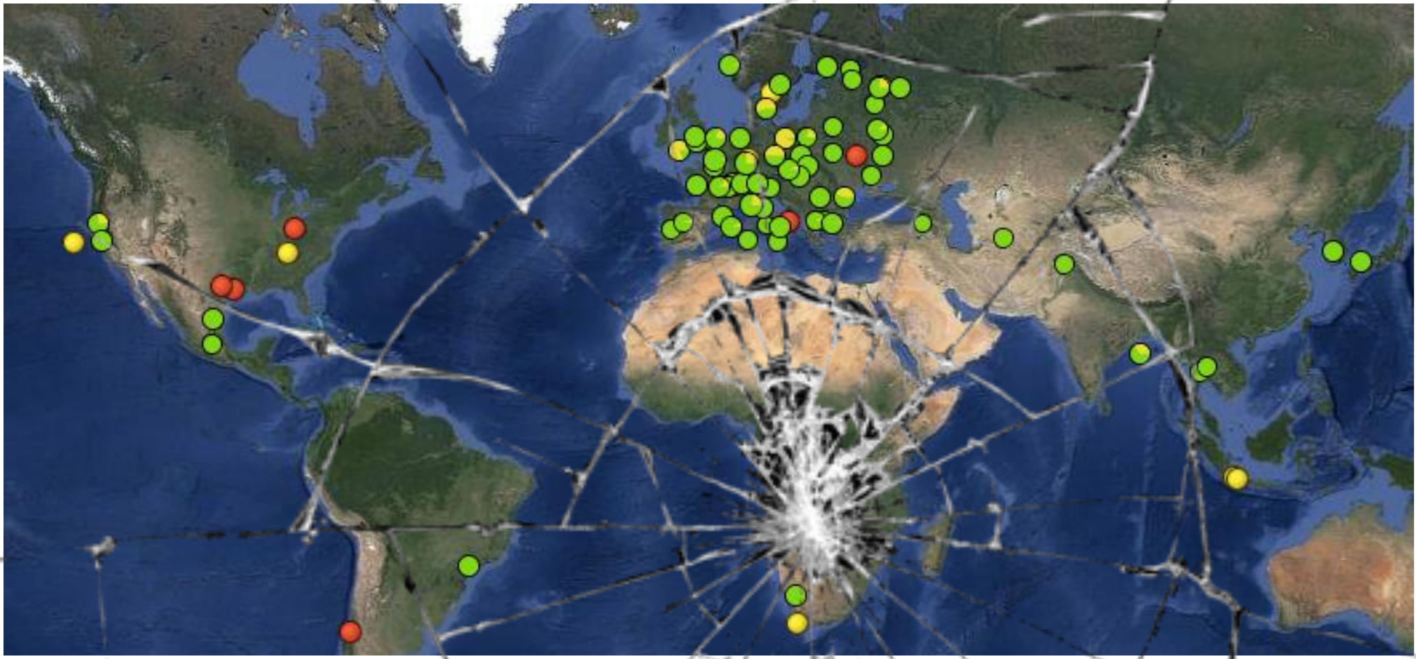
ALICE Computing Upgrade

Predrag Buncic

Computing Model in Run 1&2

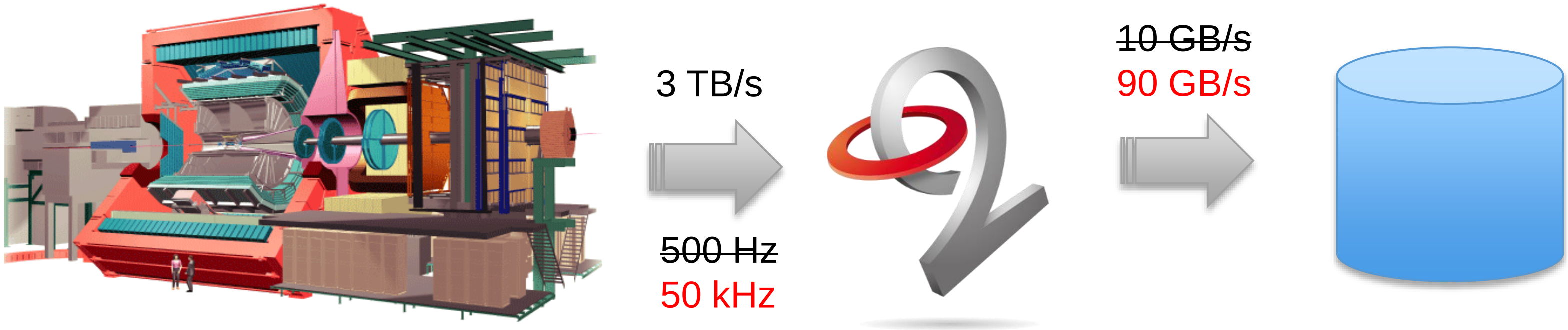


- Computing model built on top of WLCG services
  - Any job can run anywhere and access data from any place
  - Scheduling takes care of optimizations and brings "computing to data"
  - Every file and its replicas are accounted in the file catalogue
- Worked (surprisingly) well during Run 2 and Run 2

