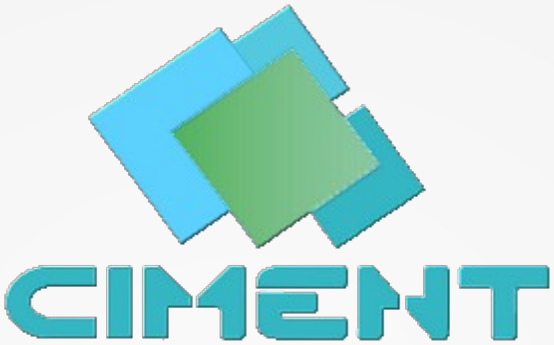


An introduction to CIMENT, HPC Center at UGA

LIG / 2015-11-12

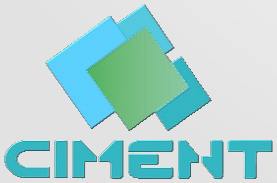
Bruno Bzeznik (CIMENT), Pierre Neyron (LIG), Laurence Viry (MaiMoSiNE)

CIMENT

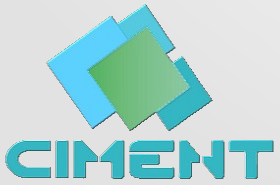
UNIVERSITÉ Grenoble Alpes

Outline

- Introduction to HPC platforms
- CIMENT, HPC center at UGA
  - Organization
  - Computing platforms and usages
  - Getting an account
  - First steps, documentation
- Support
  - Training
  - Local resources (people in the laboratories that can help you)
  - Reporting problems, requesting help
- Pole ID (Computer Science)
  - Digitalis
  - Grid5000
- Q/R



Introduction to HPC platforms



HPC today and tomorrow

With the digital revolution, **more and more computational sciences and big data require more and more computational power** (weather forecast, earthquake simulation, genomics, new media, Internet of things, etc)

This motivates **building always more powerful computational infrastructures:**

- Today's personal computer is about 100 GFlops (1 TFlop with GPU)
- Today's #1 supercomputer is about 30 PFlops
- **Tomorrow's supercomputer target (2020) is the ExaFlops**
  →e.g. believed to be the order of processing power of the human brain
  →energy consumption is now the main challenge→20MW target

While computing power grows and new scientific challenges arise, platform funding and operational costs yet meet constraints → **mutualization**

CIMENT

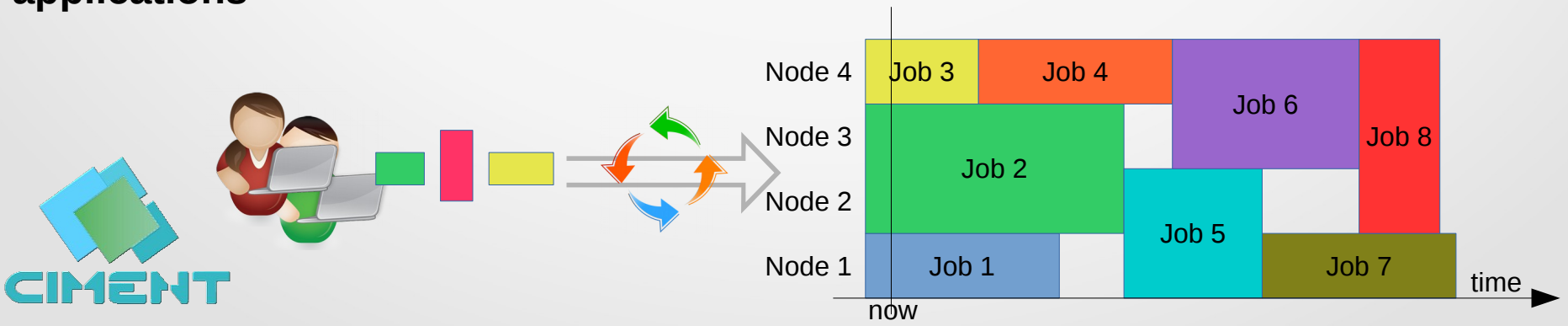
User interface to HPC platforms

Due to their important cost, HPC platforms are quite always (must be) **shared** at some level.

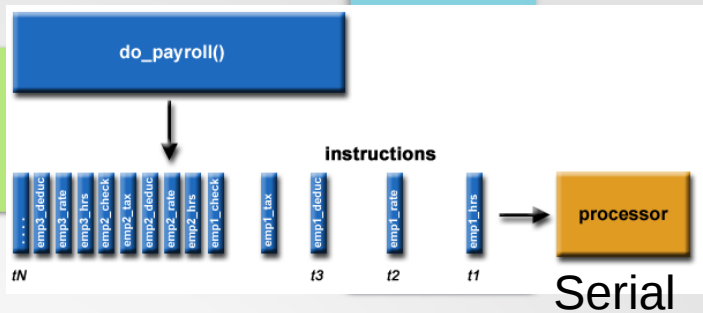
Multi-task/**multi-user** usages of those infrastructures **relies on some basic objects:**

- Resources: **Nodes or cores** (mind the granularity: computation servers, CPU cores, GPU cores,...)
- **Tasks or "jobs"** = #cores X maximum duration (walltime)
- A **scheduling policy**: maps tasks on computing resources → execution queue

Resources (compute nodes) are somehow black boxes:
→ Access to **compute nodes** is only granted to users via a **front-end** to a **Resource and Jobs Management System** (batch scheduler) which interfaces systems with **applications**

