Tier-1 Operations at CC-IN2P3

Renaud Vernet – Apr. 2nd 2012

Outline

- The CC-IN2P3 in not so brief
- « Operation » team
  - Organization
  - Activities
- Watch and monitoring tools
- Transition to Oracle Grid Engine
- Preparing the future...

The CC-IN2P3

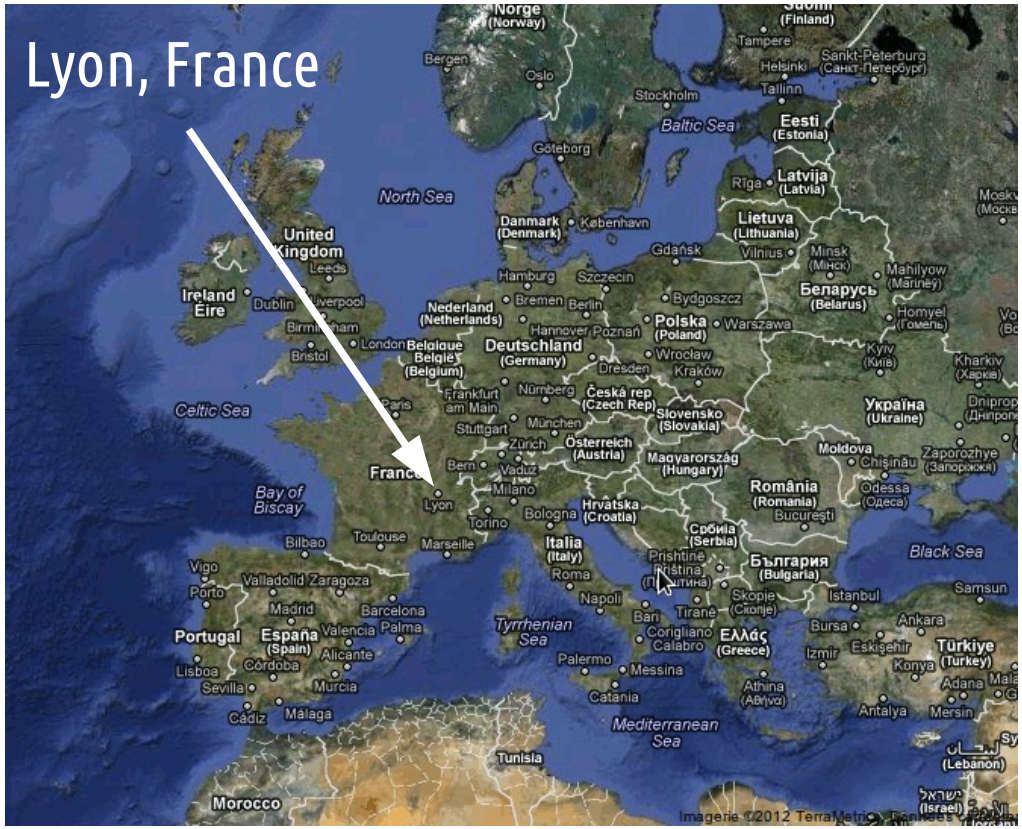
- IN2P3
  - National Institute for Nuclear and Particle Physics
- **C**omputing **C**enter originally created for physics purposes (nuclear & HEP)

→ The core of our activities

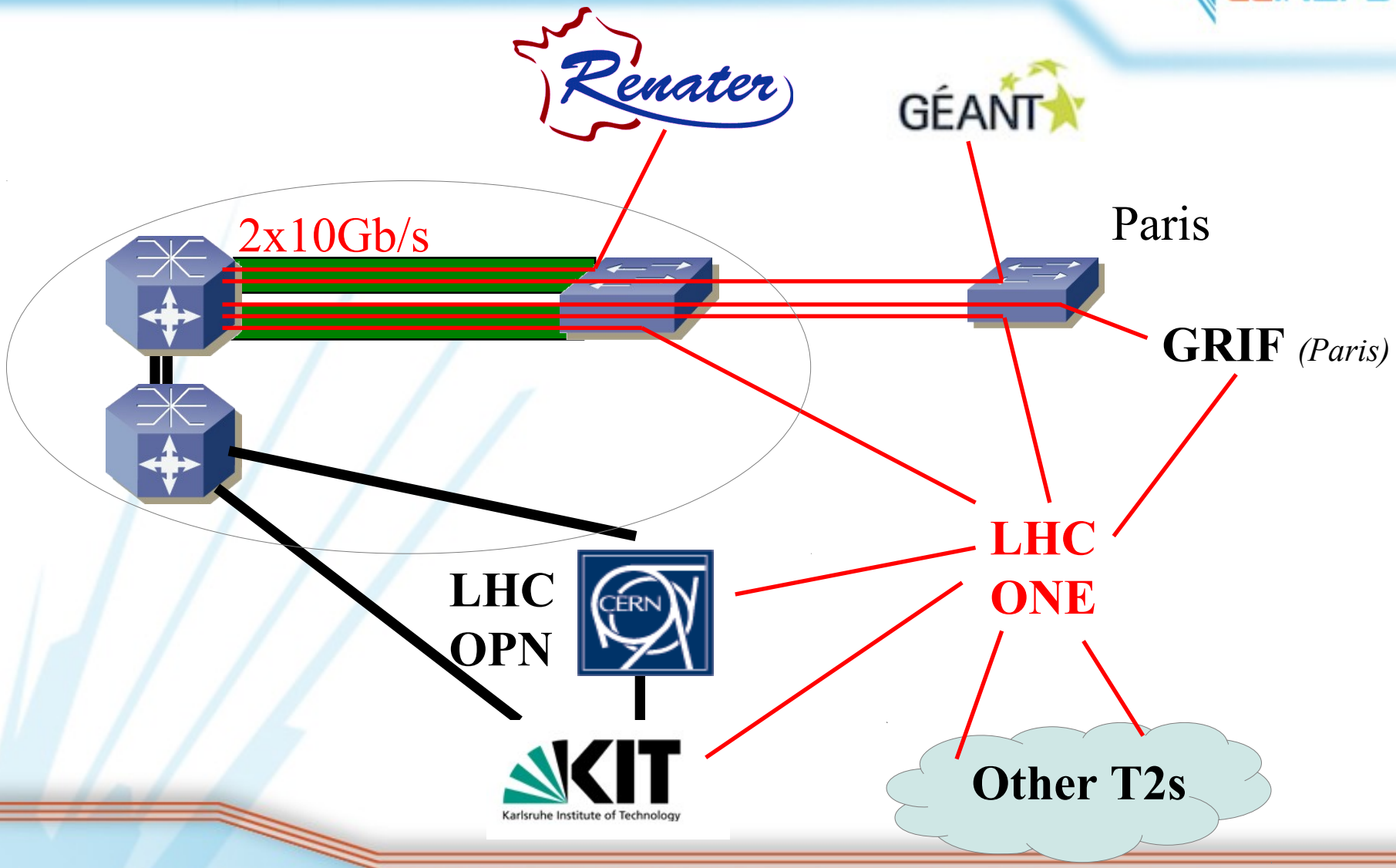
- Now serving also non-physics fields
  - Biology, medical imaging, neurosciences, humanities...
- Other activities
  - Grid development (portals)
  - Web hosting service (education purposes...)
- ~80 persons at CC
  - Some of them (user support group, 10 people) work in close contact with users/experiments

