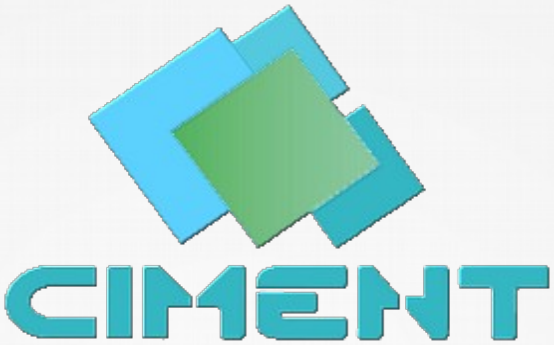


IRODS use case :
Ciment, the Univ. Grenoble-Alpes HPC center

B.Bzeznik / X.Briand

Irods users group meeting

11/06/2015

Irods is used (famous) in the French Alps since 2010!

CIMENT

Plan

- **What is CIMENT?**
  - What is provided by CIMENT?
  - The batch scheduler: OAR
- **How does iRODS work on the platforms of CIMENT?**
  - Infrastructure
  - Cigri middleware: accessing ressources
  - Configuration of the IRODS grid environment
  - Load considerations
  - Cirods and ciget: a usage of Pyrods API
- **Current scientific partnerships**
  - Seismology
  - Rosetta mission
  - Ecology
  - Particle physics

CIMENT

What is CIMENT?

CIMENT : High Performance Computing center of the univ. Grenoble-Alpes

UNIVERSITE JOSEPH FOURIER
SCIENCES TECHNOLOGIE MEDECINE

Environnement
Foehn
Physics
Nanostar
Chemistry
Ceciccluster
LUKE
FROGGY
R2d2/Fostino
Gofree
Science of de universe
Digitalis
Biology
Computer science

CIMENT

Computing platforms

HPC platform	**Data processing platform**	**Other thematic platforms**
  *Froggy*	  *Luke*	*r2d2,foehn,ceciccluster,digitalis,...*
3200 Xeon E5 cores @2.6Ghz +18 GPUS K20m	~400 cores – heterogeneous systems and continuously evolving	~3000 cores heterogeneous systems federated from 10 clusters of member laboratories
High performance distributed storage (Lustre): 90 TB	Local scratches on nodes : 450 TB	NFS filesystems: a few TB per cluster
Infiniband FDR network	10 GE network	Infiniband QDR networks

Common storage (IRODS) : 1 PB

CIMENT

Jobs scheduling : OAR

- OAR is a versatile resource and task manager (also called a batch scheduler) for HPC clusters and other computing infrastructures (→ Grid5000,...)
  - Developped a LIG/Inria by the OAR Team
  - Deployed on all CIMENT platforms

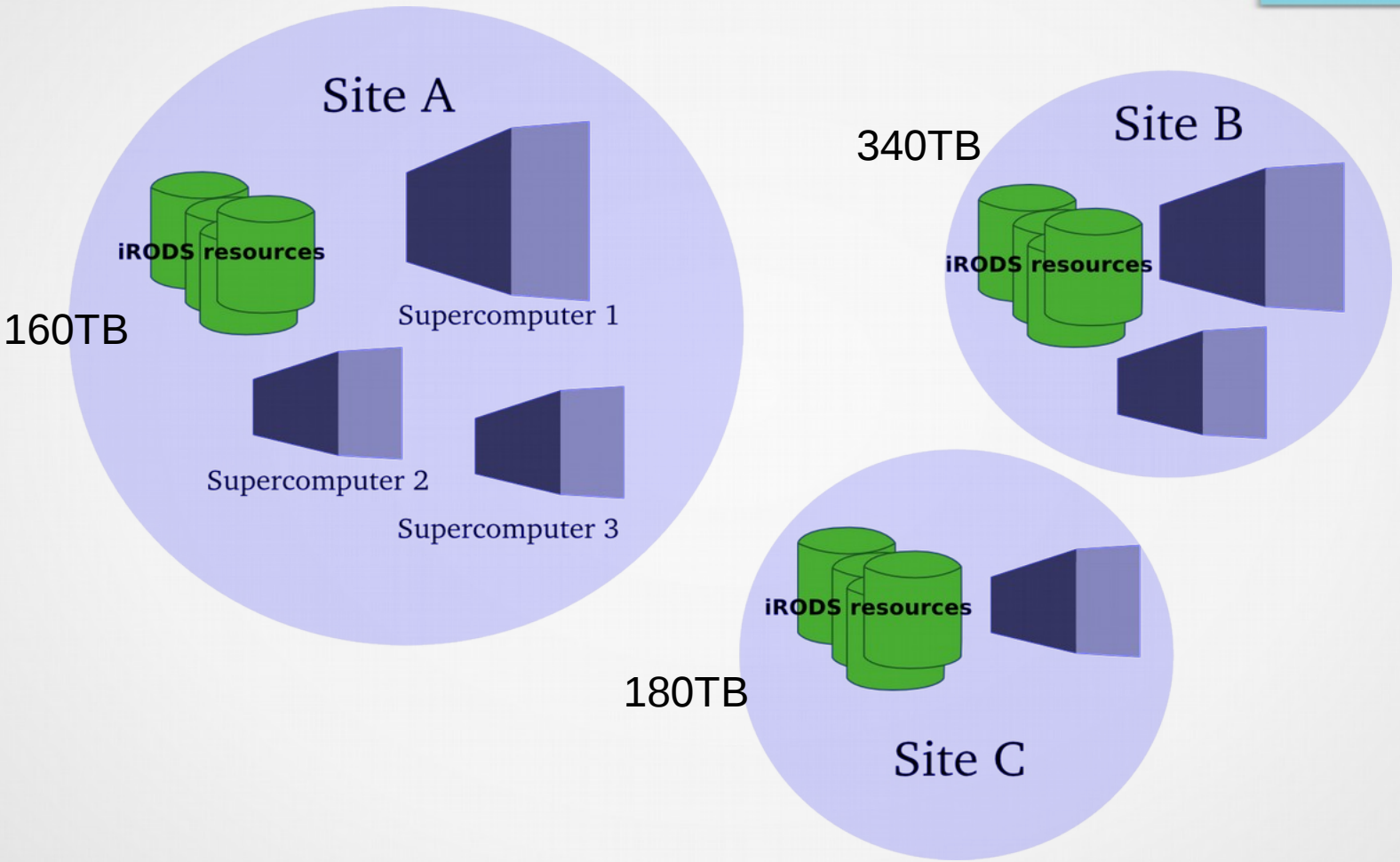
- Best-effort JOBS
  - “Opportunistic” jobs which have the lowest priority and can be killed whenever any other jobs needs the resources
  - Best-effort jobs fill the gaps let between the regular jobs by using any free resources with some kind of “elasticity”

CIMENT

How does IRODS work on CIMENT platforms?