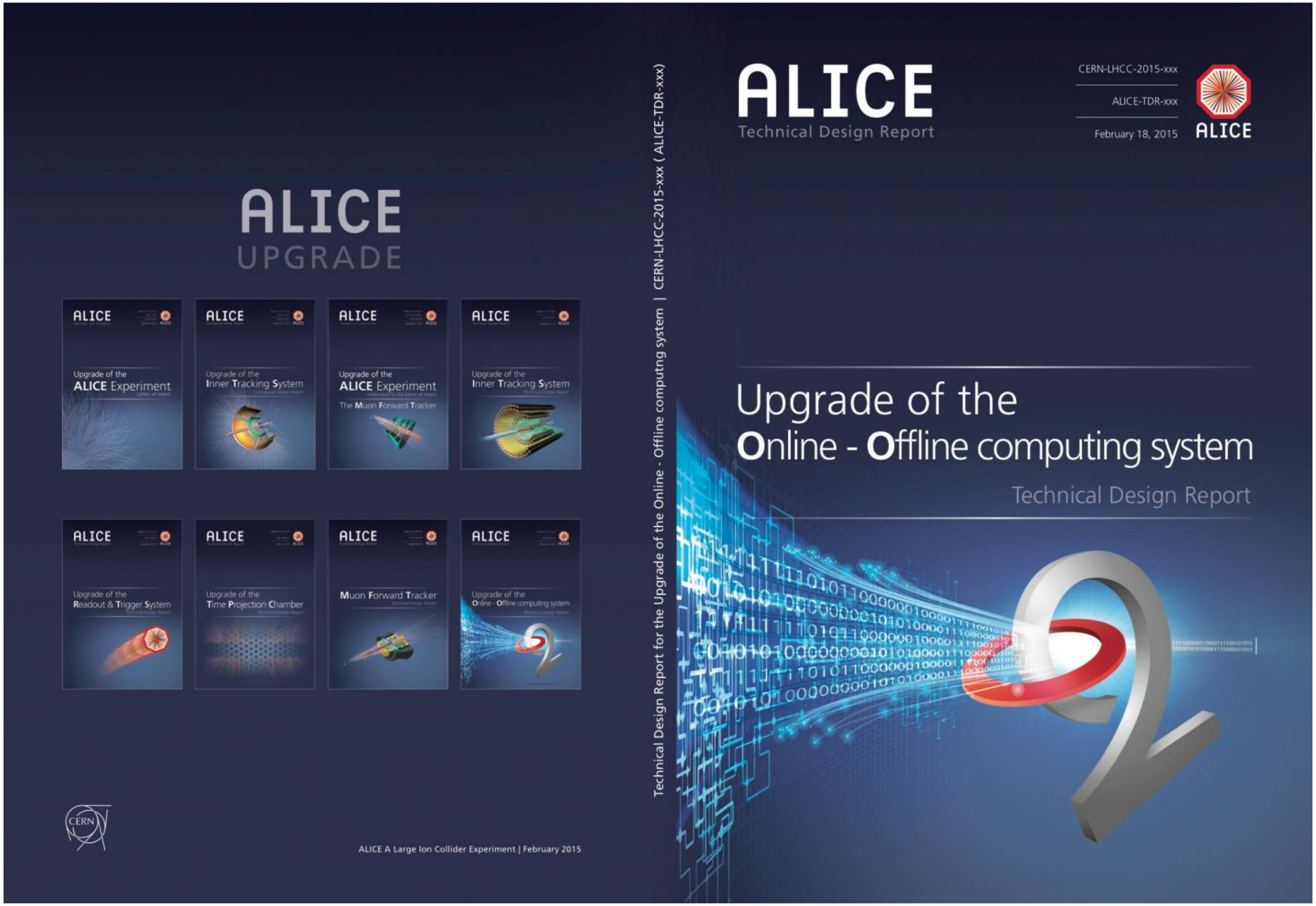
Run 3 data taking objectives

- For Pb-Pb collisions:
  - Reach the target of ~~1~~ 13 nb^{-1} integrated luminosity in Pb-Pb for rare triggers.

- The resulting data throughput from the detector has been estimated to be greater than 1TB/s for Pb–Pb events, roughly two orders of magnitude more than in Run 1



Factor 90 in terms of Pb-Pb events (x 30 for pp)



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 computng 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

http://goo.gl/T4NeUp

Run 3 Computing Model

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

+ 3000 GPUs
- To speed up the reconstruction
- 3 CPU[1)] + 1 GPU[2)] = 28 CPUs

+ 60 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

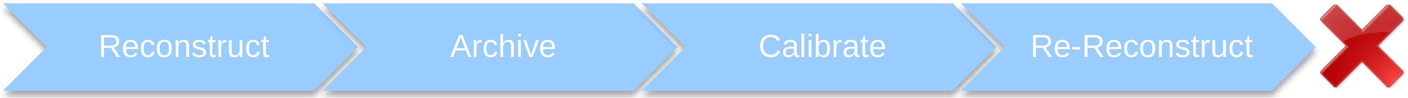
✧ Identical s/w should work in Online and Offline environments