The CC-IN2P3

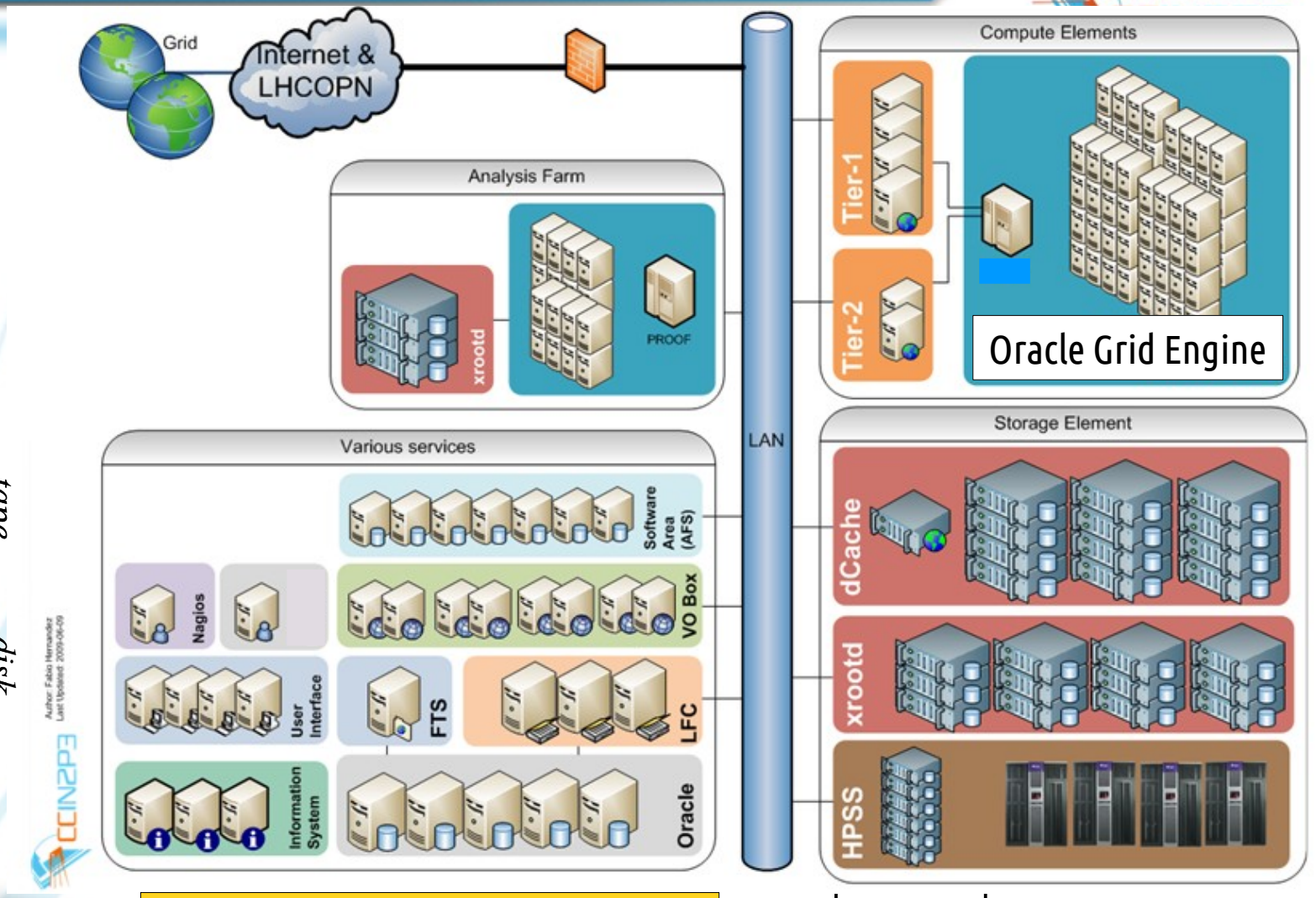
- Located in Lyon
  - ~150km from Geneva
- Hardware partners
  - Dell, Oracle, IBM, Cisco
- Network connectivity
  - LHC OPN
  - LHC ONE
  - Renater



Network


Renater
GÉANT
2x10Gb/s
Paris
GRIF (Paris)
LHC
ONE
LHC
OPN
CERN
KIT
Karlsruhe Institute of Technology
Other T2s