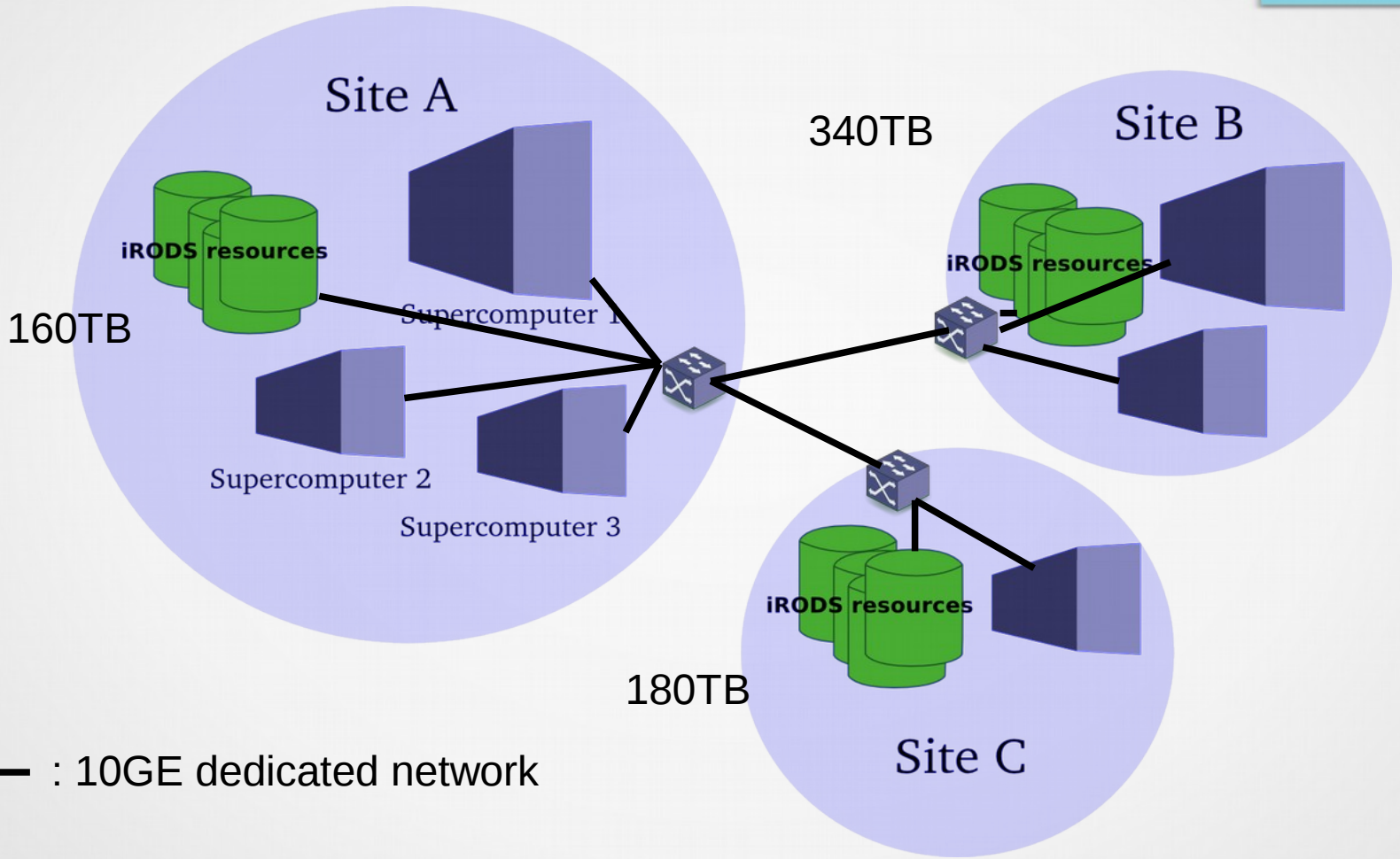
CIMENT

The IRODS infrastructure setup in Ciment



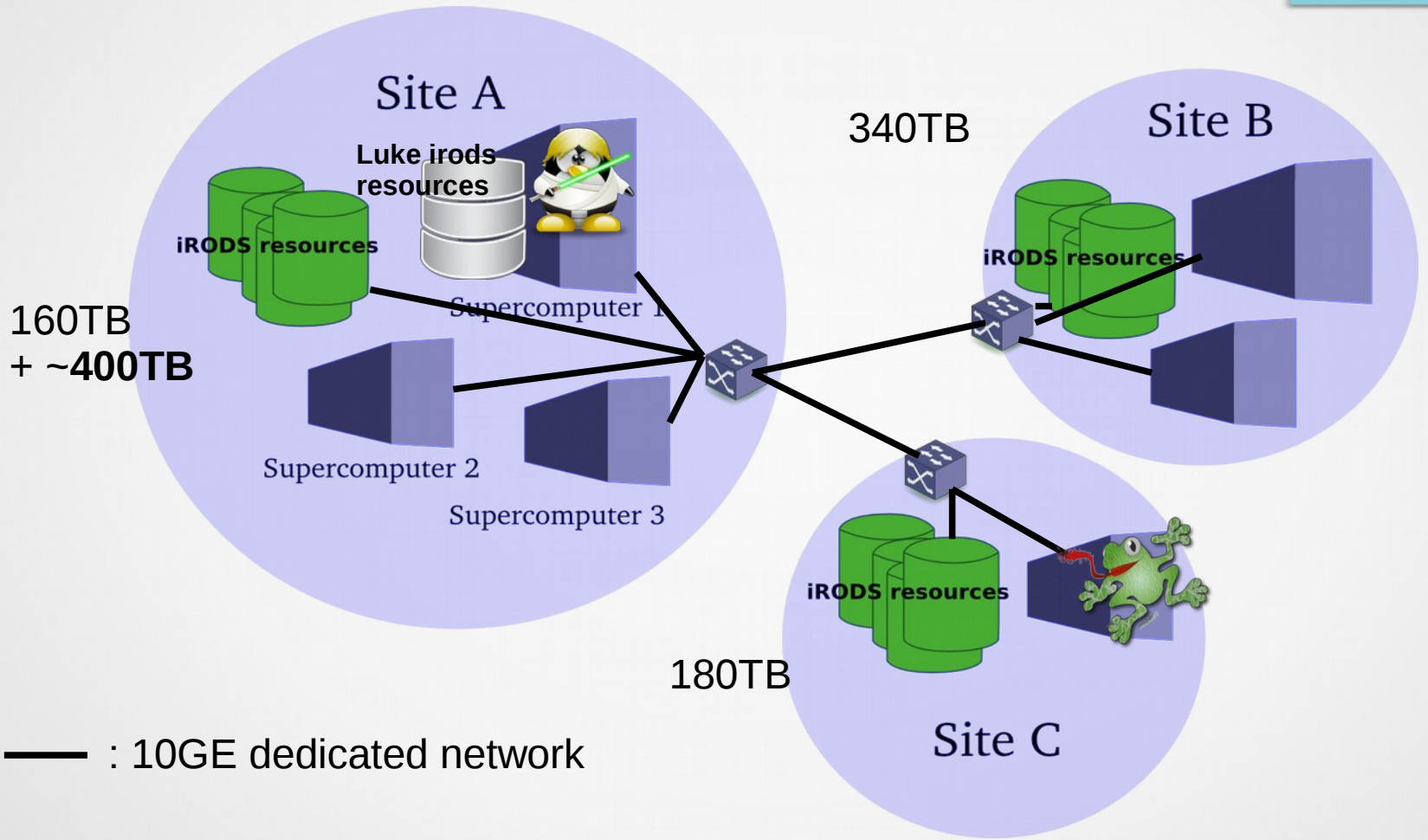
(raw storage sizes)

The IRODS infrastructure setup in Ciment



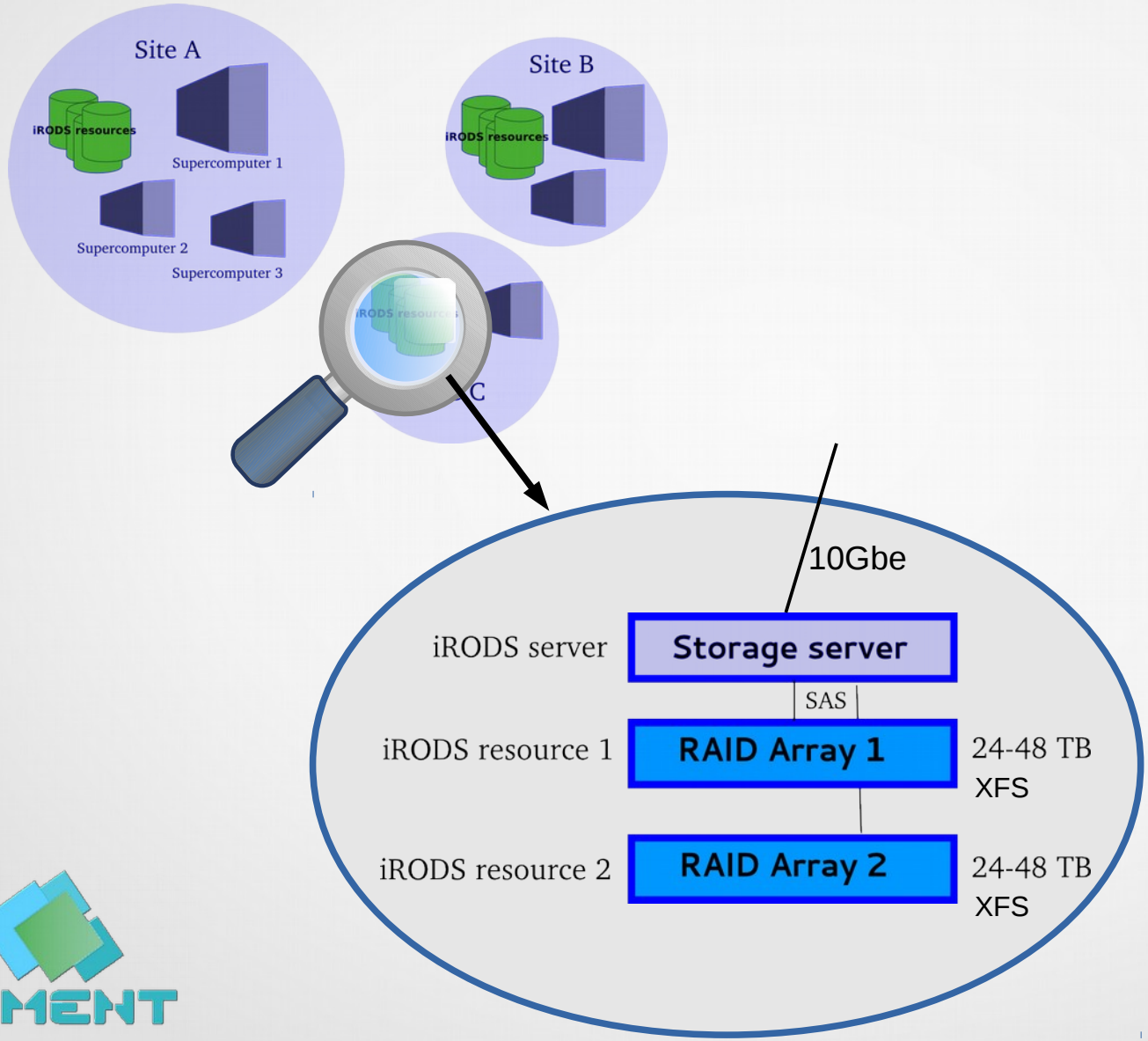
(raw storage sizes)

The IRODS infrastructure setup in Ciment



(raw storage sizes)

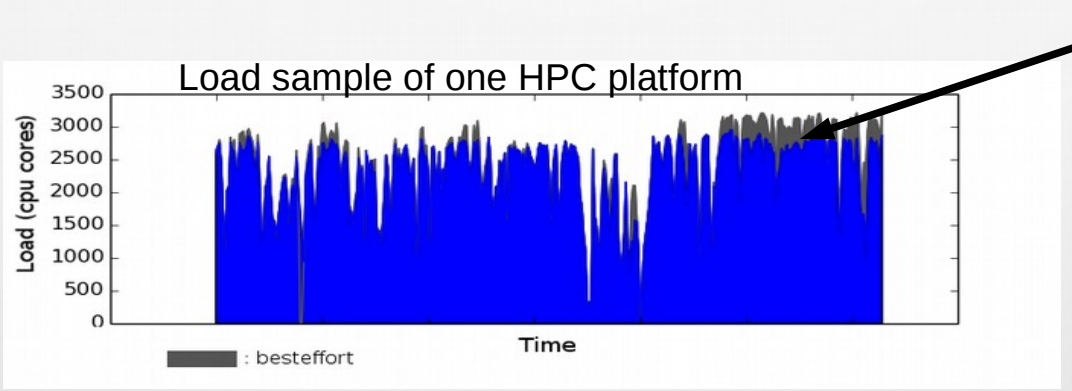
The IRODS infrastructure setup in Ciment