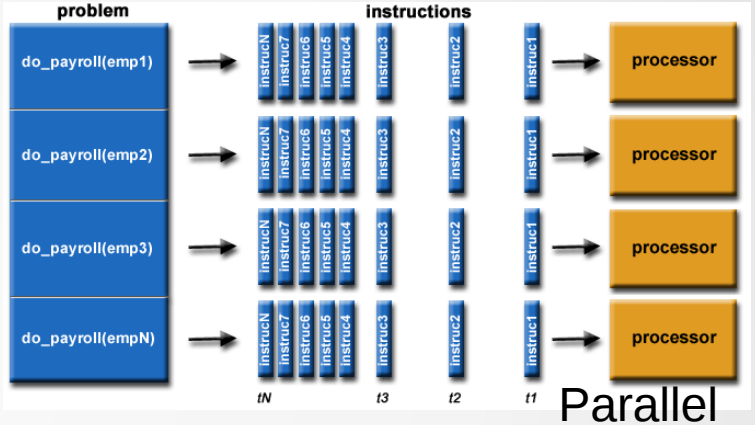


Categories of applications



Serial

Supercomputers are **parallel machines** → tasks must be mapped to the resources



Parallel

Basic categories for the applications:

- Sequential
- Multi-threaded
- Distributed
- Embarrassingly parallel

Also, applications can be qualified given their hot-spots and bottlenecks

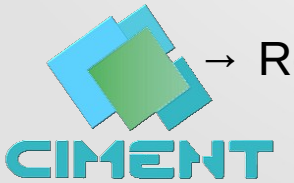
- CPU bound
- memory bound
- IO bound

Those categories leads to the choice of different types of supercomputer architectures
→ **User must find the most appropriate computing platform for his application**

CIMENT

Platform architectures

- HPC clusters: rack of servers, usually bi-socket, multi-core, high performance network/storage
- Accelerator platforms: HPC cluster equipped with accelerators (GPU, MIC, FPGA…)
- Computational Grids: federation of computation resources from different places (desktop grid, P2P grid, academic grid)
- Shared Memory systems: fat nodes / single system machines (big memory)
- HTC platforms: data centric platforms
- Cloud platforms: at infrastructure level (Amazon EC3) or application level (Hadoop)



→ Reality can be an hybrid mix of those platforms

Platform architectures

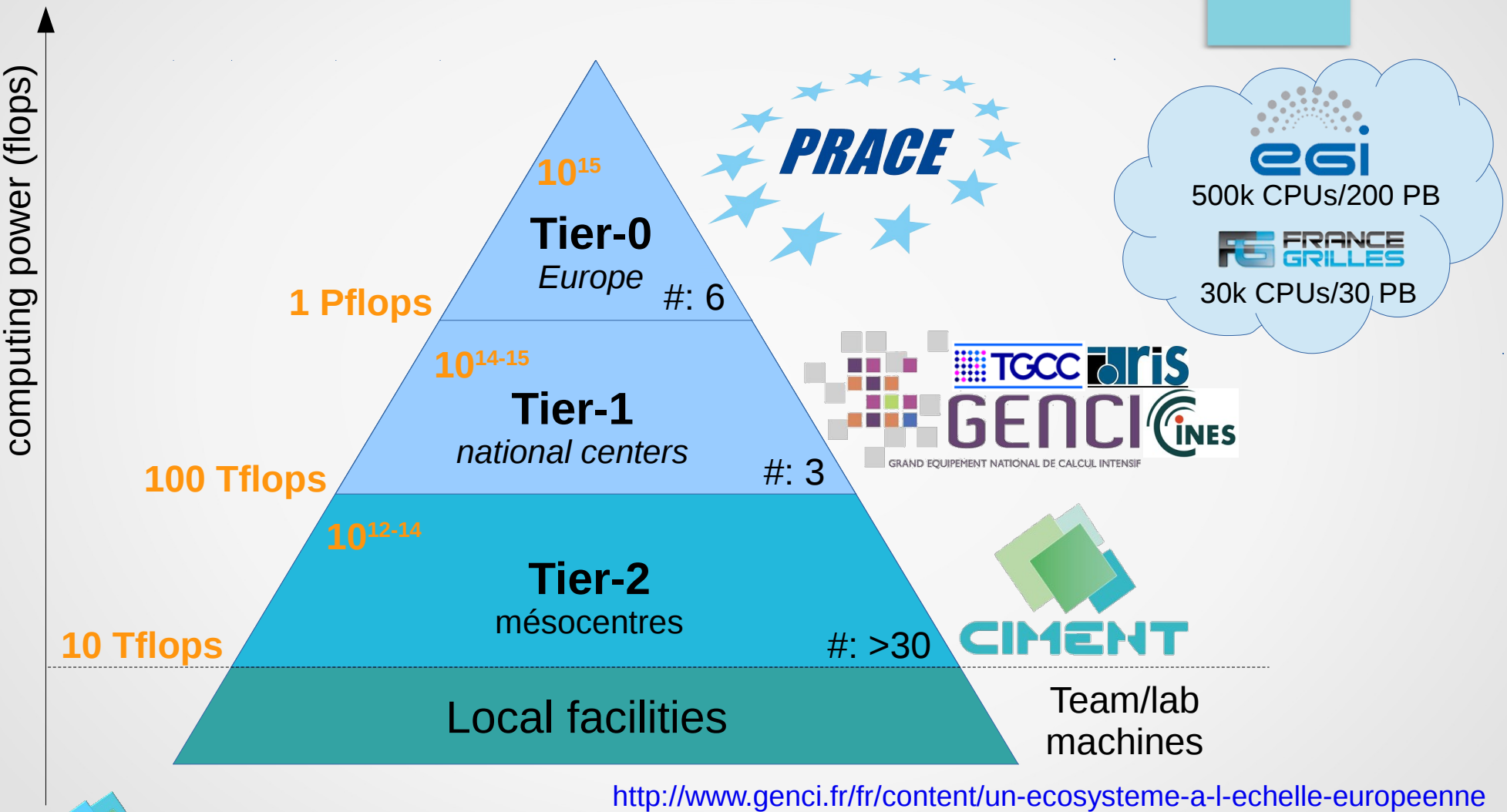
- HPC clusters: rack of servers, usually bi-socket, multi-core, high performance network/storage
  → *typically: distributed(+multi-threaded) jobs*
- Accelerator platforms: HPC cluster equipped with accelerators (GPU, MIC, FPGA…)
  → *typically: distributed jobs with offload to the accelerator (vectorization)*
- Computational Grids: federation of computation resources from different places (desktop grid, P2P grid, academic grid)
  → *typically: embarrassingly parallel jobs*
- Shared Memory systems: fat nodes / single system machines (big memory)
  → *typically: multi-threaded jobs / memory bound*
- HTC platforms: data centric platforms
  → *typically: sequential jobs / IO bound*
- Cloud platforms: at infrastructure level (Amazon EC3) or application level (Hadoop)
  → *other paradigms*