Resources overview



17 PB HPSS
1 PB TSM

8.2 PB dCache
1.9 PB irods
1.7 PB xrootd
1.2 PB GPFS
0.3 PB SRB

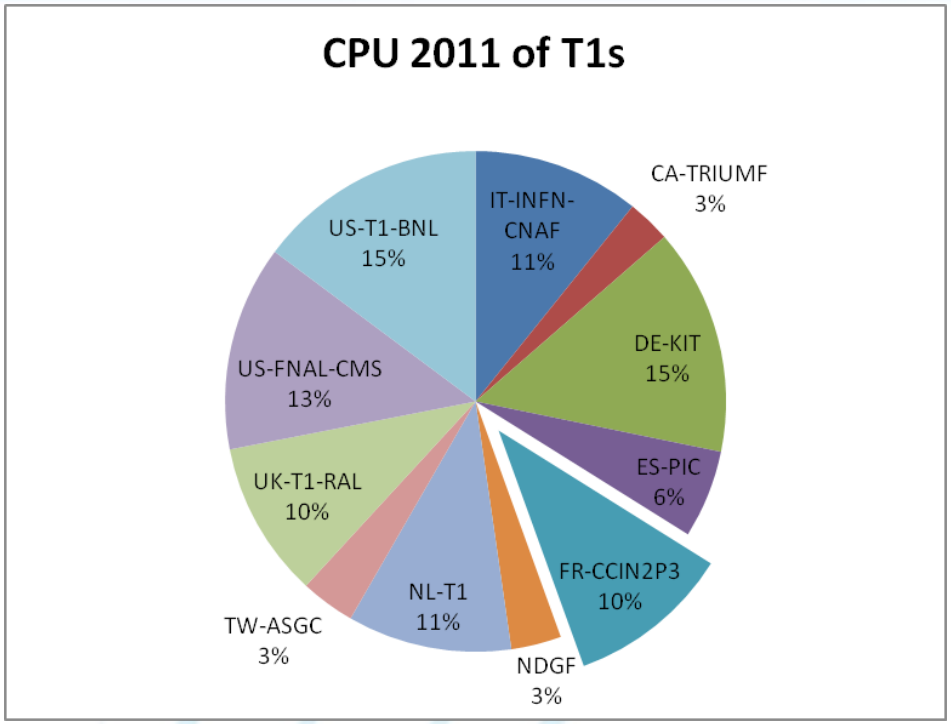
tape

disk

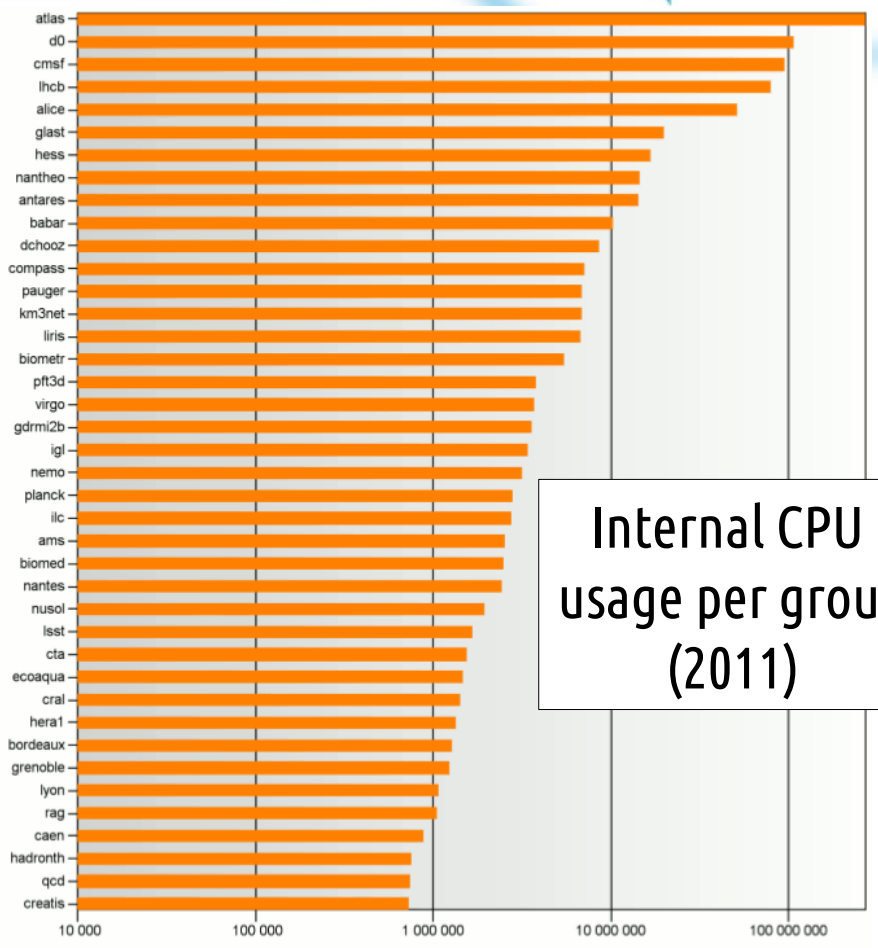
+ GPFS

~16k workpoints (essentially SL5)
~30PB storage (tape+disk)

CPU views



Share in WLCG



Internal CPU usage per group (2011)

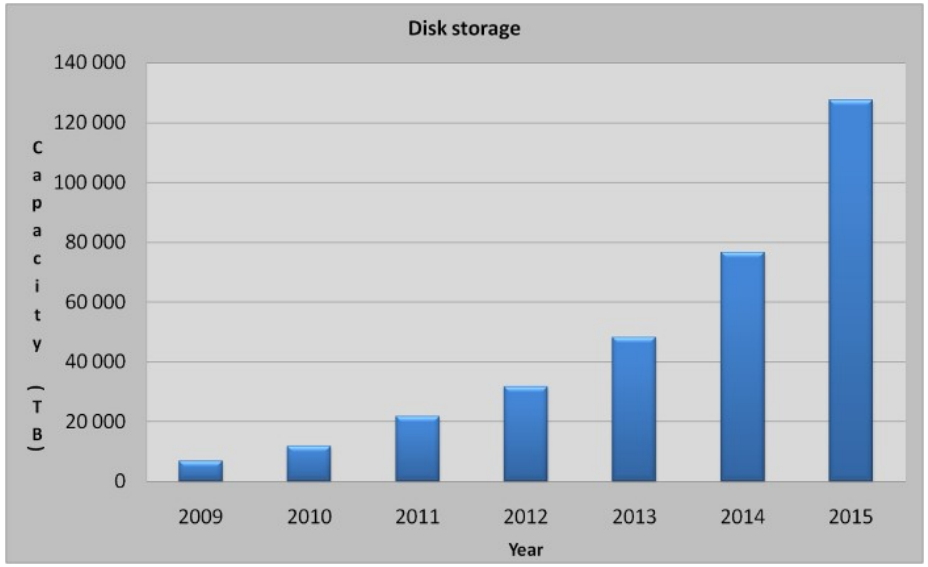
(Log scale)