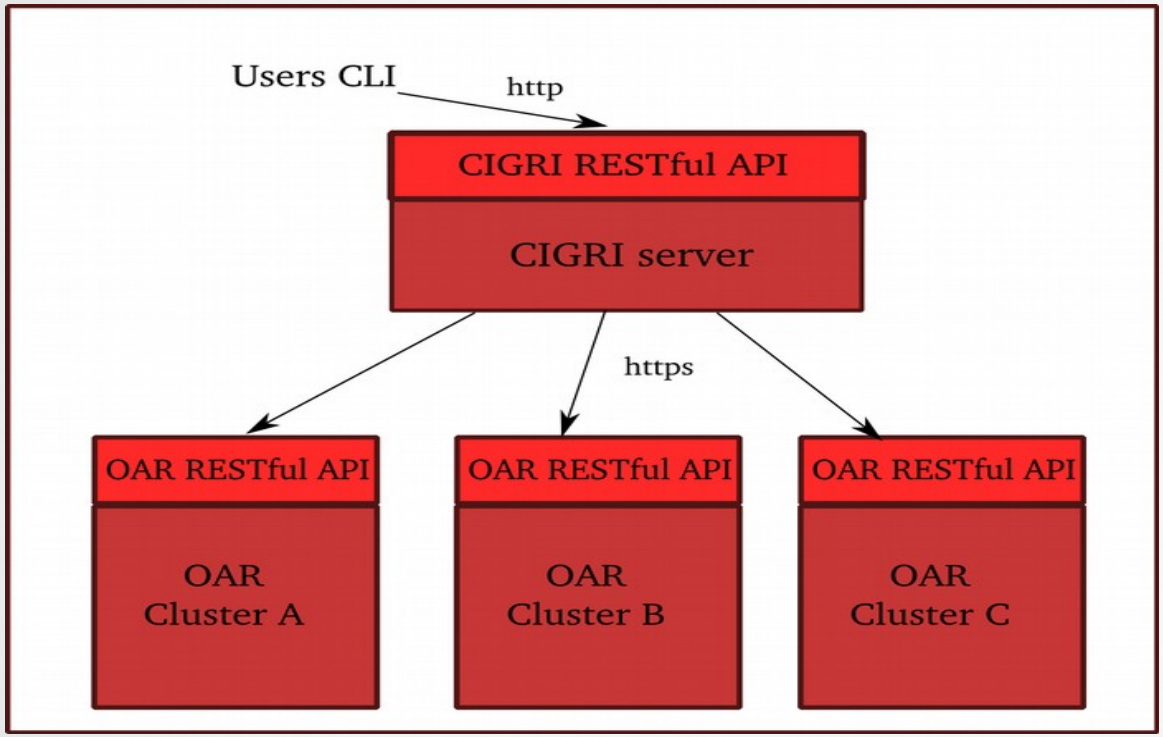
* 3 to 5
(depending on the site)

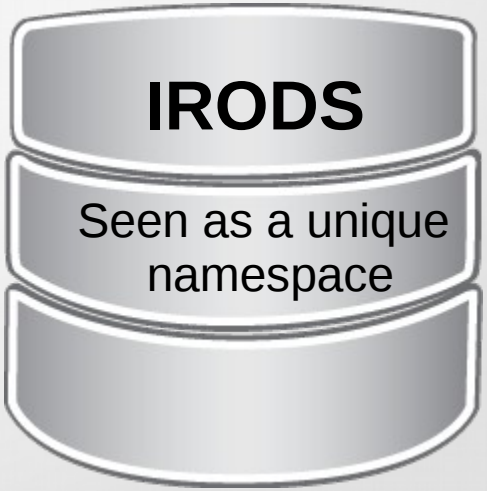
The CIGRI middleware

- Accessing all the federated resources of CIMENT
- A lightweight grid middle-ware
- Developed by the OAR team and Inria (main developer: Bruno Bzeznik)
- Focuses on “bag-of-tasks” jobs (Monte-Carlo style jobs campaigns)
- Optimized for high throughput computing: millions of small independent computations, embarrassingly parallel
- Allows an efficient use of OAR's “best-effort” mode with automatic job resubmission