CIMENT

Academic HPC platforms

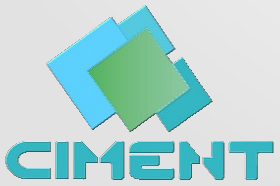
CIMENT

Academic HPC platforms pyramid

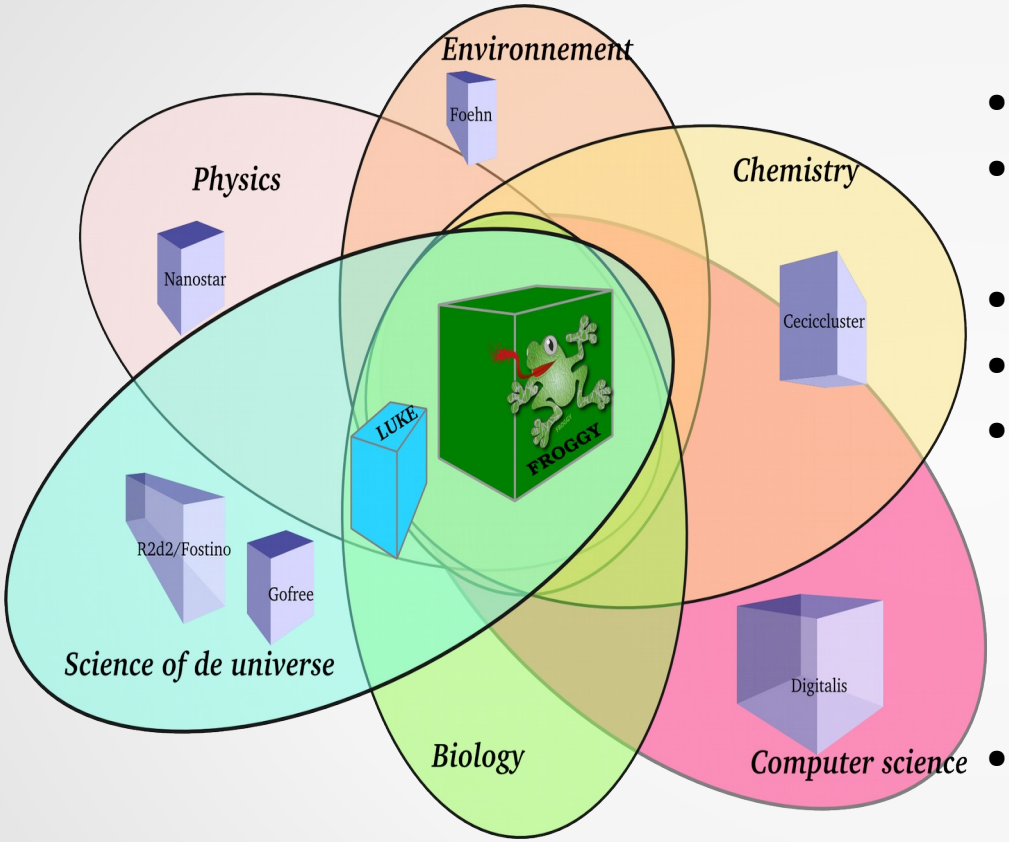


http://www.genci.fr/fr/content/un-ecosysteme-a-l-echelle-europeenne
http://www.prace-ri.eu/prace-resources/
http://www.genci.fr/fr/content/calculateurs-et-centres-de-calcul
http://calcul.math.cnrs.fr/spip.php?rubrique6
http://www.france-grilles.fr/First-step

CIMENT, the UGA HPC center



Organization of CIMENT: the “pôles”

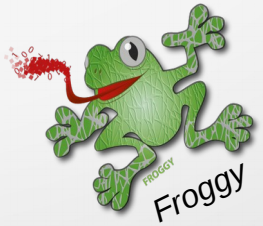


- 6 scientific “pôles”
- A charter, signed by the labs of the “pôles”
- A scientific director
- A technical director
- Each “pôle” has
  - A scientific manager
  - A technical manager
  - Some thematic platforms
  - Users!
- All managers + directors → steering committee

CIMENT

TOTAL

13 clusters
7068 cores
132 Tflops
1,4 Pbytes



3244 cores

A mutualized platform, shared resources

- All academics from Grenoble have access (partner laboratories)
- Adaptations to fit local needs/usages
- Some resources are dedicated to a local usage, but:
  - Can be used by the community when free (best-effort)
  - System administration is mutualized (possibly shared)
- Fair-sharing based scheduling (karma), not based on computing hours allocations (unlike Genci)
- Shared high performance remote visualization tools
- Shared storage infrastructure
- Unified development environment (shared licenses, ...)