HS06.h

Extension of CC-IN2P3

- Previous architecture of CC-IN2P3 could not face the upcoming growth of computing resources
  - Space, cooling reasons
- A new building was built to cope with these needs
  - Twice more space
  - New cooling technologies
  - Environmental aspects

A big design phase, a big investment for a big project :
unique opportunities in scientific data processing



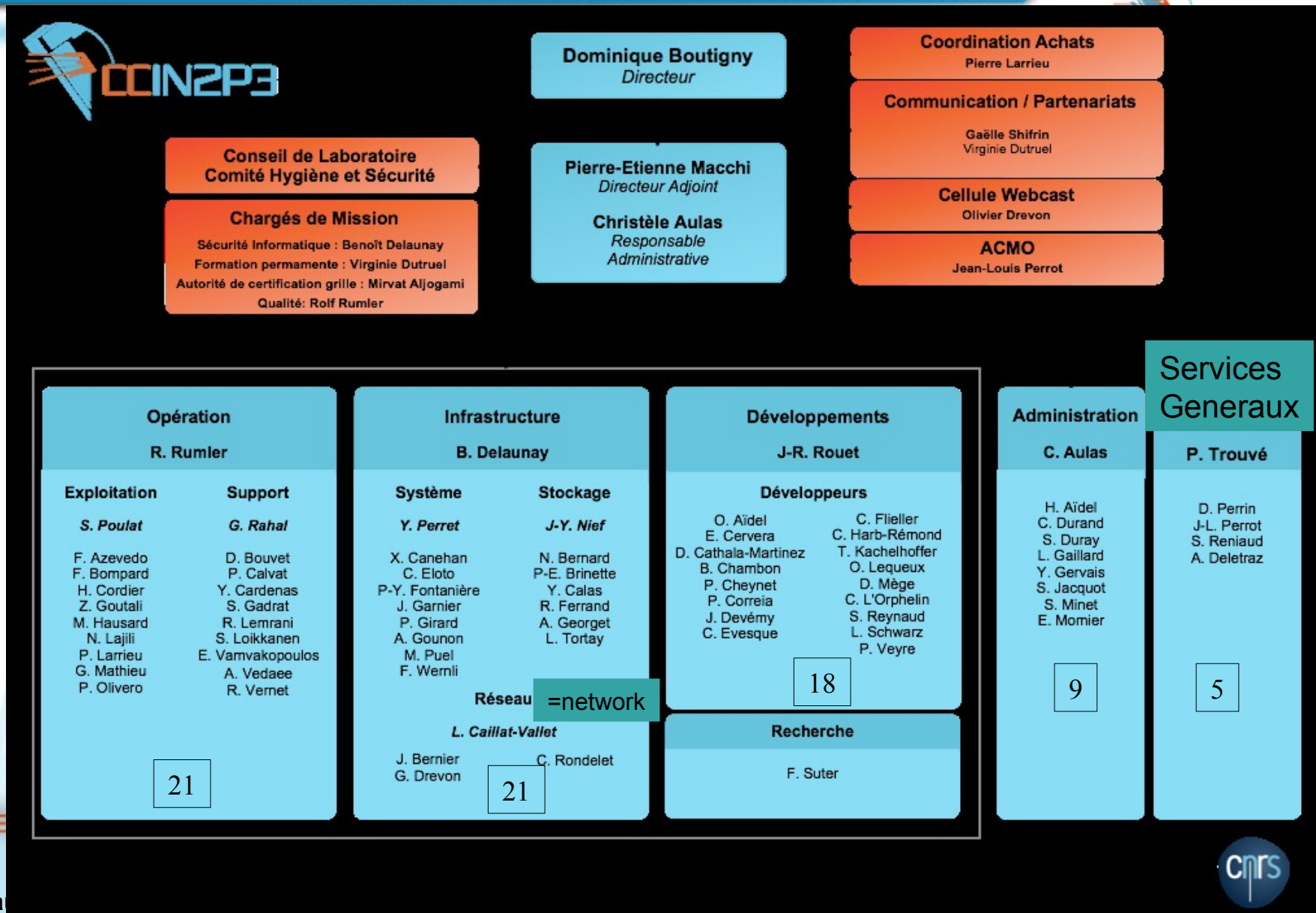
Extension of CC-IN2P3

Operations

Operation in Tier1

- Tier 1 role
  - reliability, availability, storage custody → MoU
- More complexity in its organization than for T2-T3
  - Most experts might be contacted anytime
- F. Carminati @ InGRID'12 :
  - « it's more about people than machines »

Organization



Operation team organization

- Two operation groups
  - Exploitation
    - Batch system, shares & objectives per group, farm status control, robotics, alarms...
  - User support
    - Interface with experiments, VOs, new groups, ressource requests
    - Dedicated support : LHC + Astroparticles + Biology
    - General support : all the other users & experiments
- 1 control room
  - Front line helpdesk
    - Fast reaction to tickets, problems
    - News & information to the community (downtimes, interventions...)
  - Always 1 person from each group in the c.r. during working hours
  - Weekly turnover

The control room

Main operation activities

- Ensure working state of relevant services for the experiment
  - CE's, batch system, storage, inter-siter transfers
- Inform and report about activity and problems
  - Main tools : savannah, ggus, CERN elogs, email...

- Operation duties very different depending on VO
- Different computing models mean different :
  - Services
  - Monitoring tools
  - Communication channels
  - Organization wrt other sites and T0
  - ...

