FRANCE GRILLES

UNIVERSITÉ DE BORDEAUX

CCIN2P3

cgfb
BIOINFORMATIQUE

IPHC
Institut Pluridisciplinaire Hubert CURIEN
STRASBOURG

LPSC
Grenoble

A National Data Management System based on iRODS in the French Grid Initiative

C. Biscarat on behalf of the France Grilles iRODS team

(D. Benaben, C. Biscarat, Y. Cardenas, P. Gay, B. Hiroux, G. Mathieu, J.-Y. Nief, J. Pansanel and G. Romier)

iRODS User Group Meeting, 18-19 June 2014, Cambridge, MA

Before we start

- This talk is not about
  - Big numbers
  - Complicated setups
- Rather, it is about
  - A new (modest) initiative
    - to share resources and expertise
    - across scientific fields and institutions

Outline

- The context
- Our project
- Its status
- Outlook

The Context

Scientific landscape today

- Large amount of data being produced
  → store, manage and analyse
- Data are heterogeneous (size and type)
- Collaborators spread out over different geographical locations (France, Europe, World)
- Collaborations do not (necessarily) have computing gurus

“ Large” for some could be “small” for others

Why iRODS

Suitable solution to answer users' needs:

- Analysis of large data volumes distributed in different locations
- Collaboration between sites with heterogeneous systems
- Physical organisation of data transparent to users
- Data findable by metadata search
- Fine-grained access control
- Data available from anywhere
- Data management capability (user policy)

- As a user: simple to use

- Free software

A group of people

Different backgrounds, different experience

- **Y. Cardenas & J.Y. Nief**, CC-IN2P3 (particle, nuclear and astroparticle physics)
  - Manage PB of data with SRB/iRODS since years (multi-disciplinary)
    See JY's talk today
- **C. Biscarat**, LPSC (particle, nuclear and astroparticle physics)
  - iRODS user since 2011 (CIMENT HPC centre grid)
    See Xavier's talk today
- **J. Pansanel**, IPHC (multi-disciplinary)
  - Has installed iRODS for local users (zone “IPHC”)
- **P. Gay, B. Hiroux**, MCIA (multi-disciplinary) & **D. Benaben**, CBiB (bioinformatics)
  - Plan to install iRODS (O(1PB)) together with the local HPC & grid center

Why a coordinated effort

- For scientists
  - Get small communities to benefit from the service
  - Use or test a data management solution
- For us as a group
  - Minimise costs by sharing resources and effort
  - Share and gain expertise
  - Long term: provide a reliable infrastructure