CIMENT

Platforms of the HPC center of the University of Grenoble

CiGri lightweight computing grid

OAR batch scheduler

HPC platform

Froggy

3200 Xeon E5 cores @2.6Ghz
+18 GPUS K20m

High performance distributed storage (Lustre): 90 TB

Infiniband FDR network

Remote visu nodes

OAR batch scheduler

Data processing platform

Luke

~400 cores – heterogeneous systems and continuously evolving

Local scratches on nodes
450 TB

10 Gbe network

Remote visu nodes

OAR batch scheduler

Other thematic platforms

~3000 cores heterogeneous systems federated from 10 clusters of member laboratories

NFS filesystems:
a few TB per cluster

Infiniband QDR network

Common distributed storage (IRODS) 1Po

CIMENT

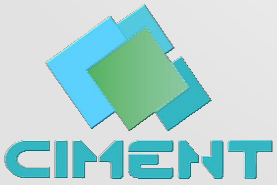
CIMENT software bricks

- **CIMENT uses the OAR batch scheduler, with a specific configuration**
  - *OAR is a versatile resource and task manager for HPC clusters, and other computing infrastructures ; developed by LIG/Inria*

- **CIMENT has a specific management of “bag-of-tasks” applications thanks to CiGri**
  - *CiGri is a lightweight grid middleware optimized for very large sets of tasks and best-effort optimization of the clusters (free computing cycles exploitation) ; developed by LIG/Inria/CIMENT*

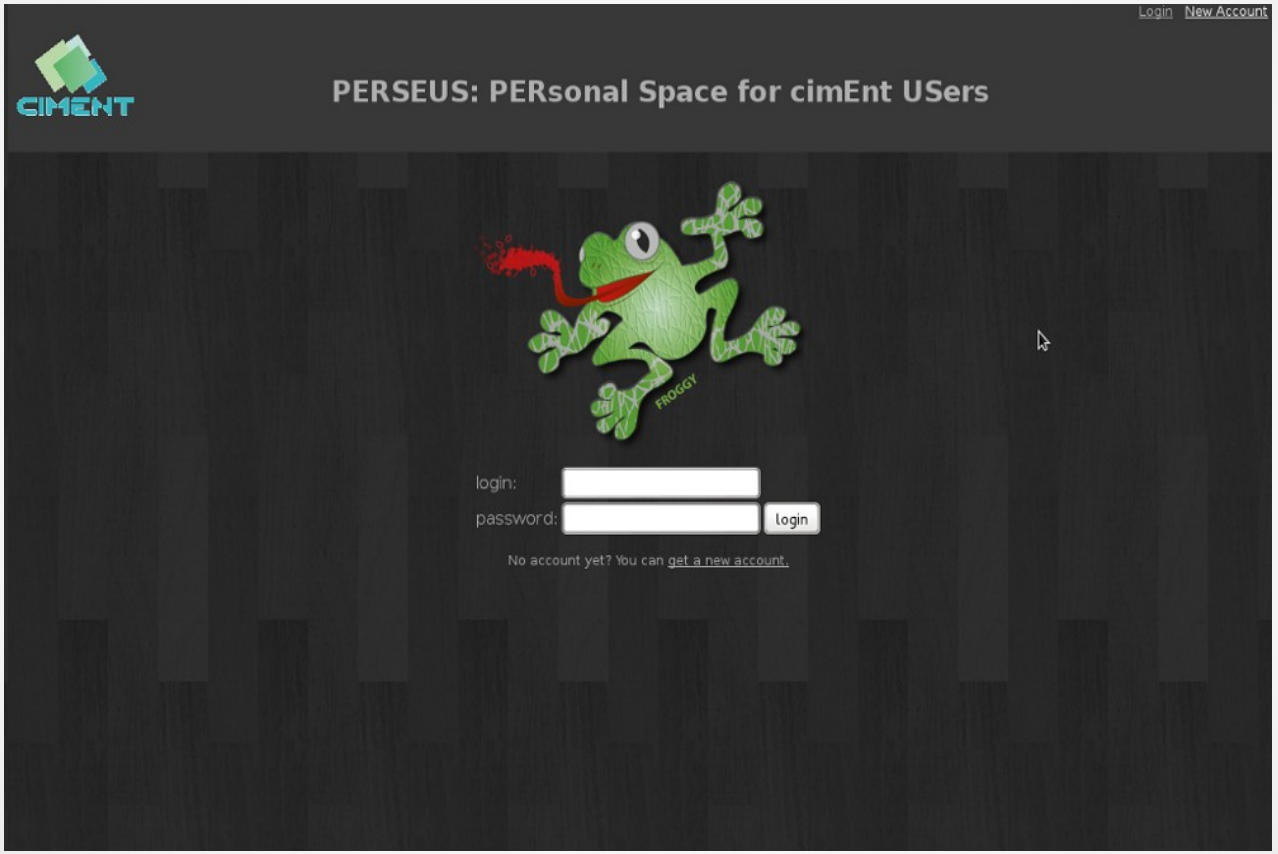
- **CIMENT provides a distributed storage powered by IRODS**
  - *IRODS is an object storage file system developed by an international consortium mainly hosted by the University of North-Carolina (UNC)*

- **CIMENT uses common “Environment Modules”**
  - *Compilers, debuggers, analyzers, softwares, and pre-installed librarie*
- **All CIMENT systems (services, computing) run GNU/Linux**



Getting a CIMENT account

https://perseus.ujf-grenoble.fr



Projects management (currently **110 active projects**)

CIMENT

Getting a CIMENT account

- Any member of a UGA laboratory can get a CIMENT account. External collaborators are also accepted.
- Access is granted by joining a validated CIMENT project
- 2 cases:
  - Creating of a new CIMENT project
    - Owner must have a permanent position in a UGA laboratory (phD students not allowed)
    - Projects creation are reviewed by the managers of the pôle.
    - Usually only takes a few days
  - Joining an existing CIMENT project
    - Membership is delegated to the project owner
    - Project owner is probably a co-worker → should be very quick

CIMENT

Getting a CIMENT account

For more informations on the CIMENT projects and accounts management, see:

https://ciment.ujf-grenoble.fr/wiki/index.php/PERSEUS:PERsonal_Space_for_cimEnt_USers

CIMENT

First steps

1) Know where to find documentation

Documentations of CIMENT are located in our **WIKI: https://ciment.ujf-grenoble.fr/wiki**

2) Find out / decide which platform to use

By default: "Froggy", but your project description or project comments may mention another platform to use. The CIMENT project form in Perseus is the place to discuss about that (comments section)

3) Setup your access

Accessing CIMENT platforms relies on the SSH protocol, and through dedicated gateways. Read the **"Accessing to clusters"** section of the WIKI.