→ basically different on « everything »

→ we employ ~0.5 FTE support person* per VO

** physicists aware of the computing activities of the experiment*

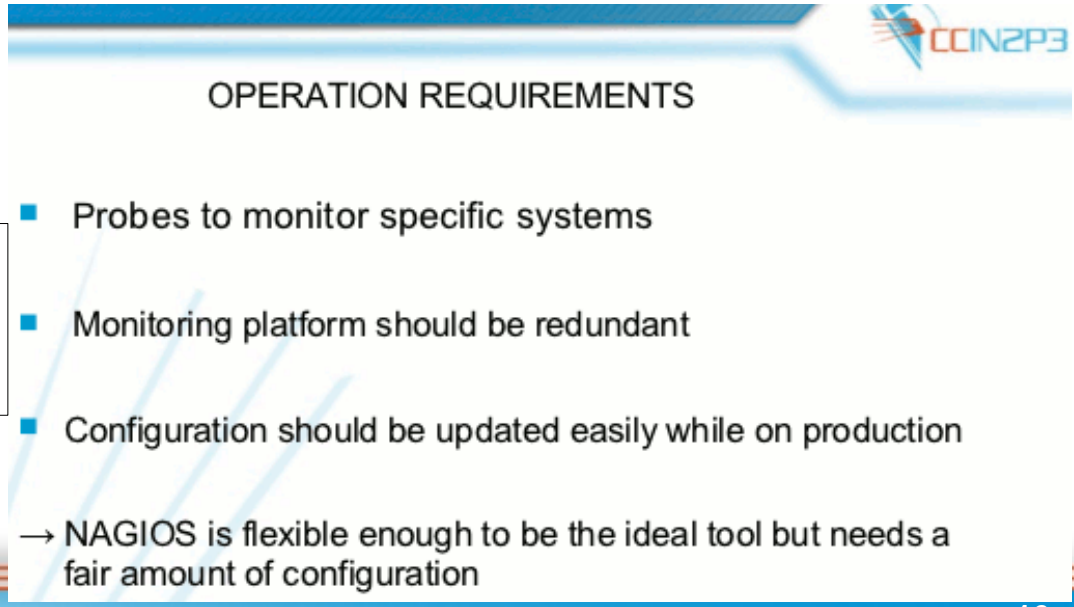
Not talking about hardware maintenance, which is done by another team

Watch & alarms

- Most services at CC-IN2P3 are monitored by Nagios®
  - (Mass) storage systems
    - dCache, xrootd, iRODS, HPSS
  - File systems
    - AFS
  - Batch system and nodes, grid elements
  - Databases

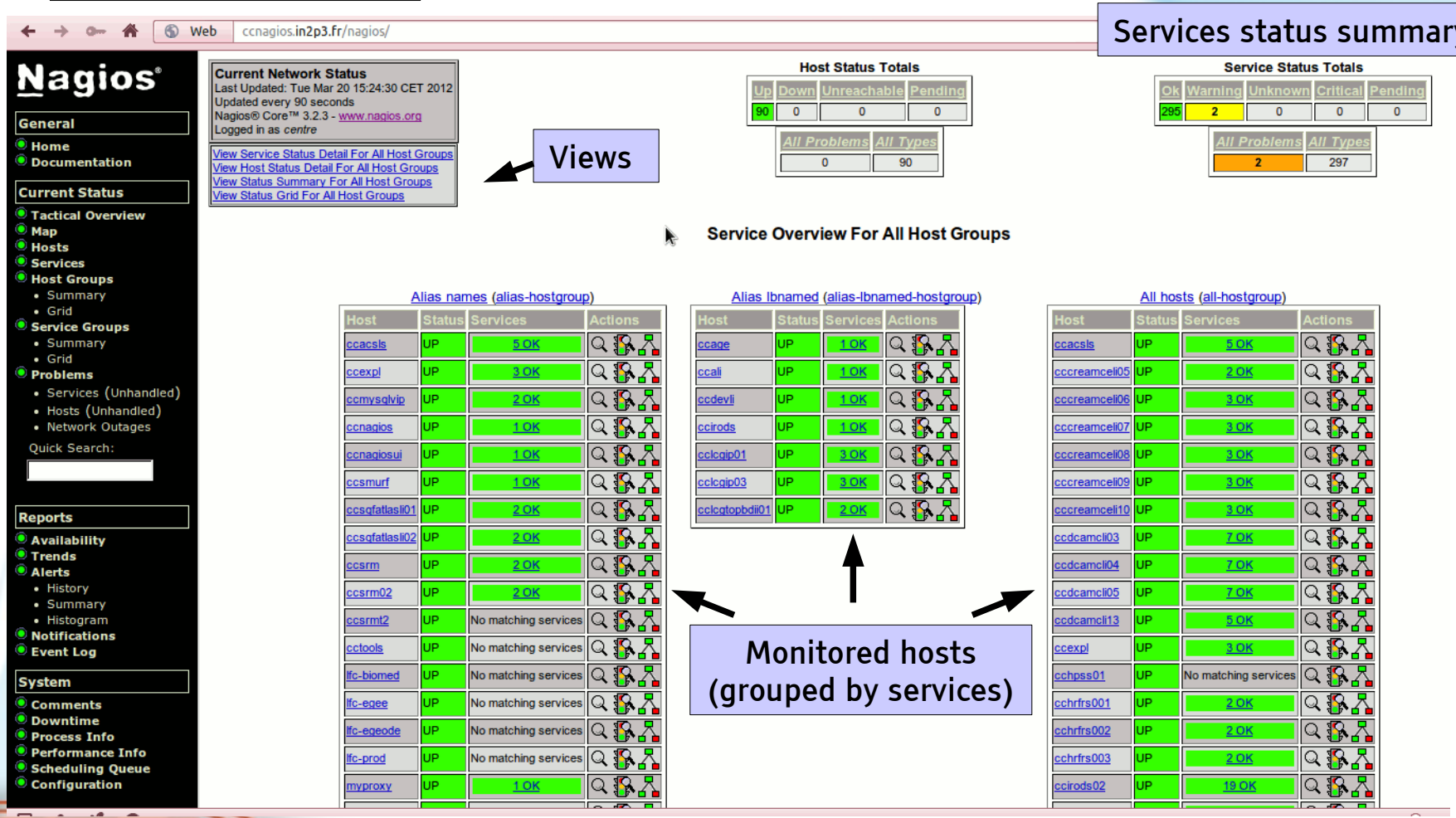
Quote from our nagiosmaster (HEPIX fall '09)

- tests are written by the service admin and given to Nagios
- 4 states : OK, warning, critical, unknown



