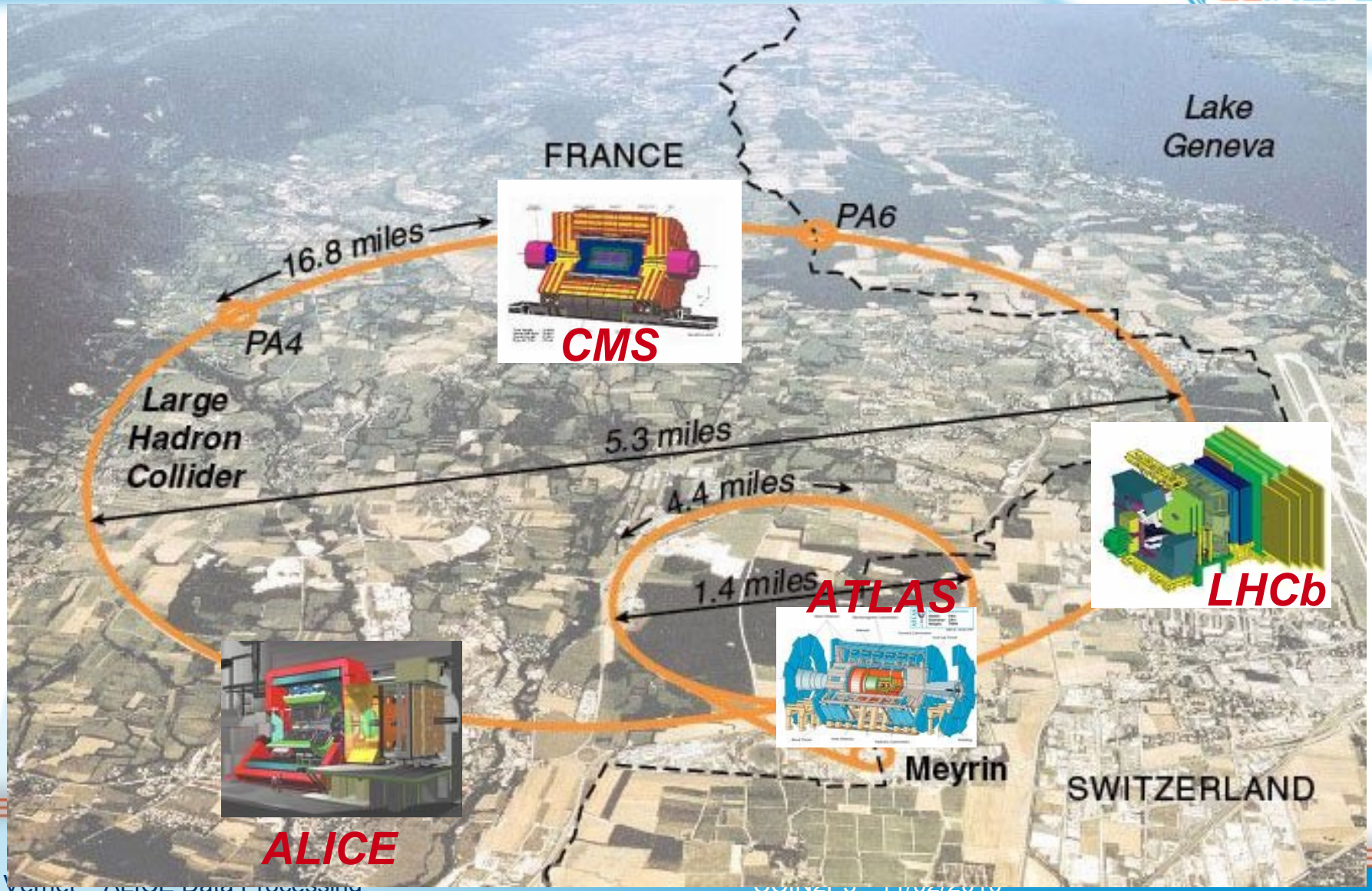
ALICE Data Processing

Outline

- ALICE experiment
- Data acquisition, processing & transfer
- AliEn, job submission, workload management...
- What should happen at CC-IN2P3

