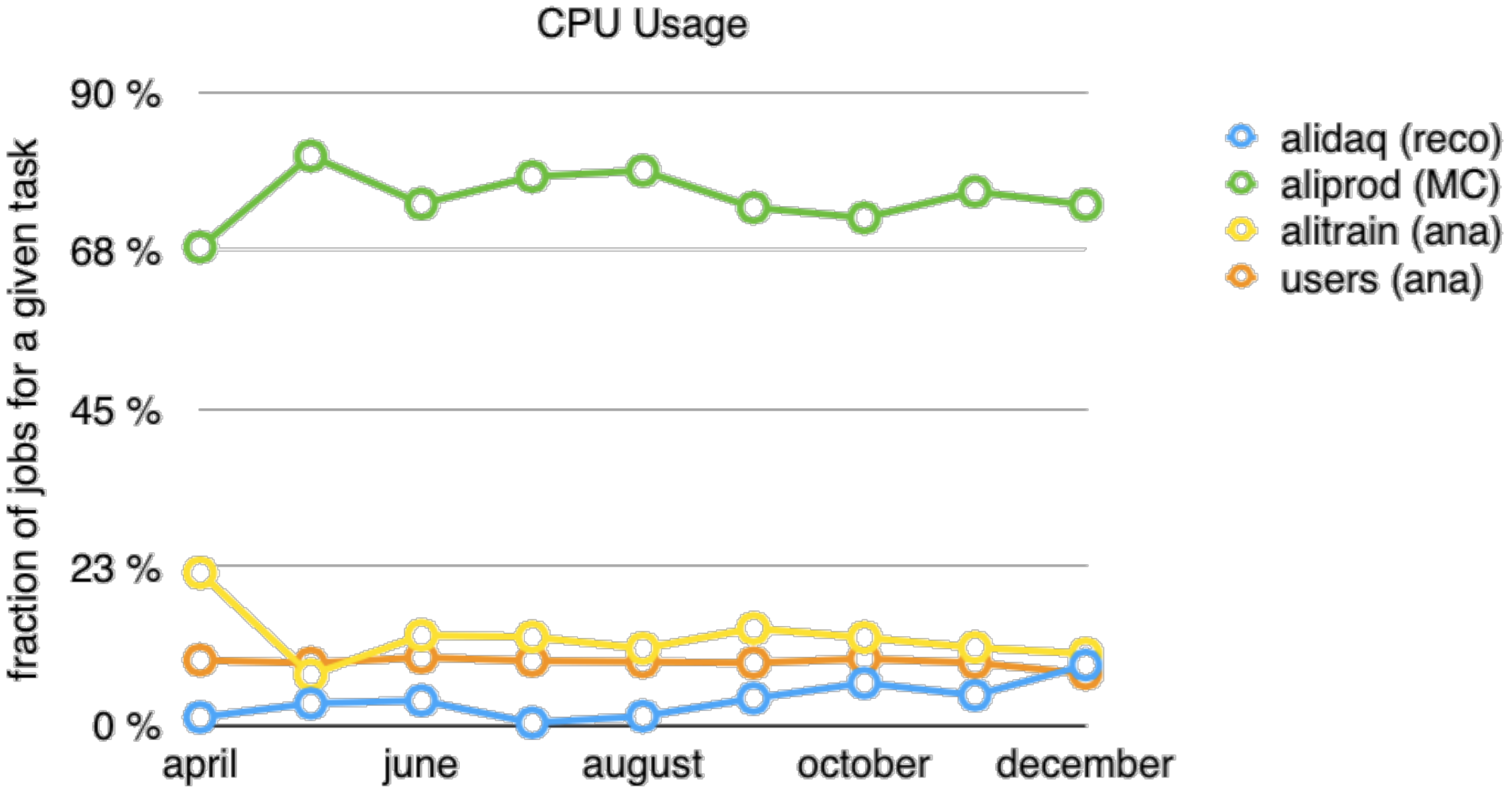


- ALICE re-commissioning
  - Test of upgraded detectors readout, Trigger, DAQ, new HLT farm
  - Full data recording chain, with conditions data gathering
  - Cosmics trigger data taking with Offline processing