France Grilles

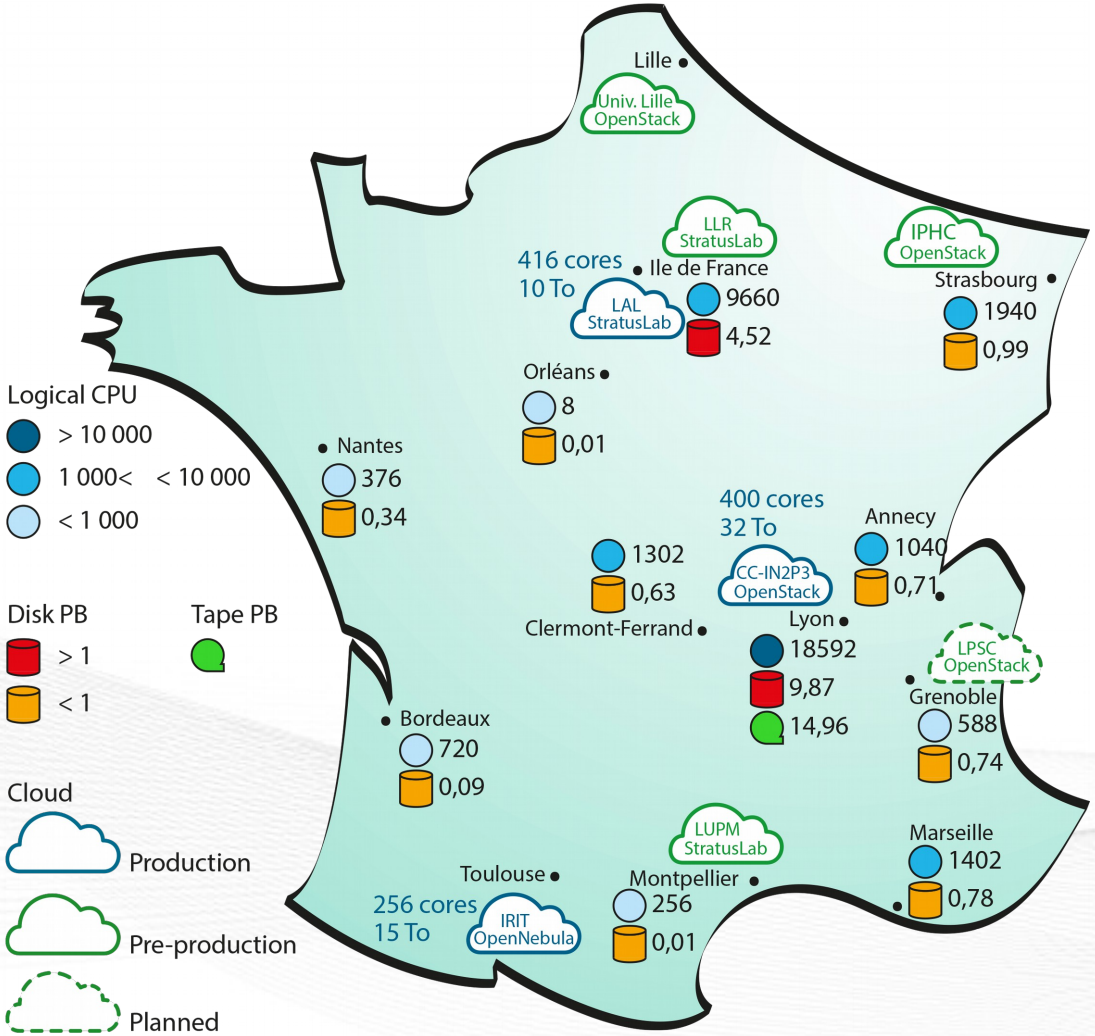
http://www.france-grilles.fr/

- France is active in Production grids since 2000 (DATAGRID, EGEE)
- Born in 2010: “France Grilles” is a "Scientific Interest Group", operated by CNRS
- Partners:

- Main objectives
  - Coordinate the operation of French production grids and clouds
  - Promote their use by Scientific communities
  - Develop a collaborative network of expertise
  - Contribute to the operations of the European Grid Infrastructure http://www.egi.eu/
    - France Grilles is the French “National Grid Initiative”

- France Grilles provides a reduced dedicated grid and a Virtual Organisation
  - Global resources are provided and owed by the organisations

An extended network of people



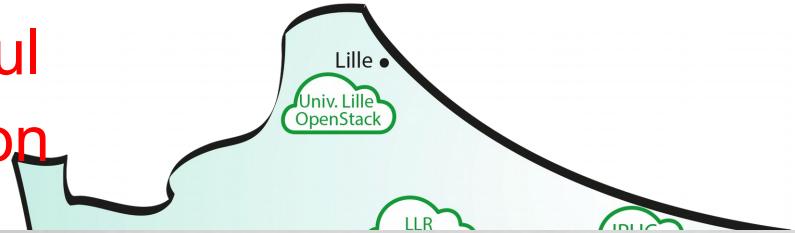
- Research institutes
- Regional centers
- Universities

- A centralised Dirac service

An extended network of people

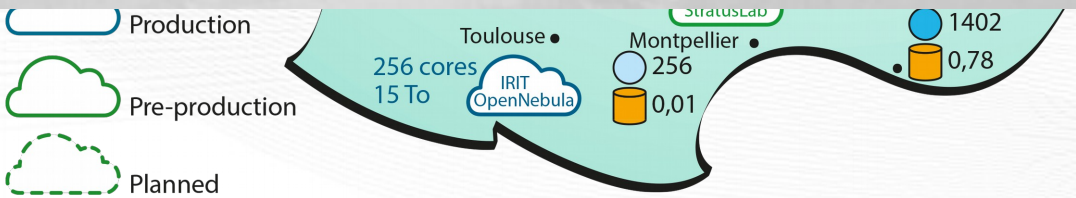
A data grid and a powerful data management solution