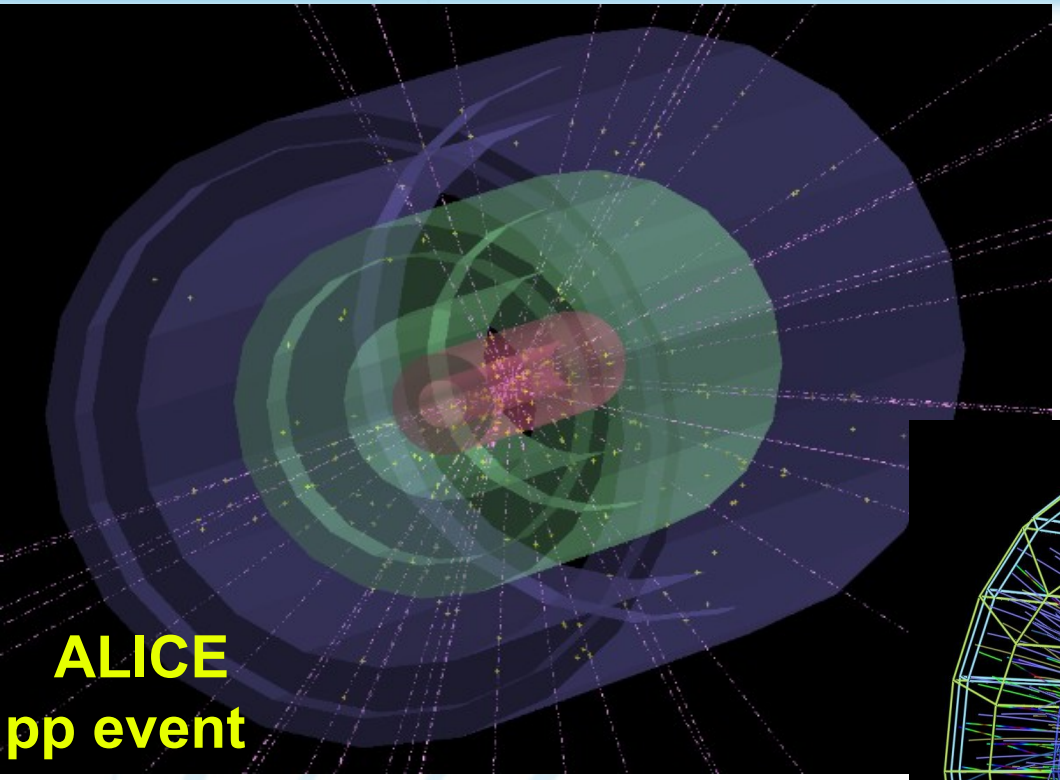
The ALICE experiment at LHC

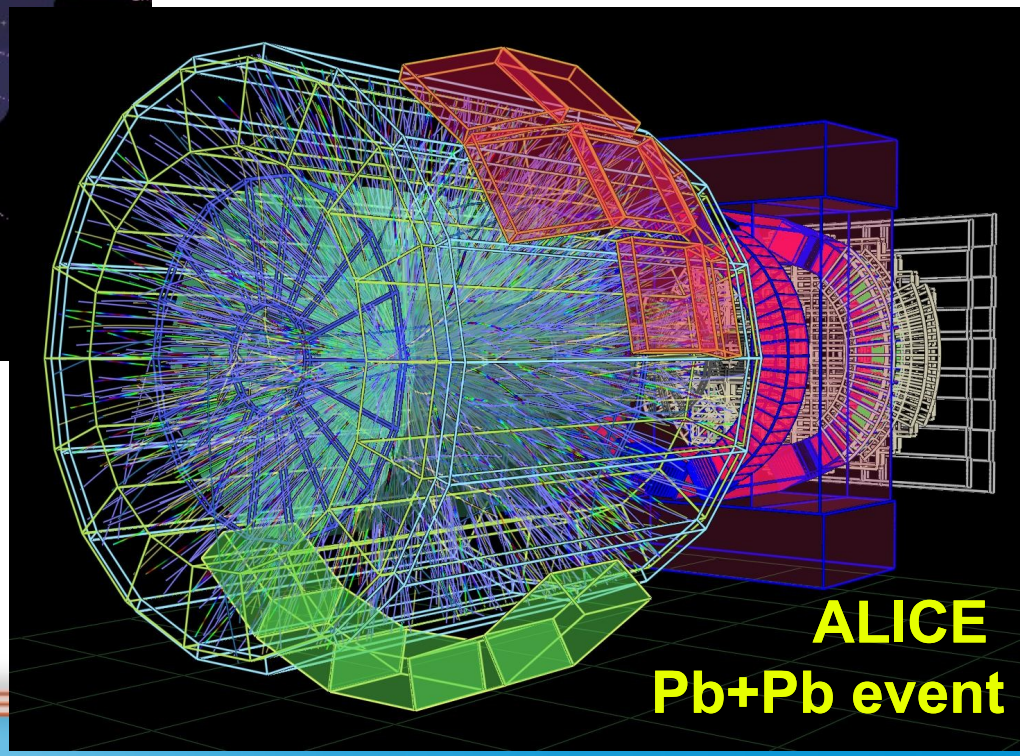


The ALICE experiment

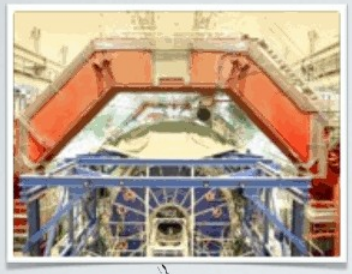
- Only experiment at LHC originally dedicated to the study of the Quark-Gluon Plasma
  - Heavy Ion collisions ! (Pb+Pb)
    - Best way to create the QGP : very large energy density available
  - Proton-proton collisions
    - Reference for heavy ions and genuine pp physics
  - Most of the time, LHC will run pp collisions (1 month of HI/year)
- ALICE computing model has been designed according to the needs of heavy ion collisions
  - Large amount of data stored per Pb+Pb collision
    - Expected mean rate: 1.25 GB/s data produced
  - Very long computing time (simulation,reconstruction,analysis)
  - Much storage space

ALICE event snapshot





Computing Model – pp



(Permanent Data Storage)

PDS

AliEN FC

(AliEn File Catalogue)

CERN T0

RAW & Calibration parameters

ODB

(Online Data Buffer)

First pass reconstruction

CAF Analysis

(PROOF)

External Copy

(Event Summary Data)