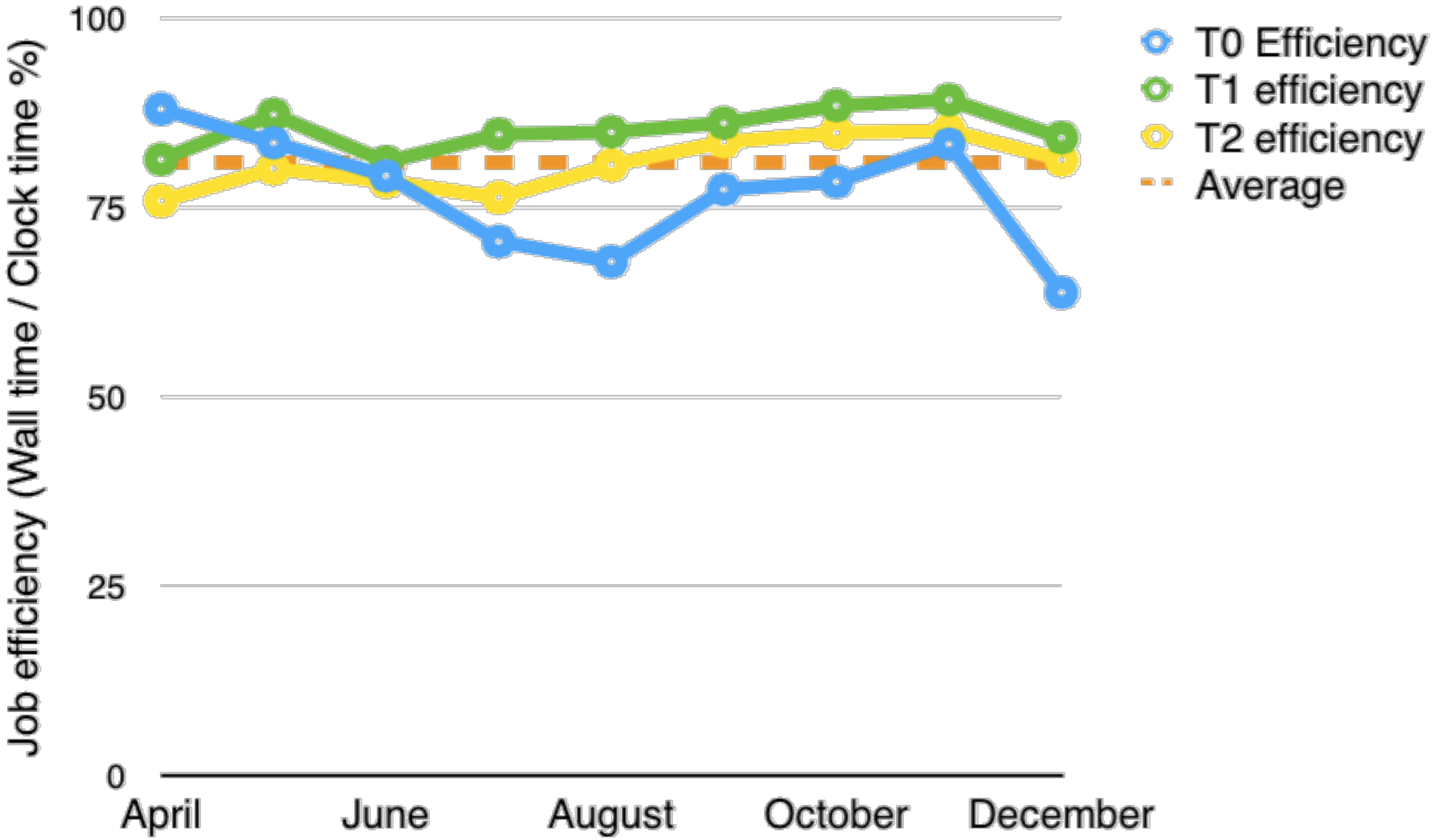
CPU Resources in 2014



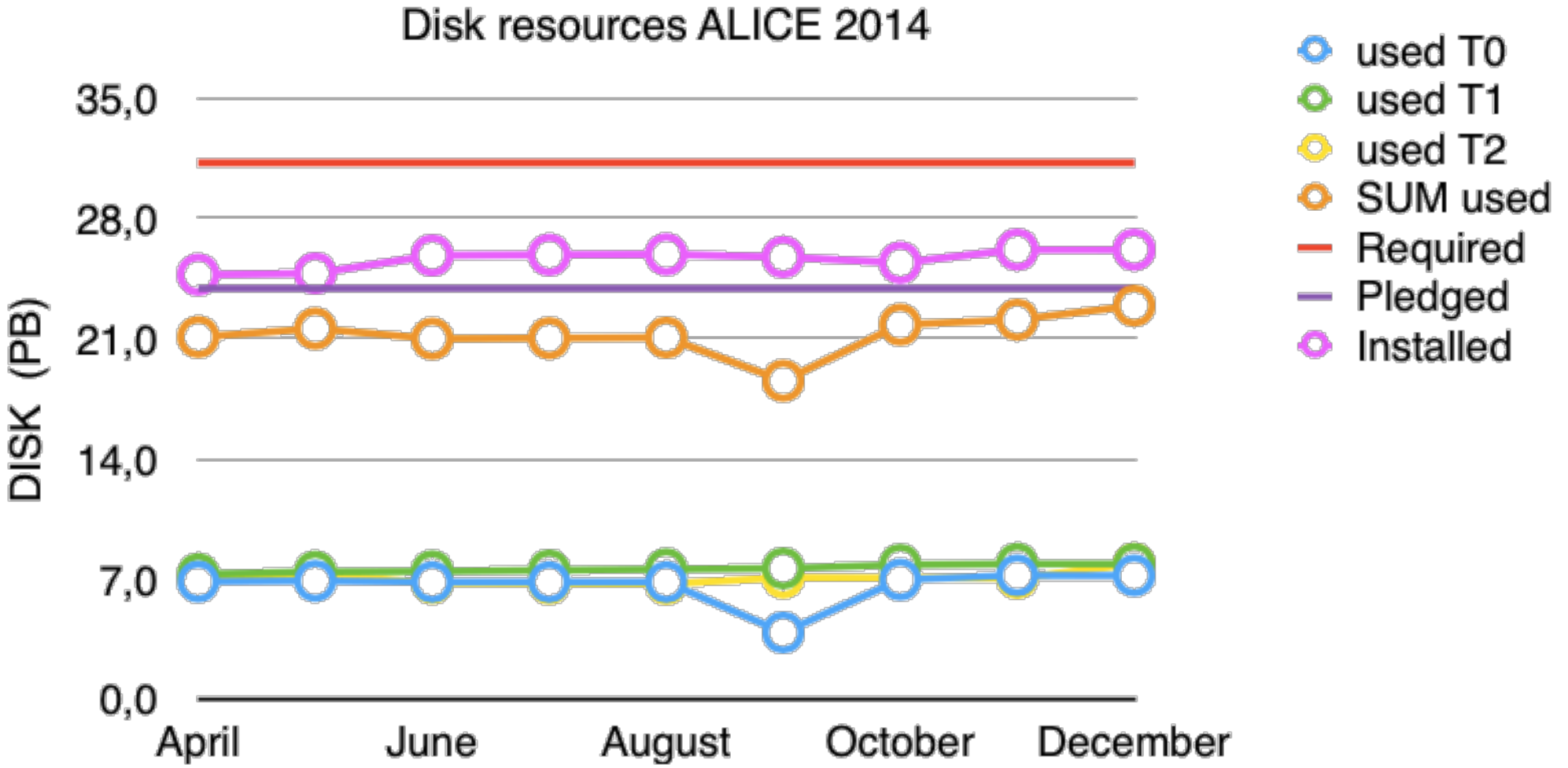
CPU Shares


CPU Usage
fraction of jobs for a given task
90 %
68 %
45 %
23 %
0 %
april
june
august
october
december
alidaq (reco)
aliprod (MC)
alitrain (ana)
users (ana)

