A Large Ion Collider Experiment

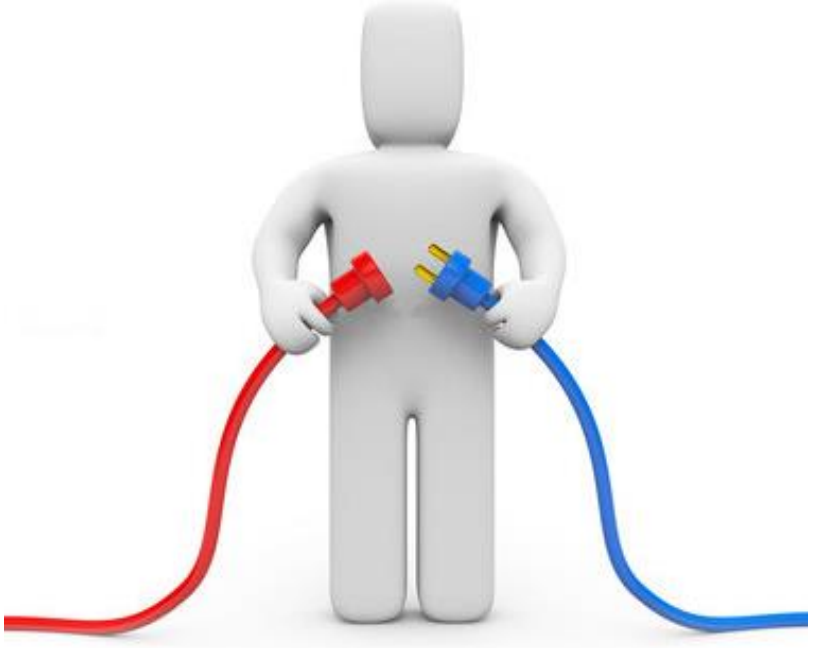
ALICE Computing Model Run2 (and beyond)

Predrag Buncic

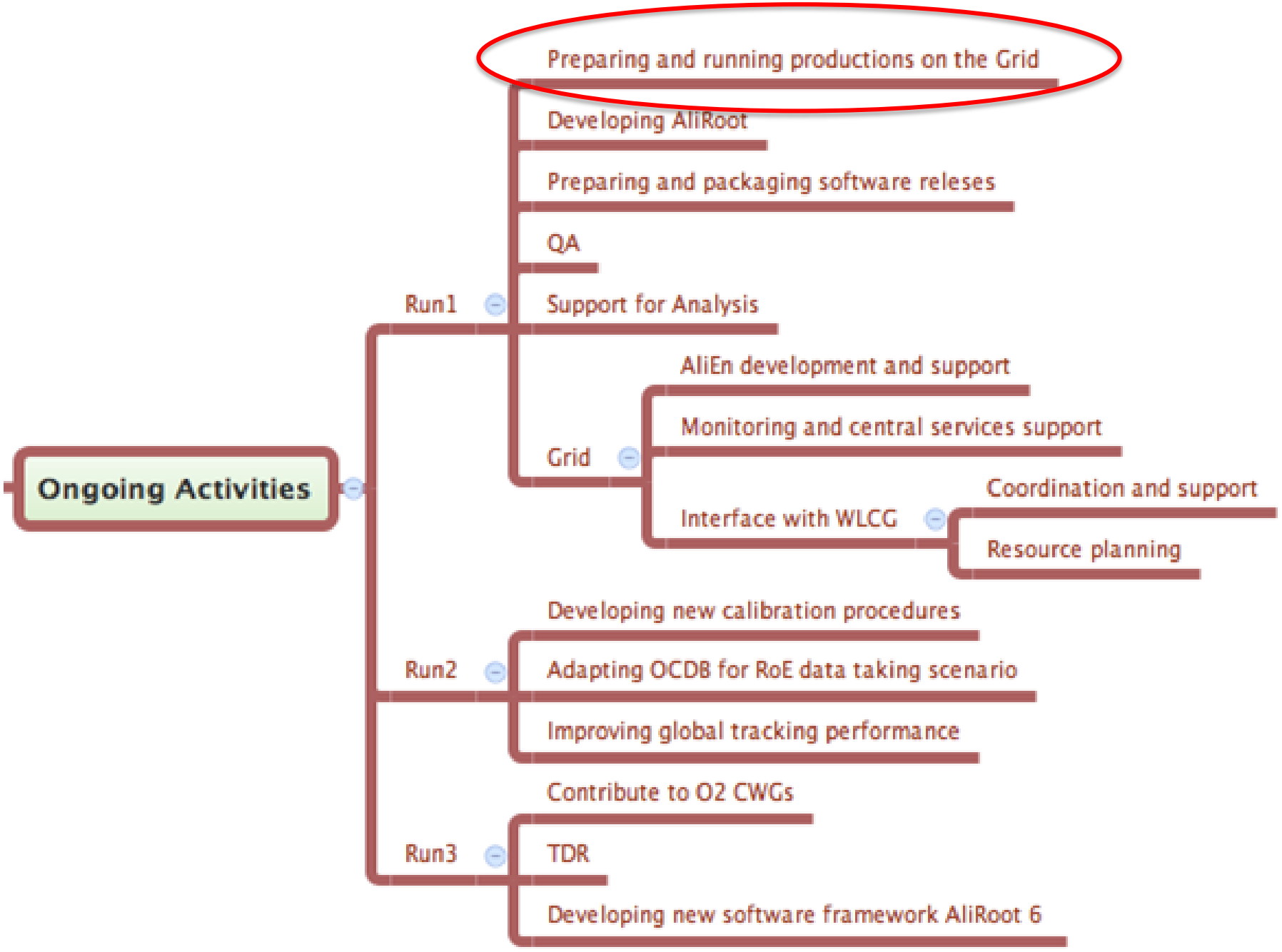
ALICE Computing Model in Run2

- … is not going to change.

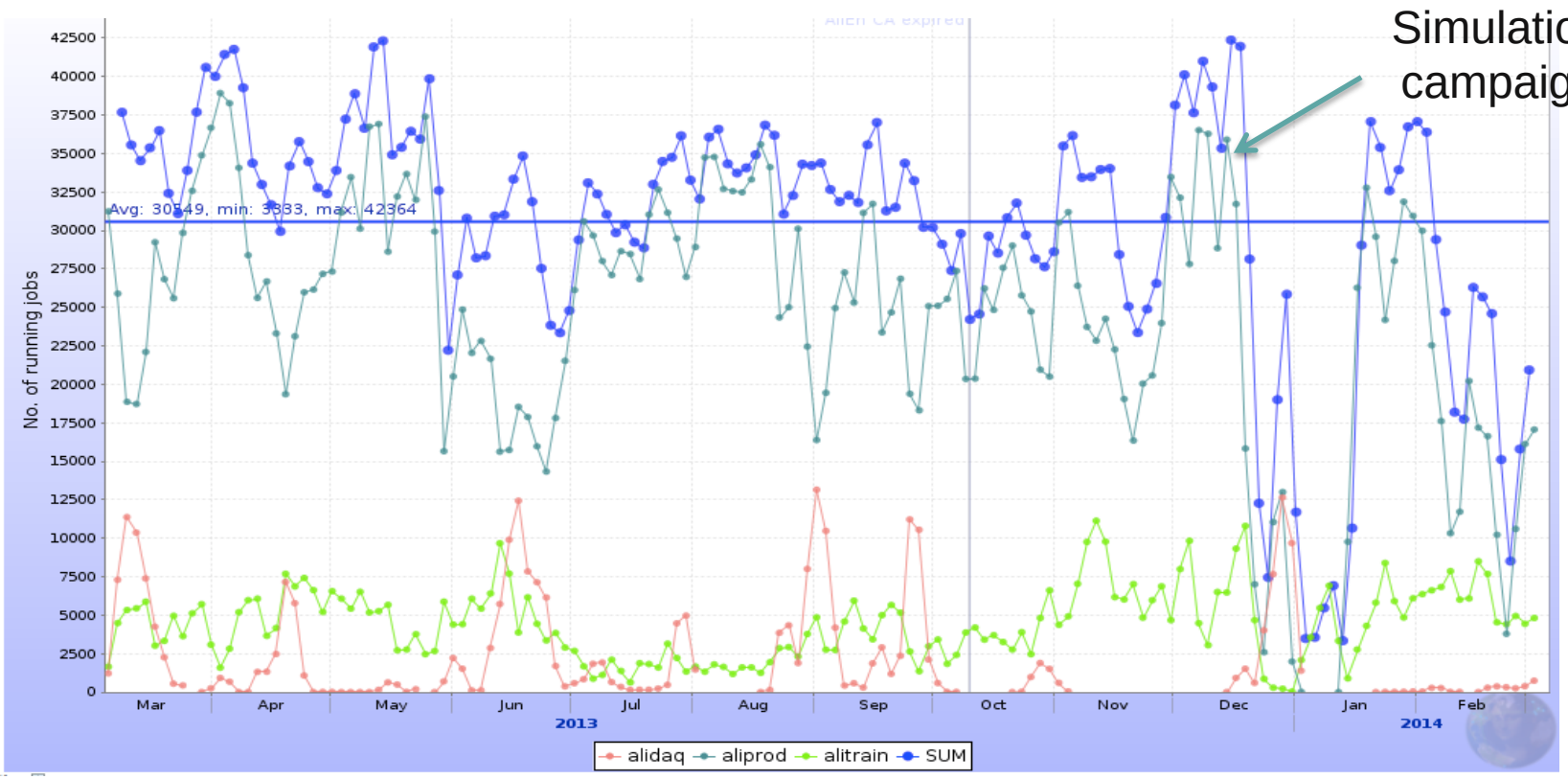
- Most of the ALICE software stack will remain unchanged
  - With exception of necessary tracking performance and calibration improvements
- This is not because
  - we think that all our problems are solved
  - we not ambitious enough
- This is simply because we do not have enough manpower to carry out at the same time Run1 support and Run2&3 preparations



What keeps ALICE Offline busy and what are the plans for future?