[1)] Intel Sandy Bridge, 2GHz, 8 core, E5-2650
[2)] AMD S9000

Changes to data management policies

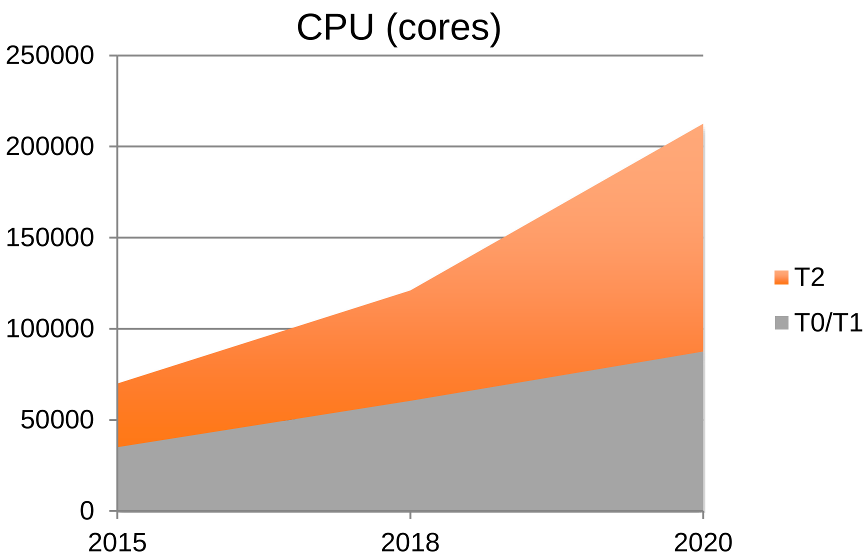
- Only one instance of each raw data file (CTF) stored on disk with a backup on tape
  - In case of data loss, we will restore lost files form the tape
  - O2 disk buffer should be sufficient accommodate CTF data from the entire period.
  - As soon as it is available, the CTF data will be archived to the Tier 0 tape buffer or moved to the Tier 1s

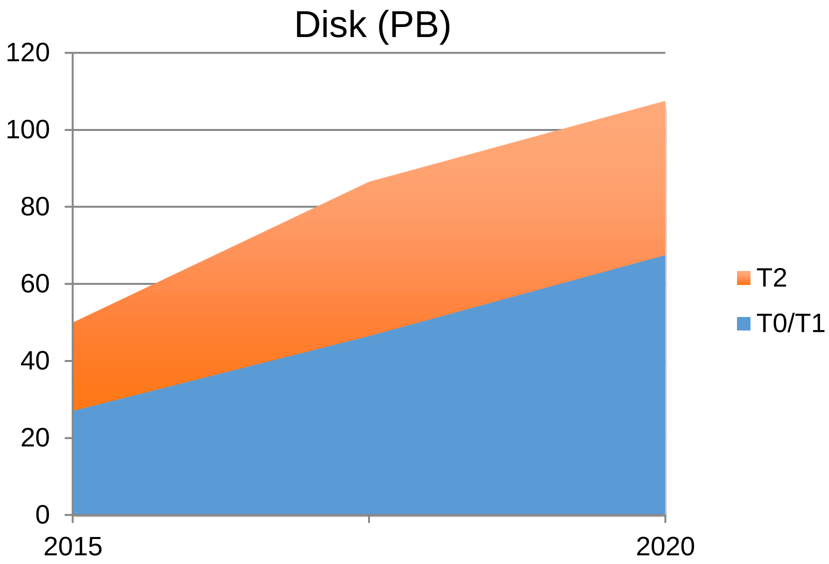
- All other intermediate data created at various processing stages is transient (removed after a given processing step) or temporary (with limited lifetime)
  - Only CTF and AODs are archived kept on disk to tape



- 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.