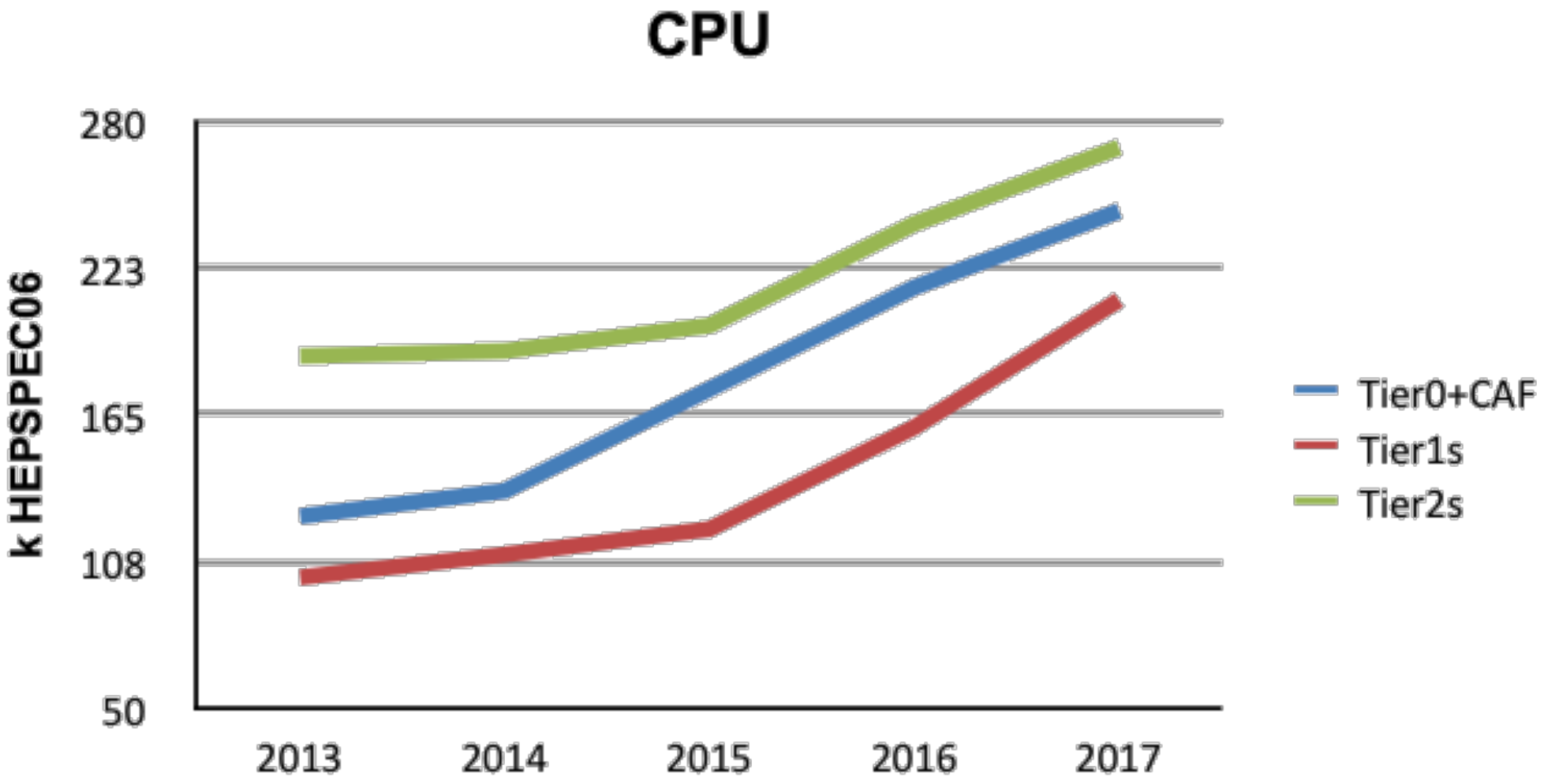
CPU Efficiency 2014



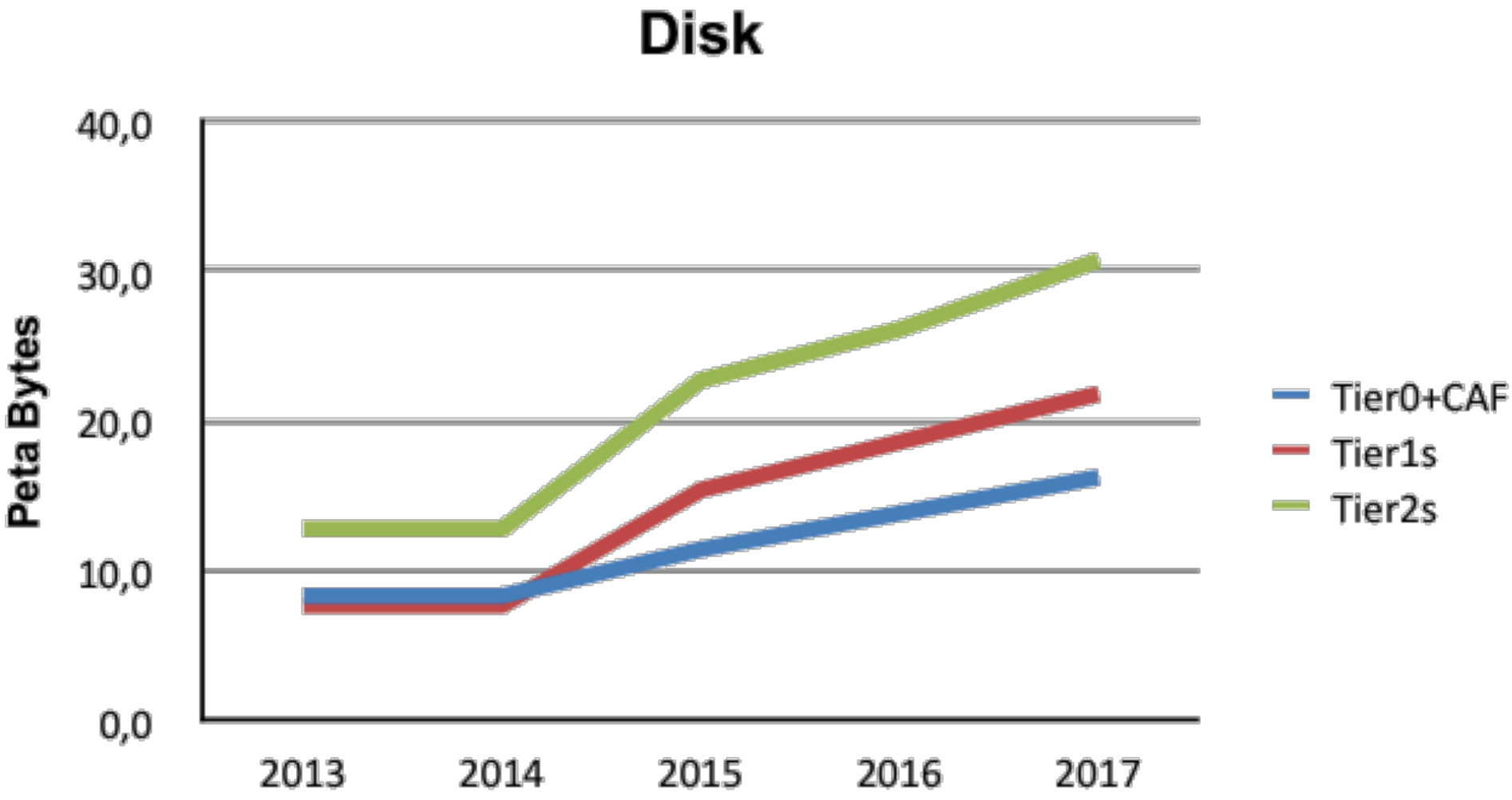
Disk Resources in 2014



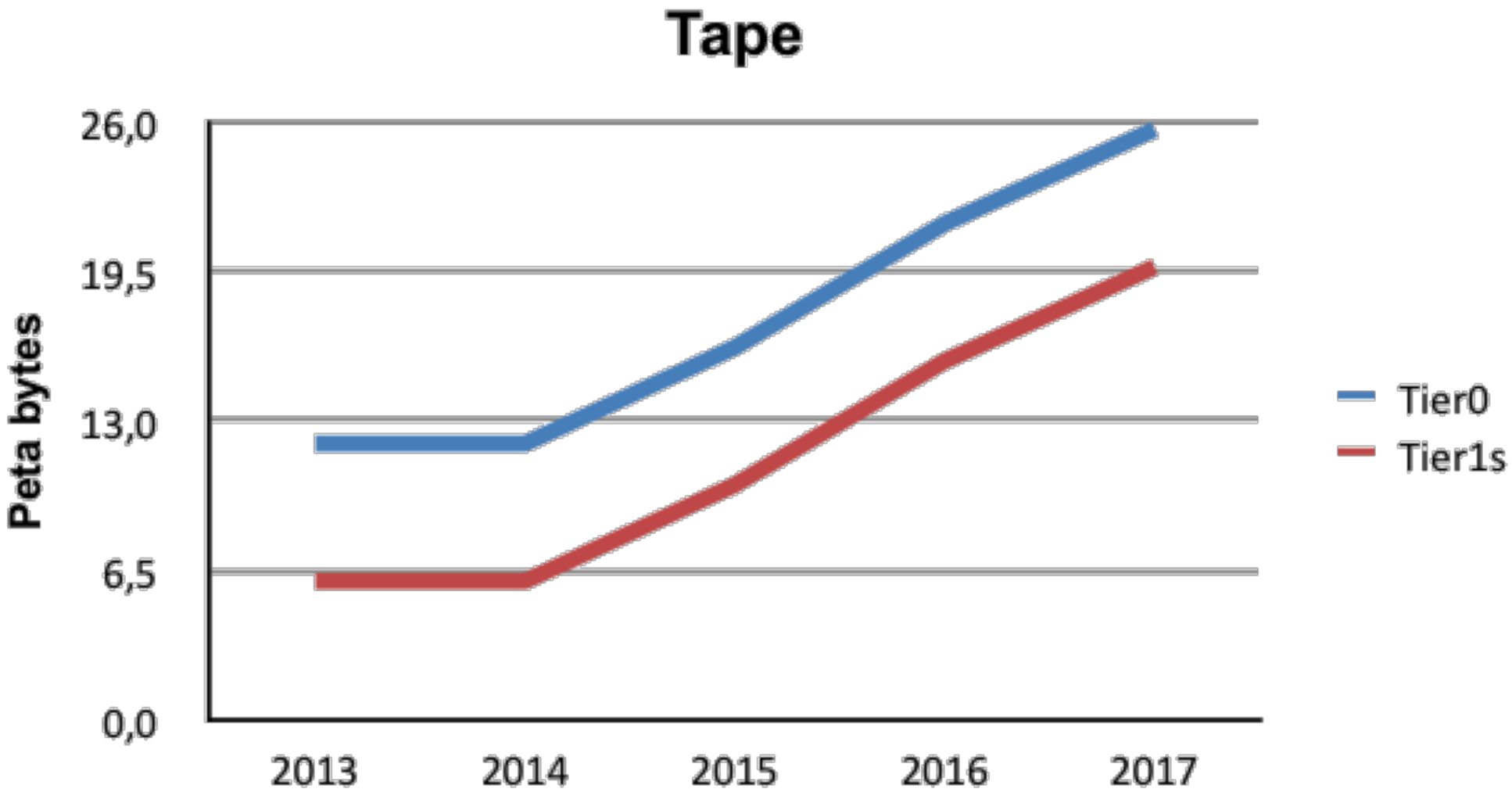
CPU Request 2013-2017



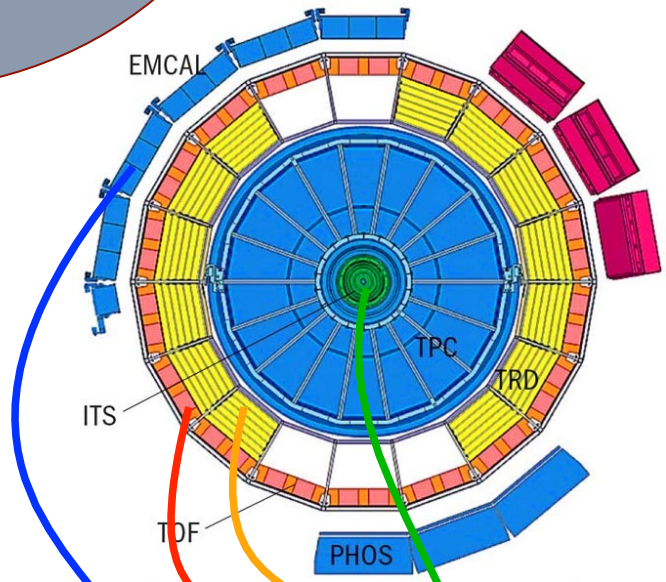
Disk Request 2013-2017

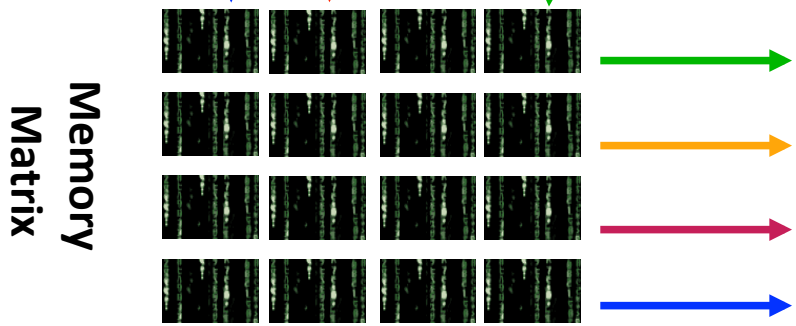


Tape Request 2013-2017

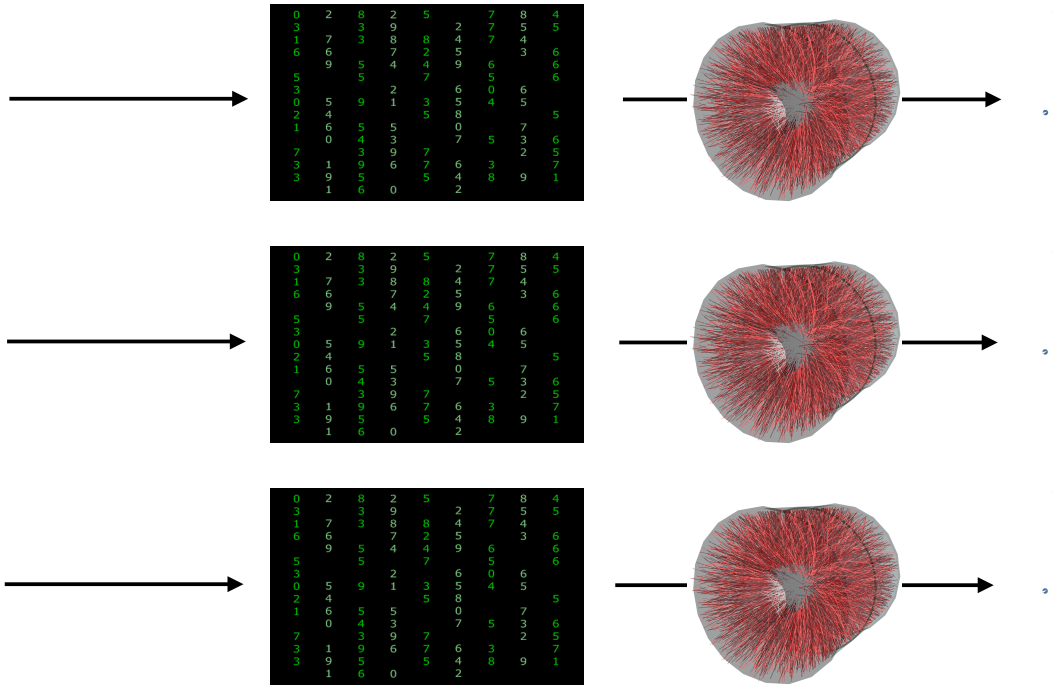


Data Acquisition (DAQ) Design Concept