ESDs

Ext T2s

Ext T1s

MC

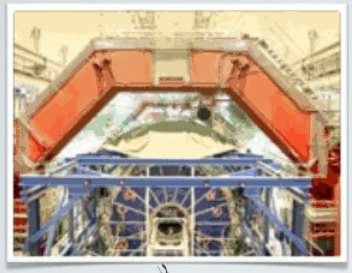
End users analysis

Pass 1 & 2 reco

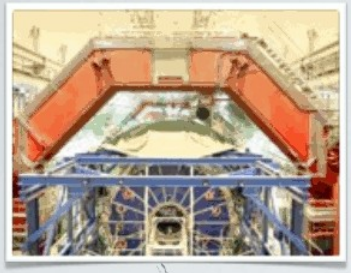
Analysis Train

WN PROOF XROOTD

Computing Model – AA



Computing Model – AA



PDS

AliEN FC

CERN T0

RAW & Calibration parameters

ODB

First pass reco

CAF Analysis

External Copy

LHC shutdown

Ext T2s

ESDs

MC

End users analysis

Ext T1s

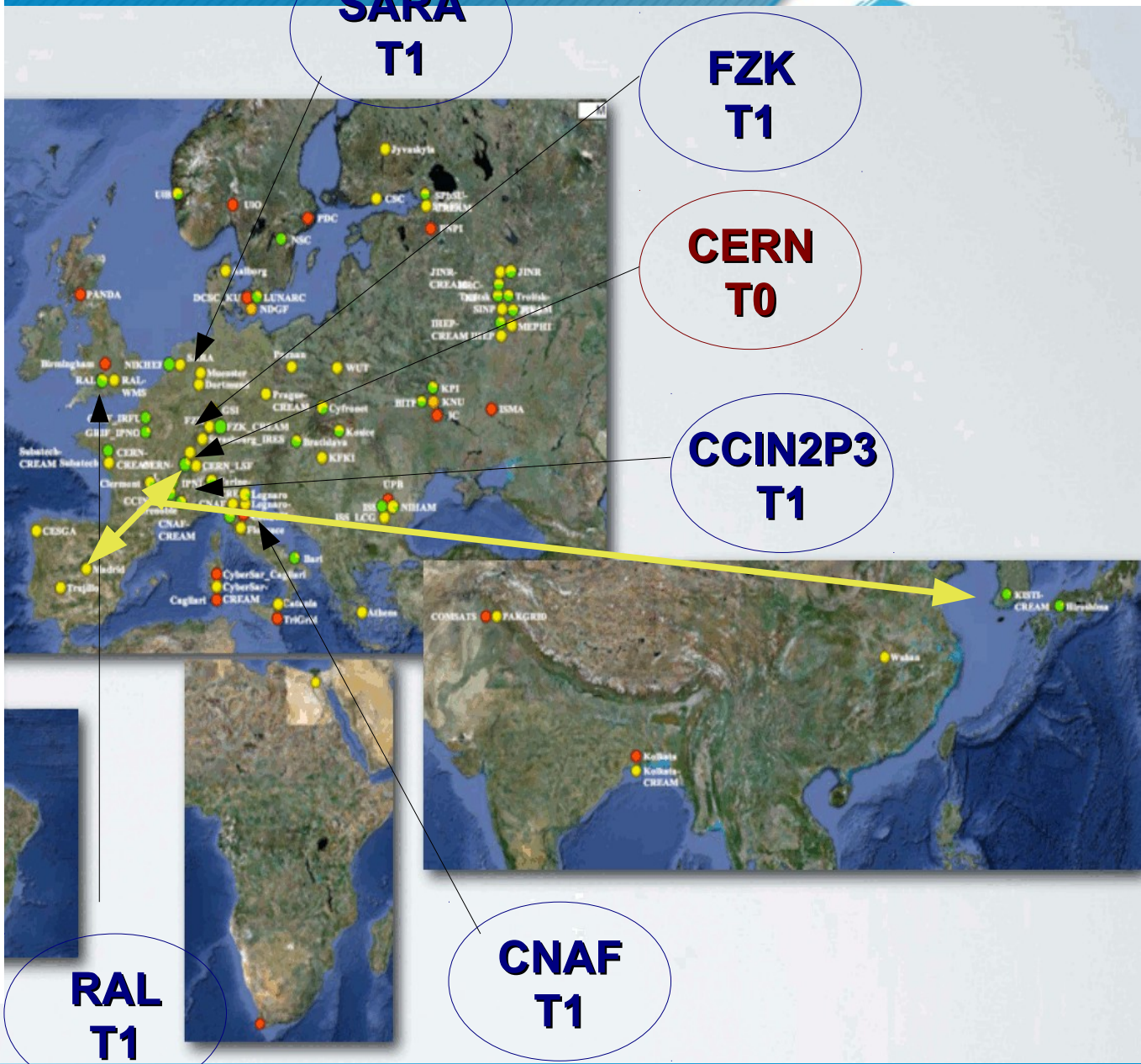
Pass 1 & 2 reco

Analysis Train

WN PROOF XROOTD

E GRID

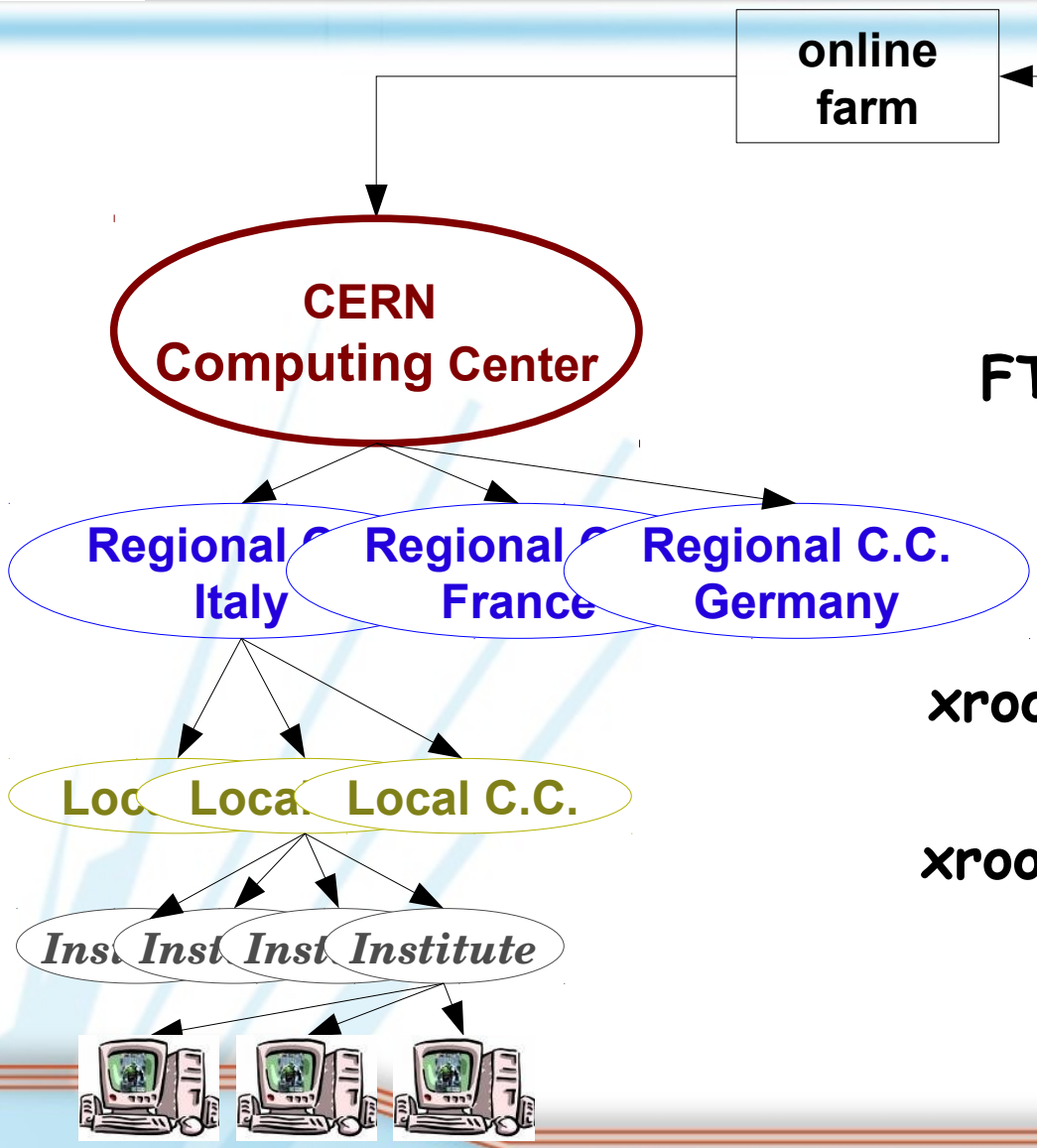
Tier 1	Tier2
CCIN2P3	French Tier2
	Sejong (Korea)
	Lyon Tier2
	Madrid (Spain)
CERN	Cape Town (South Africa)
	Kolkata (India)
	Tier2 Federation (Romania)
	RMKI (Hungary)[2]
	Athens (Greece)[2]
	Slovakia
	Tier2 Federation (Poland)
	Wuhan (china)
FZK	FZU (Czech Republic)
	RDIG (Russia)
	GSI (Germany)
	Muenster (Germany)
CNAF	Tier2 Federation (Italy)
RAL	Tier2 Federation (UK)
	Birmingham[2]
NIKHEF	SARA[2]
PDSF[1]	LLNL (USA)
	OSC (USA)
	Houston



The data transfers

- All scheduled transfers use FTD (~AliEn's FTS)
  - Transfer Queue
  - Triggered by alien 'mirror'
- T0 → T1 and T1 → T1 use LCG's FTS
  - Defined channels
  - Data go in and out the SE's via SRM interface
- T1 → T2 and T2 → T2 use xrootd
  - No predefined channel
  - Data go in and out the SE's via xrootd server

Data transfers and analysis (in brief)