Watch & alarms

CC-IN2P3 Nagios portal



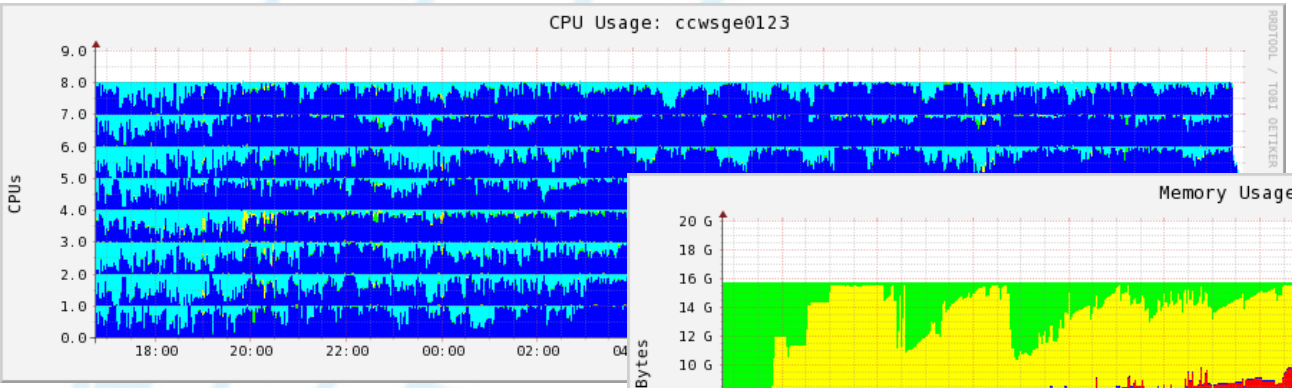
Services status summary

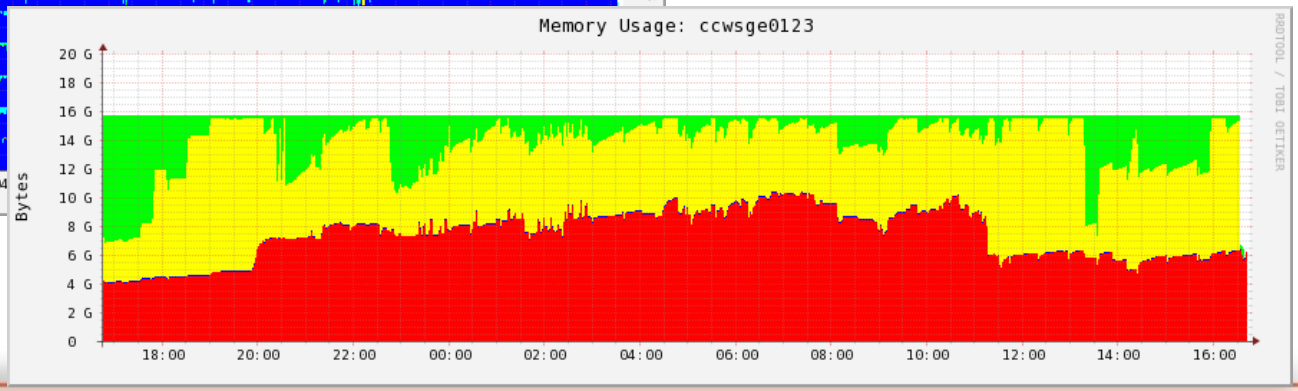
Watch & alarms

- Notification
  - email, sms, central system console
- Short automatic message describing the problem
- Link to wiki
  - each probe has a documentation and recipes in a wiki page
  - Used by operation (working hours) and shifter (otherwise)
- Nagios® is mainly for operation (services)
  - Sysadmins have got their own tools for system monitoring

System monitoring

- Home-made package based on RRDtool & perl
  - Gathers system information over time, fills DB, plots
- Manages
  - Site metadata
  - Machine info. vs time
  - Client/server interaction

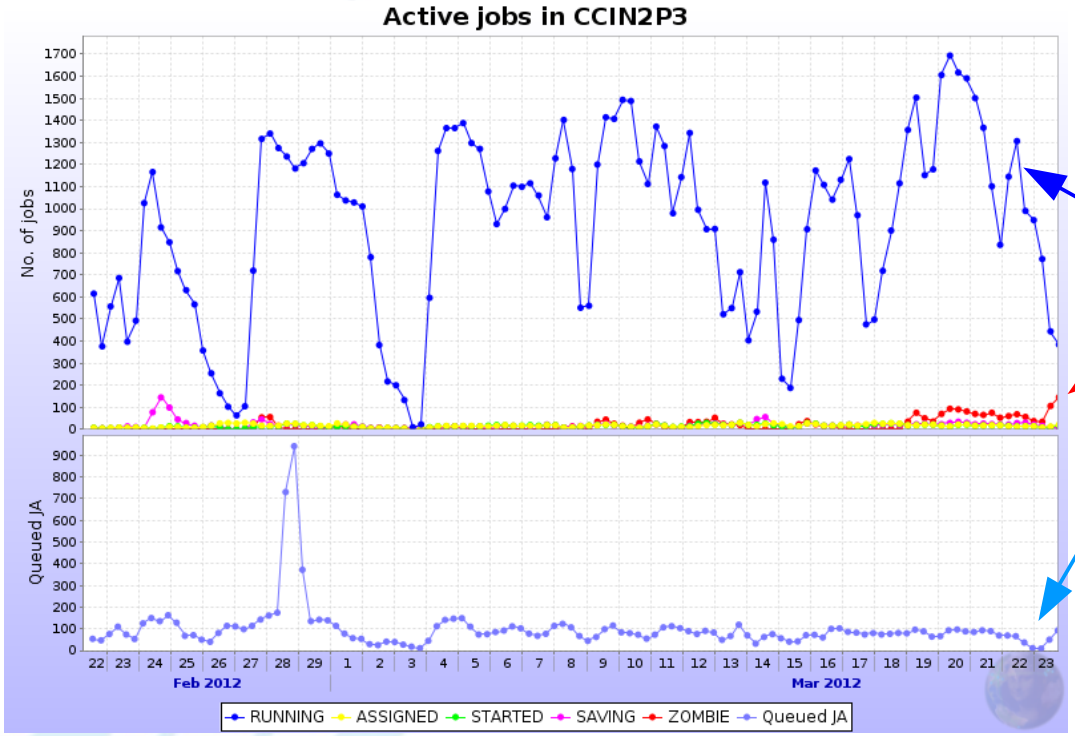




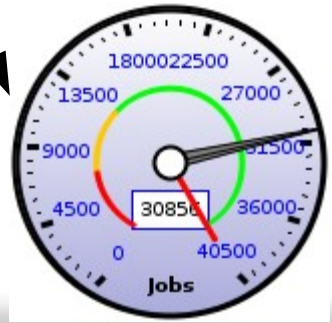
And network, disk IO,
Temperature, # processes etc...