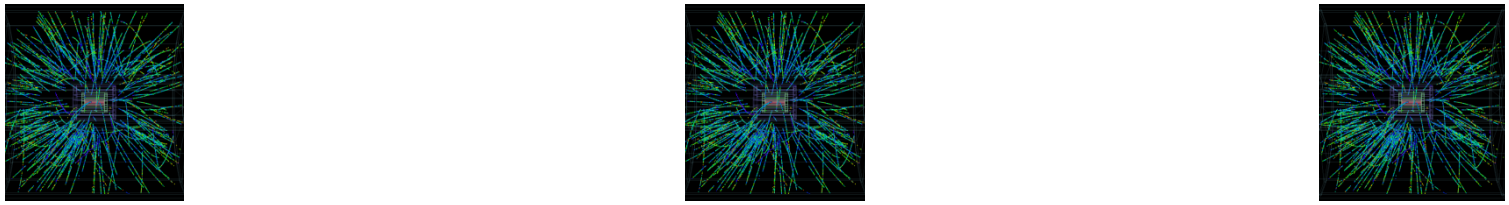
MULTIPLEXER



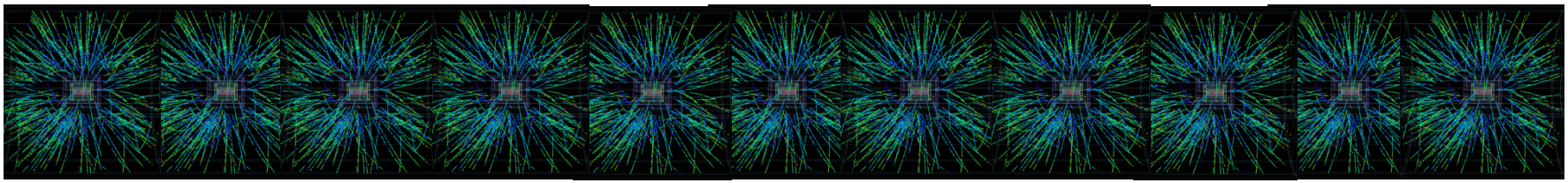
- Acquire data of tens of millions of channels
- Store them in a matrix of hundreds of memories
- Multiplex to a computer farm
- Assemble and store data from the same event

LHC: Towards Higher Luminosities

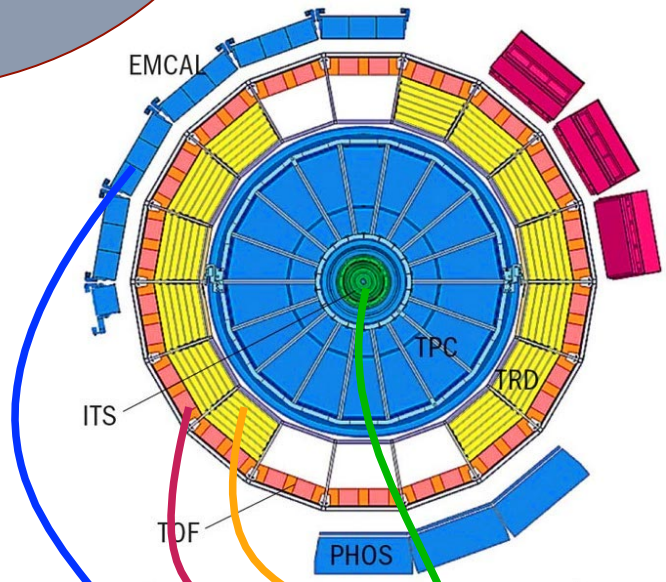
LHC proton-proton now: **Luminosity 7×10^{33} cm^{-2} s^{-1}**

