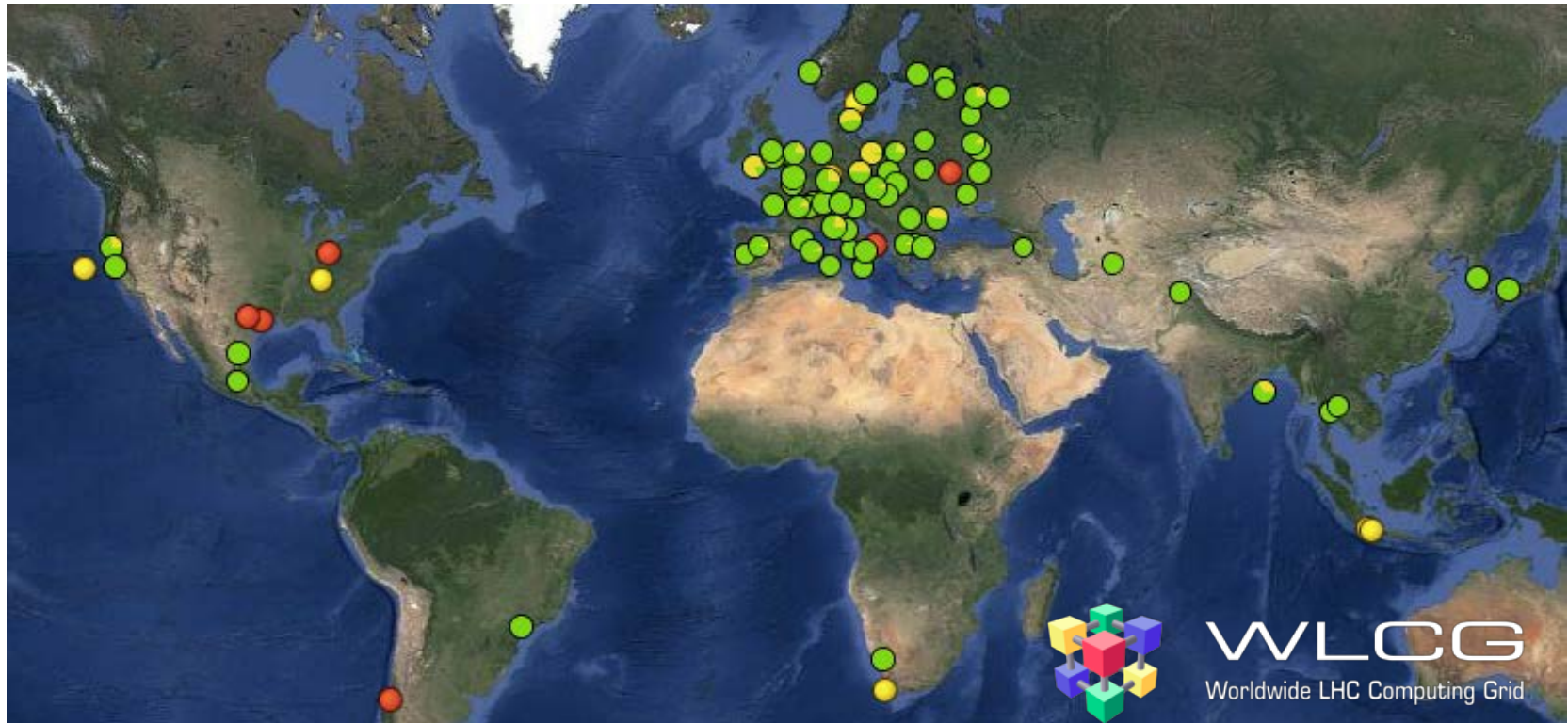




O2 Computing Model

P. Buncic

Computing Model in Run 1&2

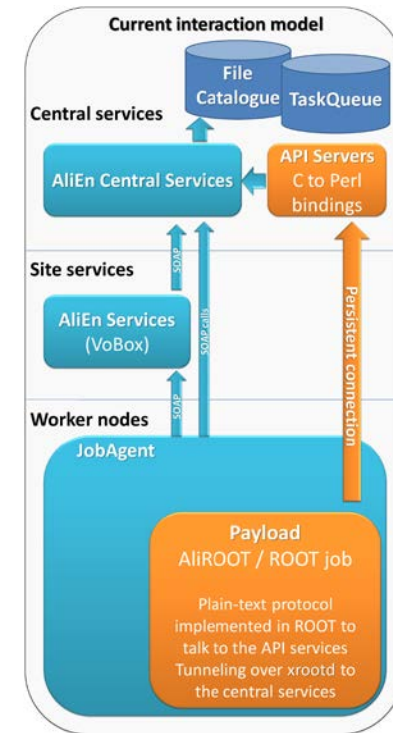


- Distributed grid computing model on top of WLCG services
 - Faithful implementation of the original grid ideas
 - Any job can run anywhere and access data from any place
 - Scheduling takes care of optimizations and brings “computing to data”
 - Every file and its replicas are accounted in the file catalogue