Ongoing Activities
Run1
Preparing and running productions on the Grid
Developing AliRoot
Preparing and packaging software releses
QA
Support for Analysis
Grid
AliEn development and support
Monitoring and central services support
Interface with WLCG
Coordination and support
Resource planning
Run2
Developing new calibration procedures
Adapting OCDB for RoE data taking scenario
Improving global tracking performance
Run3
Contribute to O2 CWGs
TDR
Developing new software framework AliRoot 6

ALICE jobs running on the grid in past 12 months



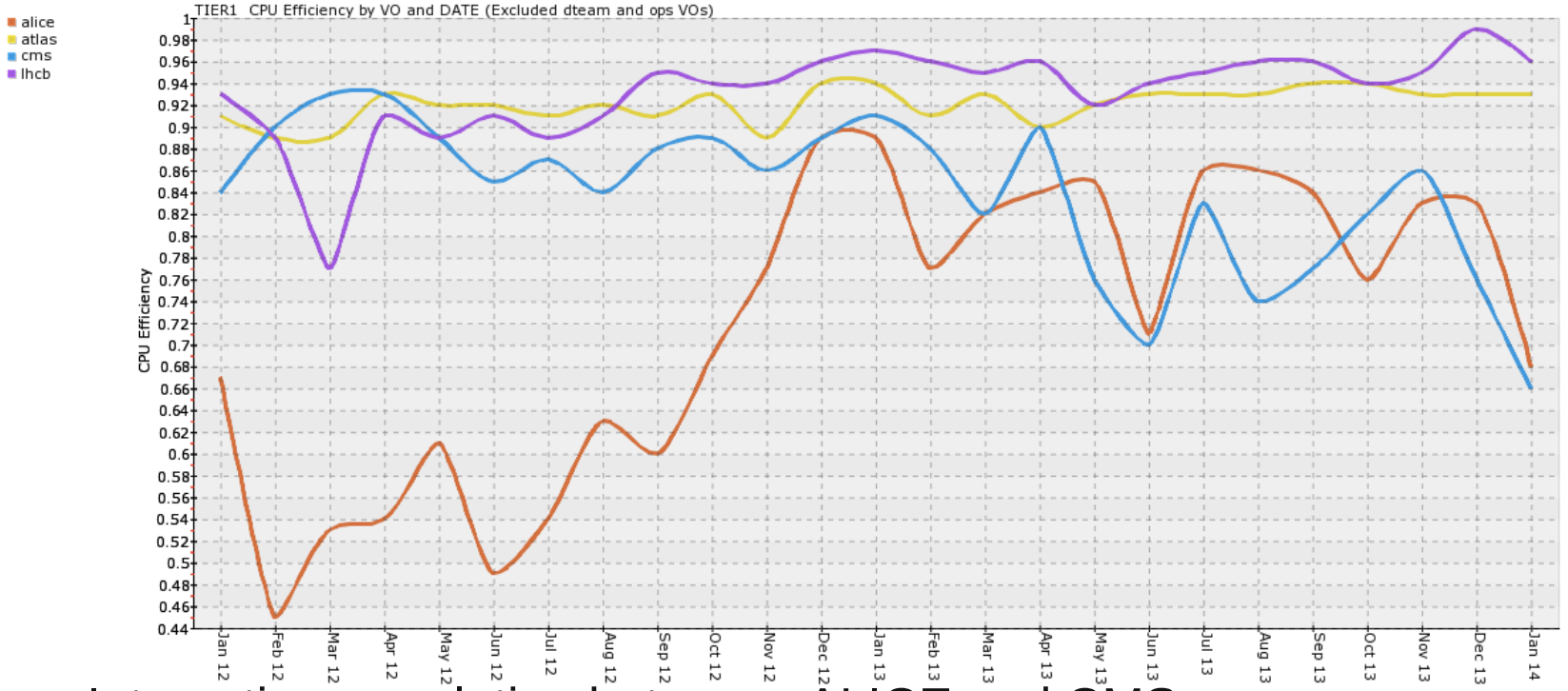
- Better than expected simulation production campaign in the run-up to Christmas
- Planned raw data re-processing postponed due to issues discovered in new software release

ALICE

XXIV QUARK MATTER
DARMSTADT 2014

- Expect to see peek of last minute requests for various kind of productions
- At the same time, analysis activities will probably reach all time high levels
- Using this tome to iron out remaining inefficiencies in computing infrastructure in addition to all ongoing Run1 related operations, Run2 preparations and Run3 activities

Job Efficiency on T1s (CPU/Wall time)