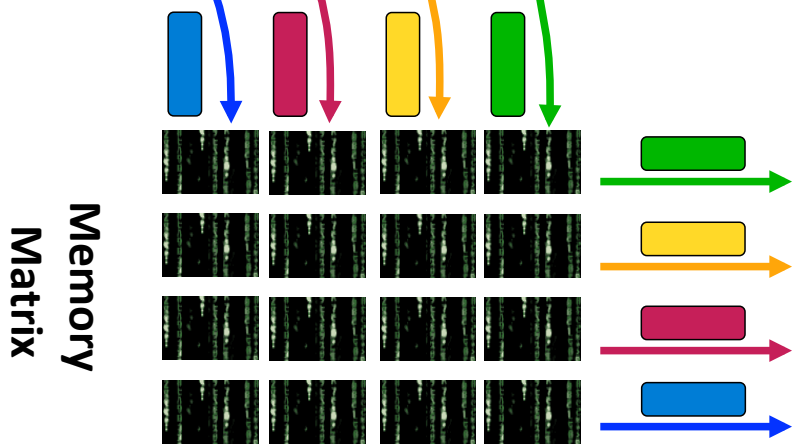


LHC proton-proton after 2018: **Luminosity 4×10^{34} cm^{-2} s^{-1}**

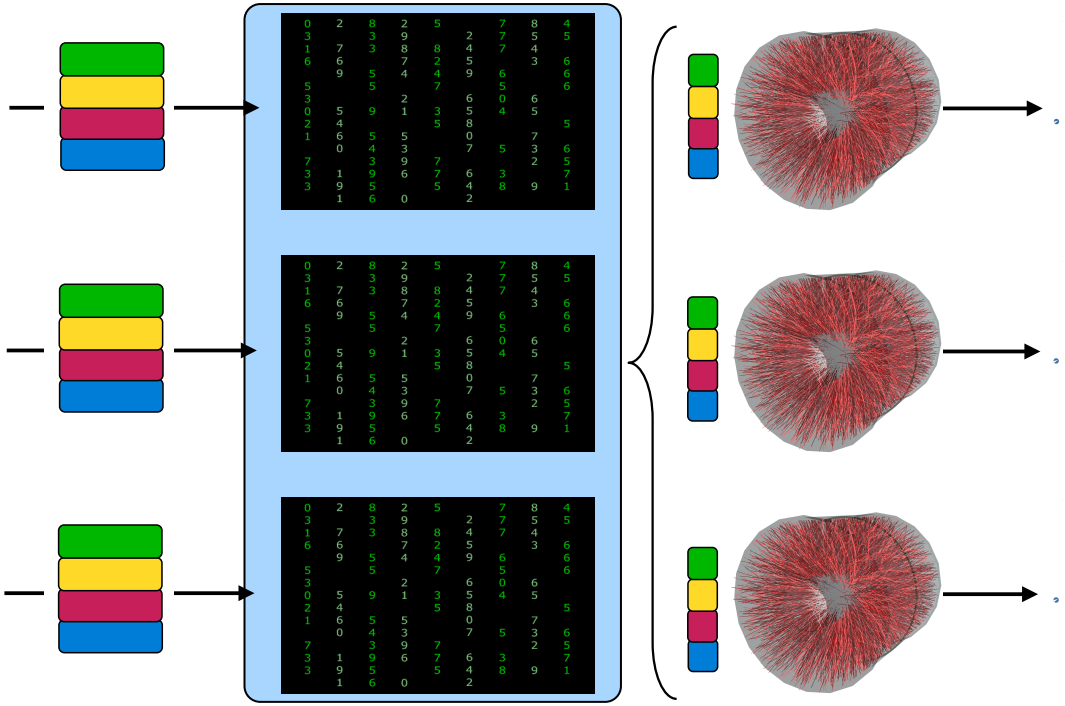


DAQ++





MULTIPLEXER



- Continuous detector reading: replace events with time windows (100 ms, ~5'000 events). Self sufficient small dataset?
- Calibrate and reconstruction online: reduce data volume & structure the data
- Prototyping with ZeroMQ and Zookeeper