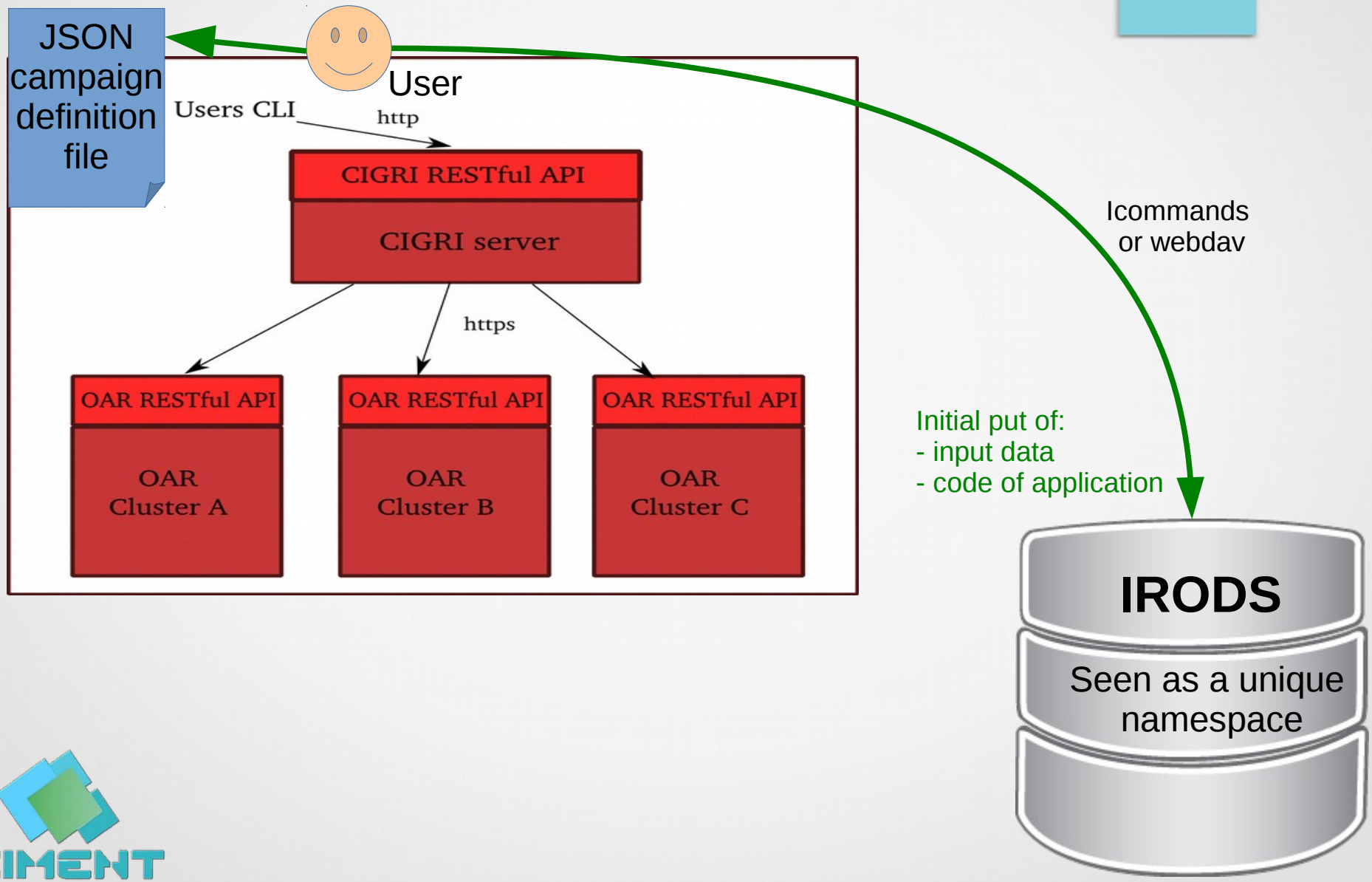


CIGRI and IRODS

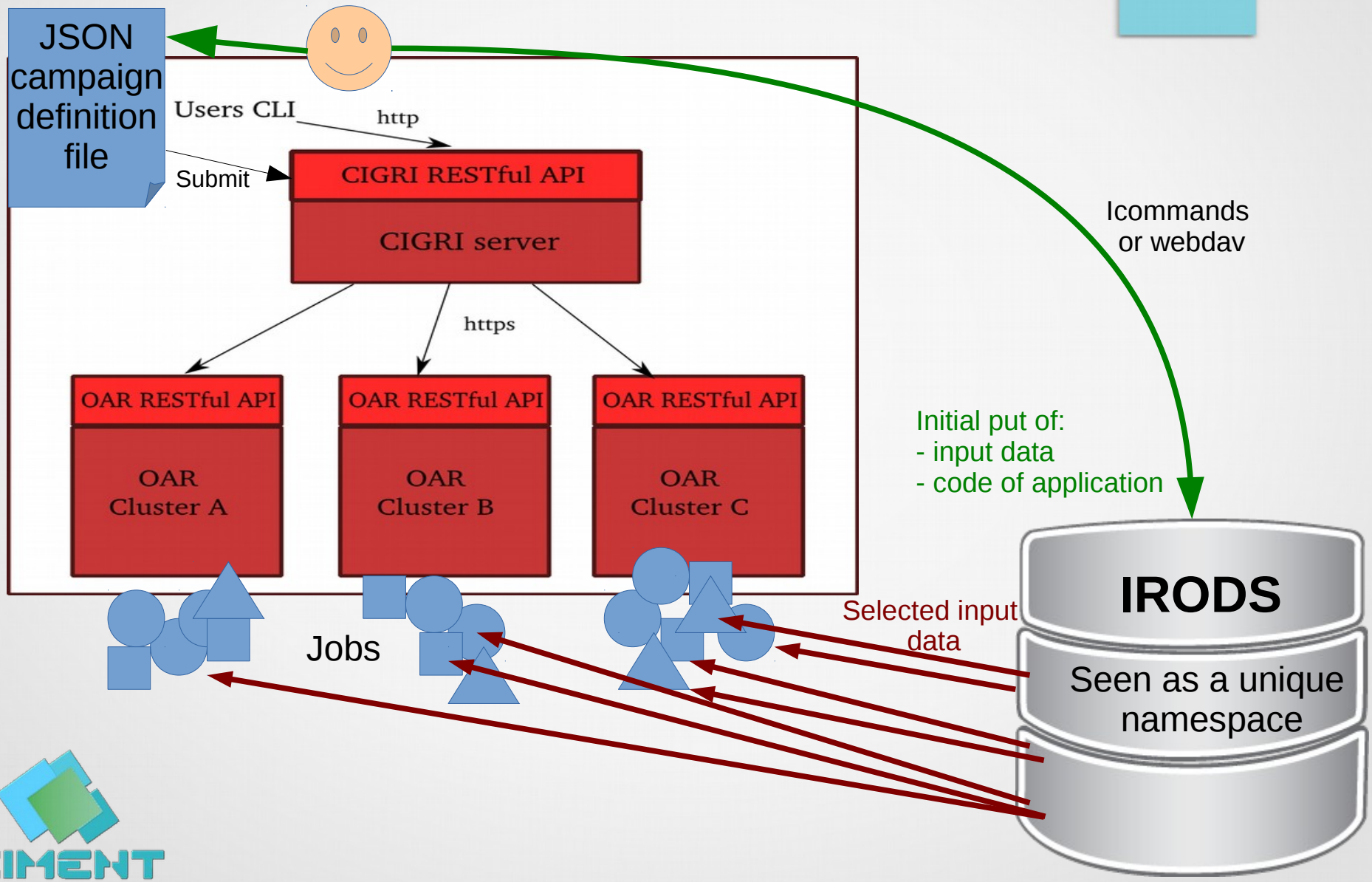




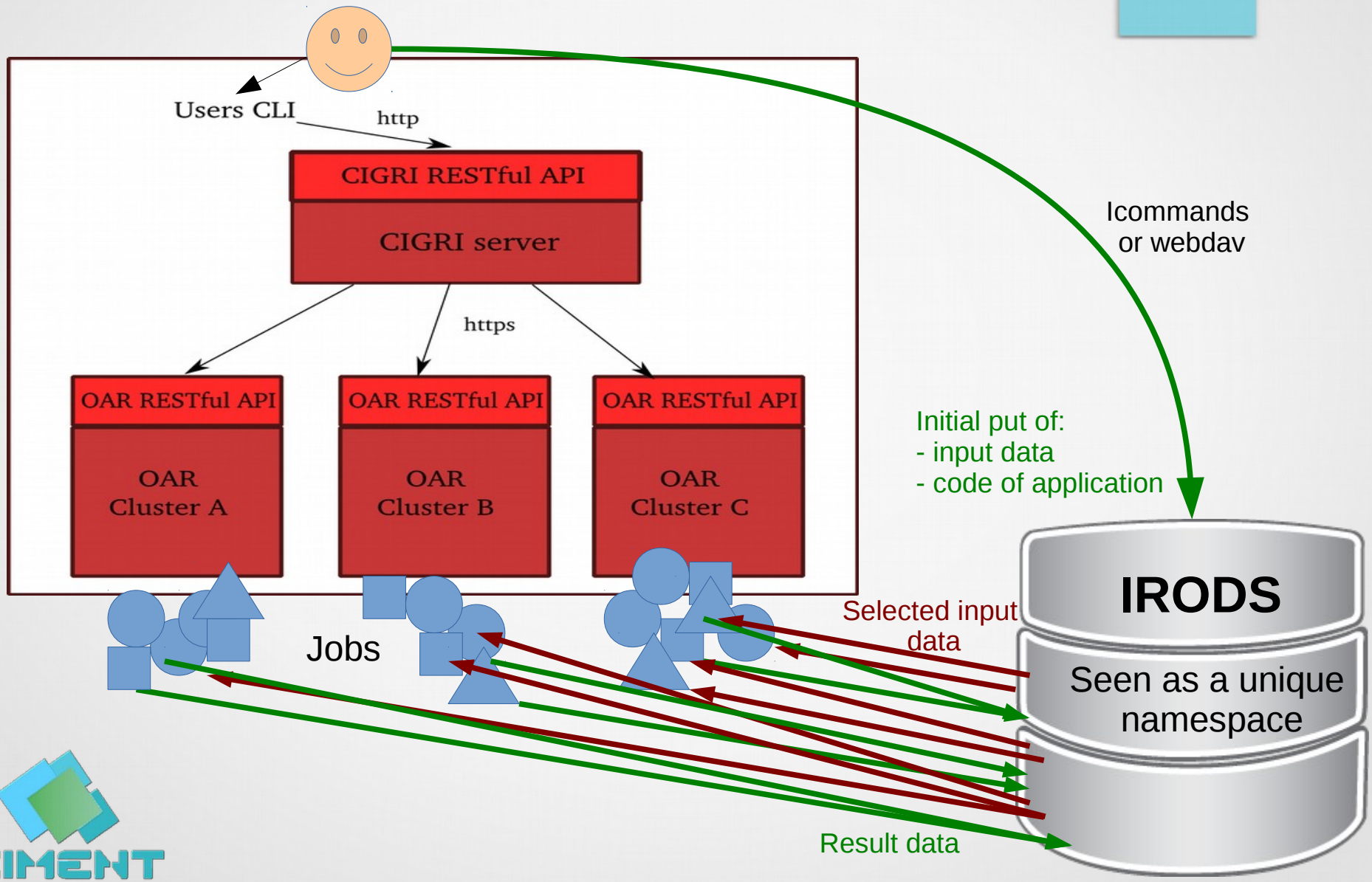
CIGRI and IRODS



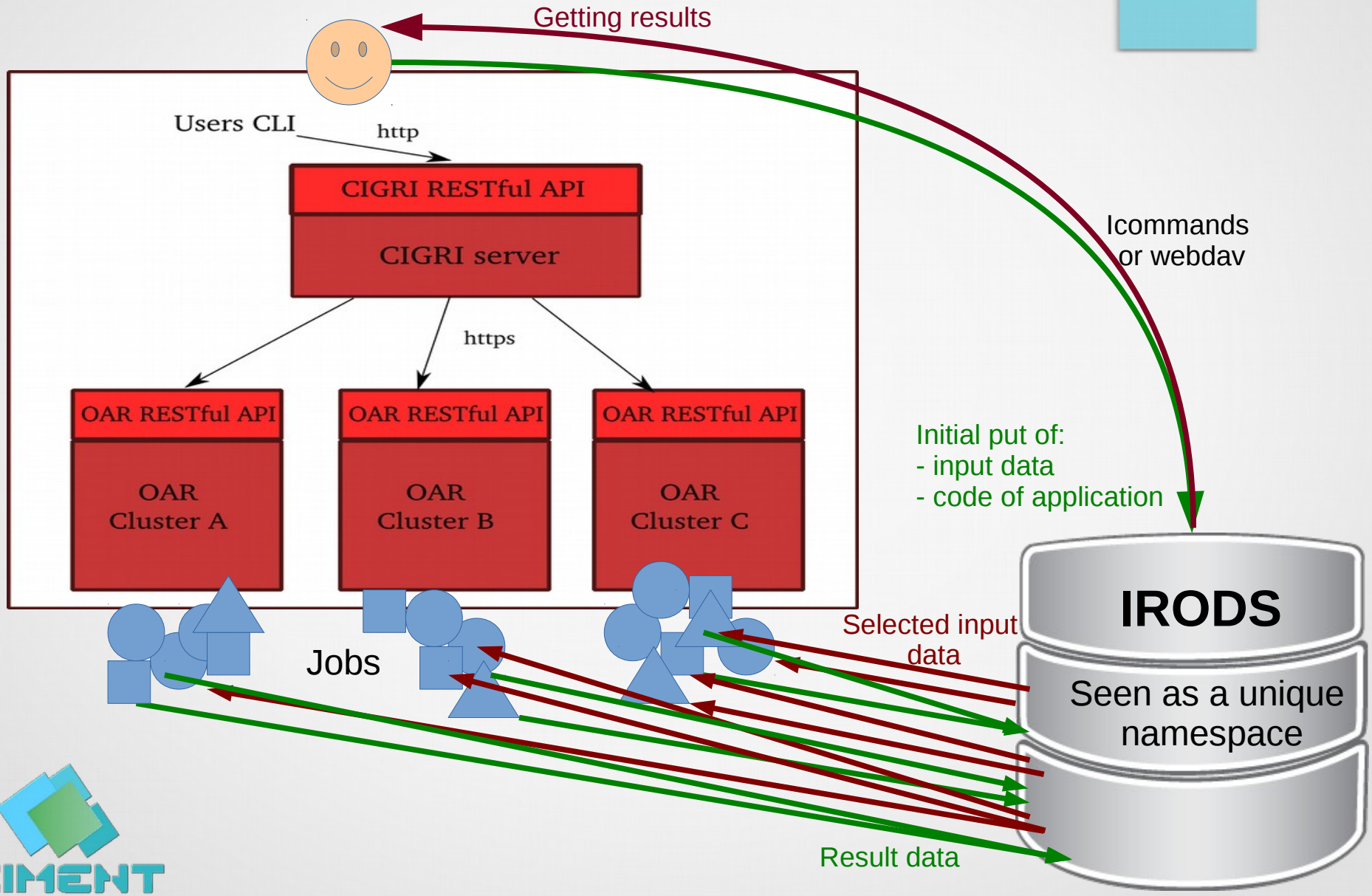
CIGRI and IRODS



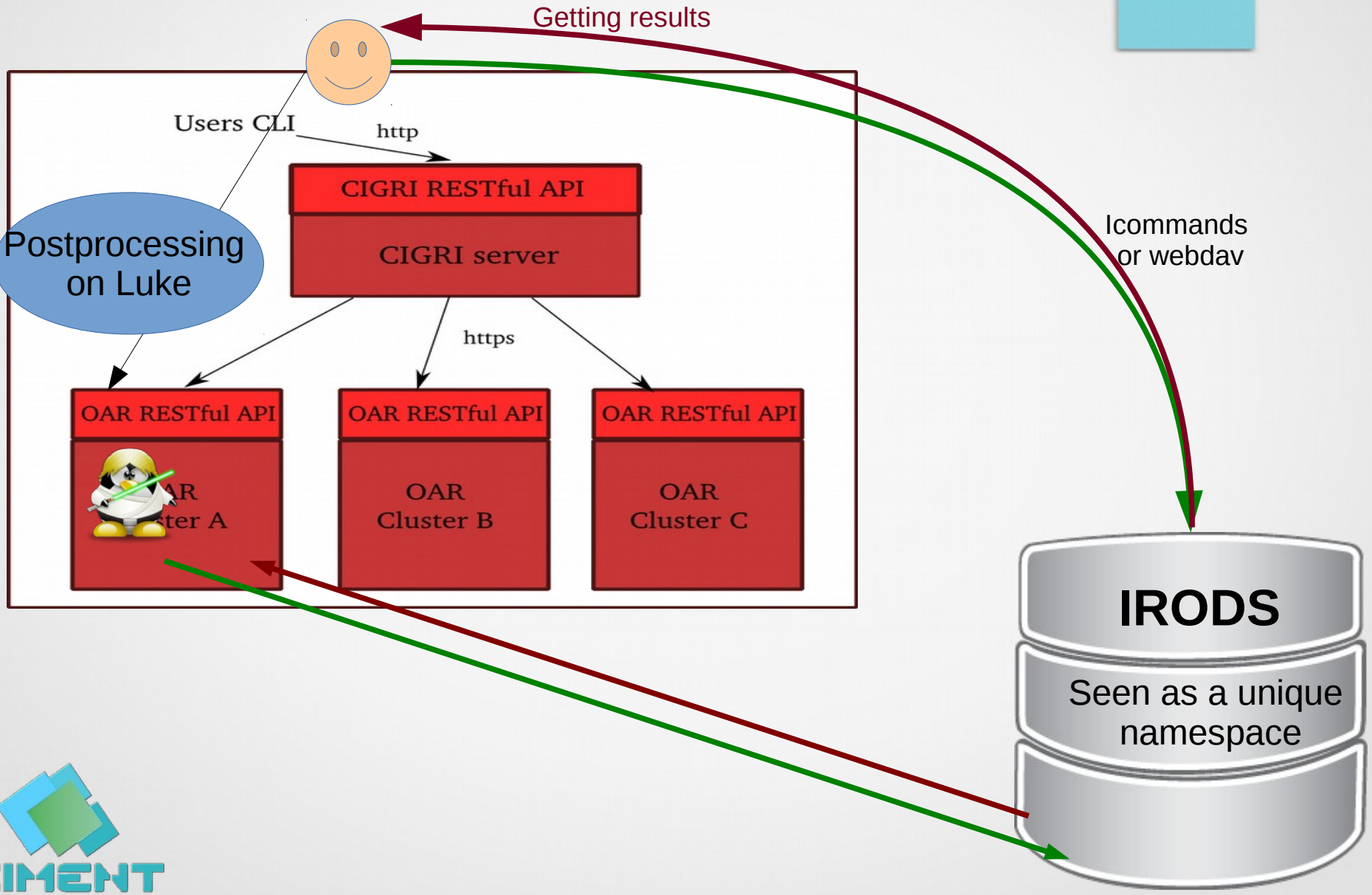
CIGRI and IRODS



CIGRI and IRODS



CIGRI and IRODS



CIMENT IRODS configuration

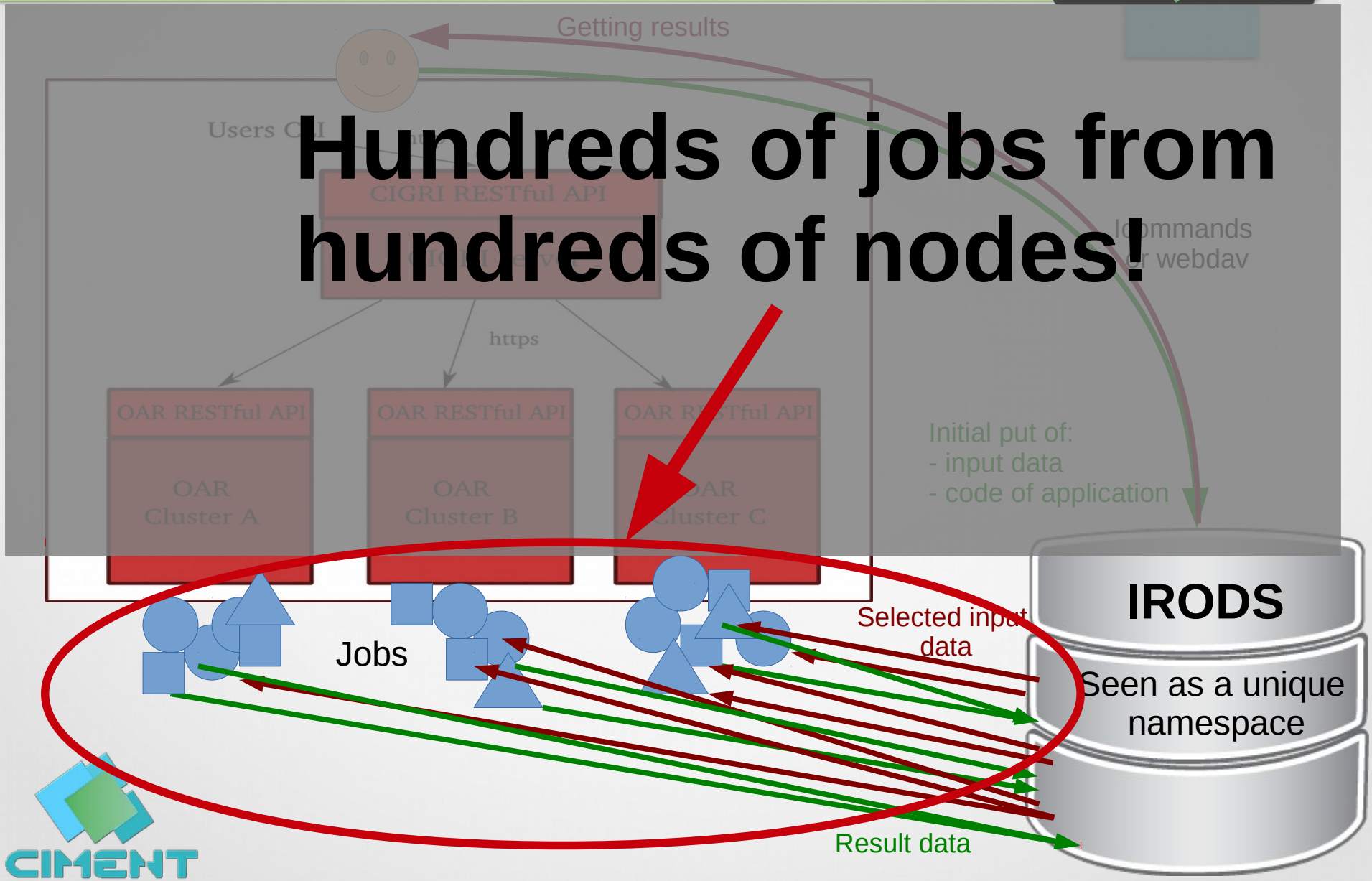
- **IRODS 3.3.1**
- Auth: LDAP custom sync → irods users
- **Computing nodes are connected to an iServ (irodsHost) from the same site**
- `.irodsEnv` files pushed into users home directories
- 1 group of resources for each site:
  - `cigri-siteA, cigri-siteB, cigri-siteC`
- Very simple rules: **files are randomly distributed on all resources of a site by default**
  - Example for site A:

```
acSetRescSchemeForCreate {msiSetDefaultResc("cigri-siteA","preferred"); msiSetRescSortScheme("random"); msiSetRescSortScheme("byRescClass");}
```

- **Webdav** (using webdavis) as a gateway for very initial and very final stages

CIMENT

Load considerations



Hundreds of jobs from hundreds of nodes!