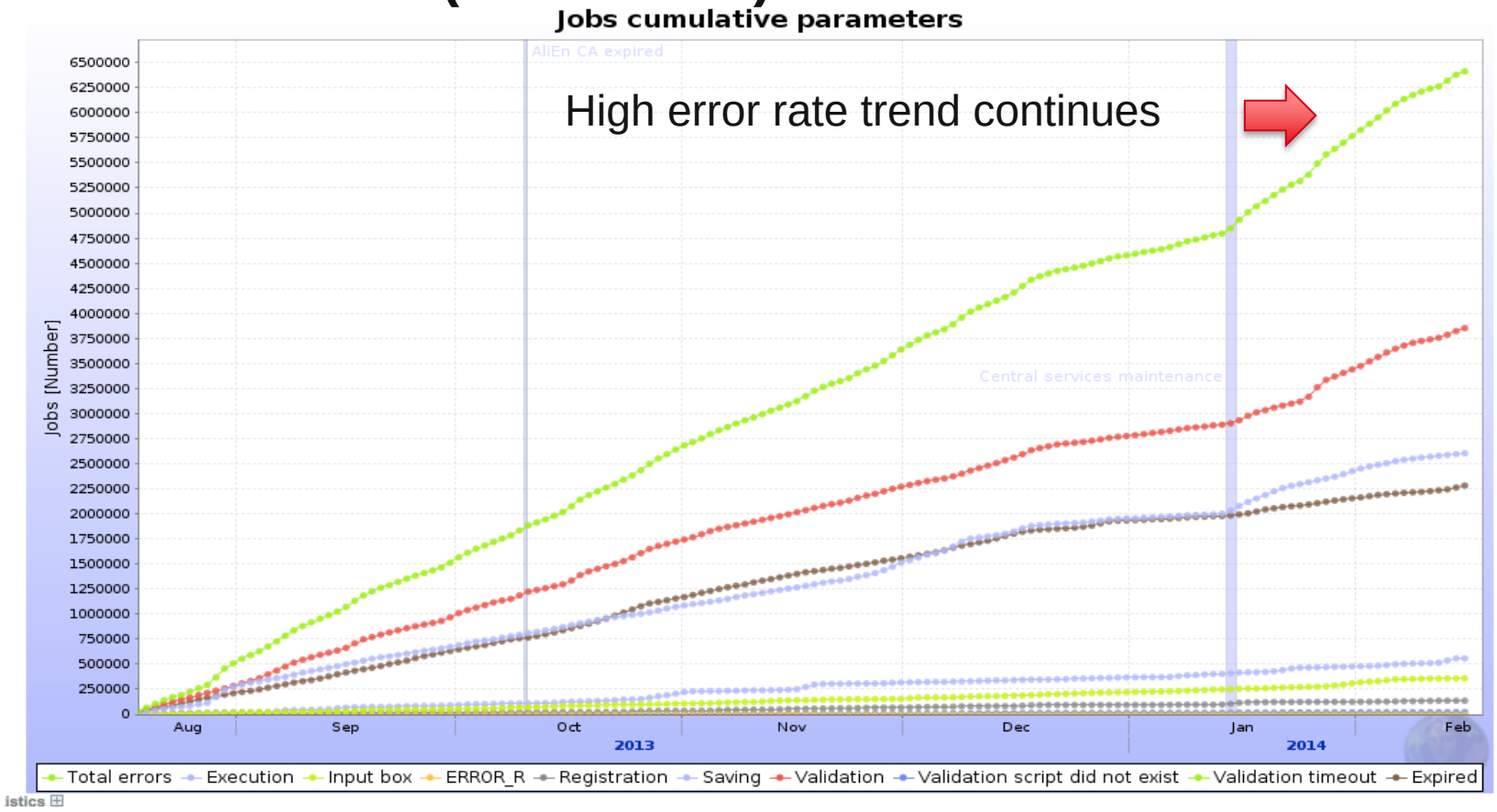


Interesting correlation between ALICE and CMS

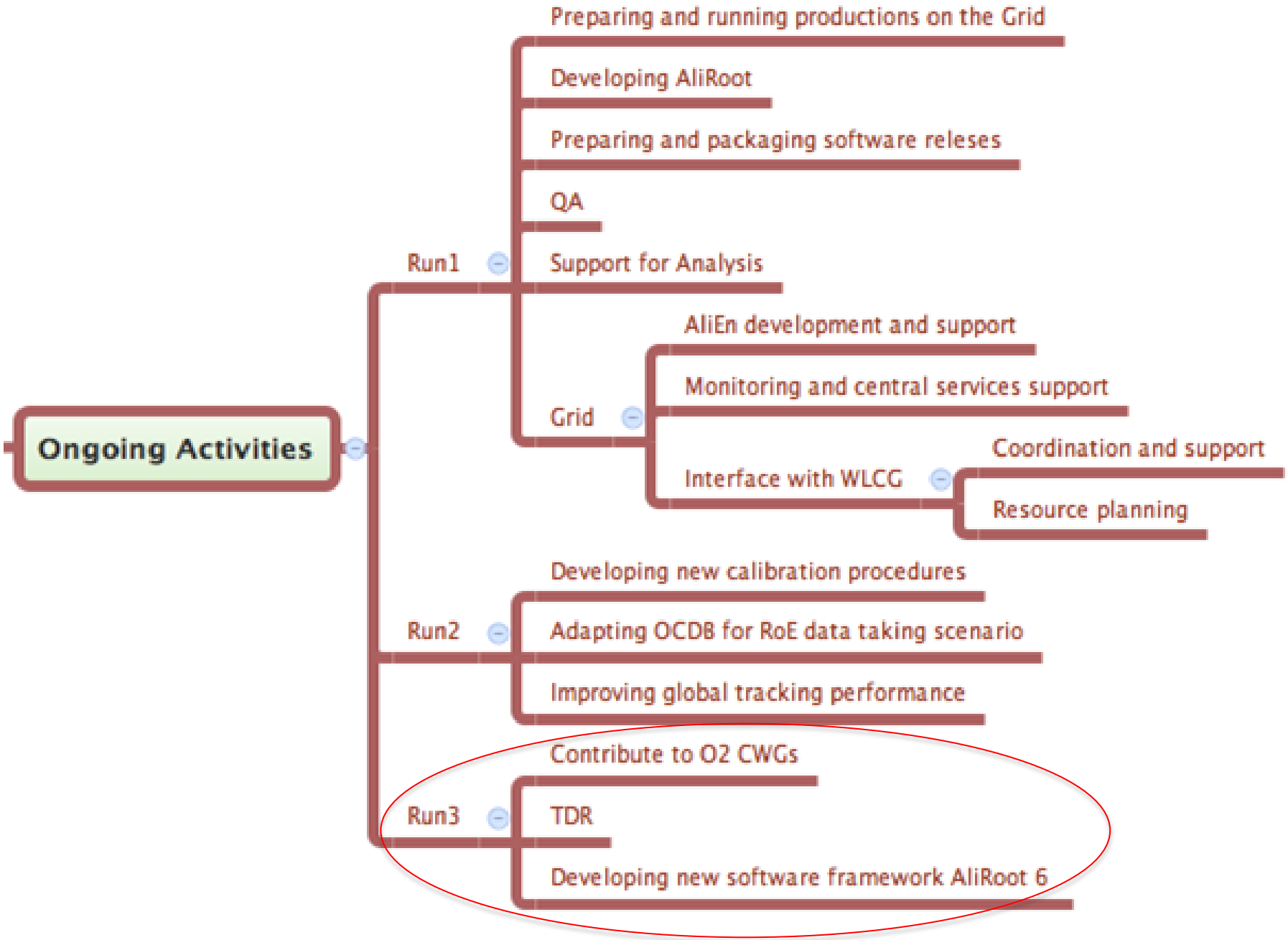
- Possibly indicating impact of analysis on CPU usage efficiency in periods before important conferences
- We still do not have full control of our job efficiencies
- Needs more work


Ongoing Activities
Run1
Preparing and running productions on the Grid
Developing AliRoot
Preparing and packaging software releses
QA
Support for Analysis
Grid
AliEn development and support
Monitoring and central services support
Interface with WLCG
Coordination and support
Resource planning
Run2
Developing new calibration procedures
Adapting OCDB for RoE data taking scenario
Improving global tracking performance
Run3
Contribute to O2 CWGs
TDR
Developing new software framework AliRoot 6

Job Error rates (cumulative)



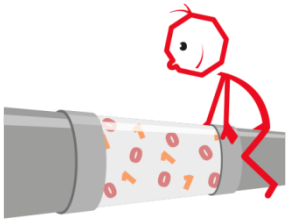
- Several problems found and fixed, still hunting for the remaining system errors
- Most of the errors are genuine job errors (validation, out of memory, wrong input parameters…)


Ongoing Activities
Run1
Preparing and running productions on the Grid
Developing AliRoot
Preparing and packaging software releses
QA
Support for Analysis
Grid
AliEn development and support
Monitoring and central services support
Interface with WLCG
Coordination and support
Resource planning
Run2
Developing new calibration procedures
Adapting OCDB for RoE data taking scenario
Improving global tracking performance
Run3
Contribute to O2 CWGs
TDR
Developing new software framework AliRoot 6