Run 3: Paradigm shift

- Now: reducing the event rate from 40 MHz to ~ 1 kHz
  - **Select the most interesting particle interactions**
  - Reduce the data volume to a manageable size
- After 2018:
  - Higher interaction rate
  - More violent collisions → More particles → More data (1 TB/s)
  - Physics topics require measurements characterized by very small signal/ background ratio → large statistics
  - Large background → traditional triggering or filtering techniques very inefficient for most physics channels
  - **Read out all particle interactions (PbPb) at the anticipated interaction rate of 50 kHz**
- **Data rate increase: x100**

The ALICE O2 Project: Architecture

Detector

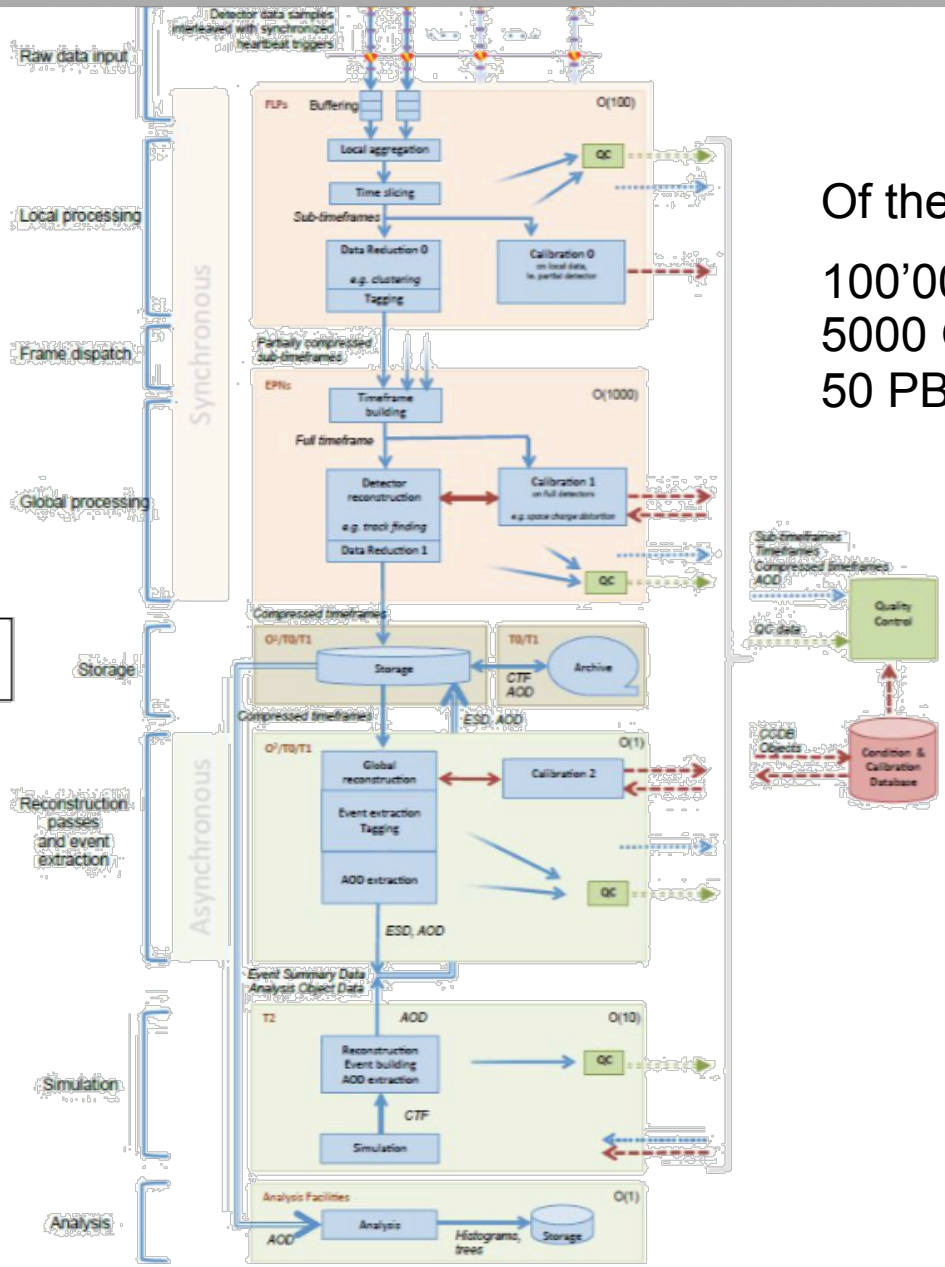
Synchronous readout, aggregation, compression
- FLP 2.5x
- EPN 8x