- Research institutes
- Regional centers
- Universities



iRODS

TAKE CONTROL OF YOUR DATA



The way we are organised

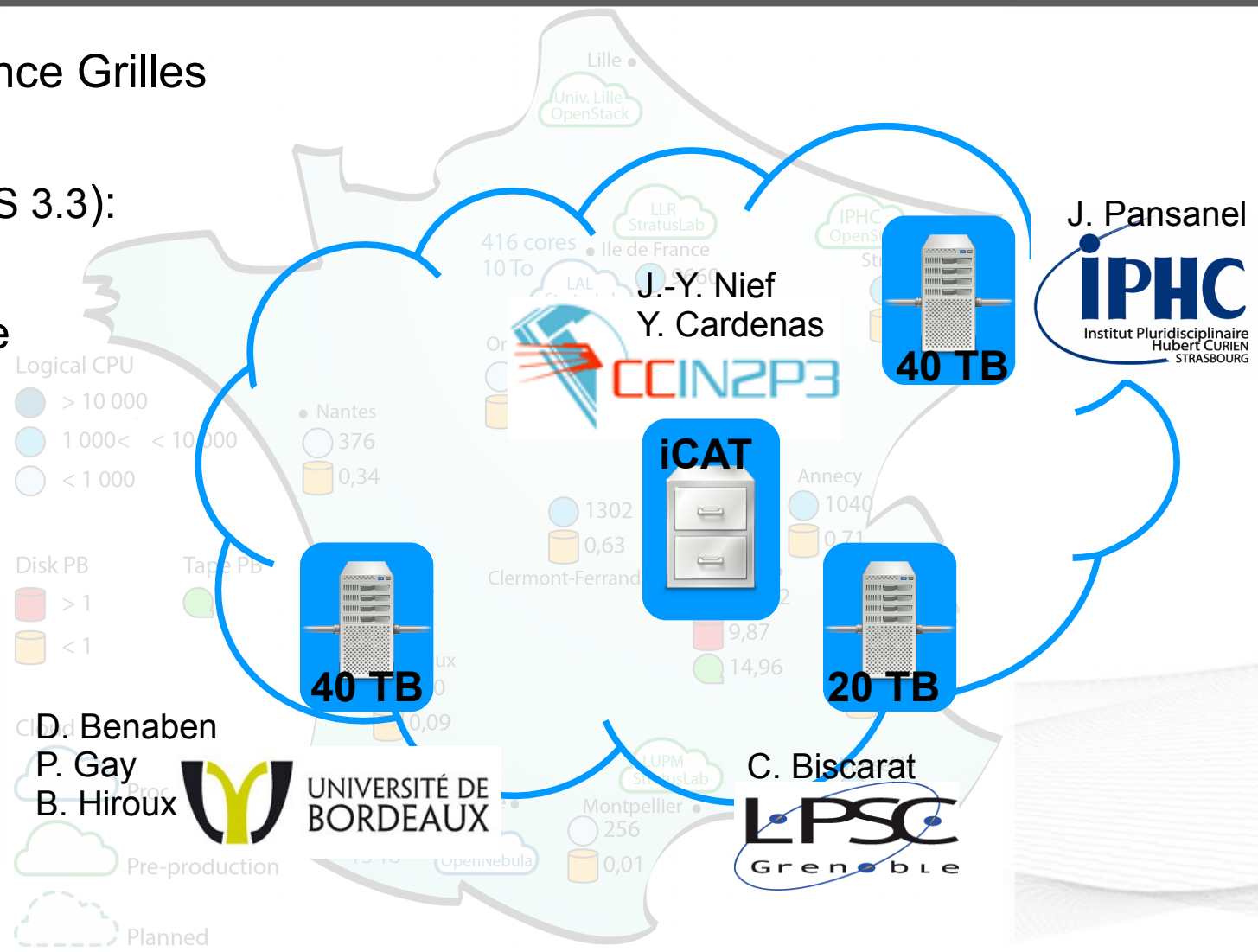
Our project

- Back in 2012
  - The French NGI started to have a look at iRODS
  - “IRODS Administration” tutorial at CC-IN2P3
- Our team gets organised in 2013
  - Participating sites identified
  - As of today: no MoU
- Our goals
  - Build an iRODS instance across organisations
  - Offer the service to any scientific fields
    - Small to medium size projects
  - Build a network of expertise