Example of ALICE support activities

- Mix of internal (CCIN2P3) and external (ALICE) tools
- ALICE tools for operation : MonALISA (MONitoring Agents using a Large Integrated Services Architecture)

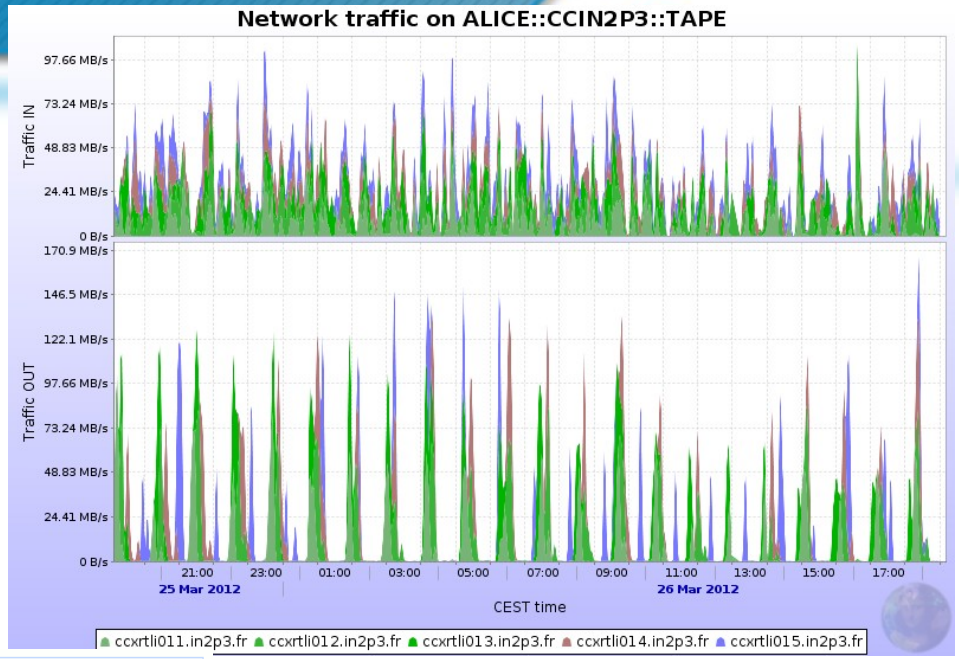


- check running jobs
- understand zombie jobs
- check queued job agents
- check global trend



Example of ALICE support activities

- Raw data transfer monitoring



Transfer requests (add new request)

ccin2p3 | Running | Filter

ID	Path	Target SE	Status	Progress	Files	Total size	Started	Ended
3409.	/alice/data/2011/LHC11a/000146686/collection	ALICE::CCIN2P3::TAPE	Running		360	1.012 TB	22 Mar 2012 07:59	
3407.	/alice/data/2011/LHC11a/000146689/collection	ALICE::CCIN2P3::TAPE	Running		1147	3.182 TB	22 Mar 2012 07:59	
3405.	/alice/data/2011/LHC11a/000146746/collection	ALICE::CCIN2P3::TAPE	Running		400	1.091 TB	22 Mar 2012 07:59	
3404.	/alice/data/2011/LHC11a/000146747/collection	ALICE::CCIN2P3::TAPE	Running		1360	3.86 TB	22 Mar 2012 07:59	
3403.	/alice/data/2011/LHC11a/000146748/collection	ALICE::CCIN2P3::TAPE	Running		560	1.567 TB	22 Mar 2012 07:59	
3402.	/alice/data/2011/LHC11a/000146801/collection	ALICE::CCIN2P3::TAPE	Running		760	2.107 TB	22 Mar 2012 07:59	
3400.	/alice/data/2011/LHC11a/000146803/collection	ALICE::CCIN2P3::TAPE	Running		160	388 GB	22 Mar 2012 07:59	
3399.	/alice/data/2011/LHC11a/000146804/collection	ALICE::CCIN2P3::TAPE	Running		3297	9.496 TB	22 Mar 2012 07:59	
3394.	/alice/data/2011/LHC11a/000146812/collection	ALICE::CCIN2P3::TAPE	Running		286	818.9 GB	22 Mar 2012 07:59	
3392.	/alice/data/2011/LHC11a/000146814/collection	ALICE::CCIN2P3::TAPE	Running		760	2.122 TB	22 Mar 2012 07:59	
3389.	/alice/data/2011/LHC11a/000146817/collection	ALICE::CCIN2P3::TAPE	Running		680	1.859 TB	22 Mar 2012 07:59	
3385.	/alice/data/2011/LHC11a/000146856/collection	ALICE::CCIN2P3::TAPE	Running		1120	3.214 TB	22 Mar 2012 07:59	
	12 requests				**10890**	**30.69 TB**		

Requests per page: 100

Use of internal system tools and host access remains however essential in case of trouble.

Feedback on OGE batch system

- Smooth transition from BQS to OGE in 2011
  - From Spring (open OGE) to Fall (close BQS)
- New nice features available
  - Interactive nodes, multicore jobs, array jobs
- Use of opensource version till feburary, then Oracle
- Difficulties we have encountered
  - Stability
  - Publication to CREAM-CE
  - Monitoring and accounting
    - Retrieve relevant info and understand it
  - Quite slower reactivity of the online support

→ quite a bit fine tuning had to be done
→ ongoing Requests For Improvement
→ still several points to fix to run smoothly

Preparing the future...