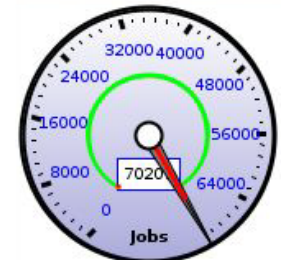


ALICE Distributed Computing

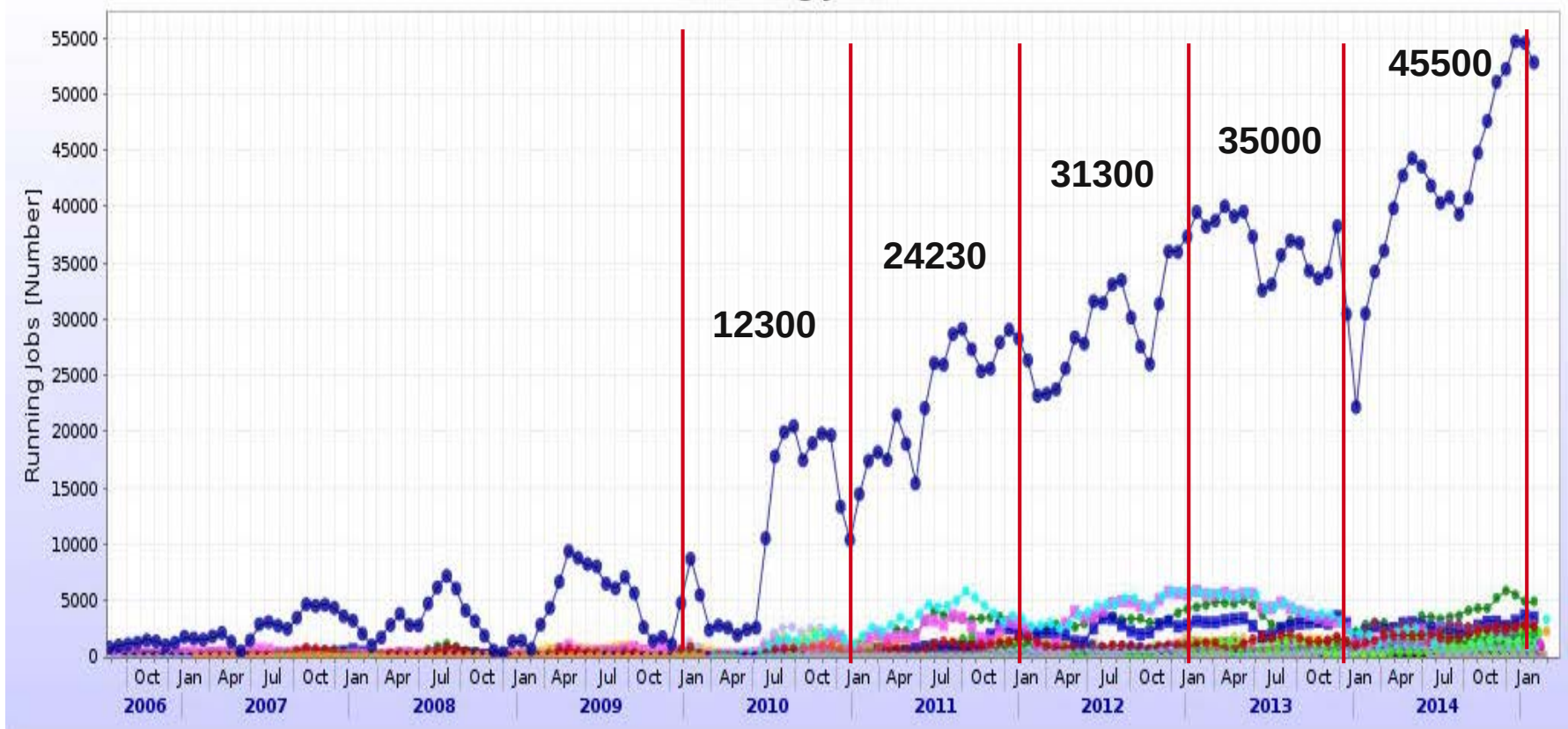
- ALICE has used Grid for the distributed production of Monte Carlo data, reconstruction and analysis at over 80 sites
 - Elaborated and well debugged job state machine
 - Used for execution of all workflows in ALICE
 - File Catalog accounting for over 10^9 files
- So far more than 280 million jobs have been successfully run worldwide from the central Task Queue (TQ), resulting in the currently active volume of 50 PB of data
 - System scaled from 500 to 80000 concurrently running jobs
- The job handling model based on pilot jobs and late binding to the real job has proven to be the solution for the large scale computing problems of the LHC experiments
 - Central Task Queue governing priorities
- Easy to adapt to new trends in computing
 - IaaS, volunteer computing, opportunistic use of HPC resources