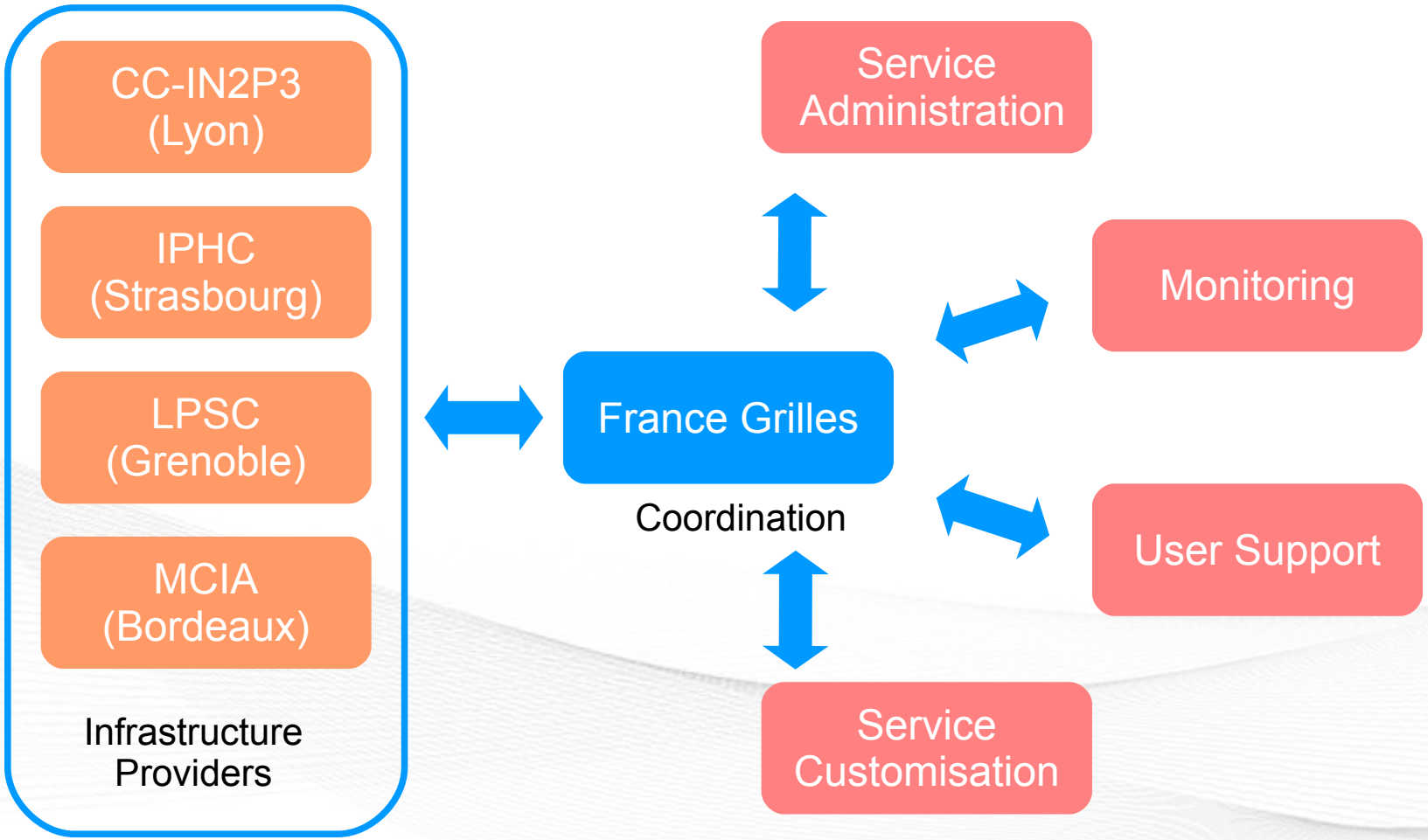
A single iRODS zone “FRGRID”