Load considerations

- **Small files** (<32MB) problems: overload of the “proxy” iServ and bad performances on “iget -r” with large collections of very small files
- **Big files** problems: overload of disks
- **Errors:** under heavy load: `SYS_COPY_LEN_ERR, SYS_HEADER_READ_LEN_ERR, USER_CHKSUM_MISMATCH`
- **Current (not always satisfaying) solutions:**
  - connectControl.config → maxConnections
  - Smart retry with **incremental delay**
    → secure_i(get|put) wrapper or user solution
  - Limiting the number of jobs
  - Cirods : our python library for the case of very small files

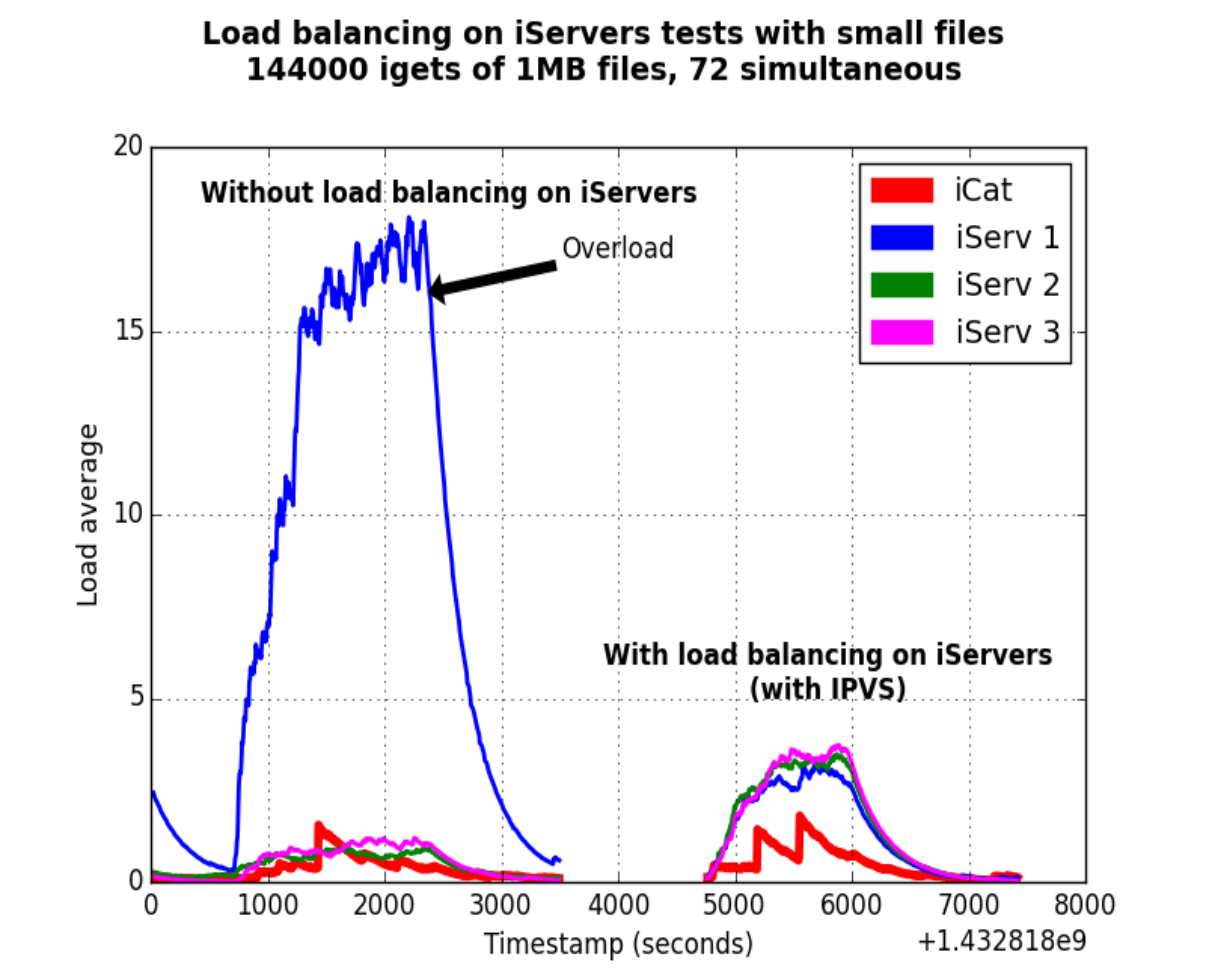
CIMENT

Overload with small files : Load balancing

- **Round robind DNS (NOT tested):**
  - Very simple to configure if you have access to the DNS server
  - Problems:
    - all the computing nodes have to use the customized DNS server
    - Cache issues (nscd, bind)
- **IPVS (Linux IP Virtual Server) for the irodsHost:**
  - Control of the load balancing with an ip address used as a proxy for a pool of other ip addresses
  - No need to modify the client configuration (except the `.irodsEnv` file)
  - Problems: ARP or routing issues

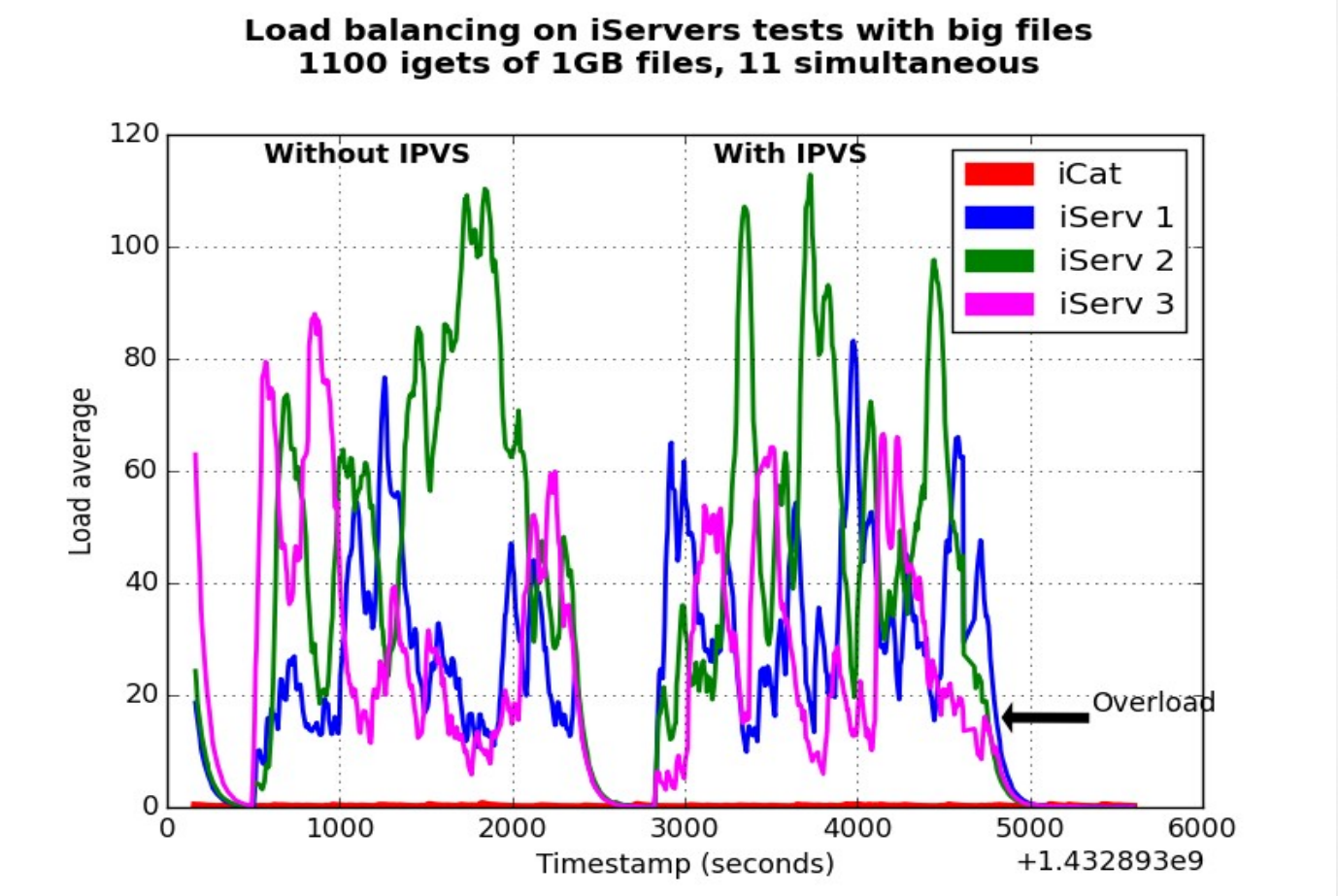
CIMENT

IPVS : Tests with small files (1MB)



- Without IPVS:
  - Rate: 85MB/s
  - 78 retries
  - Errors: `SYS_COPY_LEN_ERR`
- With IPVS:
  - Rate: 125MB/s
  - 7 retries
  - Errors: `SYS_COPY_LEN_ERR`, `SYS_HEADER_READ_LEN_ERR`