Parallel computing

- Increasing interest for parallel computing
- Already used by
  - Astroparticles (Planck)
  - Lattice QCD
  - …
- Multicore jobs
- Interest on GPU
  - ALICE
  - Lattice QCD
  - Astroparticles

MPI jobs
(1024 cores)

Dedicated infrastructure
to be provided very soon
(Dell, chassis with 4 GPU blades)

Cloud at CC-IN2P3

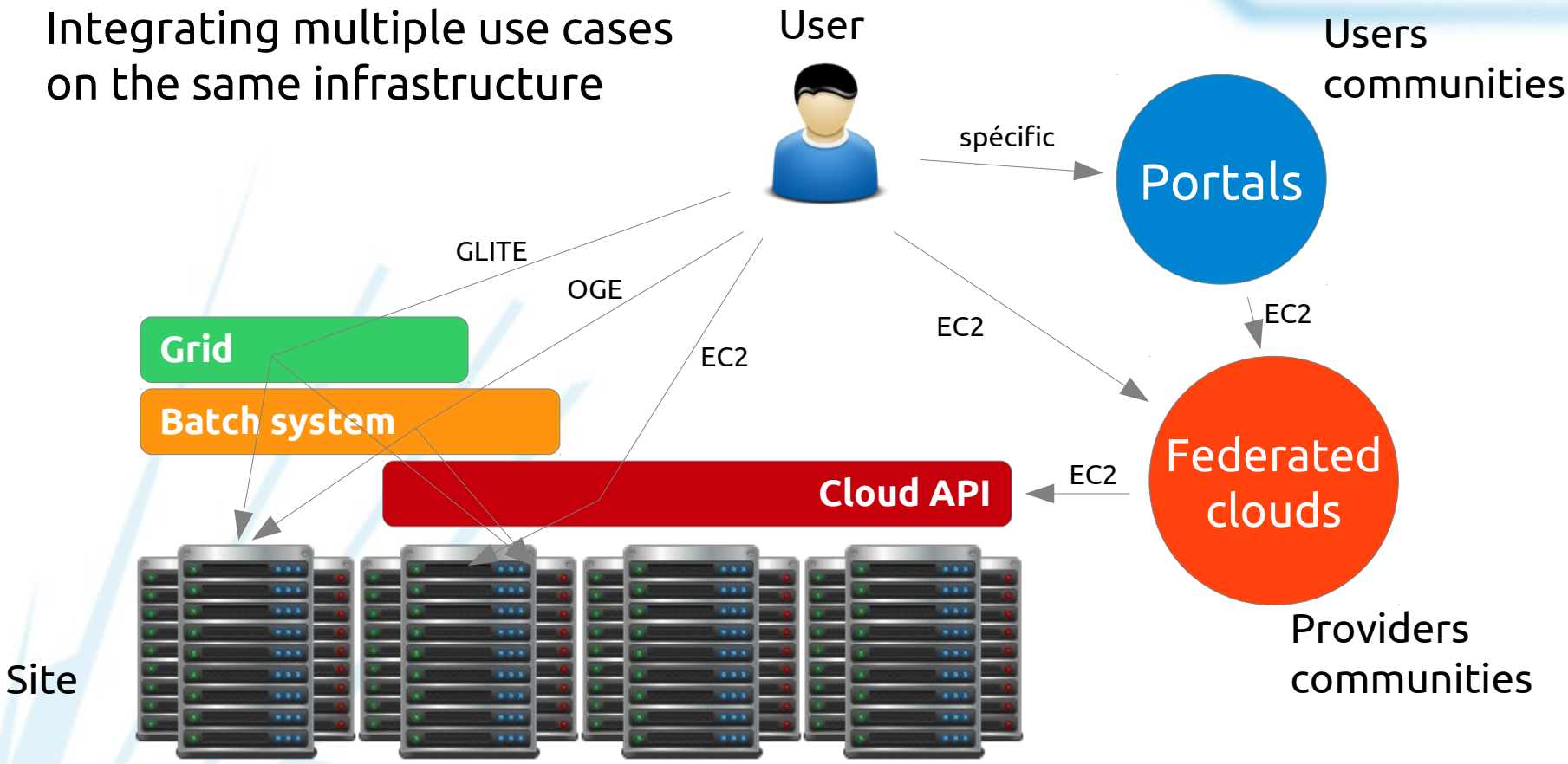
- Main goals
  - Build an academic community cloud, integrated to wider federations
  - HTC use case is no option
- Motivation & user needs
  - Users need flexibility (elasticity)
  - Another way of achieving distributed computing
  - Satisfying new use cases (servers on demand...)
- Steps
  - Offer IAAS ressources through generic interfaces (EC2/OCCI)
  - Integrate national/european/worldwide academic federations

Cloud at CC-IN2P3

- Work in progress
  - Evaluation of existing/upcoming technologies
    - Proprietary : IBM, Dell/Canonical, VMWare, Oracle, RedHat
    - OSS : OpenNebula, Openstack, Nimbus...
  - Reuse experience
  - Identify hot spots
    - Security, storage, performance, networking
- And then
  - Open to new communities
    - Other scientific fields, academic institutes, industry ?
  - High level tools for users (to PAAS/SAAS)
  - Branch to the batch system
  - How will the users follow these new technologies ?

Computing and cloud interfaces

Integrating multiple use cases on the same infrastructure



Visit us !

Lyon is a very nice city, with famous gastronomy and wines !

backup

Example : VO software deployment

- ALICE
  - Previously putting VO software on AFS
    - One common place accessible by any WN
    - Heavy load when many competing jobs
  - Transition to bittorrent last summer
    - Software installed on each worker sandbox
    - Sizeable overhead due to download → fixed in new AliEn
- LHCb, ATLAS
  - CVMFS (same motivation as for ALICE)
- CMS
  - AFS (working fine for this VO)

→ operation team and sysadmins have to maintain and deal with different services

A cloudy approach to use cases

- IAAS

  has to conform to adopted standards (EC2/OCCI/CDMI...)

  offer VMs instanciation

  offer storage cloud resources (ala S3)

- PAAS

  Example : provide support for Grid Engine cluster instanciation

- SAAS

  integrate workflow management solutions (sysfera)

  eg : support for jobs submission through web portals

Testbed infrastructure

- 16 Dell Poweredge C6100 hosts
- On each node
  - 2 Xeon 24cores X5675 @ 3.07GHz
  - 96GB RAM
  - 2TB raid 10 local storage (4 SAS 7.2 krpm)
- Catalog of images (1.4 TB)
- Public IPv4 subnet → full VM network isolation
- Powered by Openstack
- Available interfaces :
  - EC2 and Nova API at the moment
  - Ongoing work on OCCI

Cloud : identified use cases

- EGI FCTF testbed (HEP, traditional grid users)

  provide IAAS ressources to former grid users, computing

- Webimatics (neurobiology)

  medical image data analysis

- Etriks (academic and pharmaceutical companies)

  data analysis

- Sysfera

  SAAS approach to workflow management

- Dirac

  Yet Another grid scheduler