Coordinated by France Grilles

In production (iRODS 3.3):
- Central catalogue
- Distributed storage



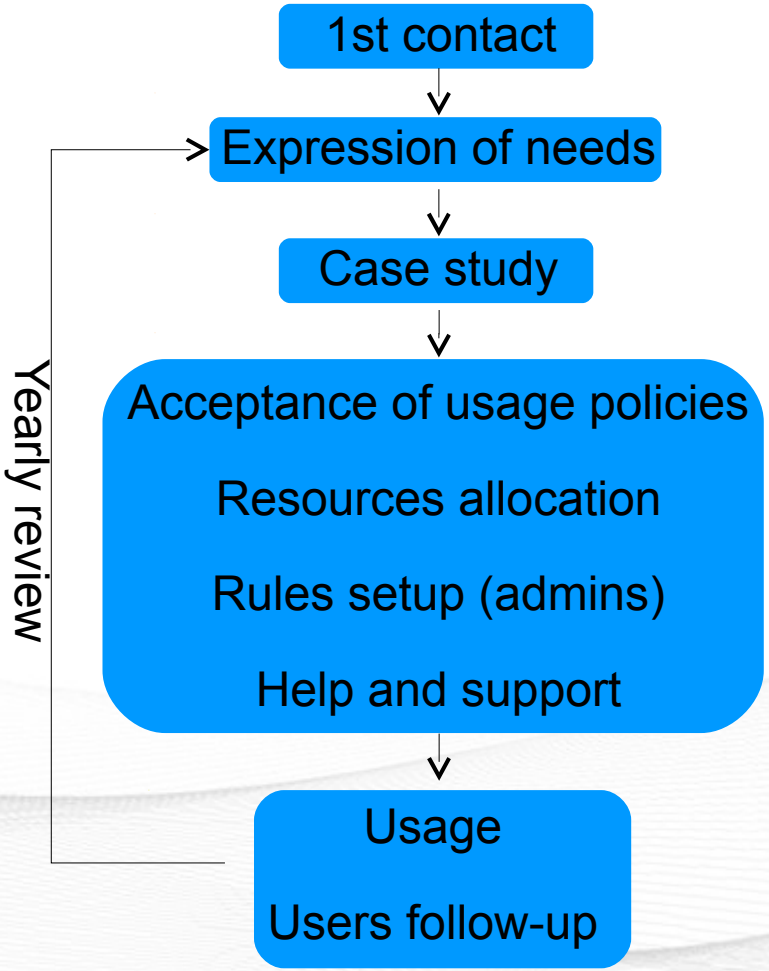
Beyond the infrastructure



User hosting

Scientific data, any domain

- Needs
  - Data distribution
  - iRODS system evaluation
  - Coupling with the EGI computing grid
- Authentication
  - Identifiers
  - Certificate



Usage policy & data life cycle

Prerequisite for users

- To own an electronic certificate from an academic organisation
- To belong to the national VO “vo.france-grilles.fr”

→ users have to sign up the VO "Acceptable Use policy", an extract:

- Projects are accepted on one year basis and until the autumnal yearly review.
- The data stored in iRODS must be scientific and related to the project.
- When the agreed period of the iRODS usage ends, the user should get his data back within a one month period. Afterwards, the data are removed.

Where we stand

Our team provides

Service customisation

Service customisation

Goal: support Windows, Unix and MacOS users

- Prepare specific tools and make them public
  - iRODS icommands packaging with GSI support (deb, rpm)
  - VM appliance to access the grid (User Interface) and icommands
- Couple our iRODS infrastructure with the grid
  - icommands installed on grid sites supporting the national VO
- “iRODS Browser” available

Our team provides

User Support

User support

Goal: help the user to get started and follow-up

- Help the user to get started
  - Tutorial for Users (February 2014, this autumn)
  - Collaborative documentation (wiki)
- Personalised advice of the user specific application
  - Rules to be setup on demand

A fresh start

Technically

- iCAT and resources in production
- Real synergy between the administrators
- Ready to get started with users hosting

On the user end-point

- First user training has been performed
- Our very first users have subscribed

Extending the network

- France: our team has been contacted to discuss existing/future iRODS systems
- Europe: our project was presented in this year EGI Community Forum

Our very first (beta) users

- #1 Life Science & Bioinformatics – two users
  - Needs: analysis of the data on the grid (500 GB)
    - Specific libraries have been deployed
  - iRODS is easier than the pure EMI commands

- #2 Life Science & Bioinformatics – two users
  - Needs: regular exchange of data between three labs in France (1 TB/year)
    - In the framework of the National proteomic infrastructure (ProFI)
    - Evaluation of the iRODS solution
  - Automatic replication
  - Metadata usage
  - Graphical interface

→ Time Scale: this summer

What's next

Short and long terms

- Get feed-back from our users
  - Identify what's to be improved
- Regarding the technical infrastructure
  - Develop a centralised monitoring solution (24/7 planed)
  - Test the S3 plugin
- Regarding the project
  - Find new financial resources to ensure the sustainability of the infrastructure
  - Extend the storage pool with new resource providers
  - Make more users benefit from the infrastructure
  - Share expertise with other groups

Concluding remarks

- IRODS is powerful data management middleware
  - We start with simple things
- We still have some points to clarify
  - To be fully compatible with GSI: is there any Graphical User Interface with an authentication by certificate ?
  - A user starting on our infrastructure is bound to the iCAT
    - What if the user wishes to move from this iCAT to his own?

Questions ?

To contact us:

France Grilles	info@france-grilles.fr
People involved	irods-admin-l@france-grilles.fr
David Benaben	david.benaben@u-bordeaux2.fr
Catherine Biscarat	biscarat@in2p3.fr
Yonny Cardenas	cardenas@cc.in2p3.fr
Pierre Gay	pierre.gay@u-bordeaux1.fr
Jean-Yves Nief	nief@cc.in2p3.fr
Jérôme Pansanel	jerome.pansanel@iphc.cnrs.fr