Upgrade activities

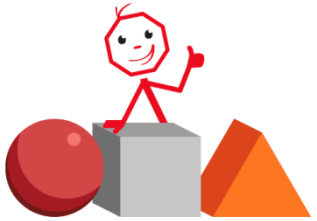
O^2 Technical Design Report

Architecture



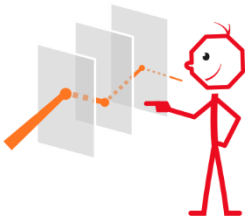
Data flow

Data model



Computing platforms

Tools



Simulation

Calibration

Reconstruction

DQM

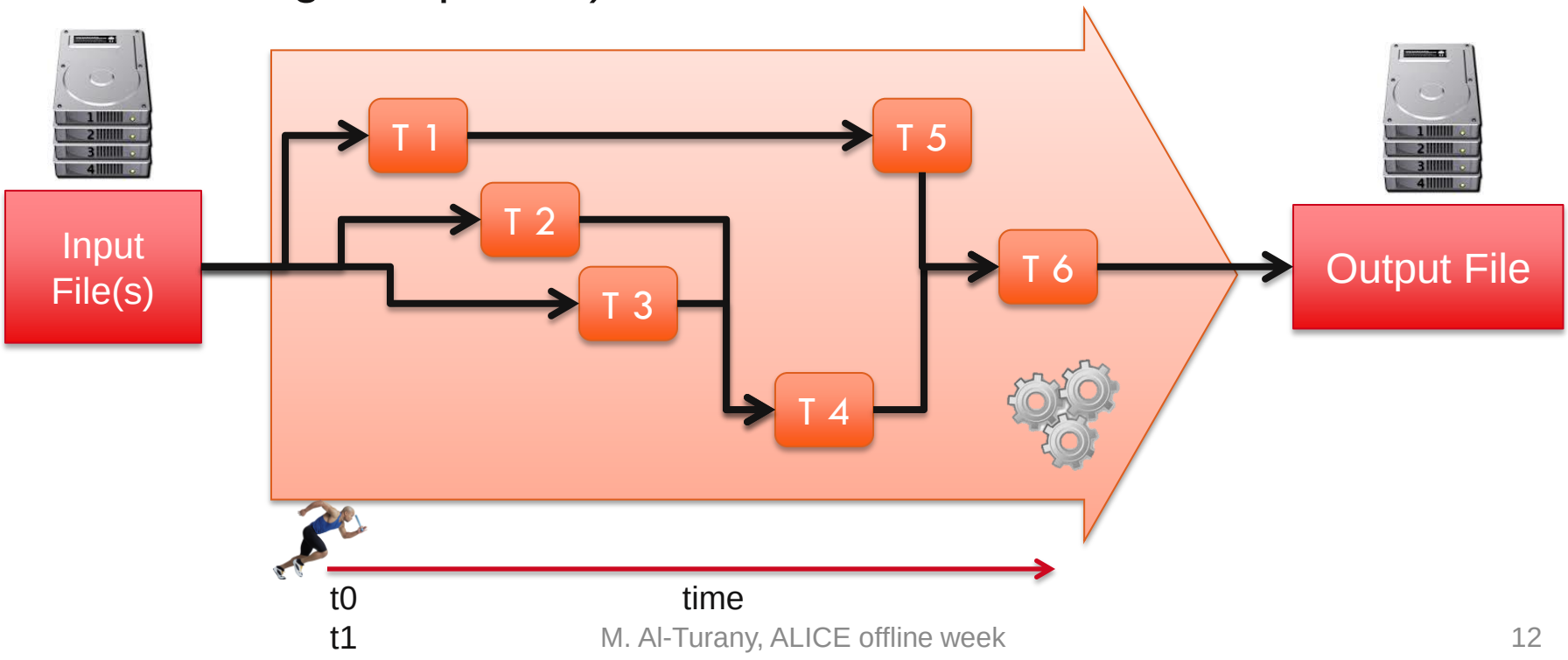
Control Configuration Monitoring

Software Lifecycle

Software Framework

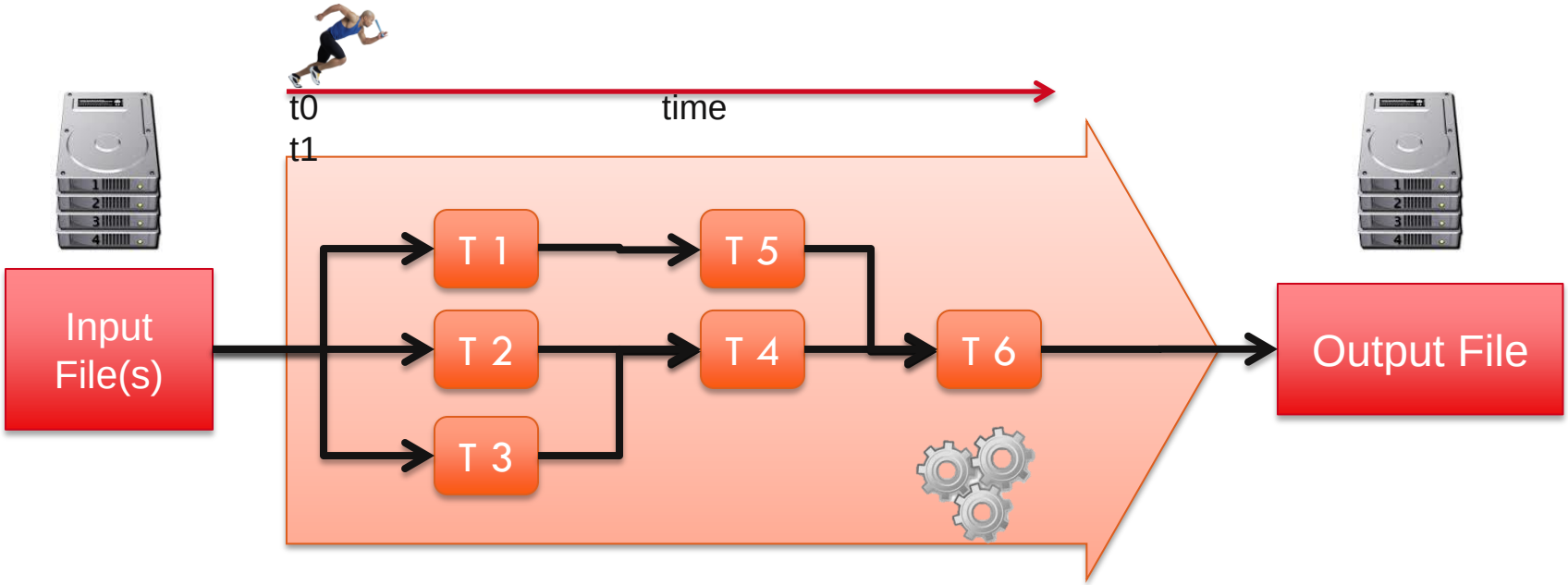
Current Model

- ROOT event loop
- User code in Task hierarchy
- Task hierarchy runs sequentially in one process
- Tasks implement only algorithms (can be exchanged/replaced)



Future

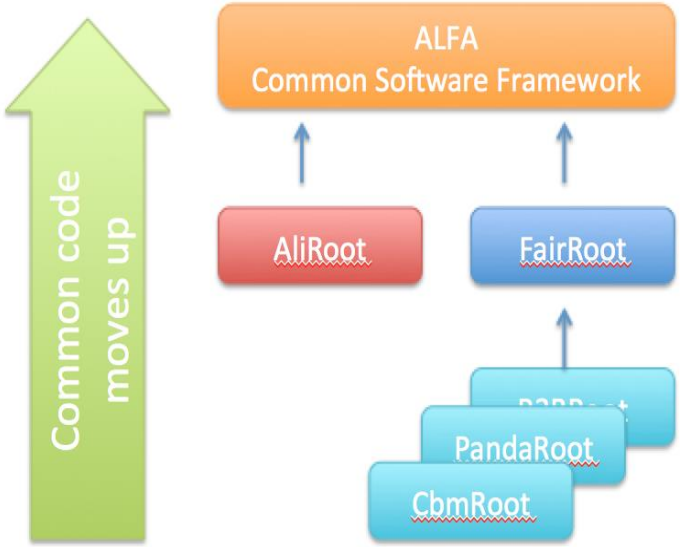
- Each Task is a process (can be Multi-threaded)
- Message Queues for data exchange
- Support multi-core and multi node



Multi-processing vs. Multi-threading

- Different processes are insulated from each other by the OS, an error in one process cannot bring down another process.
- Error in one thread can bring down all the threads in the process.
- Inter-process communication can be used across network
- Inter-thread communication is fast

Software Framework II

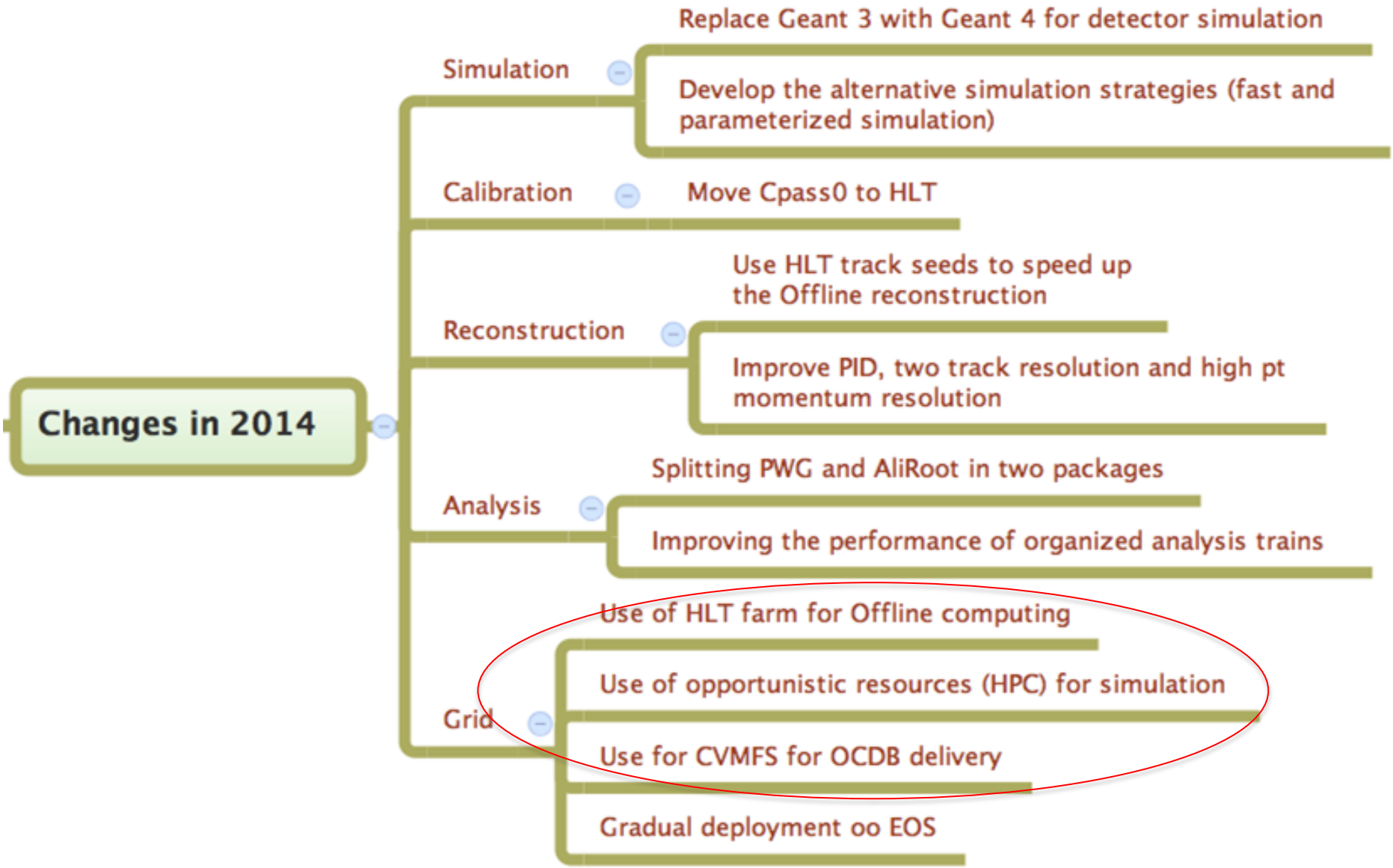


Alice + Fair = α(AlFa)

the basis of AliRoot 6.0

Work on ALFA already started

- GSI group works on factoring out common components from FairRoot and puts them to Github
- ALICE group works on importing ITS simulation for ALICE Upgrade in FairRoot environment
- Aim is to have ITS and TPC simulation in new framework ready by September


Changes in 2014
Simulation
Replace Geant 3 with Geant 4 for detector simulation
Develop the alternative simulation strategies (fast and parameterized simulation)
Calibration
Move Cpass0 to HLT
Reconstruction
Use HLT track seeds to speed up the Offline reconstruction
Improve PID, two track resolution and high pt momentum resolution
Analysis
Splitting PWG and AliRoot in two packages
Improving the performance of organized analysis trains
Grid
Use of HLT farm for Offline computing
Use of opportunistic resources (HPC) for simulation
Use for CVMFS for OCDB delivery
Gradual deployment oo EOS

CVMFS deployment

Jan 2013
……...
……
Jun 2013
……
July 2013
……
Aug 2013
……
……
……
Dec 2013
~~Apr 2014~~

- ✓Setup Stratum 0 for ALICE
- ✓Deploy ALICE S/W on CVMFS
- ✓Migrate ALICE Repository to Stratum 1s
- ✓Test, test, test
- ✓Start deployment process on all ALICE sites
  - ✓Deploy CVMFS repository but do not use it in production
- ✓Run AliEn from CVMFS on selected site(s)
  - ✓Validate and evaluate stability, performance
- ✓ Run AliEn from CVMFS on all site(s)

DONE. THANK YOU!