Tests with big files

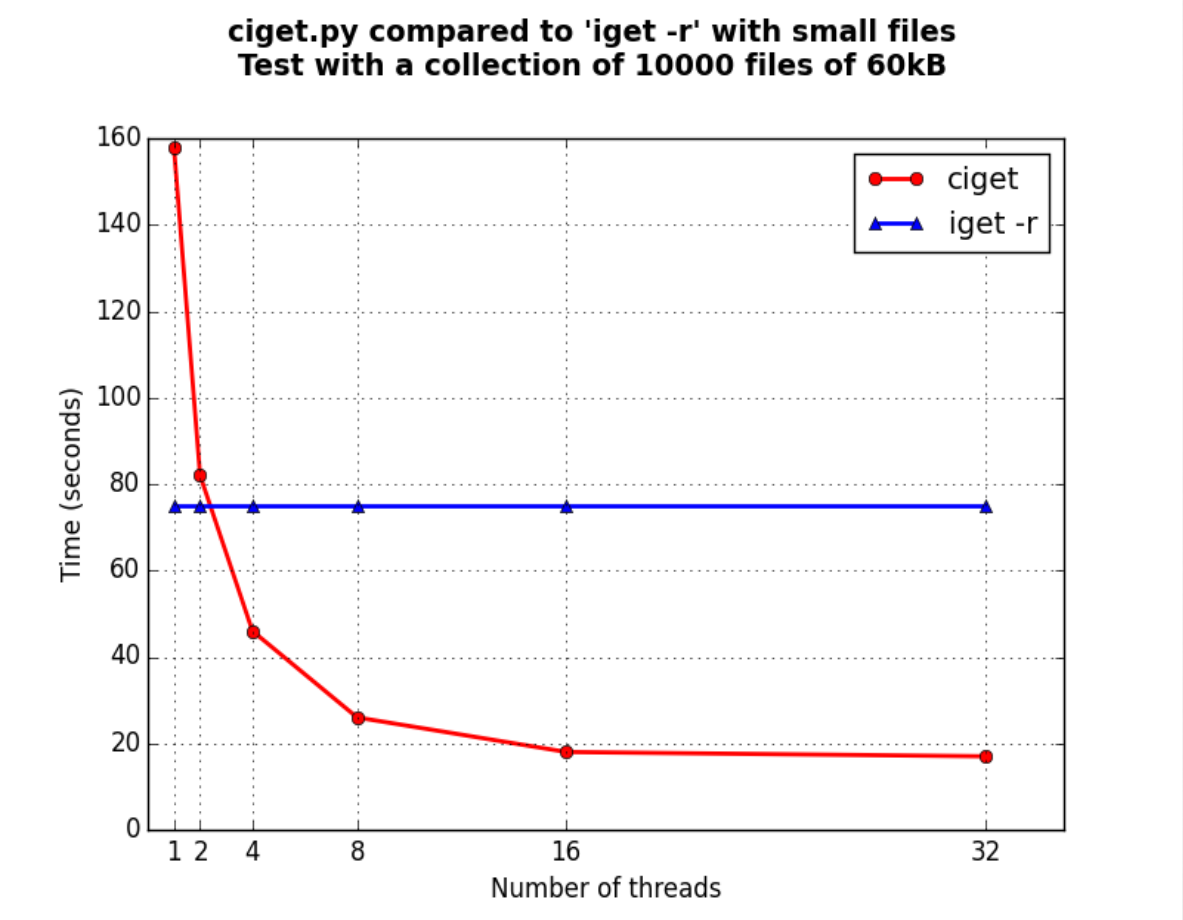


Rate:
~700 MB/s

This illustrates the problem of overload on disks with big files

Cirods : a small library based on pyrods to parallalize operations on large sets of files < 100kB

- “ciget.py” vs “iget -r”



- Parallelization on files:
  1 thread → 1 file
- Meta-data creation test:
  - ~2600 / s
  - From an input file (ease of use)

Load considerations : conclusion

- In our normal usage, iCat never shows overload, but targeted iServers (irodsHost) does
- connectControl is a nice feature but the –retry option of icommands is not satisfactory for the actual need
- IPVS load balancing with rr on iServers is a solution with small files but some errors still occur (`SYS_COPY_LEN_ERR`) and new errors appeared (`SYS_HEADER_READ_LEN_ERR`)
- IPVS will not help in the case of big files; we need some kind of global rate limiting to prevent the overload of resources
- The python API is a solution to efficiently manage large sets of very small files (Cirods)

CIMENT

Current scientific partnerships

CIMENT

CIMENT IRODS in numbers

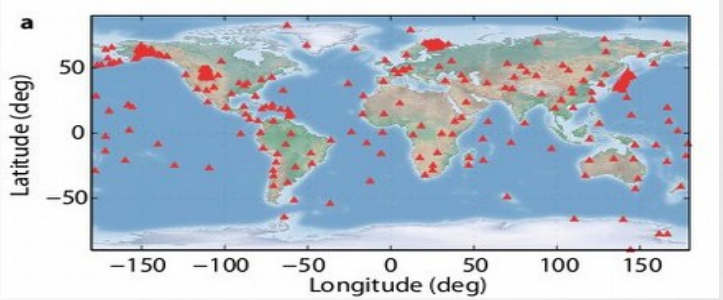
- June 2015 status:
  - 174.398.593 files
  - 6.744.478 collections
  - 351 TB used / 210 TB free (+ ~320 TB on Luke)
  - 27 resources (20-40 TB RAID arrays)
  - 12 iServers (+ 5 on Luke)
- Last 5 months (January 2015 – May 2015):
  - Number of Cigri jobs: 2,7 millions
  - Number of IRODS transactions: 6,6 millions
    → average ~ 43000 transactions / day

CIMENT

Seismology : Whisper, European seismological project

Project: Detect sligth changes of properties in the solid Earth

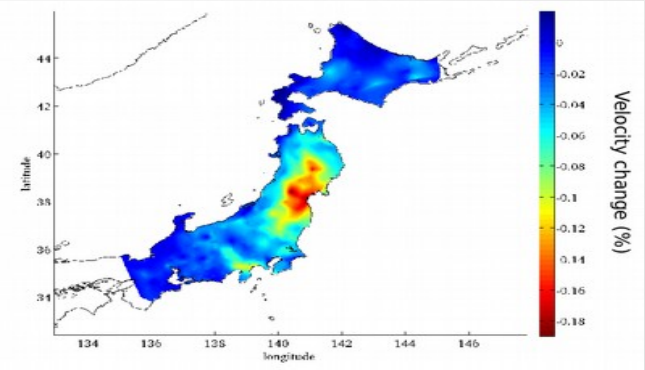
Data : Noise Continuously recorded by seismic stations worldwilde.
The computations produce even more data
More than 200 TB managed at the same time



Data Intensive processing :
Use intensively the data grid environement.
Specific python library is devellopped
Lot of feedback on cigri and irods

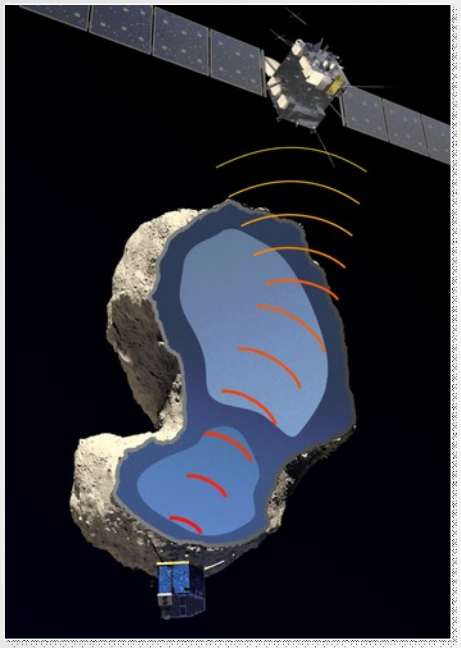
Scientific :
Many papers, posdocs and students use the data grid environment for whisper

About varation of velocity change
of the tohoku earthquake in Japan (Science)



CIMENT

Rosetta / CONSERT

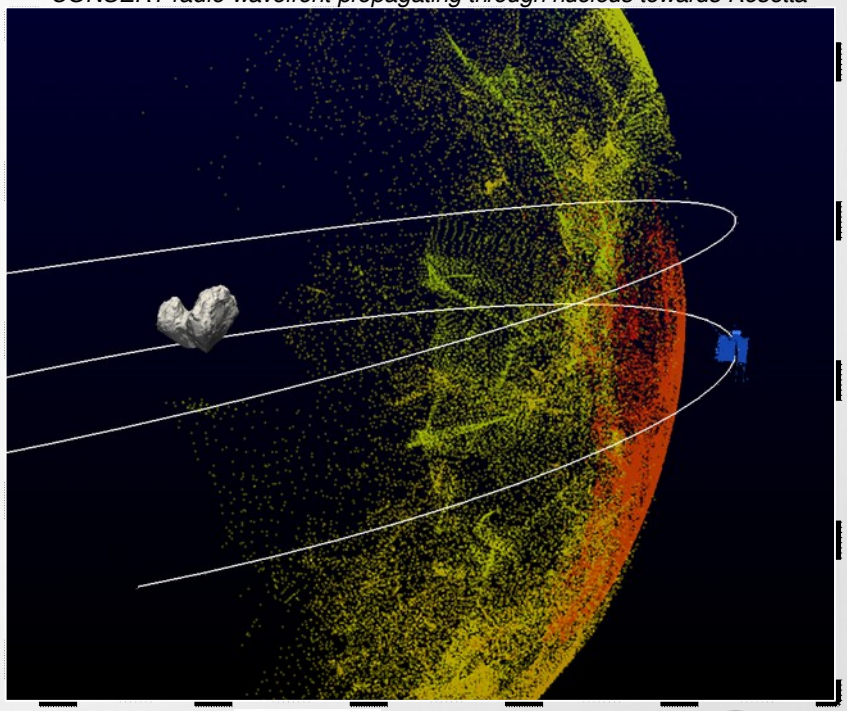


COmet Nucleus Sounding by Radiowave Transmission

An experiment on-board **Rosetta** of the European Space Agency

Performing radar tomography of the comet nucleus
of 67P/Churyumov-Gerasimenko

CONSERT radio wavefront propagating through nucleus towards Rosetta

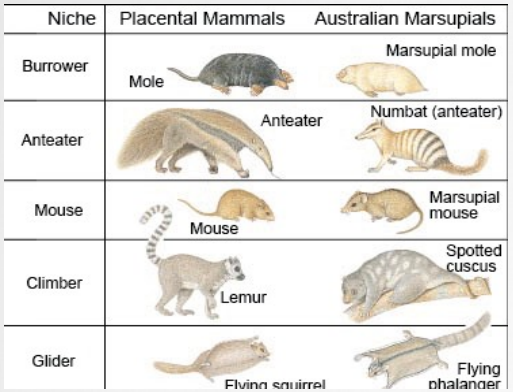


CIMENT with iRods were used for:

- preparation of space operations, and especially for Philae landing (12 Nov. 2014),
- inversion of dielectric properties, deriving better knowledge on composition and structure.