FTS

- raw data
- 1st reconstruction pass

Tier-0

- raw copy
- 2nd /3rd reconstruction
- scheduled analysis
- MC copy

Tier-1

xrootd

- "physics" data
- user analysis
- MC production

Tier-2

xrootd

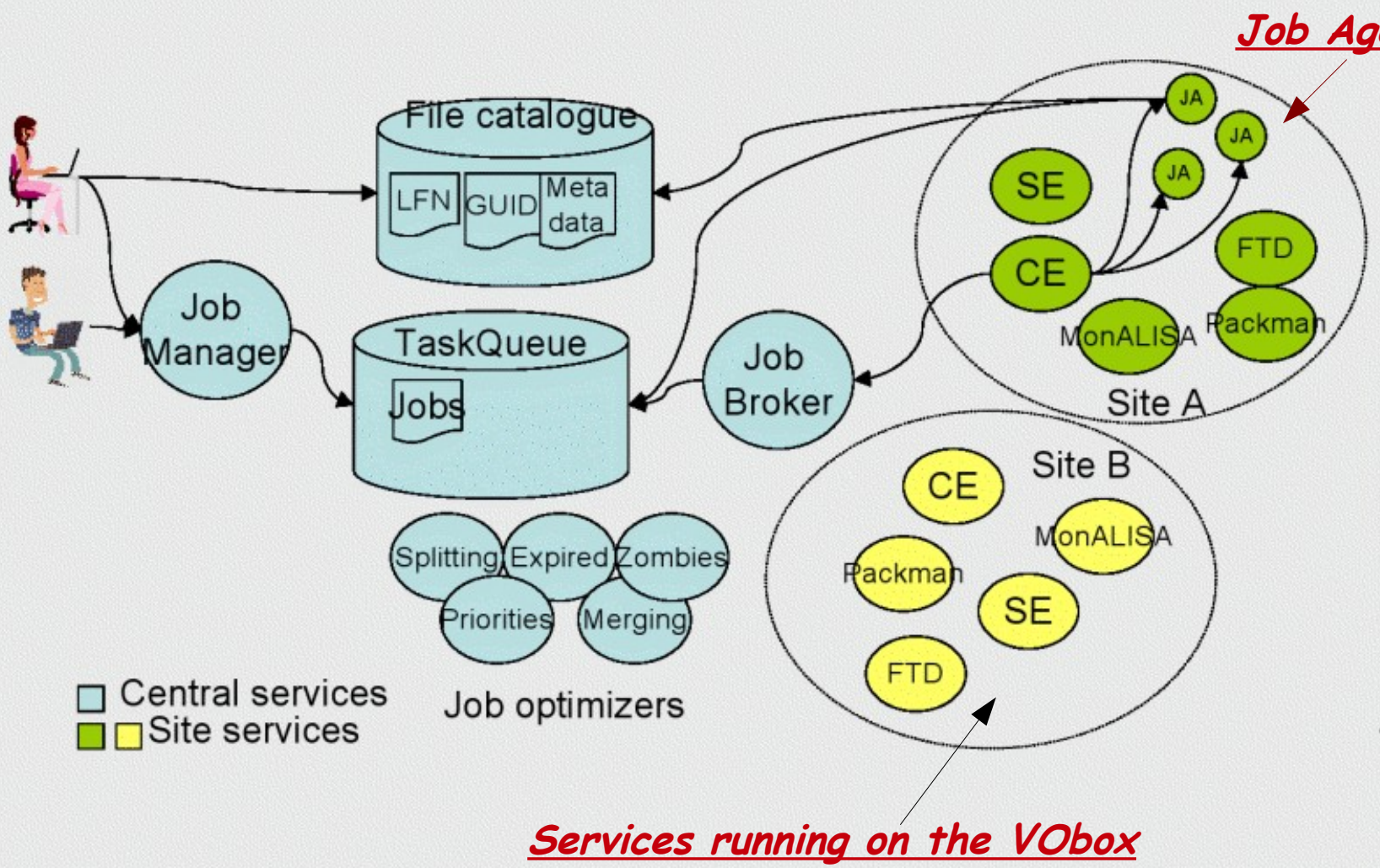
- locally cached data
- local analysis facility

"Tier-3"

AliEn

- GRID middleware for ALICE
- AliEn provides
  - Site services (CE/SE interfaces, package manager...) on VObox
  - File catalogue
    - All ALICE data (simulation+reconstruction)
    - User data
  - Client services
    - Catalogue browsing
    - Job submission
    - Job monitoring
    - Data transfers : between SE's + local→SE + SE→local

JOB EXECUTION MODEL



Pablo Saiz's talk @ CHEP07

AliEn (user view)