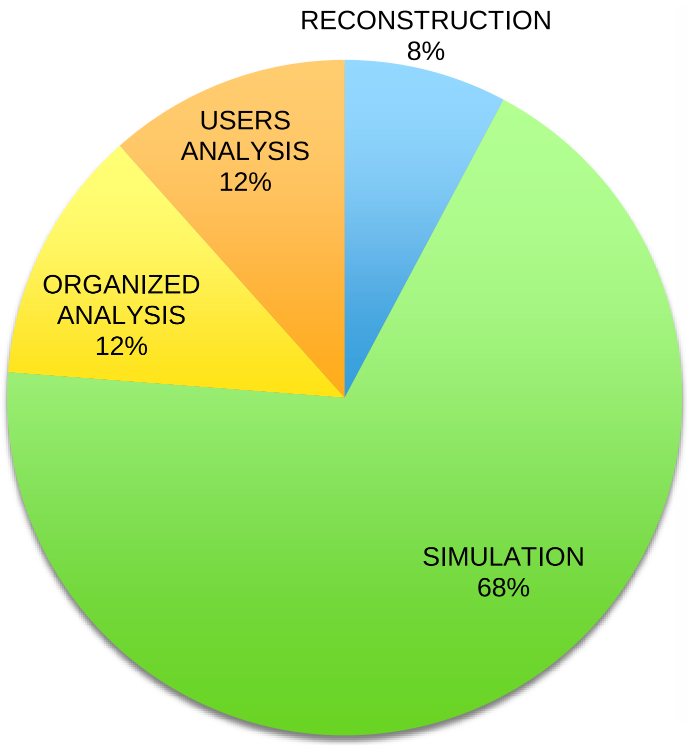
Rebalancing disk/CPU ratio on T2s

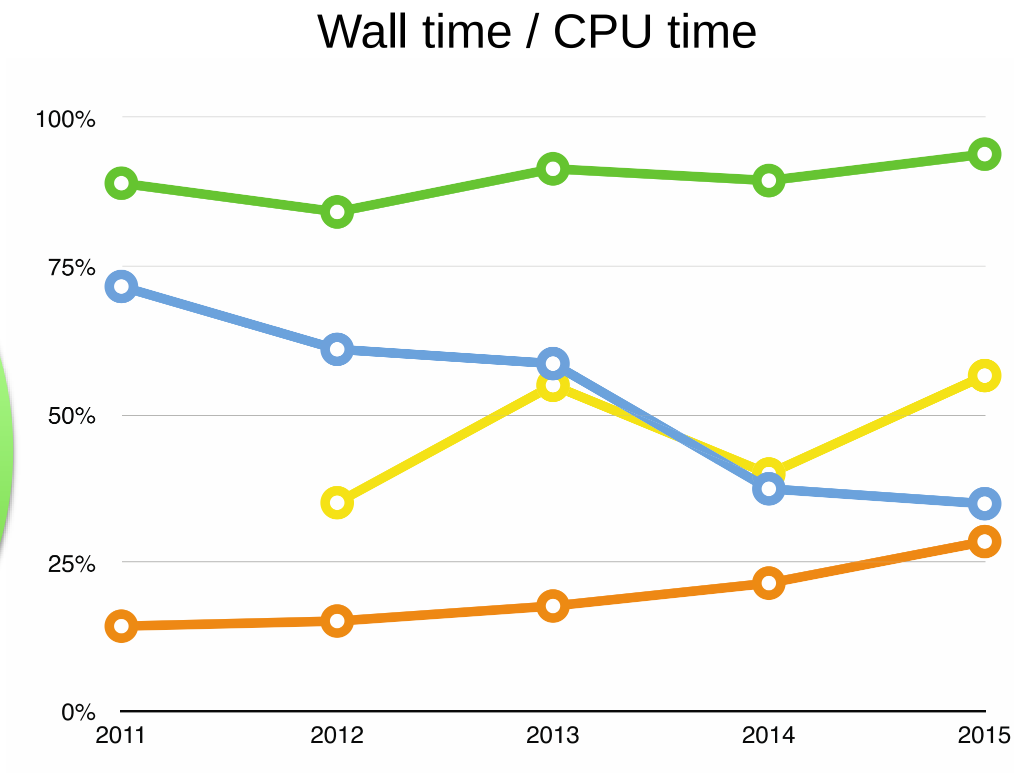




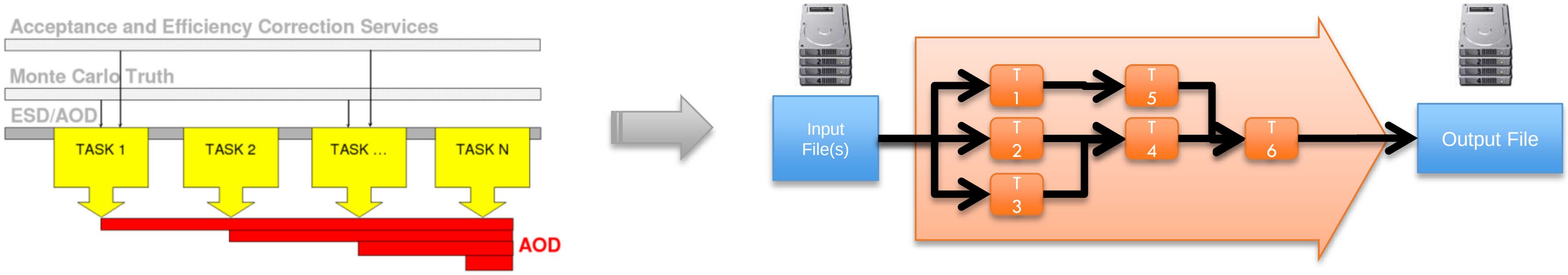
- Expecting Grid resources (CPU, storage) to grow at 20% per year rate
  - Large number of disk will be used by Run 1 and Run 2 data
- Since T2s will be used almost exclusively for simulation jobs (no input) and resulting AODs will be exported to T1s/AFs, we expect to significantly lower needs for storage on T2s