Running jobs trend

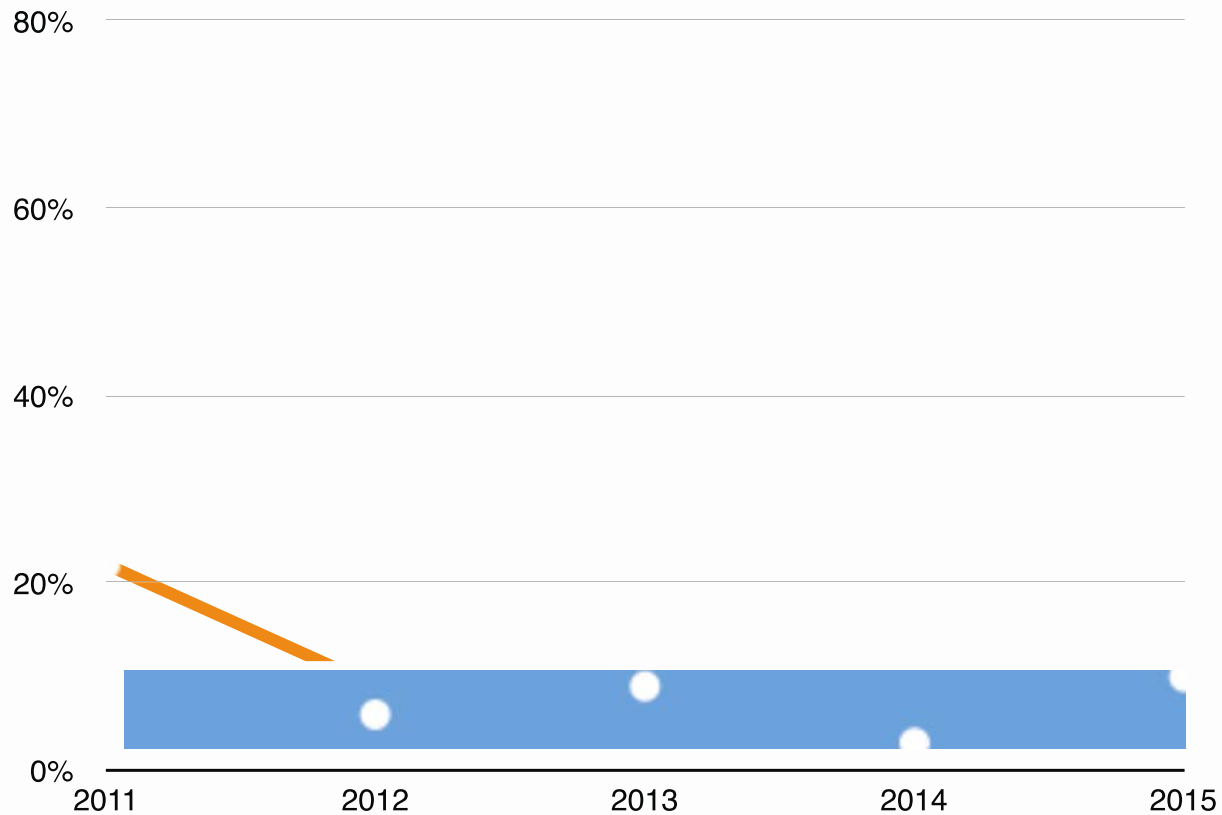


ALICE Grid evolution



- 22% average per year – slightly above the WLCG projection
- We assume that the trend will continue (20% year on year growth)

ALICE Workflows on the Grid



New in Run 3: O2 facility

- + 463 FPGAs
 - Detector readout and fast cluster finder
- + 100'000 CPU cores
 - To compress 1.1 TB/s data stream by overall factor 14
- + 5000 GPUs
 - To speed up the reconstruction
 - 3 CPU¹⁾ + 1 GPU²⁾ = 28 CPUs
- + 50 PB of disk
 - To buy us an extra time and allow more precise calibration



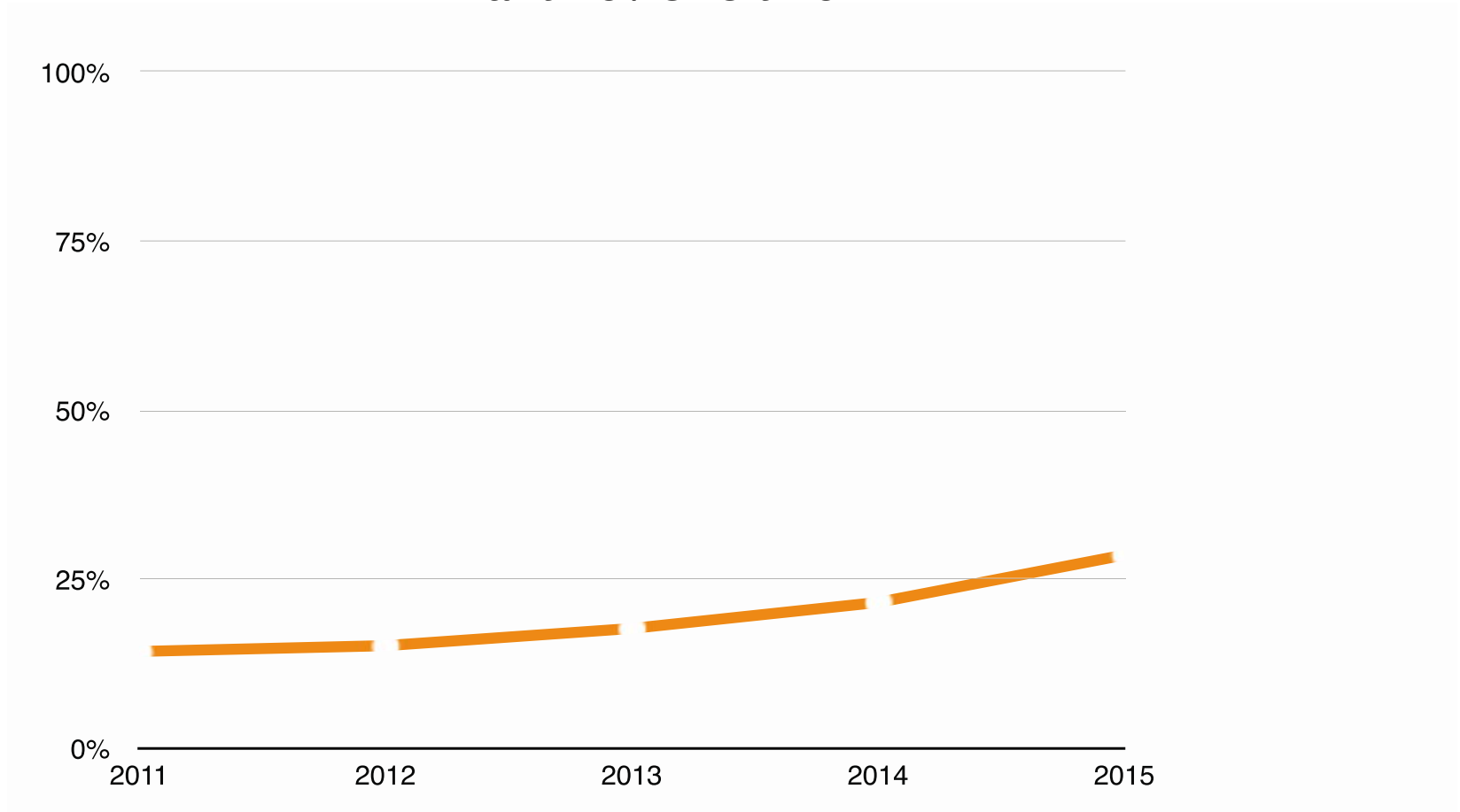
-
- = Considerable (but heterogeneous) computing capacity that will be used for Online and Offline tasks
- ✧ Identical s/w should work in Online and Offline environments