Storage

Asynchronous reconstruction
75 % EPN / 25 % T0-T1

Simulation

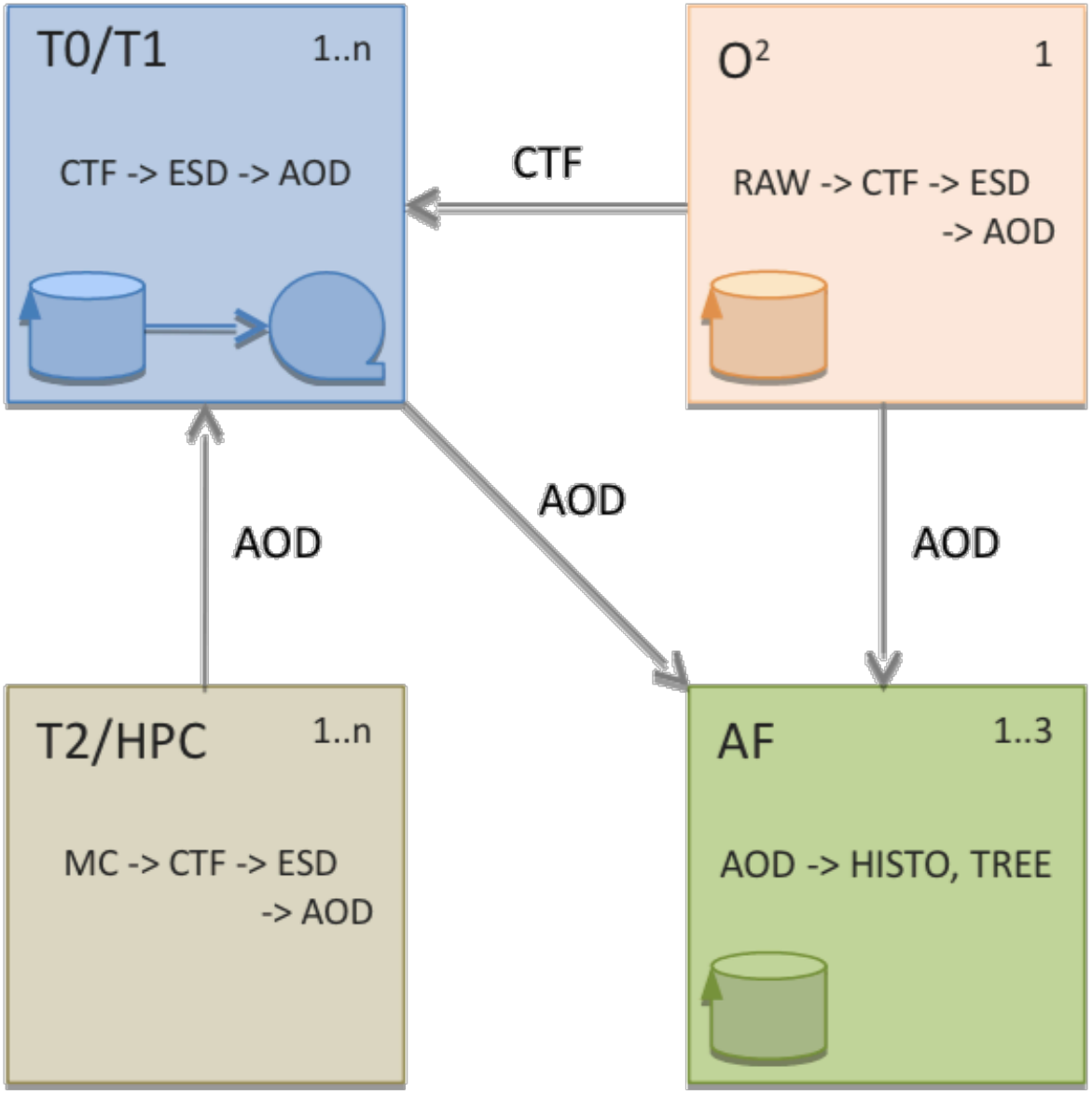
Analysis



Of the order of:

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

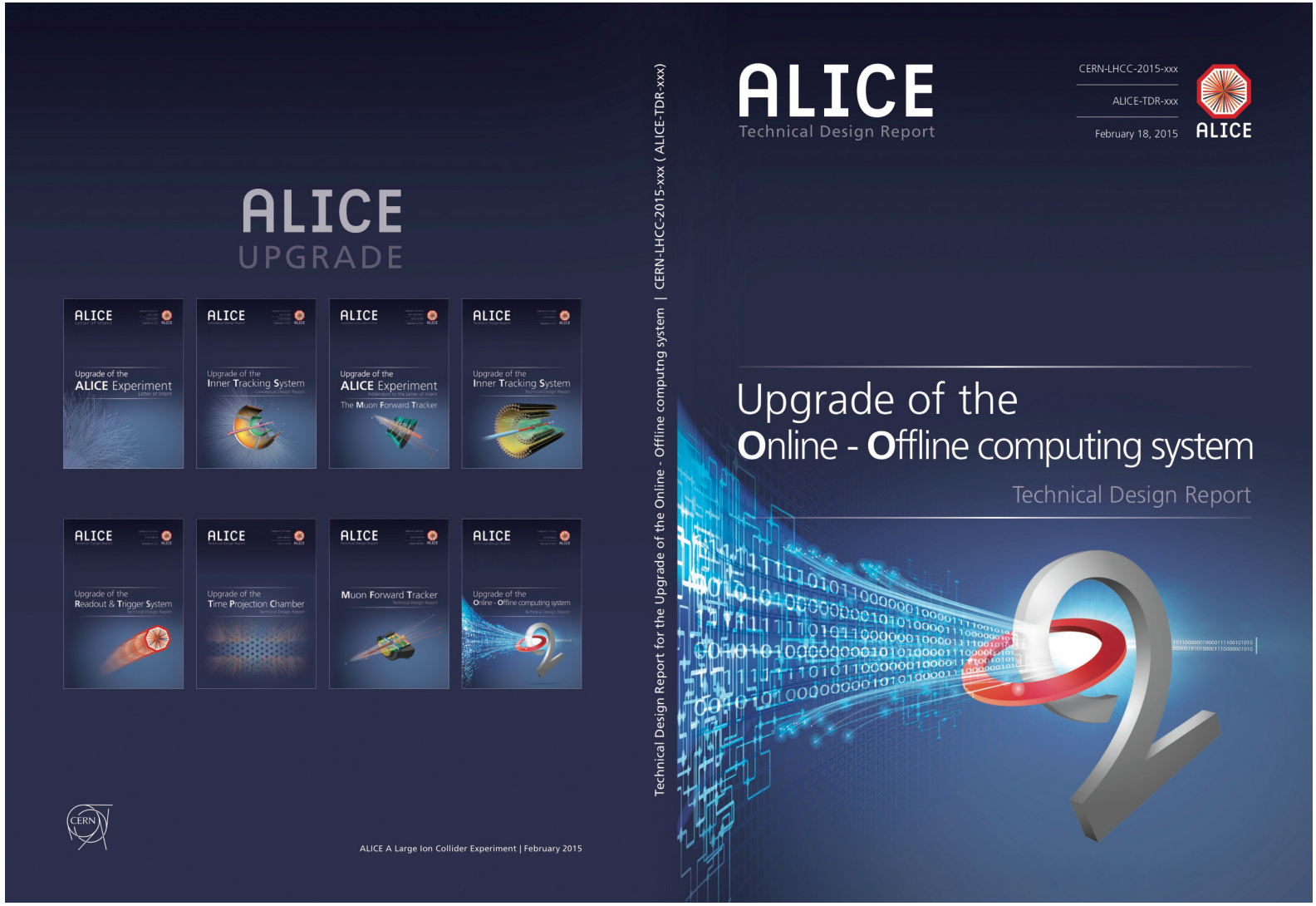
A fraction of any of these would do…

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrade
Name/Planned Installation	**Edison**	**TITAN**	**MIRA**	**Cori 2016**	**Summit 2017-2018**	**Aurora 2018-2019**
System peak (PF)	2.4	27	10	>30	150	>150
Peak Power (MW)	3	8.2	4.8	<3.7	10	~13
System memory per node	64 GB	38 GB	16 GB	64-128 GB DDR4 16 GB High Bandwidth	> 512 GB (High Bandwidth memory and DDR4)	TBA
Node performance (TF)	0.460	1.452	0.204	>3	>40	>15 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & mulitple Nvidia Voltas GPUS	TBA
System size (nodes)	5,200 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	~50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	TBA
File System	17.6 PB, 168 GBs, Lustre®	32 PB, 1 TB/s, Lustre®	GPFS™	28 PB, 744 GB/sec , Lustre®	120 PB, 1 TB/s, GPFS™	TBA

Impact on Grid T1/T2 sites

- Expectations for Grid resource evolution
  - 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 resource
  - Data management is going to be the biggest issue, we need a uniform solution

O2 TDR



Soon in your INBOX!