Optimizing Efficiency



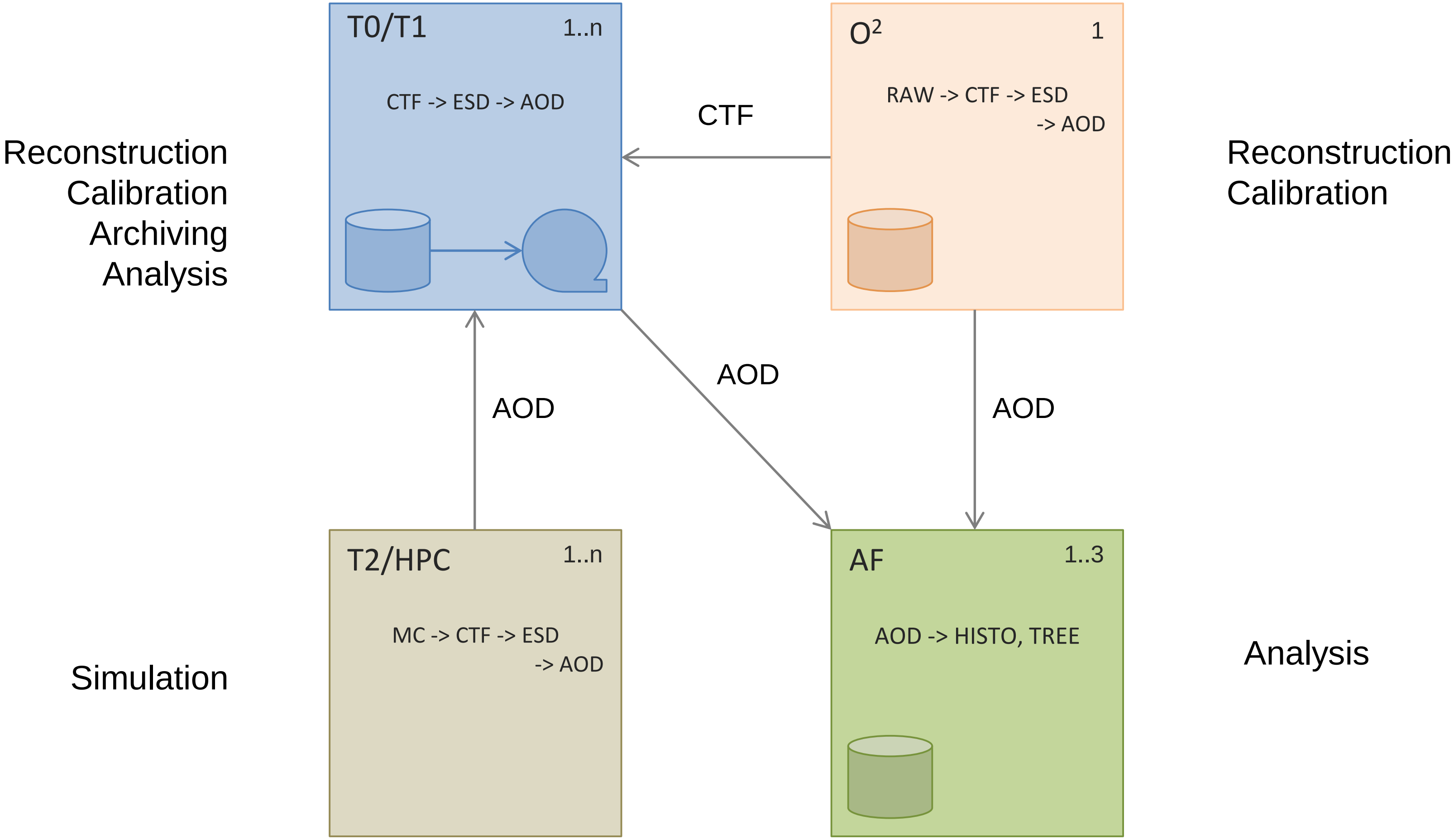


Analysis Facilities



- Motivation
  - Analysis remains /O bound in spite of attempts to make it more efficient by using the train approach
- Solution
  - Collect AODs on a dedicated sites that are optimized for fast processing of a large local datasets
  - Run organized analysis on local data like we do today on the Grid
  - Requires 20-30’000 cores and 5-10 PB of disk on very performant file system
  - Such sites can be elected between the existing T1s (or even T2s) but ideally this would be a purpose build facility optimized for such workflow