¹⁾ Intel Sandy Bridge, 2GHz, 8 core, E5-2650, 22.5 HEPSPEC06 per core

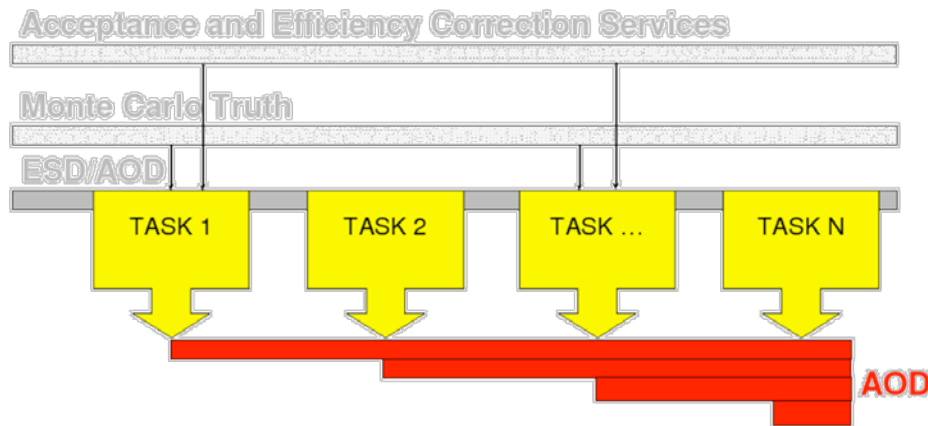
²⁾ AMD S9000

Job efficiency in Run 1

Wall time / CPU time

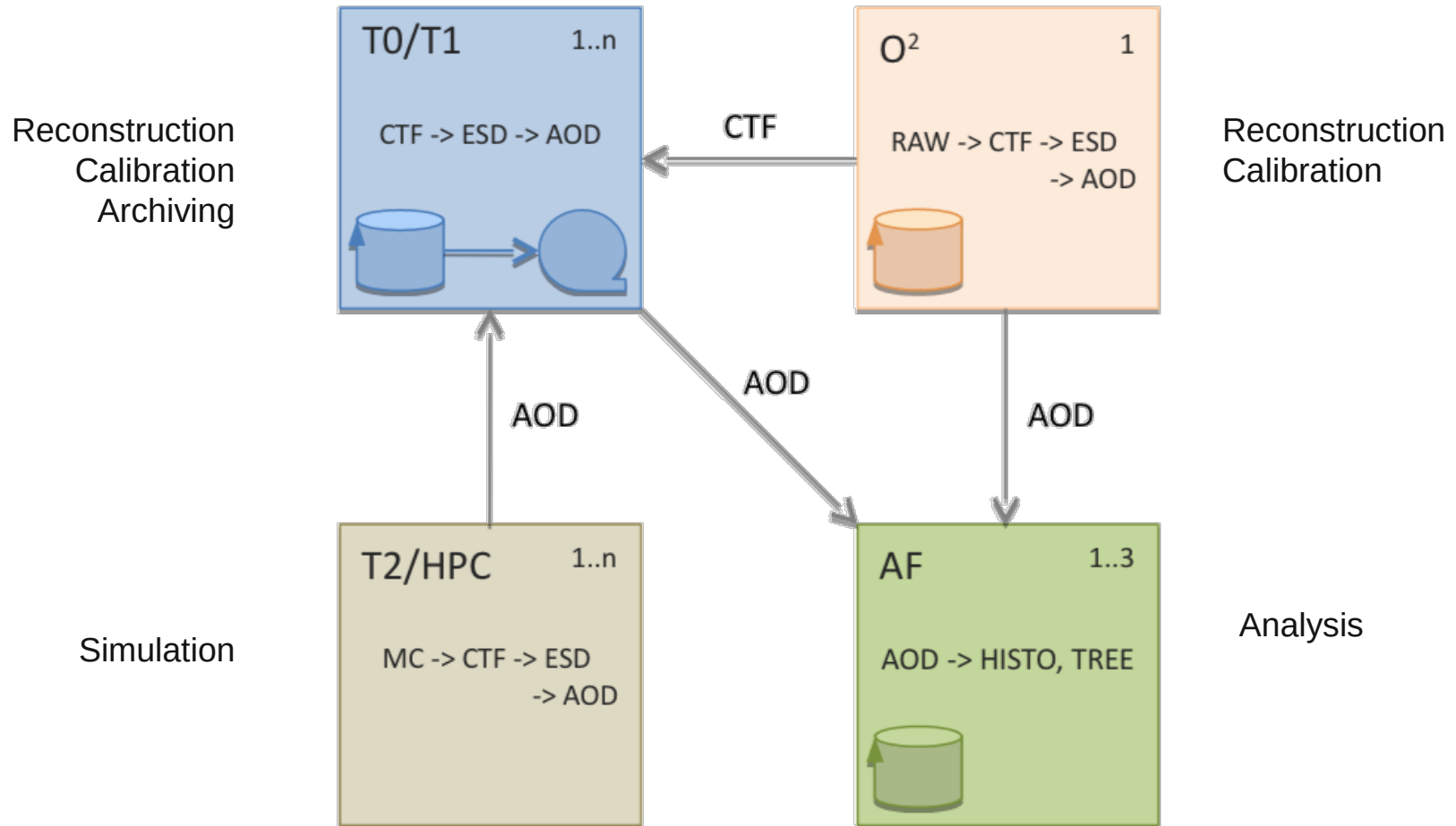


New in Run 3: Analysis Facilities

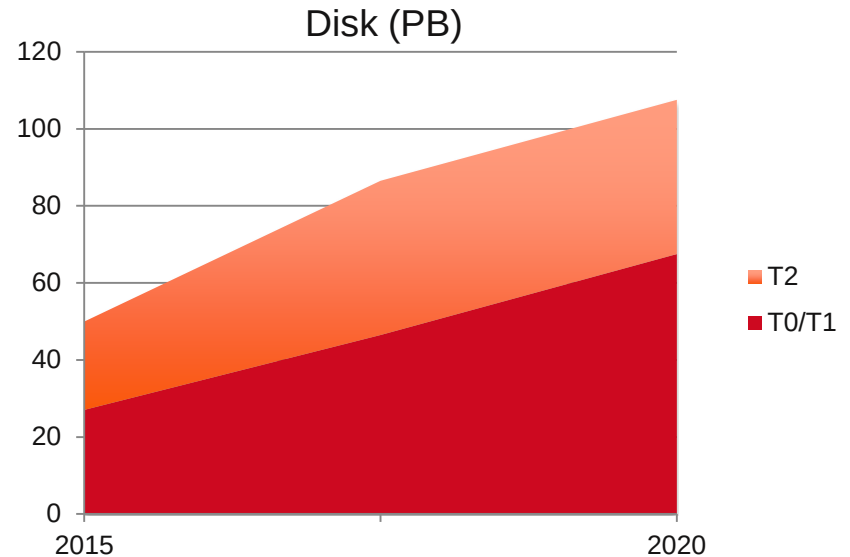
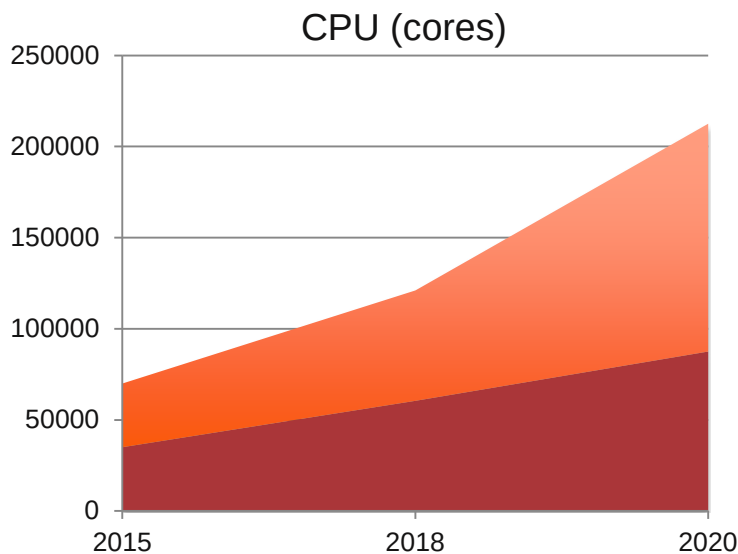


- Motivation
 - Analysis is the least efficient of all workloads that we run on the Grid
 - I/O bound in spite of attempts to make it more efficient by using the analysis trains