4) Read the quickstart for the targeted platform

1) **Learn how to submit computing jobs**
   Ciment uses OAR as its resources and jobs management system. See **OAR tutorial** in the WIKI.
   Ex: `oarsub --project test -l /nodes=2 ./my_test_job.sh`
2) **Write your launcher script to run your parallel program (or a sequential one, but then, you are probably in a grid (CiGri) use case)**

CIMENT

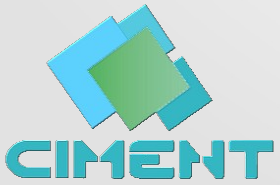
First steps

The “test” project: the beginner's sandbox project

- For a very first test, no need to create a CIMENT project: just join the “test” project
- 3000 hours / 3 months max
- Access to Froggy only
- Please: as soon as your tests are ok, create a dedicated project (do not wait for the end of your trial period)

CIMENT

Support



Wiki

- Quickstarts
- Tutorials
- Docs
- Projects pages linked to Perseus (automatic formatting of bibliography)

https://ciment.ujf-grenoble.fr/wiki

CIMENT

Training

Formations CED MaiMoSiNE/CIMENT (2015-2016)

Les formations en collaboration avec le collège doctoral, MaiMoSiNE et CIMENT sont orientées principalement vers les doctorants mais sont ***ouvertes aux chercheurs et ingénieurs des instituts de recherche et de l'université.***

- Calcul Scientifique – HPC
  - Introduction à LINUX (9h : B. Bzeznik, F. Audra)
  - Environnement de développement d'application de calcul scientifique (15h : F. Pérignon, B. Bzeznik, C. Bligny, L. Viry)
  - Introduction au calcul parallèle (33h : F. Roch, B. Bzeznik, F. Pérignon, P. Begou, C. Biscarat, L. Viry)
  - Programmation sur accélérateurs (12h : L. Génovèse, B. Videau)
- Statistiques
  - Bases des statistiques et logiciel R (30h : F. Letué, R. Drouilhet, A. Samson-Leclercq, L. Viry)
- Fouille de données
  - Recherche d'information – Fouille de données : principes de base (36h : M. Clausel, M-R Alimi)
  - Recherche d'information – Fouille de données : applications à des données complexes (27h : M. Clausel, M-R Alimi)

Trois modules ont été ajoutés cette année : introduction à LINUX et deux modules sur les fouilles de données.
Affichées sur le site du groupe calcul et sur le site de MaiMoSiNE.

CIMENT

Training
Séminaires - Workshops

- Séminaires groupe calcul grenoblois
  - Apport d'expertises, retour d'expériences
  - Ils peuvent être issus de la demande d'utilisateurs ou de propositions d'experts du calcul

- Workshop – Ecoles thématiques
  - 18th VI-HPS Tuning Workshop (UGA, Grenoble, PRACE)
  - ...

- Affichage des séminaires/formations sur le site de MaiMoSiNE et le site du groupe calcul
  http://grenoble-calcul.imag.fr/ et http://www.maimosine.fr/

- Groupe calcul grenoblois
  - Ouvert à tous les acteurs du calcul et toute personne en interaction avec les modélisateurs grenoblois.
  - Mise en commun des connaissances et des expérimentations (méthodes numériques, bibliothèques et outils de développement, architectures parallèles, paradigmes de programmation,...)
  - Favoriser les interactions entre les ingénieurs calcul grenoblois et les ingénieurs qui gèrent les ressources de calcul.

- Outils :
  - Liste de diffusion : **grenoble-calcul@imag.fr**
  - Site Web : http://grenoble-calcul.imag.fr/

CIMENT

Local support resources

- CIMENT has a network of technical representatives designed by the labs which signed the charter. The representative can drive you to the first steps.

- CIMENT has an “operational committee” with people that can also be into your lab.

Reporting problems / asking for help

- CIMENT helpdesk platform (GLPI tickets): https://virgule.imag.fr/glpi/
- Ticket creation via the mailing list: Sos-ciment@imag.fr

CIMENT

Have a project involving heavy computations or data processing?

- Feel free to contact us at the very beginning
- The better we know your needs, the better we can find solutions, together, mutualized or not.