Roles of Tiers in Run 3



Reconstruction
Calibration
Archiving
Analysis

T0/T1 1..n

CTF -> ESD -> AOD

CTF

O^2 1

RAW -> CTF -> ESD
-> AOD

Reconstruction
Calibration

AOD

AOD

AOD

Simulation

T2/HPC 1..n

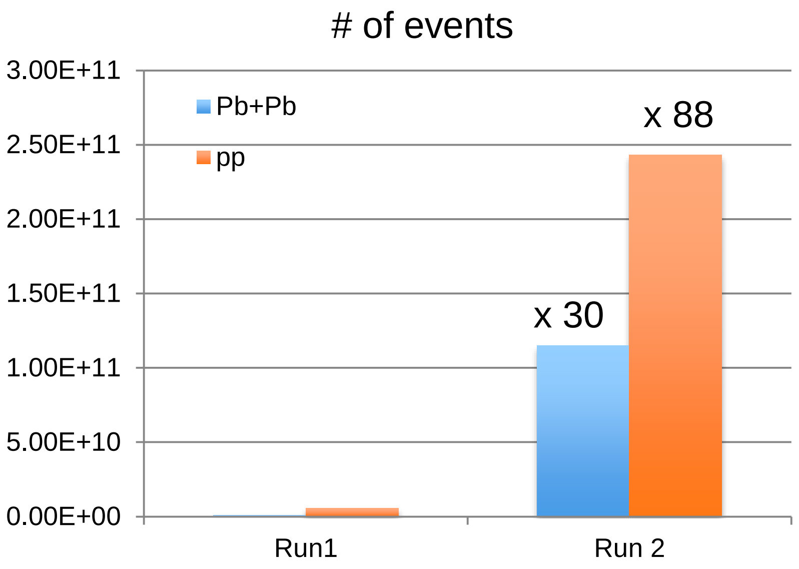
MC -> CTF -> ESD
-> AOD

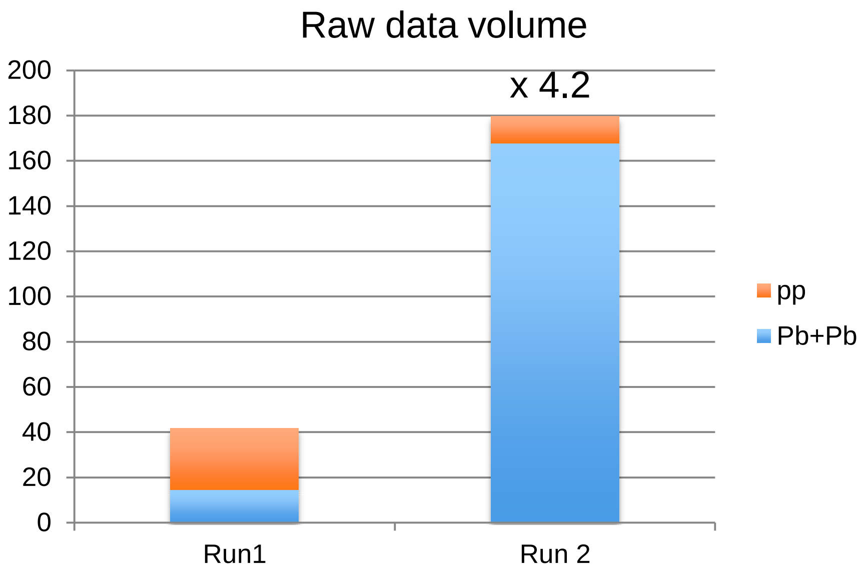
AF 1..3

AOD -> HISTO, TREE

Analysis

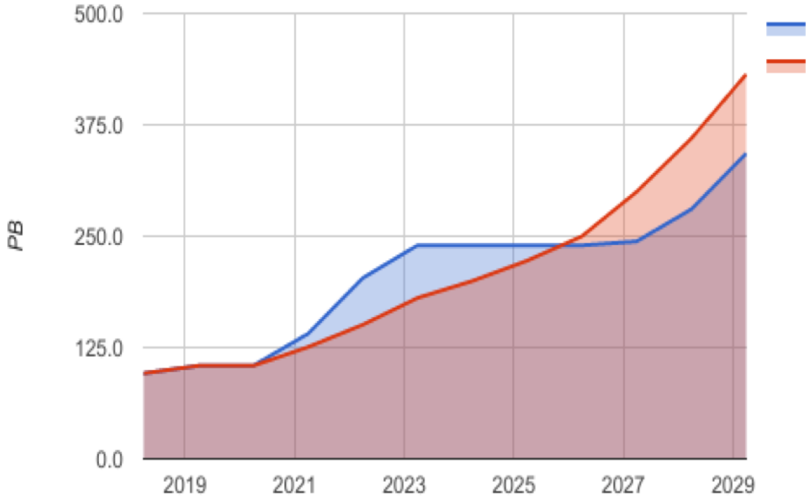
Run 2 vs Run 3: Data volume and # of events

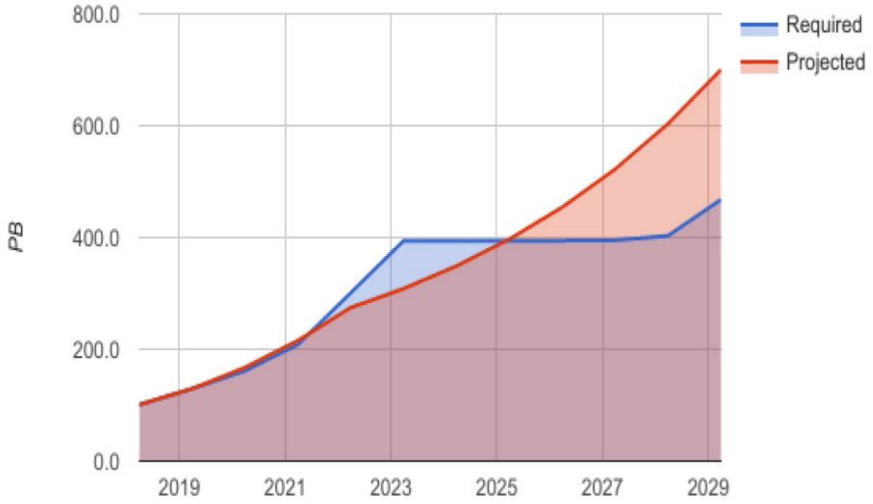


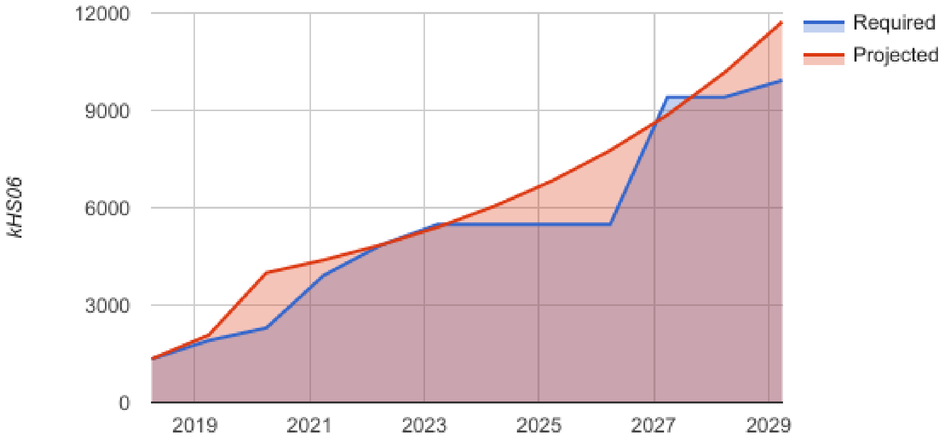


- While event statistics will increase by factor 30in Pb-Pb (x88 in pp), data volume will increase by factor 4.2
  - Thanks to data reduction in O2 facility
    - Online tracking that allows rejection of clusters not associated with tracks
    - Large effect in case of pileup (pp)