 - Requires merging of partial outputs which is a complex and error prone in Grid environment
 - Increased data volume will only magnify the problem
- Solution
 - Collect AODs on a dedicated sites that are capable of locally processing quickly large data volume
 - Requires 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

Roles of Tiers in Run 3

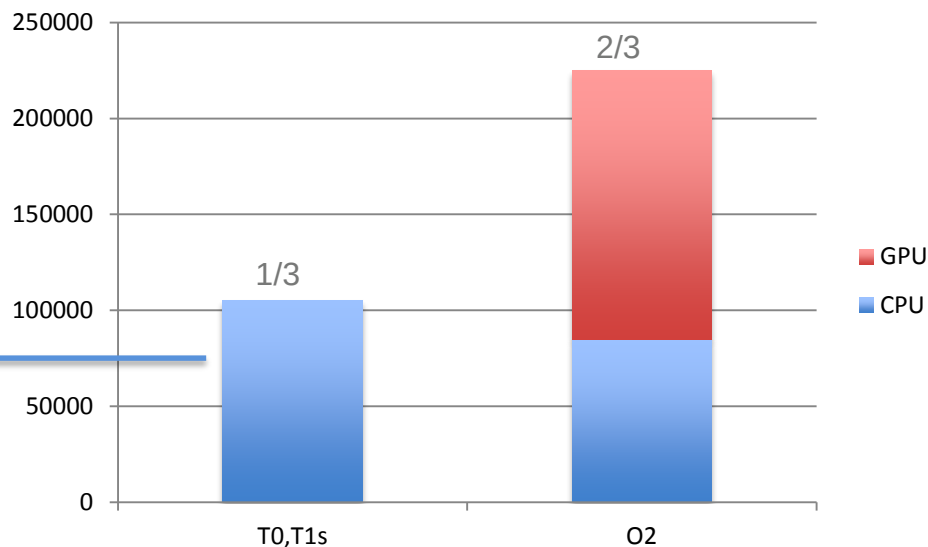
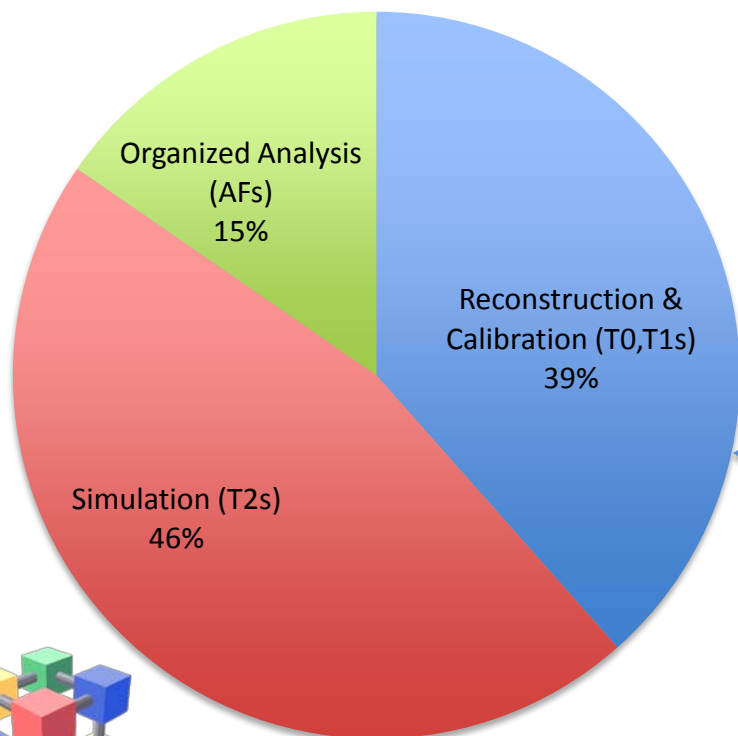