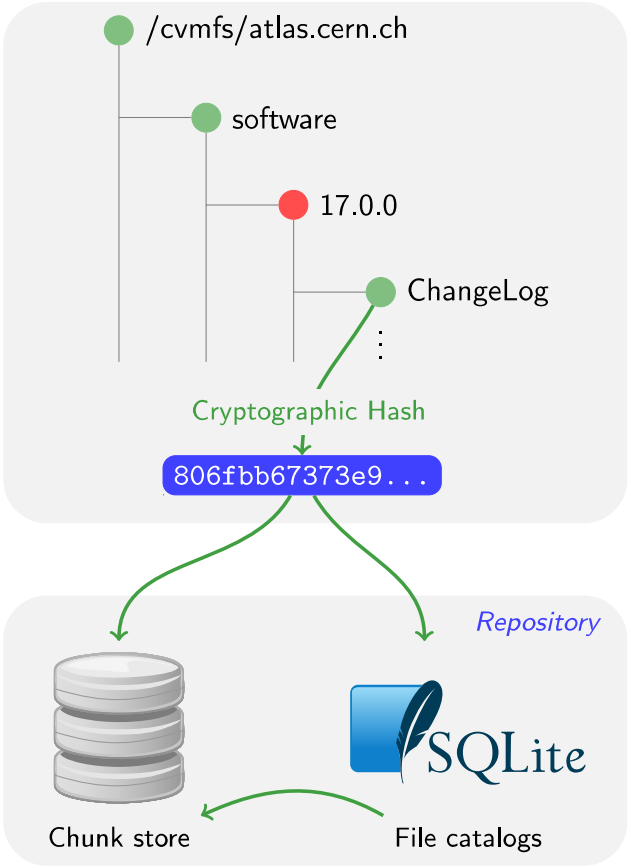
What can we do now with CVMFS?

1) Login to lxplus, test develop and debug the code in exactly the same environment as on the Grid

   /cvms/alice.cern.ch/bin/alienv enter AliRoot[/<version>]

2) We can deploy OCDB on CVMFS to leverage on already deployed proxy/cache infrastructure for efficient delivery of OCDB data files using HTTP protocol
3) On unsupported platforms, download CernVM 3.0 from http://cernvm.cern.ch, start it and then do the same as in 1)
4) On a private Cloud (such as CERN's OpenStack Cloud), deploy a cluster of CernVMs and use it to validate software releases
5) On any Cloud, deploy a cluster of CernVMs, start AliEn JobAgent and extend our Grid capacity
6) Deploy OpenStack middleware on ALICE HLT and do the same as in 5)
7) Access CVMFS repository using parrot anywhere (including places where CVMFS is not pre-installed and pre-configured)
8) Run unmodified ALICE software on Cray XK7
9) Let volunteers start preconfigured CernVM that will automatically join ALICE grid and run MC jobs

How that stuff works?



Data Store

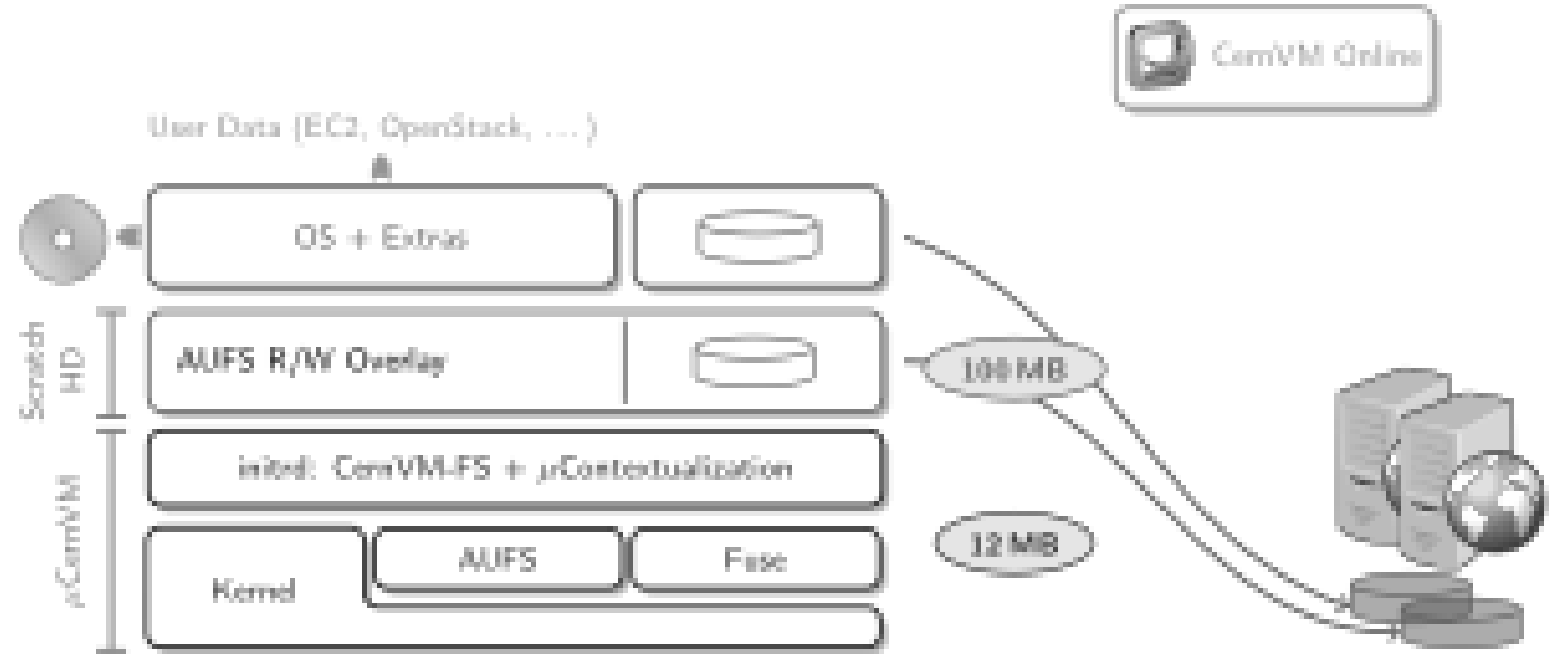
- Eliminates duplicates
- Never deletes, archiving

File Catalog

- Directory structure, symlinks
- Content hashes of regular files
- Digitally signed
- Plain files

The *root hash* (40 characters) defines a file system snapshot (similar to git)