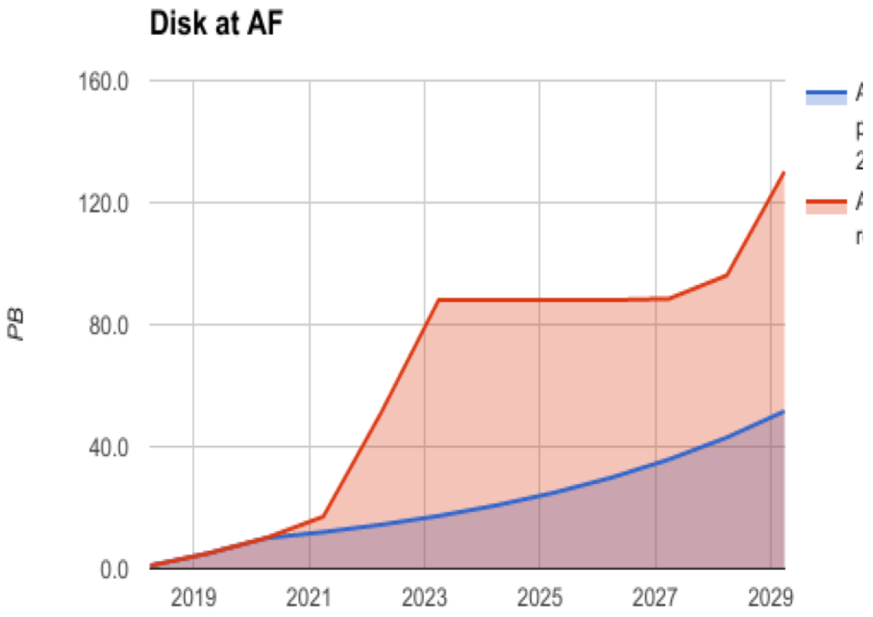
Flat budget scenario

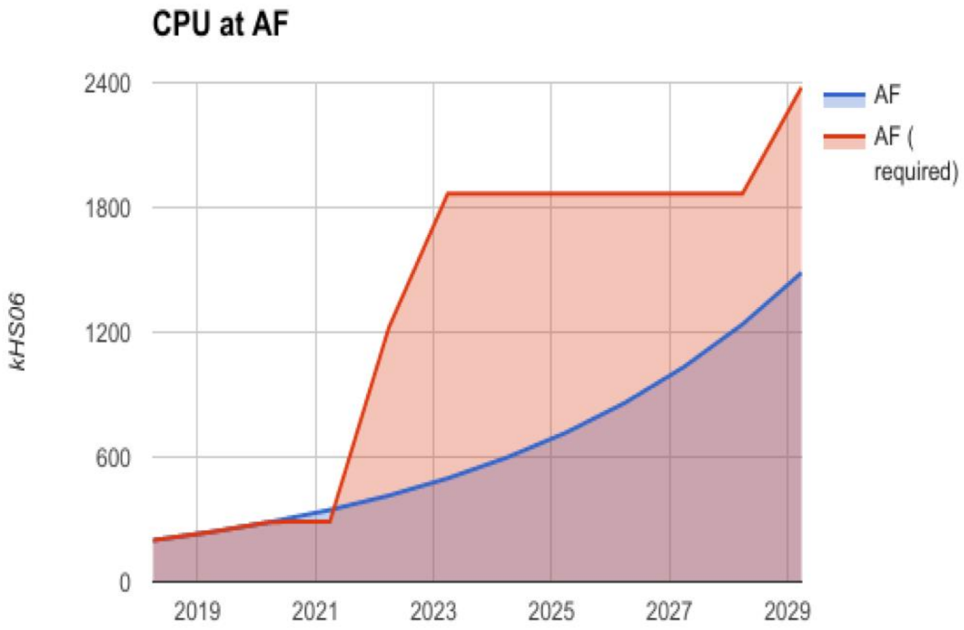

Total Tape


Total Disk


Total CPU

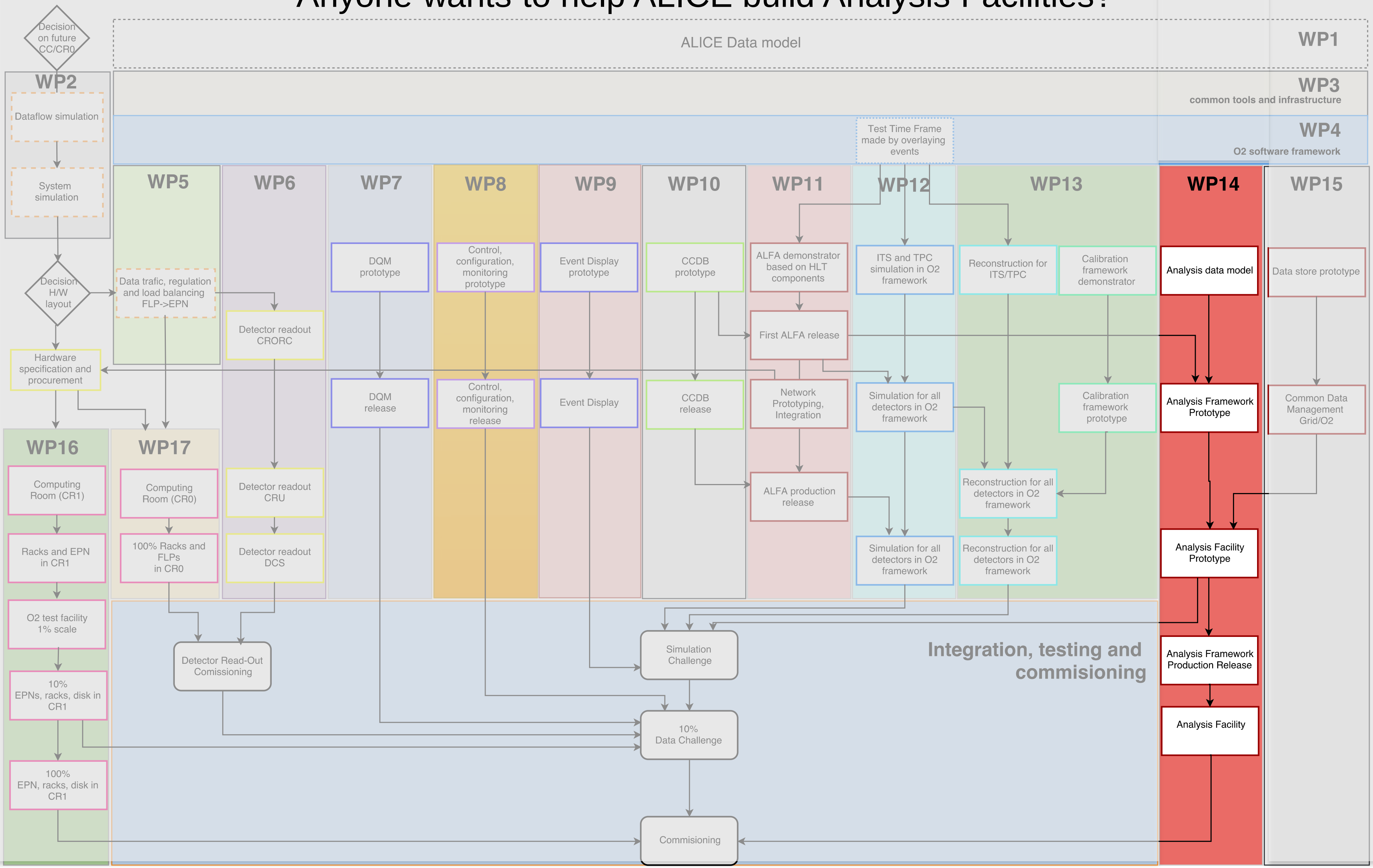
- Overall, we seem to fit under projected resource growth curves
  - Still doable under the flat budget scenario





- Analysis Facilities are new in our Computing Model
  - So far we have secured one at GSI
  - Assumed growth in disk and CPU: 10%
- We need to look for more AF facilities or re-purpose some of the T1s for that purpose

Anyone wants to help ALICE build Analysis Facilities?


Decision on future CC/CR0
ALICE Data model
WP1
WP2
Dataflow simulation
System simulation
WP3
common tools and infrastructure
WP4
O2 software framework
Test Time Frame made by overlaying events
WP5
WP6
WP7
WP8
WP9
WP10
WP11
WP12
WP13
WP14
WP15
Decision H/W layout
Data trafic, regulation and load balancing FLP->EPN
DQM prototype
Control, configuration, monitoring prototype
Event Display prototype
CCDB prototype
ALFA demonstrator based on HLT components
ITS and TPC simulation in O2 framework
Reconstruction for ITS/TPC
Calibration framework demonstrator
Analysis data model
Data store prototype
Detector readout CRORC
First ALFA release
Hardware specification and procurement
DQM release
Control, configuration, monitoring release
Event Display
CCDB release
Network Prototyping, Integration
Simulation for all detectors in O2 framework
Calibration framework prototype
Analysis Framework Prototype
Common Data Management Grid/O2
WP16
WP17
Computing Room (CR1)
Computing Room (CR0)
Detector readout CRU
ALFA production release
Reconstruction for all detectors in O2 framework
Racks and EPN in CR1
100% Racks and FLPs in CR0
Detector readout DCS
Simulation for all detectors in O2 framework
Reconstruction for all detectors in O2 framework
Analysis Facility Prototype
O2 test facility 1% scale
Detector Read-Out Comissioning
Simulation Challenge
Integration, testing and commisioning
Analysis Framework Production Release
10% EPNs, racks, disk in CR1
10% Data Challenge
Analysis Facility
100% EPN, racks, disk in CR1
Commisioning

O2 Plenary | ALICE Week | March 2017 | Predrag Buncic

http://goo.gl/QZUE1d