Ecology : ***The geography of evolutionary convergences***

Principle

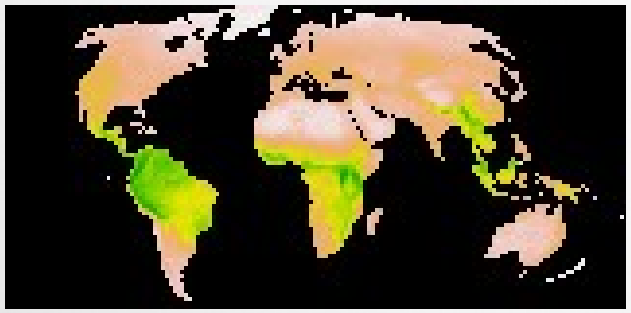


Computations

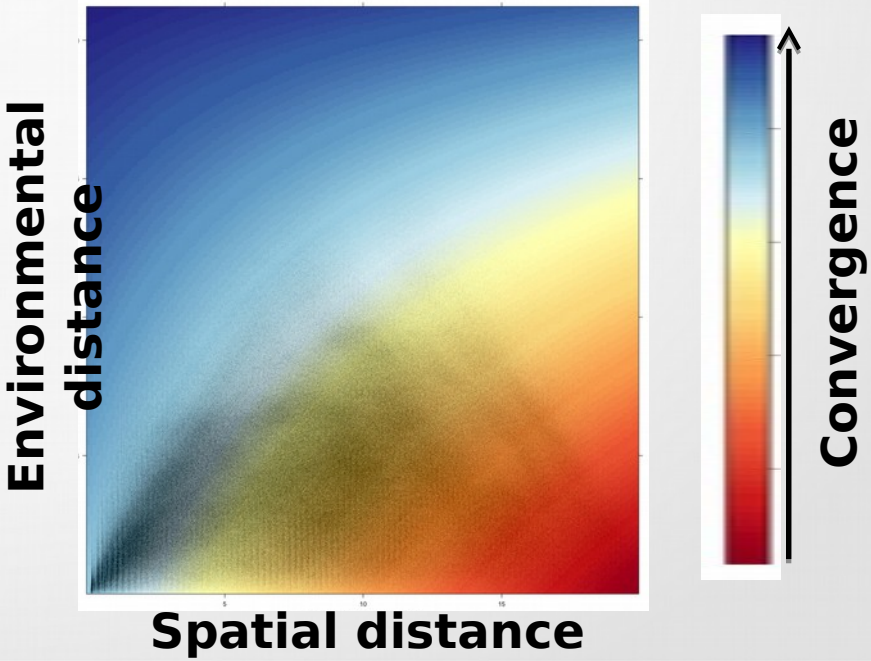
- Measure of morphological and species similarities between sites
  ⇒ 6 000 000 values

- Detect assemblages that morphologically ressemble each other but contain very different species

Data : 3600 pixels / 5000 mammals



Results



Funding

Particle Physics - LHC

First stable proton collisions at 13 TeV
June 3rd 2015

ATLAS EXPERIMENT

Computing model in Particle Physics with Colliders

- Event by event computation → grid computing is ideal
- The LHC experiments use a grid of ~ 160 computing centres around the world (WLCG)
- CIGRI+iRODS : used as a local farm for ATLAS analyses lead in Grenoble (LPSC/CNRS)

An new area has just began, an un-preceded high energy Physics goal: hunt for exotic particles

Analysis on CIGRI for ATLAS

- Search for extra dimensions in di-photon final states
- Event cross section computation
- "CIGRI is an asset"

- Already used for the earlier phase of the LHC (Run 1)


New Journal of Physics
The open access journal at the forefront of physics

This is to certify that the article

Search for extra dimensions in diphoton events from proton–proton collisions at $\sqrt{s}$ = 7 TeV in the ATLAS detector at the LHC
by The ATLAS Collaboration

has been selected by the editors of *New Journal of Physics* for inclusion in the exclusive 'Highlights of 2013' collection. Papers are chosen on the basis of referee endorsement, novelty, scientific impact and broadness of appeal.

Professor Eberhard Bodenschatz
Editor-in-Chief
New Journal of Physics
www.njp.org

DPG | IOP Institute of Physics

NJP 15, 043007 (2013)

At the Tevatron (US)

Phys. Rev. Lett. 108, 151804

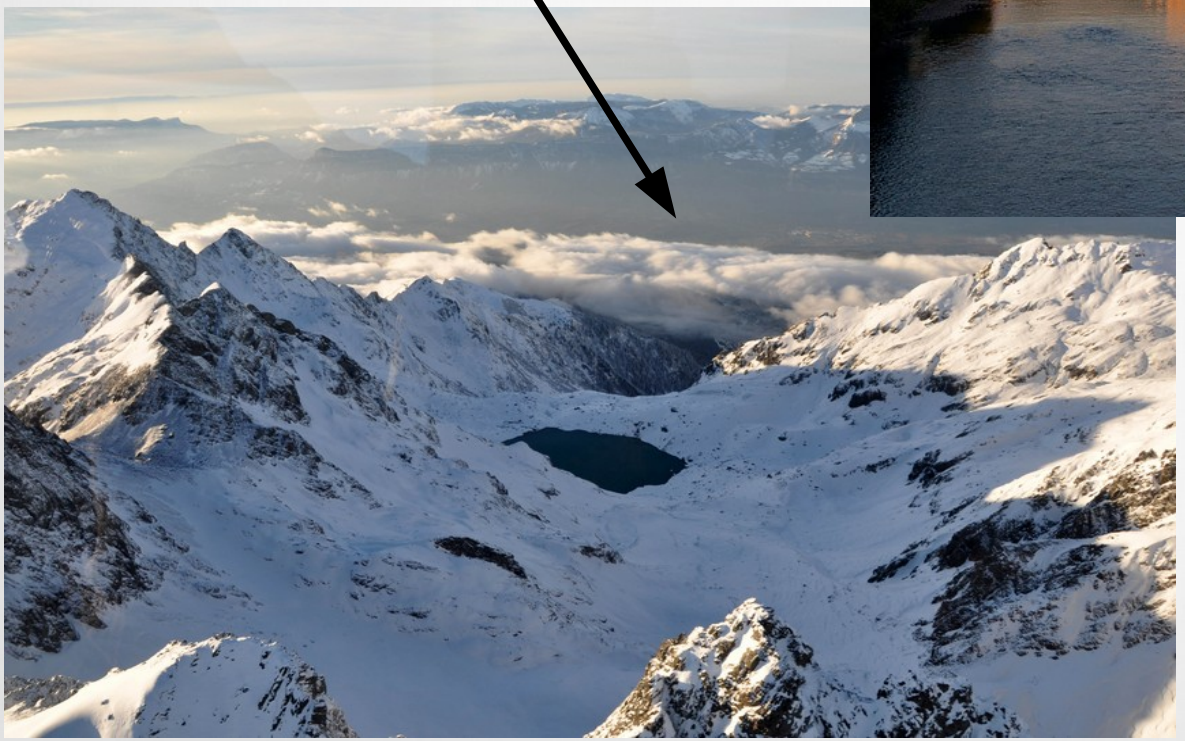
Conclusion

- IRODS is a very satisfying solution for the management of the data of the federated computing resources of CIMENT
- Whenever needed, we just have to add new servers to add more storage
- Cigri and Irods complement one another
- Some stability issues under heavy load but we can deal with them
- A lot of scientific results thanks to Cigri + Irods since 2010!
- Future works:
  - Install Irods 4.1 and adapt Cirods
  - Find solutions for better reliability and control of the load (will irods 4.1 help?)
  - Better integration within Cigri (→ overload events)
  - Add support for storage resources in our batch system?

CIMENT

Thank you !

Grenoble



http://ciment.ujf-grenoble.fr
http://oar.imag.fr
http://cigri.imag.fr

Load experiments

- 1 iCat
- 3 iServers containing 2 resources of 40TB each (12HD,RAID5)
- Small files tests:
  - An irods collection containing 5000 files of 1MB
  - 18 compute nodes, 4 jobs on each compute node, randomly downloading the 5000 files
- Big files tests:
  - An irods collection containing 100 files of 1GB
  - 11 compute nodes, 1 job per node, randomly downloading the 100 files

CIMENT

Cirods : python library to optimize the use of large set of small files and custom meta-data

- Work of an intern during the summer 2014 in the context of a geological scientific project (NERA):
  - Making a python library using pyrods to easily manage a large set of small files (several millions, 60KB):
    - Import millions of custom meta-data on a large set of files
    - Efficiently get a subset of files corresponding to a request on meta-data
  - Make performance comparisons with the use of iCommands

CIMENT