aterm

```
rvernet@ccvere:/$ aliensh
 [ aliensh 2.1.18a (C) ARDA/Alice: Andreas.Joachim.Peters@cern.ch/Derek.Feichtinger@cern.ch]
aliensh:[alice] [1] /alice/cern.ch/user/r/rvernet/ >ls -l Analysis/
drwxr-xr-x   rvernet  rvernet            0 Jan 17 17:14    .runAnalysis.jdl
-rwxr-xr-x   rvernet  rvernet          387 Jan 17 16:29    CreateChainFromXML.C
-rwxr-xr-x   rvernet  rvernet          396 Jan 17 23:51    mergerootfile-sequential.jdl
drwxr-xr-x   rvernet  rvernet            0 Jan 17 17:47    output
-rwxr-xr-x   rvernet  rvernet          632 Jan 17 17:14    runAnalysis.jdl
-rwxr-xr-x   rvernet  rvernet         1746 Jan 17 13:52    sample.xml
-rwxr-xr-x   rvernet  rvernet         1797 Jan 17 16:12    strangeAna.C
aliensh:[alice] [2] /alice/cern.ch/user/r/rvernet/ >submit Test/cream_aliroot.jdl
Submit submit Test/cream_aliroot.jdl
submit: Your new job ID is 41931762
aliensh:[alice] [3] /alice/cern.ch/user/r/rvernet/ >ps
   rvernet 41552696   __   D                        aliroot
   rvernet 41552697   __   D                           date
   rvernet 41552698   __   D                        aliroot
   rvernet 41552699   __   D                           date
   rvernet 41552927   __   D                        aliroot
   rvernet 41552928   __   D                           date
   rvernet 41552929   __   D                        aliroot
   rvernet 41552940   __   D                           date
   rvernet 41891047   __   D                        aliroot
   rvernet 41894985   __   D                        aliroot
   rvernet 41931761    0   I                        aliroot
   rvernet 41931762    0   I                        aliroot
aliensh:[alice] [4] /alice/cern.ch/user/r/rvernet/ >whereis /alice/data/2009/LHC09d/000104160/ESDs/pass2/09000104160018.10/root_archive.zip
Feb 11 12:54:00  info   The file LHC09d/000104160/ESDs/pass2/09000104160018.10/root_archive.zip is in
                SE => ALICE::CCIN2P3::SE  pfn =>root://ccxrdsn038.in2p3.fr:1094//07/23620/8e0b747a-f759-11de-a909-0018fe730ae5

                SE => ALICE::FZK::SE  pfn =>root://f01-120-123-e.gridka.de:1094//07/23620/8e0b747a-f759-11de-a909-0018fe730ae5

                SE => ALICE::Subatech::SE  pfn =>root://nanxrdmgr01.in2p3.fr:1094//07/23620/8e0b747a-f759-11de-a909-0018fe730ae5

                SE => ALICE::JINR::SE  pfn =>root://lcgxrdr01.jinr.ru:1094//07/23620/8e0b747a-f759-11de-a909-0018fe730ae5

aliensh:[alice] [5] /alice/cern.ch/user/r/rvernet/ >
```

Job submission

- Jobs submitted via AliEn shell
  - Job optimizer selects the best sites that satisfy the request
  - Performs splitting into subjobs
  - The subjobs will run where the data is and only there !! *(in principle)*
    - NB: several sites may contain the same data (mirror)
- Job Agents are launched on differents sites
  - Check availability of site
  - Check the required software
    - If not, the package manager will install it
  - Pulls 'real' jobs and runs them on the WN
- Which VObox (group of CEs) to use?
  - If not specified, the Job Optimizer decides itself
  - This can be specified in the JDL
    - `Requirements=other.CE==”Alice::CCIN2P3::CCIN2P3-CREAM”`

What happens at CCIN2P3?

- T1+T2 site → we do many different things
  - Raw data backup (HPSS)
  - Reconstruction (jobs '`aligrid`')
  - MC production (jobs '`aligrid`')
  - Scheduled analysis (from production, jobs '`aligrid`')
    - Creation of "Physics" data → ESD, AOD
  - Chaotic (end-user) analysis
    - Analysis of ESD,AOD (jobs '`alicexxx`')
- 2 computing elements
  - LCG CE → VObox `cclcgalice01` submits on it
  - CREAM CE → VObox `cclcgalice02` submits on it
    - More promising, will be the default at some point
    - `cclcgalice01` will become a backup VObox
- Storage
  - Tape (HPSS) : raw data transfers (FTS) from T0
  - Disk (xrootd) : output from reconstruction done in T0 (→ send via FTS) and done in T1 (CC)
- Data staging Tape → Disk
  - Feature not yet implemented
  - Wish to use TReqs !

Job behaviour @ CC

- Common software for all jobs in
  - /afs/in2p3.fr/grid/toolkit/alice
  - 1 AFS volume → stress ?
  - Resource used with BQS: `u_stress_afs_alice`
    - Limit the incoming job flow
- Memory issue on production jobs
  - Simulation & reconstruction require large amount of memory
    - Min. 2GB
  - We reached a value of 4.5GB for some jobs
    - Need to use the BQS J class for these jobs (hopefully temporary issue)
  - AliRoot offline core goal is to make it get down to 2GB again
  - However we should take an action soon

Storage @ CC

- RAW data replication
  - Should do CASTOR@CERN → HPSS@CCIN2P3 via FTS
  - Staging HPSS → xrootd not in place yet
  - ⇒ no raw data at CCIN2P3 at the moment
- Subsequent reconstruction passes have been stored on xrootd@CCIN2P3
  - → user analysis can be performed on ESD/AOD
- Simulation
  - Performed partly at CCIN2P3
  - → everything on xrootd
- SE Monitoring
  - Files are regularly sent and stored in SE's
  - Provides “working status” of storage services