Expected evolution of ALICE Grid share



- Expecting Grid resources (CPU, storage) to grow at 20% per year rate
 - Large number of disk will be used by Run 1 and Run 2 data
- 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 needs for storage on T2s
 - Use available funding to buy more CPUs

Shares of various workflows on the Grid

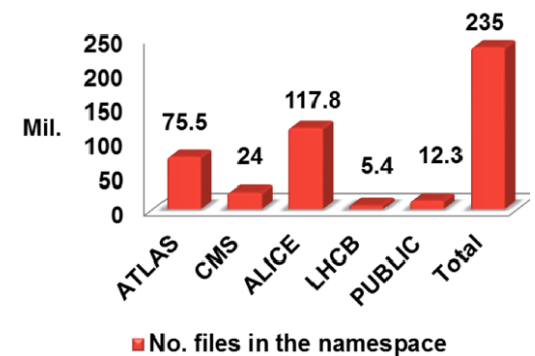
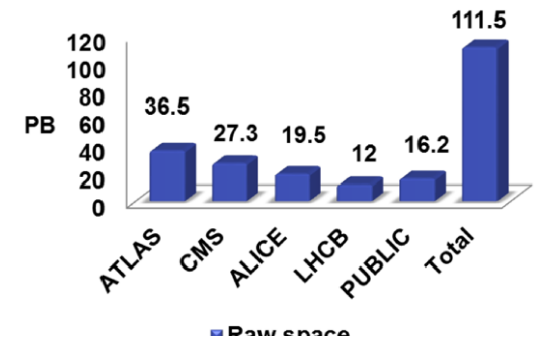


- Relative importance of O2 vs Grid may change with time if Grid resources continue to grow as expected