Conclusions

- Resources for Run 2 are sufficient and we are showing that we can used the efficiently
- The upgrade will be challenging due to a large data volume
- 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 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
- We expect the Grid to grow by 20% per year and at present the resources are sufficient
- Dedicated AFs for analysis need to be funded externally

Conclusions

Backup slides

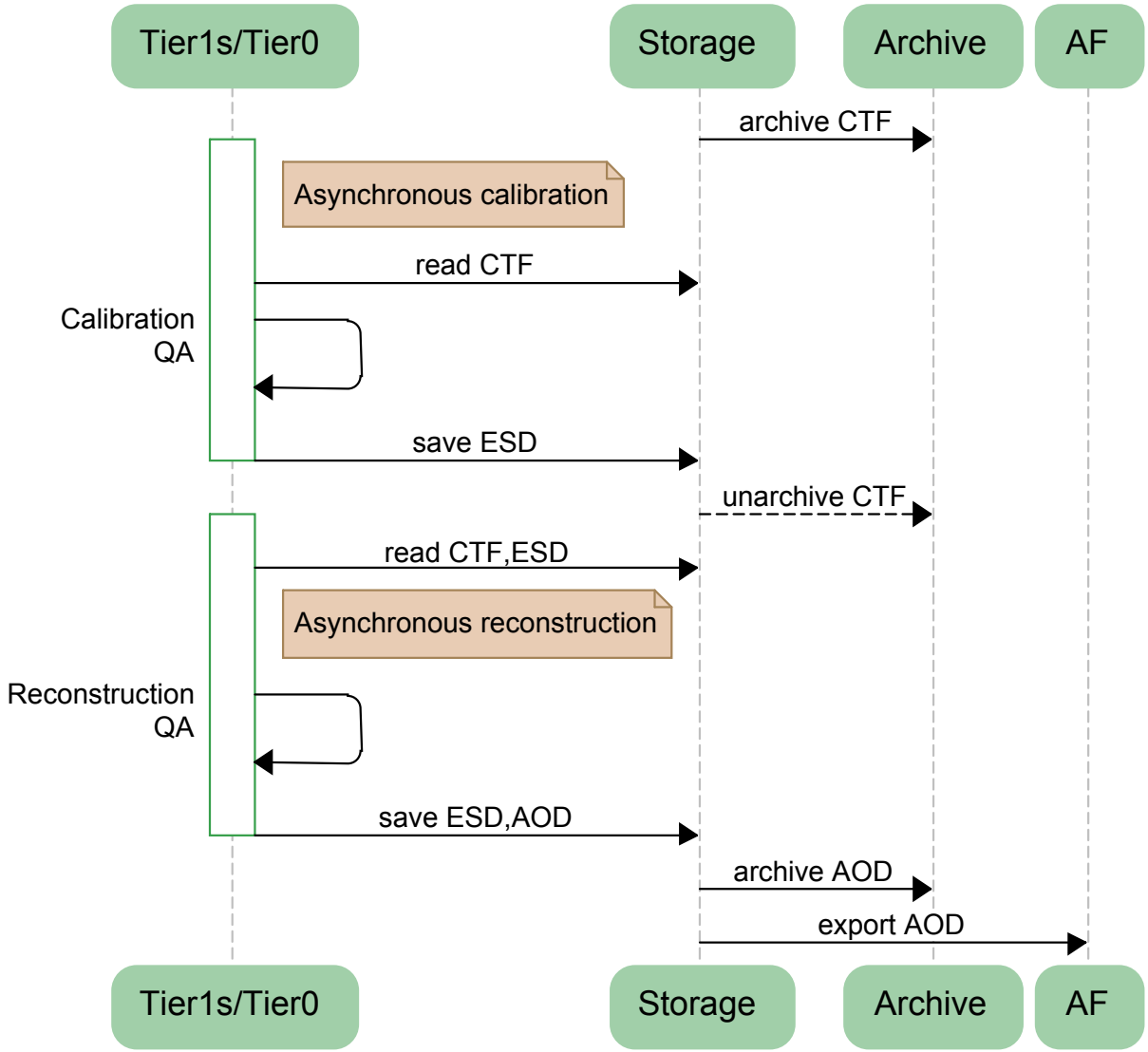
Running scenario

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}	$2.3 \cdot 10^{10}$

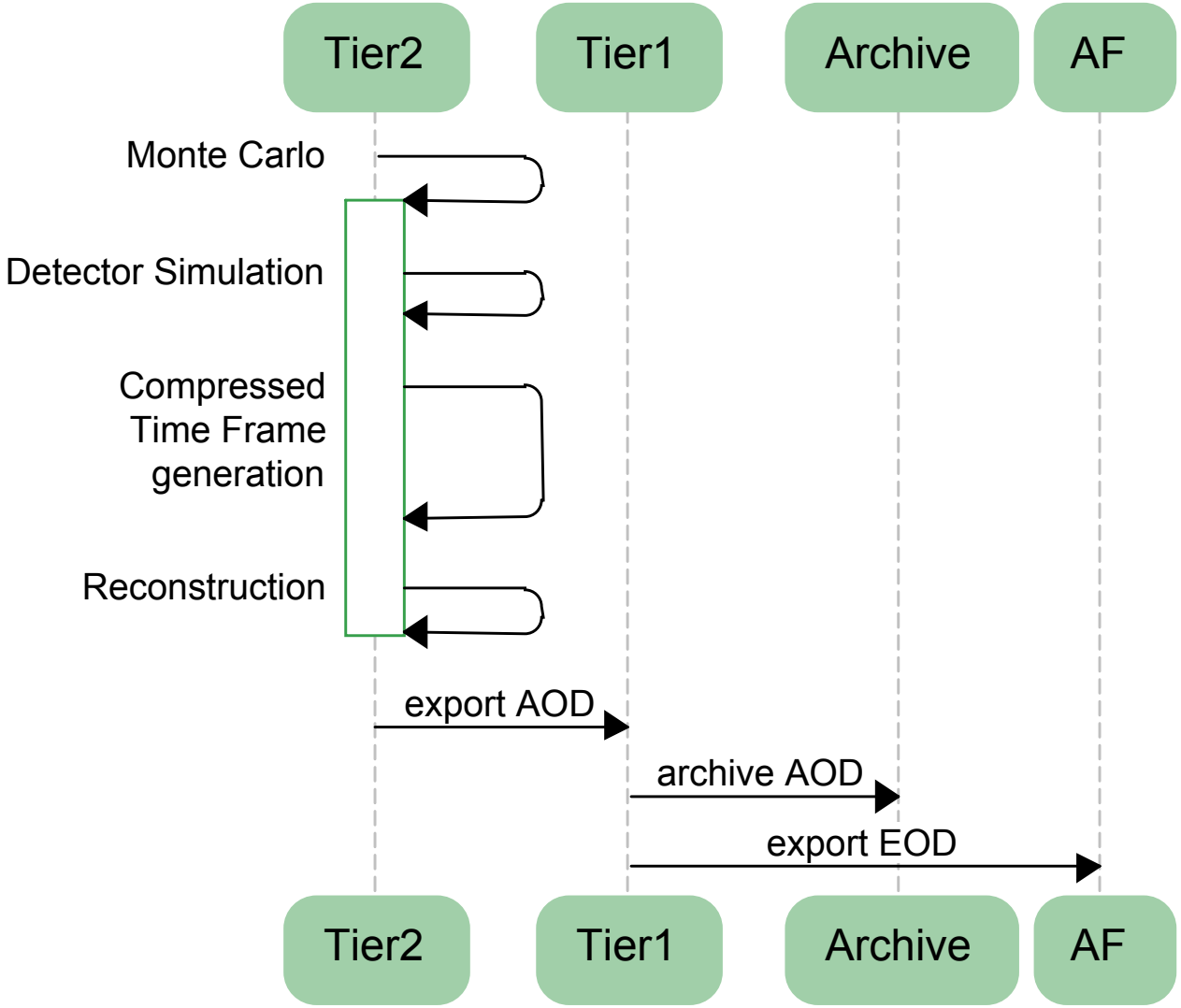
Data types

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