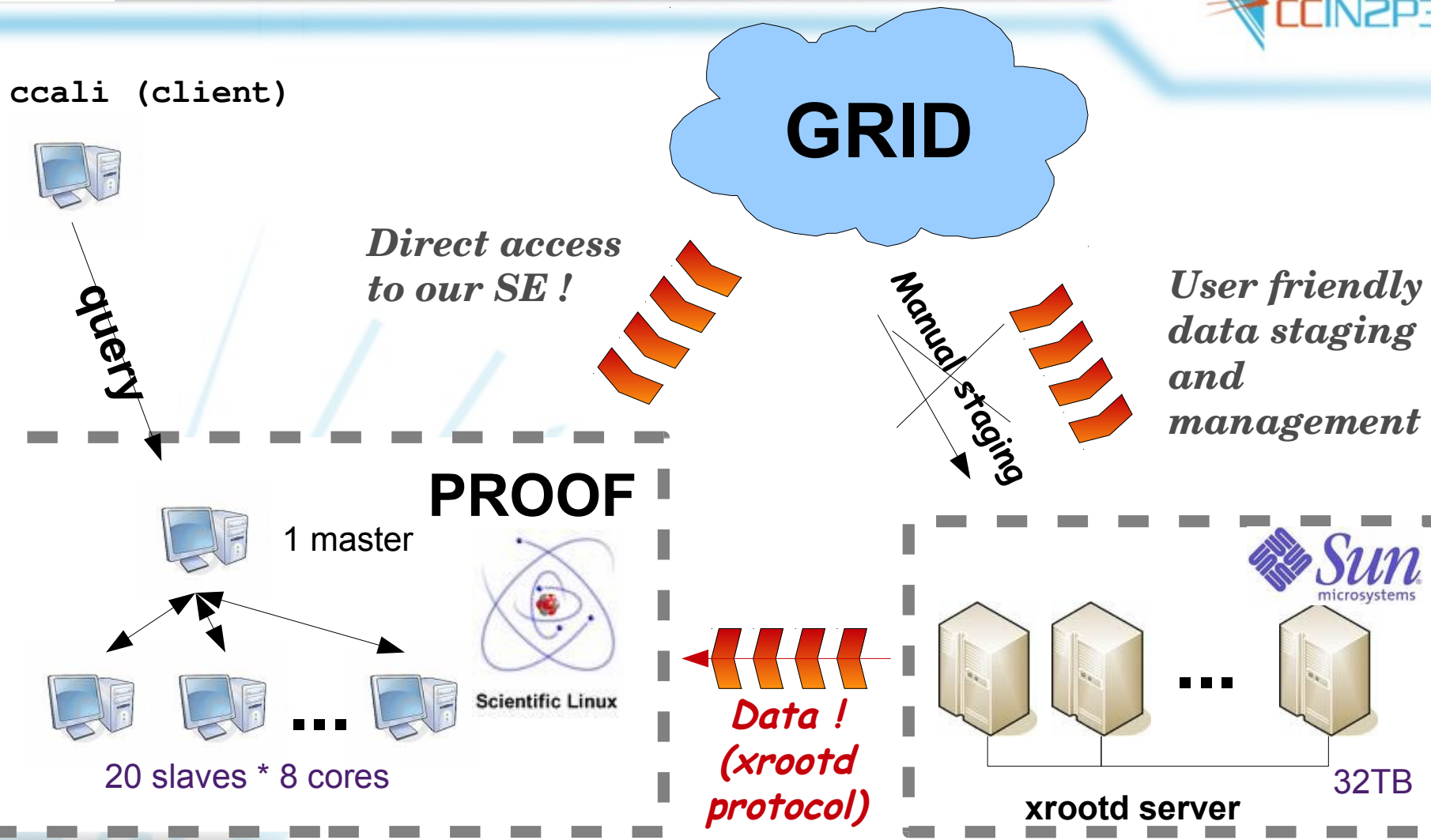
Storage issues

- Deployment procedure of xrootd software with ALICE authentication library
- Effective data transfer rate from xrootd servers
  - “GRID SE” slower than “LAF” server
  - Is this expected ?
  - Software or bandwidth issue ?

PROOF cluster (LAF)

- Computing model of ALICE and Analysis Framework designed to work on PROOF
  - Exact same code must run on local, GRID and PROOF
- See LAF tutorial @ CC (last week) for more info

Current status and prospects



questions?

FTS vers SE xrootd (december 17th -18th)