CernVM 3.0 = SLC6 via CVMFS



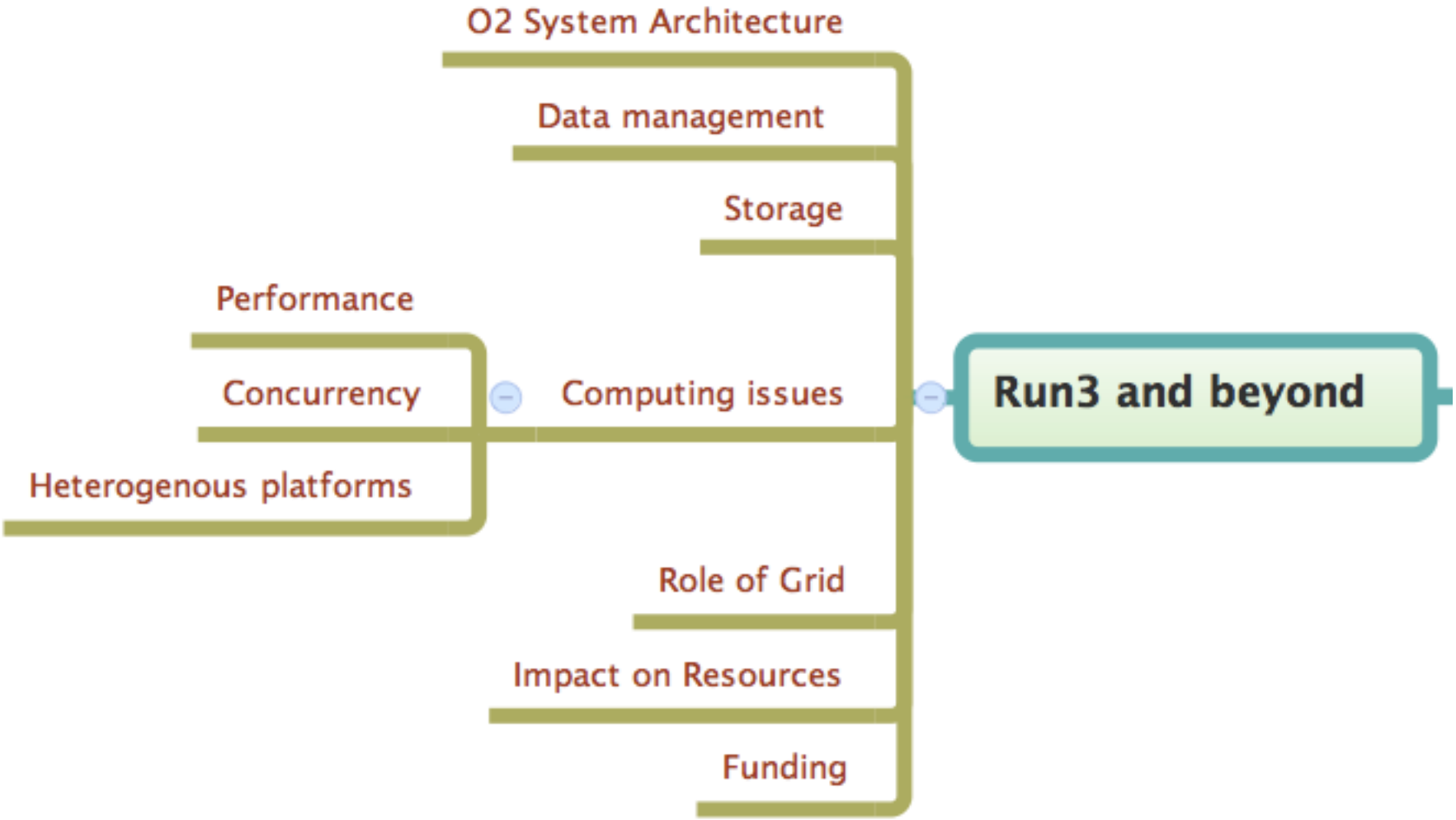
Twofold system: μCernVM boot loader + OS delivered by CernVM-FS

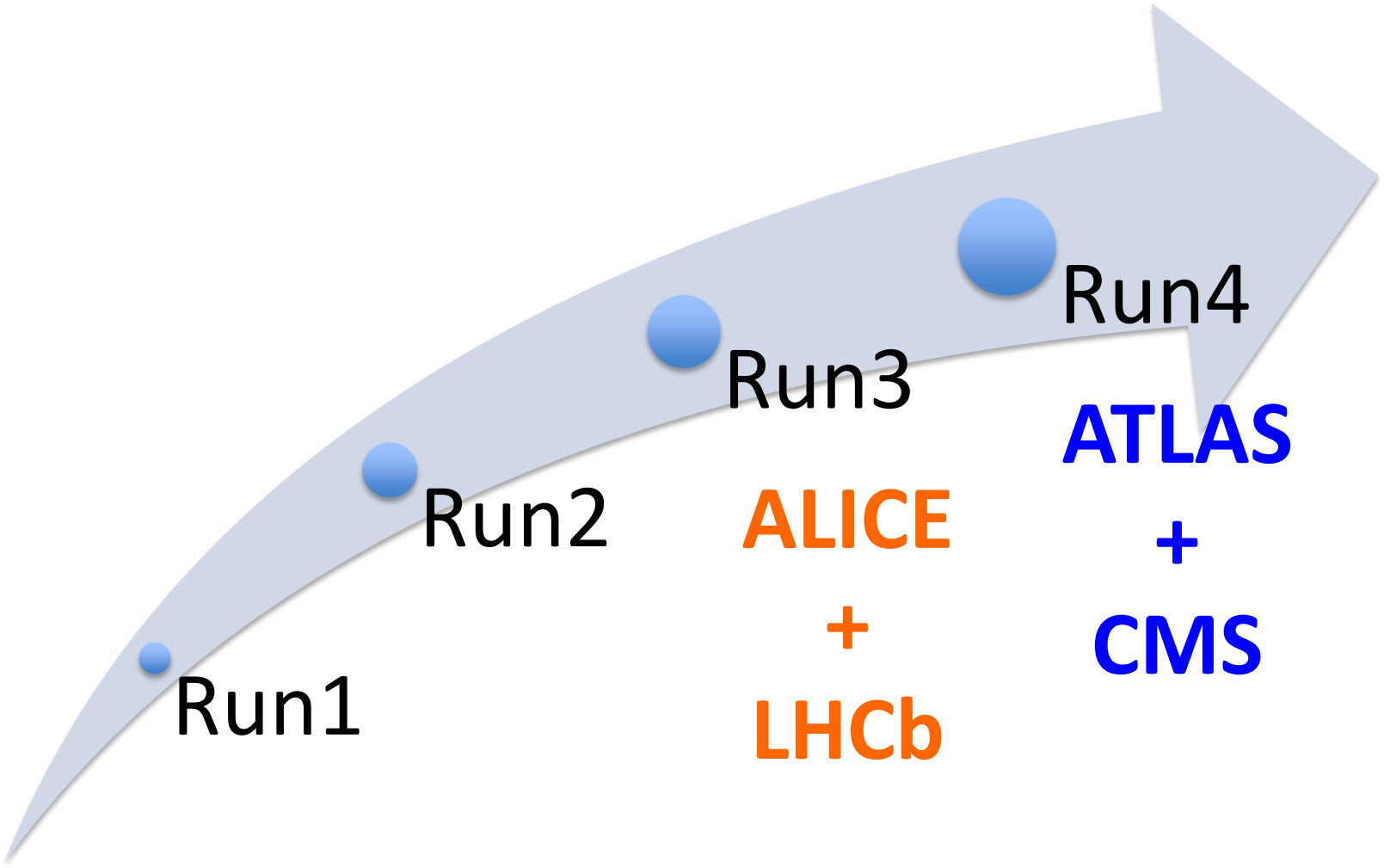
- The very same image can be contextualized to run Scientific Linux 4 32bit as well as the latest Scientific Linux 6 64bit
- Solution for Long Term Data Preservation problem?

ALICE

The Virtual Analysis Facility

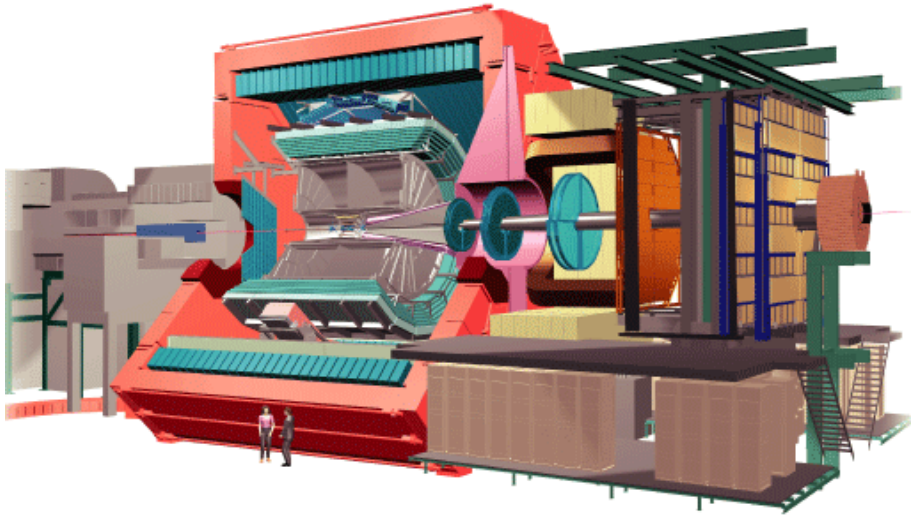
- A cluster of original unmodified CernVM virtual machines
  → *all configured during contextualization*
- Cluster context: one head node + scalable num. of workers
  → *available on http://cernvm-online.cern.ch*
- Portability and usability
  → *both for users and system administrators*
- One PROOF deployment for all LHC experiments


Run3 and beyond
O2 System Architecture
Data management
Storage
Computing issues
Performance
Concurrency
Heterogenous platforms
Role of Grid
Impact on Resources
Funding



- CPU needs (per event) will grow with track multiplicity (pileup) and energy
- Storage needs are proportional to accumulated luminosity

ALICE @ Run 3



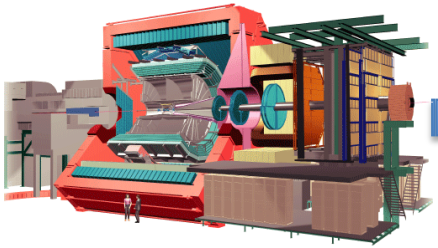
50 kHz

Online/Offline Facility

50 kHz (1.5 MB/event)