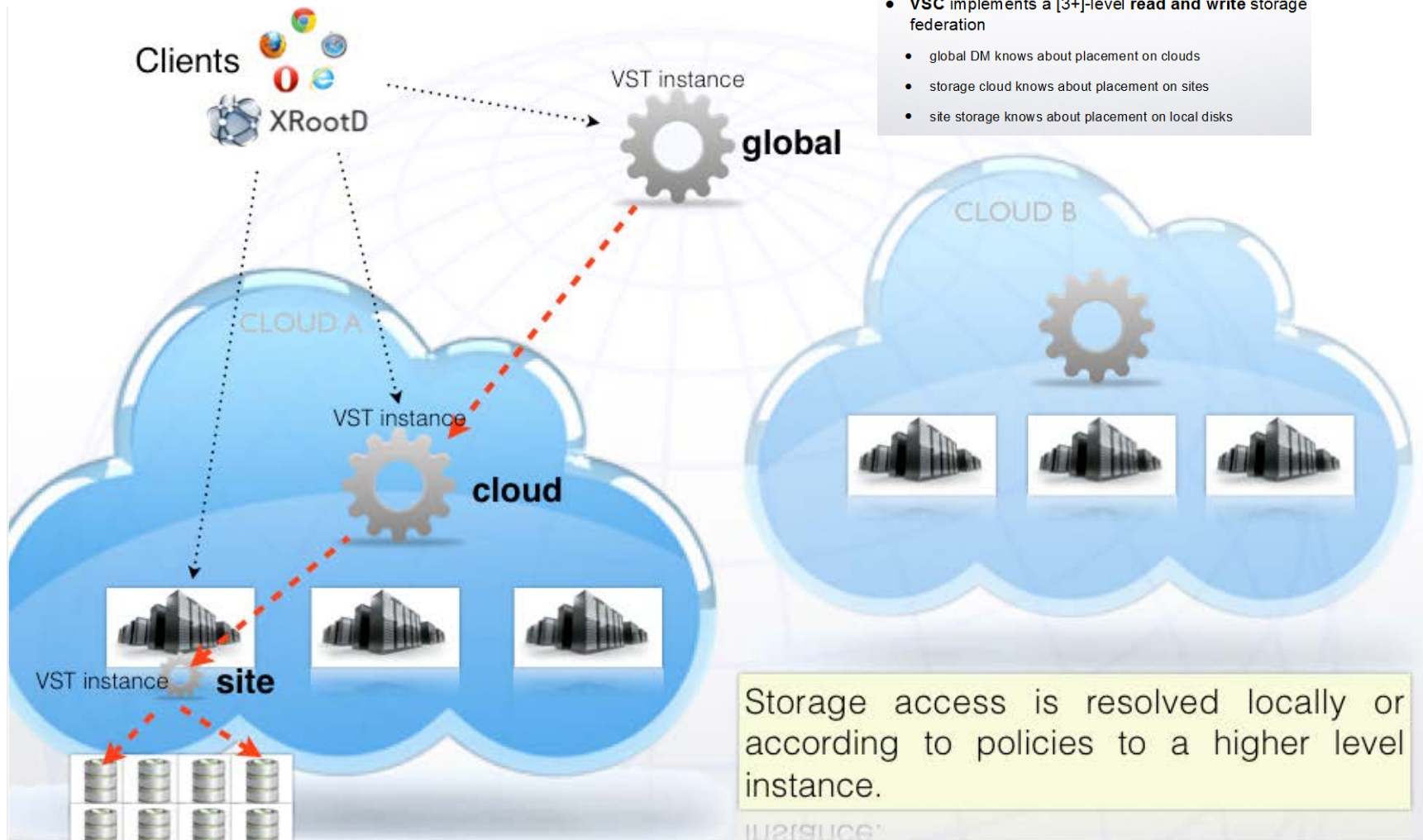
Data management

- In RUN 3, it will have to scale to much higher levels in terms of number of files, overall size and I/O performance
 - We need a uniform solution supported across the Grid
- Our approach would be to evolve the current data management model (based on a network of xrootd Storage Elements) to a generalized global namespace based on the EOS system
 - Currently used to manage about 111.5PB of RUN 1 data for all four large LHC experiments.
 - 37400 disks
- This system is itself based on the next iteration of xrootd technology and allows for gradual and non-disruptive transition



Complexity management

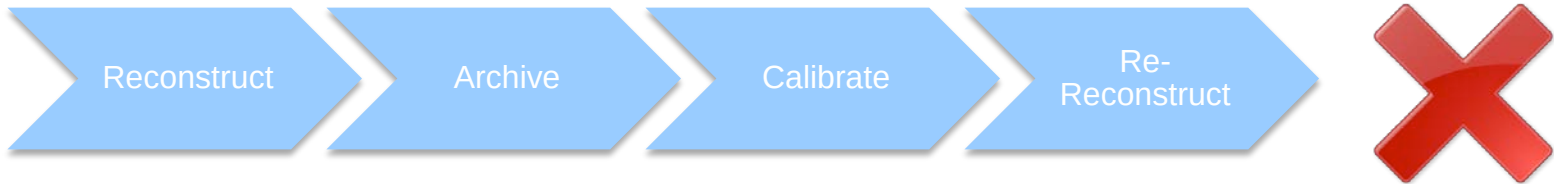
- instead of scheduling individual disks, **schedule sites and clouds** (clouds of clouds...)
- a site or a cloud is represented by a virtual storage node called **VST**
- **VSC** implements a [3+]-level **read and write** storage federation
 - global DM knows about placement on clouds
 - storage cloud knows about placement on sites
 - site storage knows about placement on local disks



(No)Replication policy

- Due to a substantial increase in data volume, in RUN 3, there will be only one instance of each raw data file (CTF) stored on disk with a backup on tape
 - EOS with CEPH backend with erasure encoding provides 20% overhead in available storage space with 100% availability in large installations such as CERN
 - In case of data loss, we will restore lost files from the tape
 - The size of O2 disk buffer should be sufficient to accommodate CTF data from the entire period.
 - As soon as it is available, the CTF data will be archived to the Tier 0 tape buffer or moved to the Tier 1s
- => Expecting 20 Gbps to Tier1s during data taking