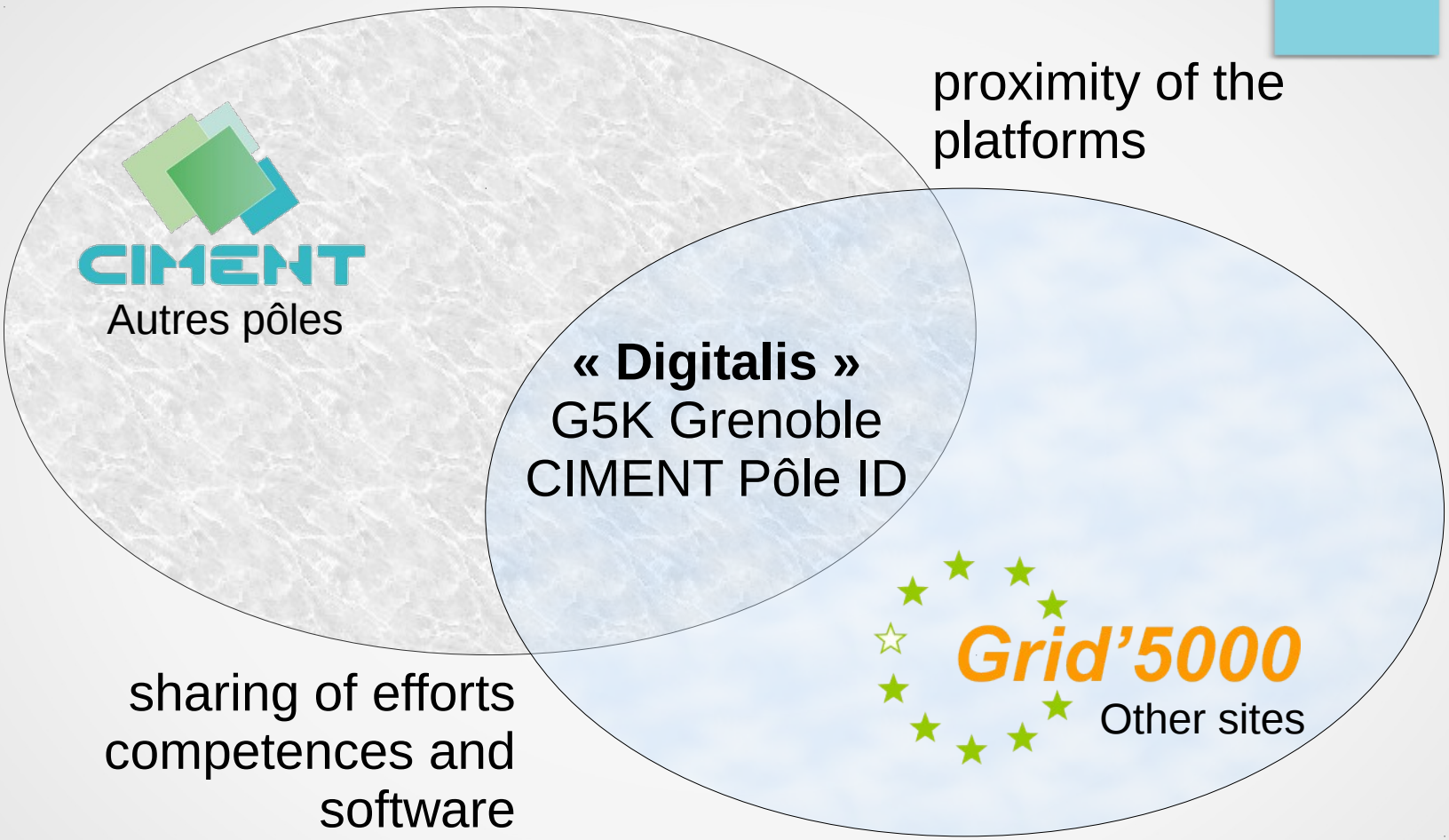
  → CIMENT's organization is based on the proximity to users

CIMENT

Pôle ID (Computer Science)

CIMENT

Pole ID: CIMENT ∩ Grid'5000 → Digitalis



proximity of the platforms

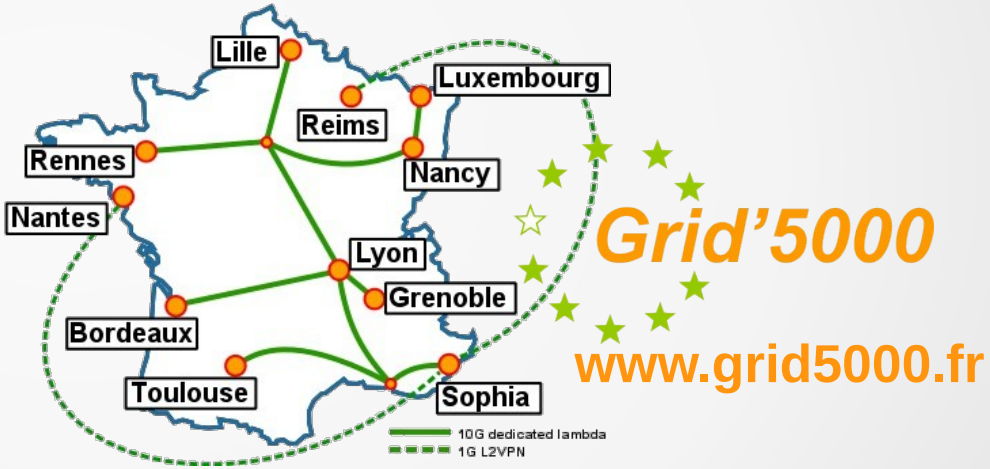
sharing of efforts competences and software

+ some experimental HPC machines:
multi-GPU, manycore SMP, Xeon Phi, ARM64,...
http://digitalis.imag.fr