Storage

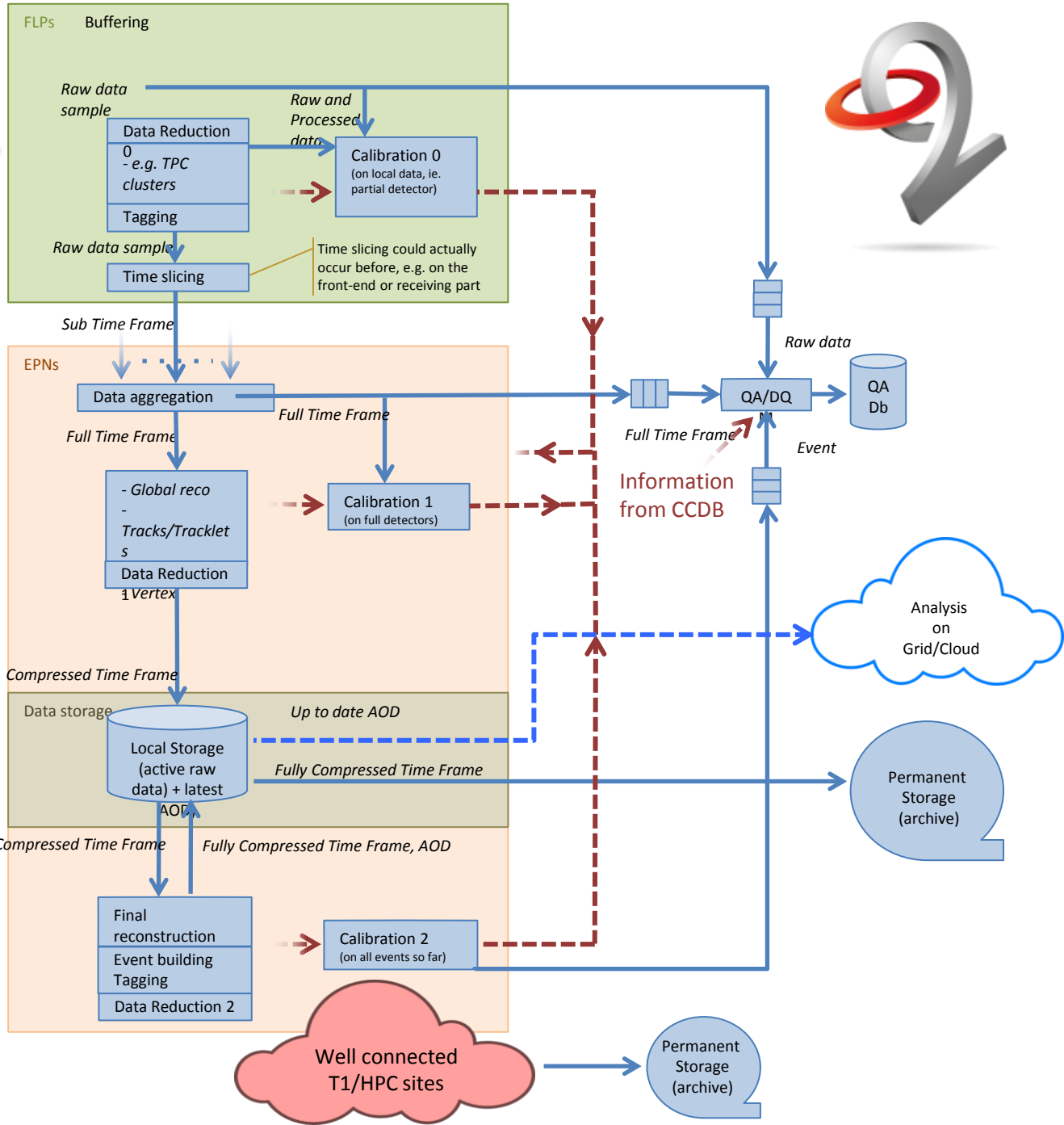
75 GB/s

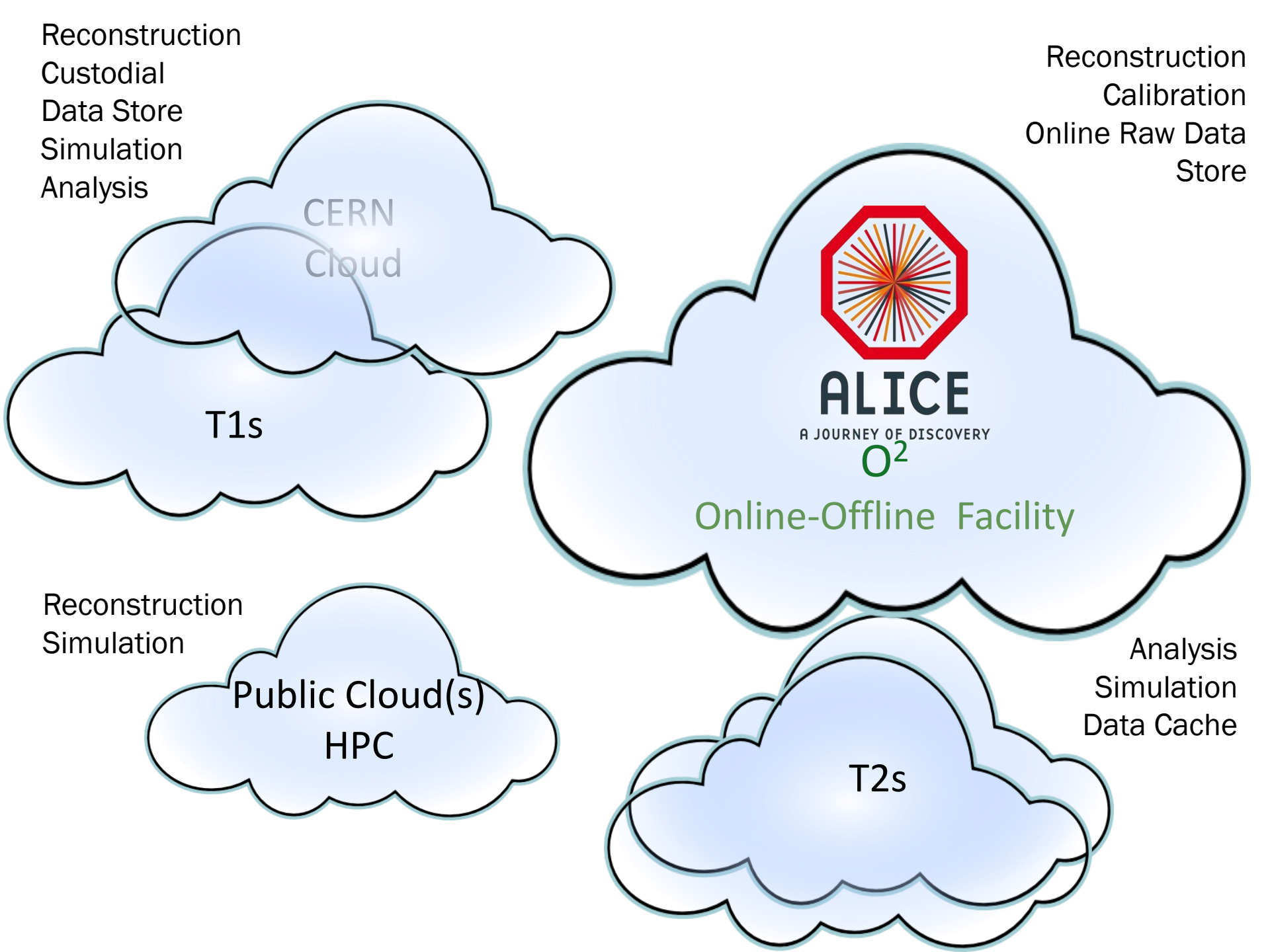


1 G2014

50 PB disk buffer

4 G2014



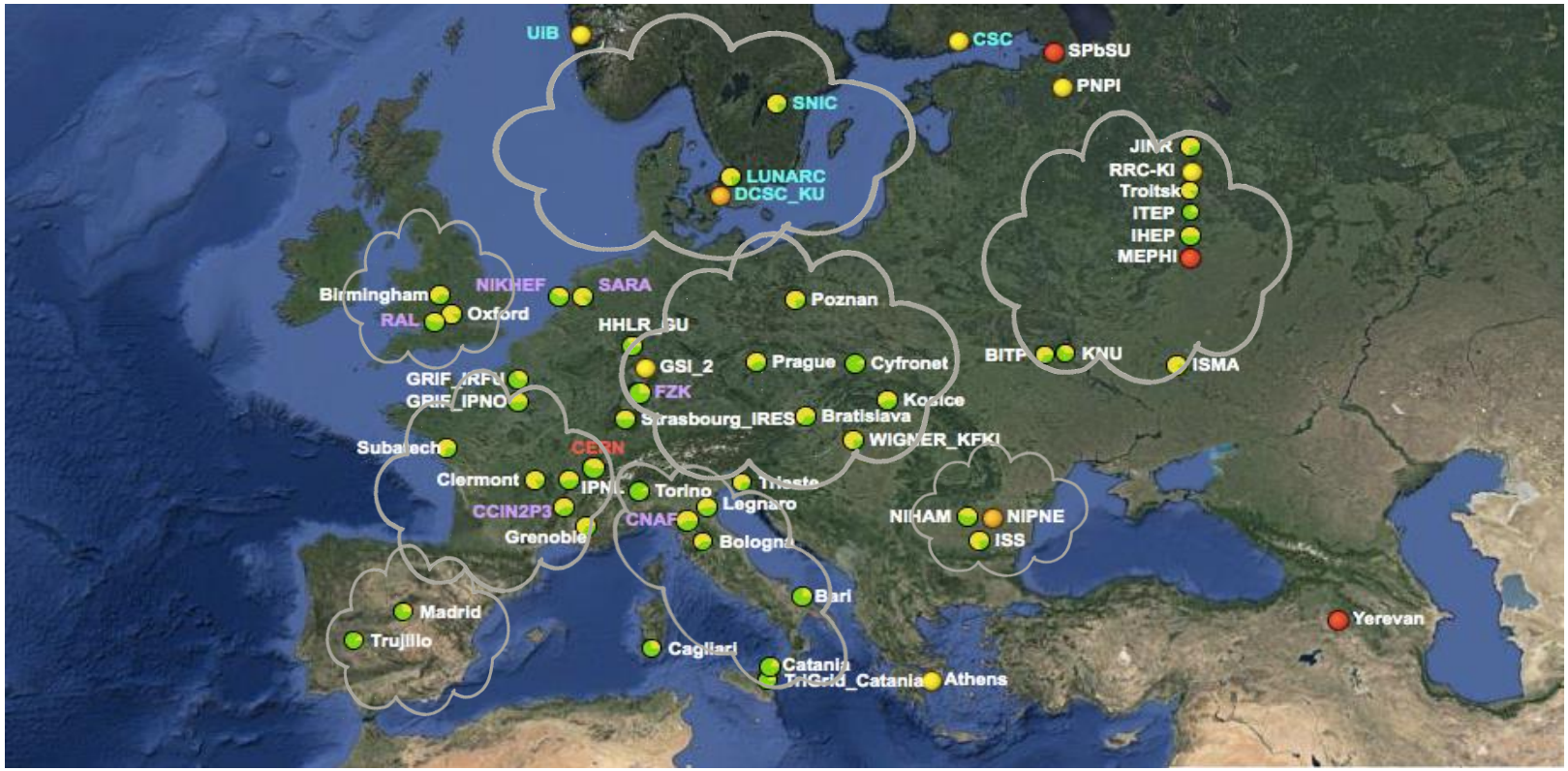

Reconstruction
Custodial
Data Store
Simulation
Analysis
CERN
Cloud
T1s
Reconstruction
Calibration
Online Raw Data
Store
ALICE
A JOURNEY OF DISCOVERY
O2
Online-Offline Facility
Reconstruction
Simulation
Public Cloud(s)
HPC
Analysis
Simulation
Data Cache
T2s

Simulation as a service

- Currently, simulation represents 70% of all our CPU time spent
- In general, we should try to reduce the simulation requests to absolute minimum
- Running simulation in the same way as any other job introduces many overheads

- Consider running simulation as a service
  - HPC resources to simulation data sources
  - Similar to raw data from experiment
  - The result only needs to be registered in a common name space and storage pool

From Grid to Cloud(s)



- In order to reduce complexity national or regional T1/T2 centers could transform themselves into Cloud regions
  - Providing IaaS and reliable data services with very good network between the sites, dedicated links to T0

Goal: Reduce complexity

- Deal with handful of clouds/regions instead of individual sites
- Each cloud/region would provide reliable data management and sufficient processing capability
  - What gets created in a given cloud, stays in that cloud
- This could dramatically simplify scheduling and high level data management

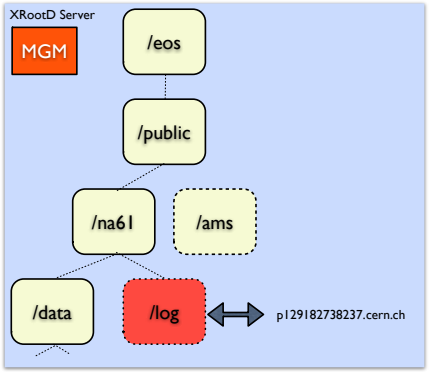
- Again, data management is the key

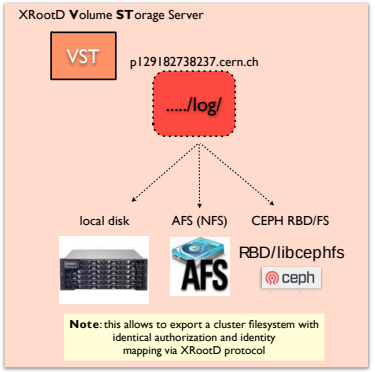
CITRINE VST

Infinity ∞

- EOS *Infinity*

AFS-like attached volumes hosting data+meta data of a subtree
- small/many file use cases
- allows to attach any mountable FS tree into EOS namespace
- allows to have extended attributes on file and directory level for meta data tagging



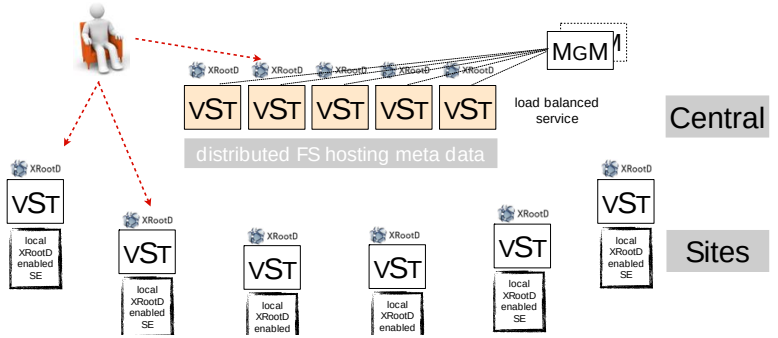


CITRINE VST

Unity

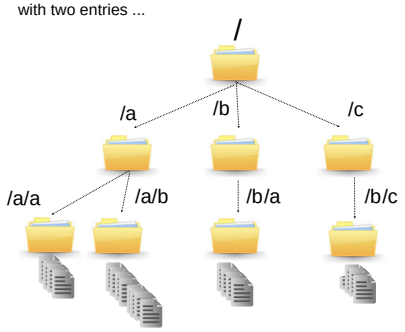
- EOS *Unity*

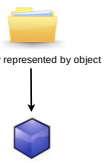
Today's Federations provide a redundant functionality via a read-only overlay network. A complete storage federation should have also placement capabilities, honor replication policies and a global reliable namespace. We can use a group of VSTs to host the global logical namespace redirecting read and write requests to VSTs hosting a logical or physical namespace (sites). A site VST is just a redirection and report gateway to any regular XRootD enabled SE or a local EOS setup. For placement and file access we can extend the already existing geo placement/scheduling capabilities of EOS used for the CERN/Wigner CC setup.