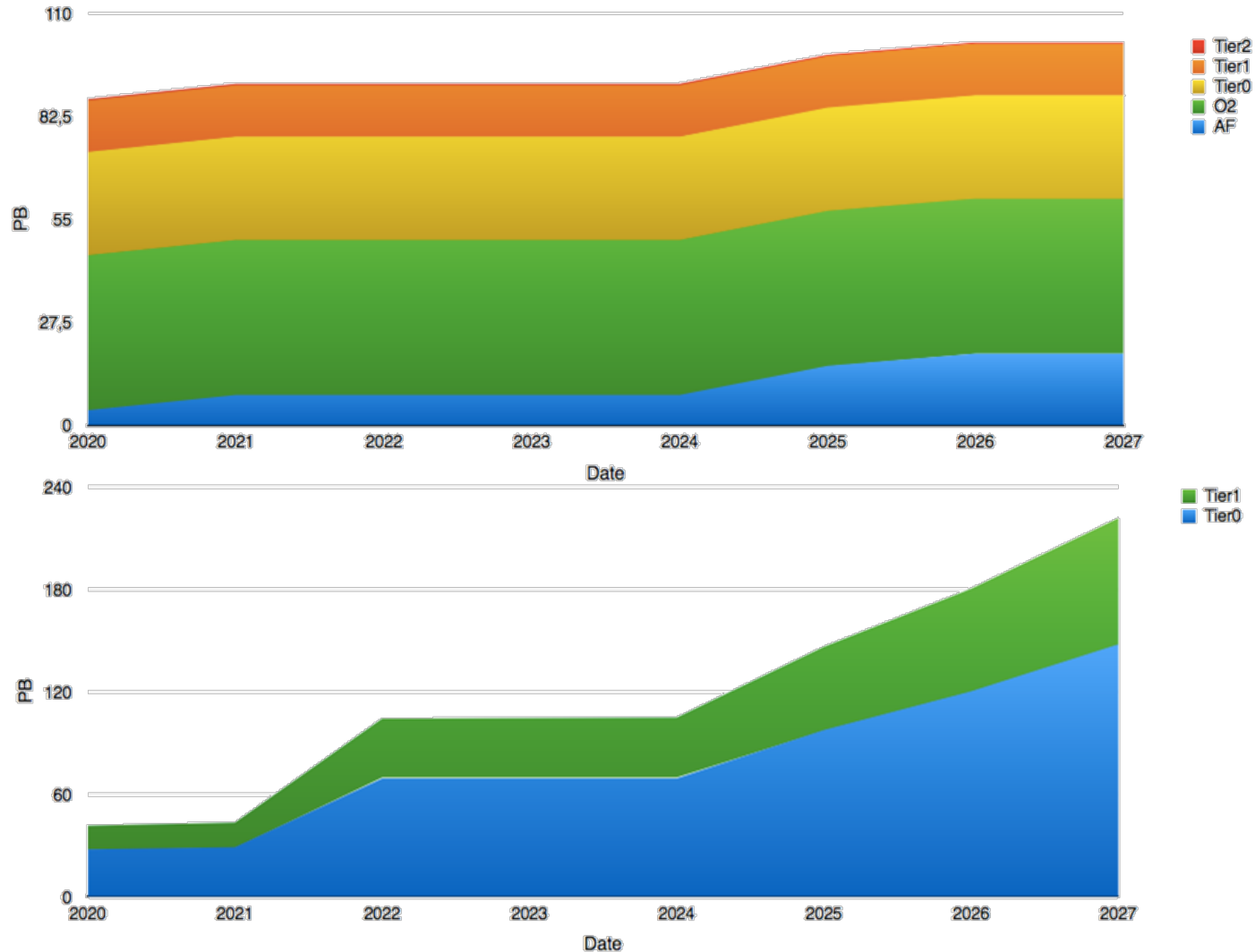
Deletion policy



- With the exception of raw data (CTF) and derived analysis data (AOD), all other intermediate data created at various processing stages is transient (removed after a given processing step) or temporary (with limited lifetime)
 - CTF and AODs are archived to tape
- Given the limited size of the disk buffers in O2 and Tier 1s, all CTF data collected in the previous year, will have to be removed before new data taking period starts.
- All data not finally processed during this period will remain parked on tapes until the next opportunity for re-processing arises: LS3



Expected tape and disk usage



Conclusions

- We aim to minimize amount of data moving between Tiers and carry out most of processing on local datasets
 - The primary goal if O2 facility is data compression
- We expect the Grid to continue growing by 20% per year
 - 2/3 of raw data processing to be done in O2 facility and 1/3 on T1s
 - Limited by the network and storage/archive capacity at T1s
- Since T2s will be used mostly for simulation and we can rebalance CPU/disk ratio in favor of more CPUs
 - Reduces operational cost at T2
 - Increases CPU capacity available for simulation
- Analysis Facilities are new component in our Computing Model that are meant to improve the analysis efficiency and assure guaranteed turnaround time for organized analysis



Backup slides

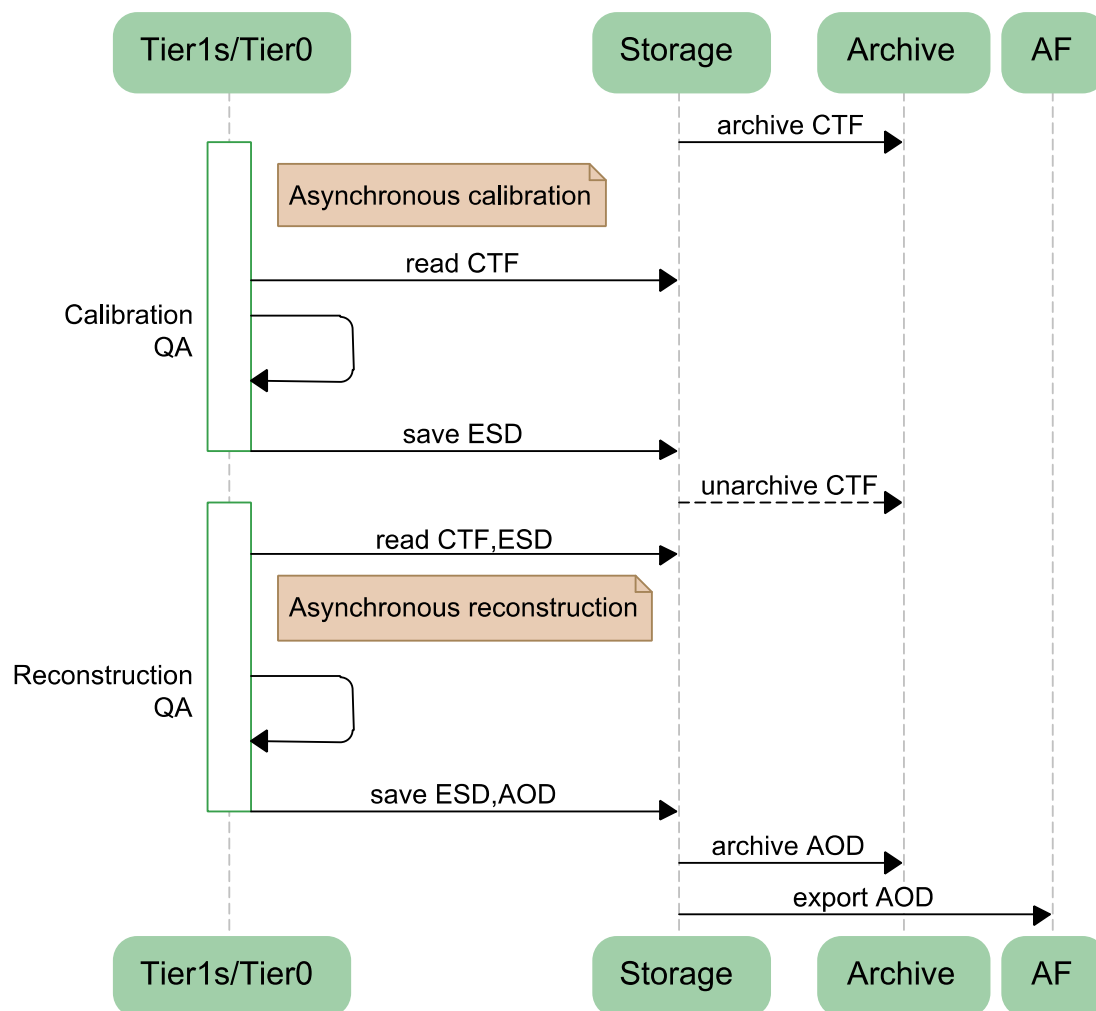
Running scenario

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



Tier 2 Workflow