CIMENT

Grid'5000

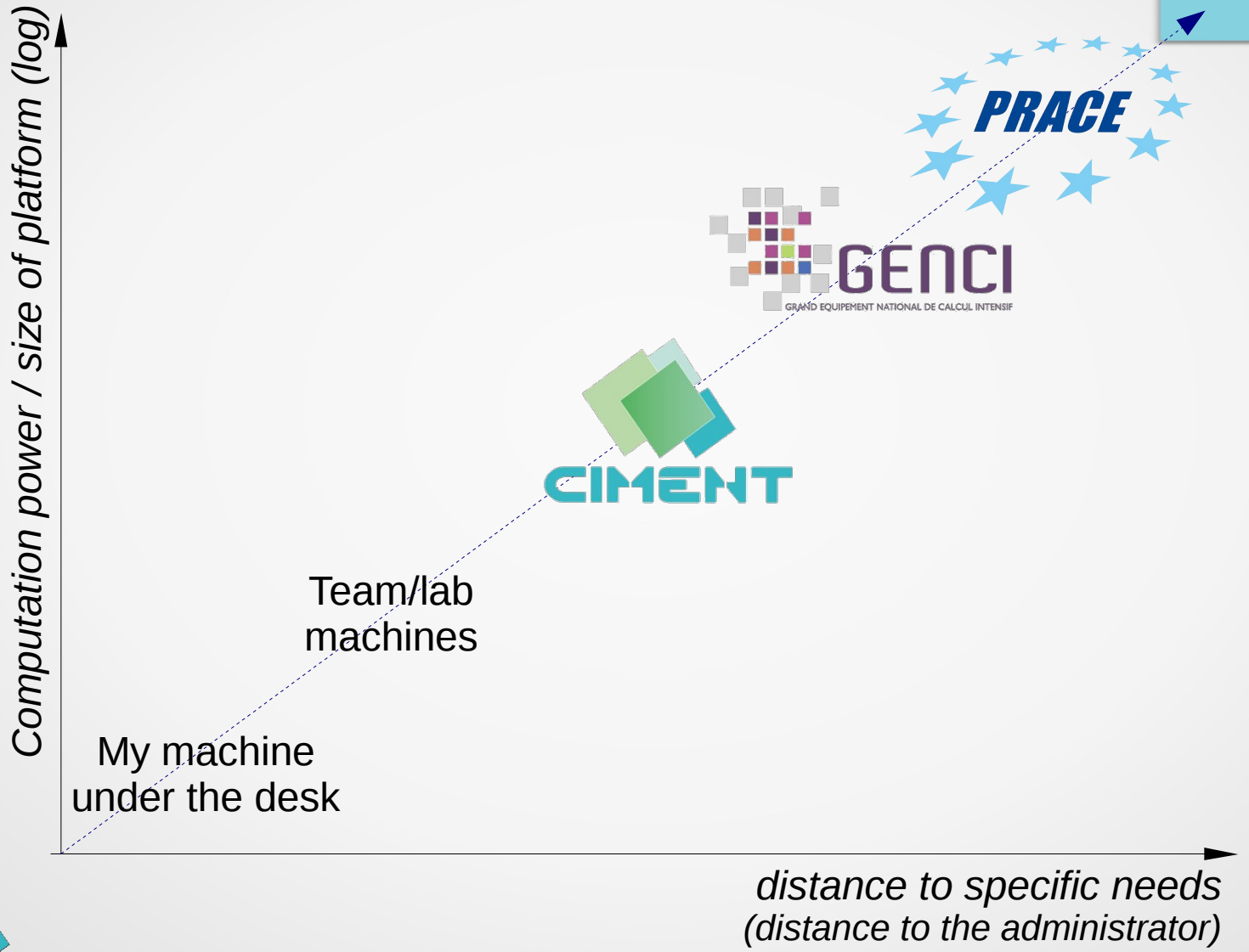
- « A **large-scale** and **versatile** testbed for **experiment-driven research** in all areas of computer science, with a focus on **parallel** and **distributed computing** including **Cloud, HPC** and **Big Data** »
  → Experimental validation of models, algorithms...



- **Reconfigurability & deep control**
  Users deploy their own experimentation platform!
  → (**Hardware-aaS**)
  - Change/setup/tune your own HPC stack
  - Control/monitor your own cloud computing infrastructure
  - Test/benchmark your own BigData storage protocols