Diamond R&D

- trivial idea: store a namespace in a scalable object store
  - we can represent data in a *hierarchical structure* using directories and files and we *don't need* to group an infinite amount of files into a single directory
  - each *file* is a *list entry* with meta data in a directory
  - each *directory* is represented as an *object* in an object store
  - to circumvent central locking we can allow a conflict if two files get created with the same name and different contents and make it visible in the namespace like a conflict in DropBox with two entries ...





dir.attributes

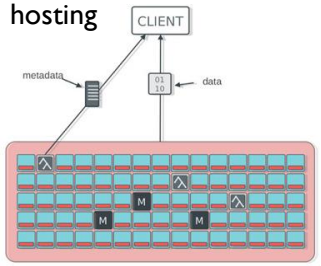
owner	acl	xattr
root root	xyz	user.x sys.y

file table

Name	Size	Cks	Locatio	UUID
a	1	0xa	1:2	A
b	2	0xb	2:3	B
c	3	0xc	3:4	C
d	2	0x4	4:5	D
e	1	0x5	5:6	E

Diamond R&D Scalable Object Store/Namespace using CEPH

- ceph is an open source implementation of an object store providing features like *dynamic resizing, self-healing, guaranteed consistency, low read latency, async object IO, extended attributes + key-value map per object, object notifications*
- IT-DSS provides now a ceph (rados) object store **service** with 1 PB capacity [x3] (~50 nodes) - initially for VM hosting



EOS+

- EOS is already tested to the scale required for O2 internal buffer some extras might be needed
  - Media aware caching (SSD, fast disk, shingled disk…)
  - Sophisticated disk pool monitoring, visualization
- Scalable global name space
  - Replacement for file catalog
- Storage federation
- Integration of foreign file systems for specific purposes

- More in talk by Andreas Peters

Summary

- Run1 operations
  - Gearing up for QM2014 followed by already delayed data re-processing with improved software chain
- Run2 preparations
  - Focusing on improving calibration procedures and software performance
- Run3 activities
  - Work on AliRoot 6 has started, collaboration between ALICE and FAIR
- Lots of challenges and potentially interesting computing related research projects
  - Storage, data management, O2 facility..
- Existing manpower is already overcommitted
  - Trying to mobilize the collaboration to provide extra manpower to carry out various computing related tasks