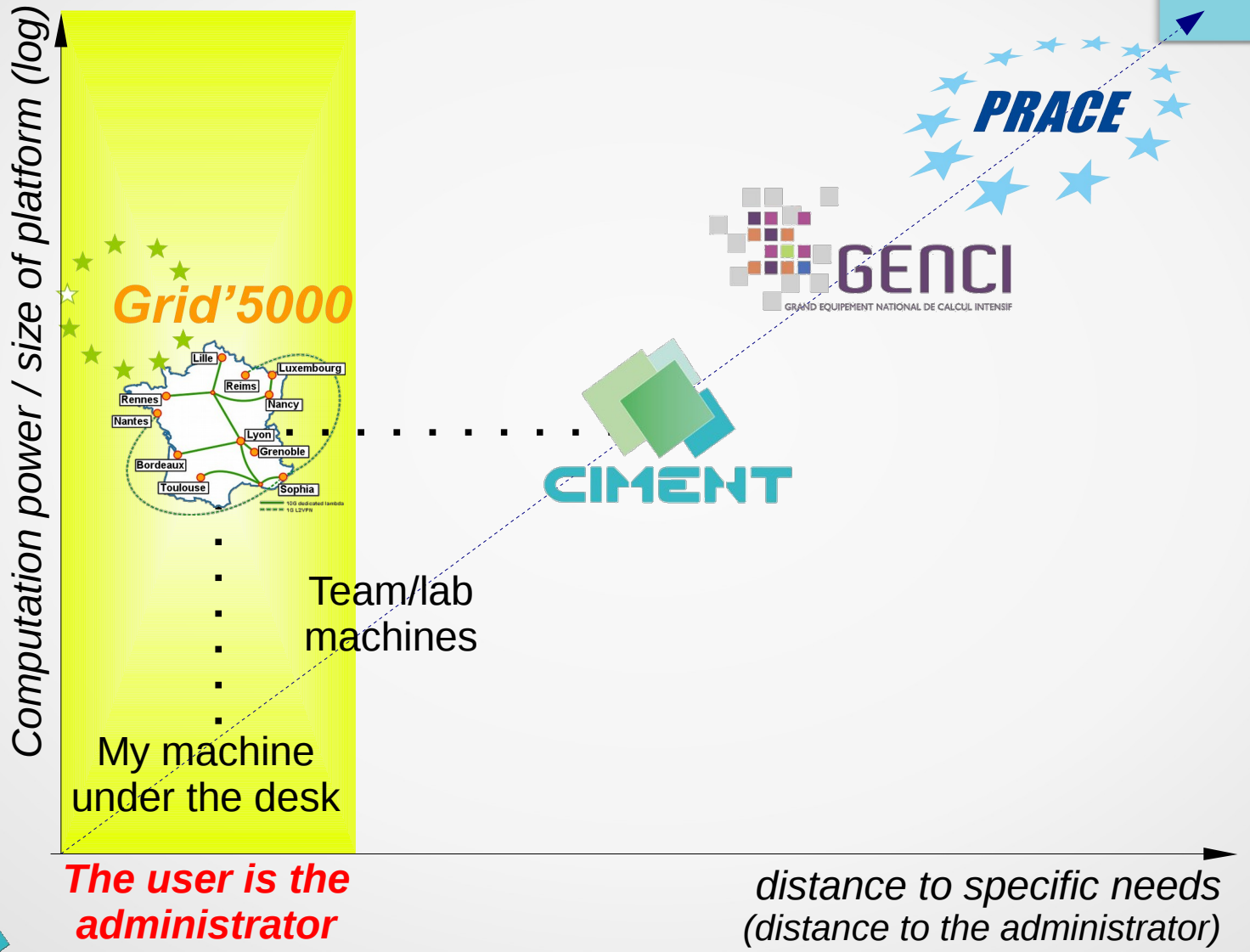
CIMENT

Trade-off: proximity vs. more resources



CIMENT

Grid'5000 vs. classic HPC platforms



CIMENT

Grid'5000 winter school

February 2-5 2016

Grenoble,
France

10 years after the Grid'5000 winter school and after the successful 2009, 2010, 2011, 2012 and 2014 editions (respectively 100, 72, 80, 67, 75 and 50 registered participants), Grid'5000 practitioners and future users are invited to gather, learn and share experience around the usage of Grid'5000 as a scientific instrument.

Grid'5000

Hosted by Inria Grenoble Rhône-Alpes, from February 2nd to February 5th 2016, this 7th edition of the Grid'5000 school will bring together, but is not limited to, Grid'5000's newbies as well as expert-users, technical team and executive committee for 4 days of tutorials and talks focusing on best-practices and results. Presentations and practical sessions will cover basic and advanced usages of the platform as well as lessons on experiment control at large-scale. A challenge to showcase tools and environments demonstrating the deployment of distributed systems on Grid'5000 (including large-scale middleware, parallel and cloud applications, etc.) will be held this year for the fourth time.

MORE INFORMATION:

https://www.grid5000.fr/mediawiki/index.php/Grid5000:School2016

Organized by GIS G5K, LIG Laboratory with the financial support of Inria

December 4th, 2015:

full-submission deadline for presentation and practical sessions proposals

CIMENT

Conclusion and perspectives

CIMENT

Success stories

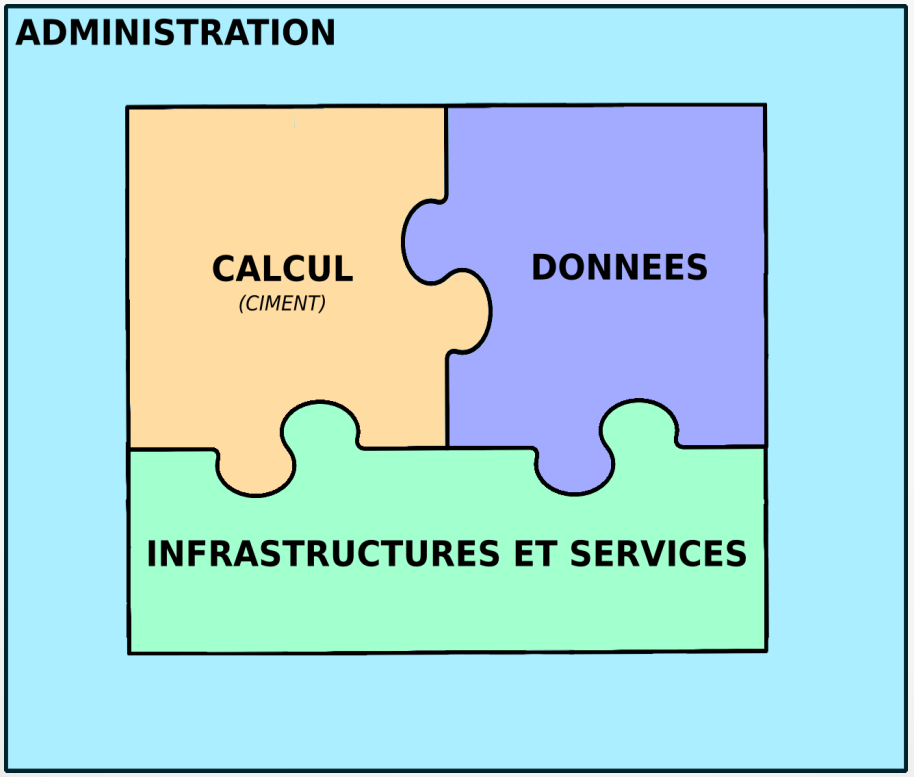
- 2014 → **101 publications** involving CIMENT computations
- Rosetta / Consert (IPAG):
  - Localization of the philae lander in a 23 x 42 m² area [Kofman, 2015] using **Luke and CIGRI**
- D0 / Particle physics (LPSC):
  - Measurement of the W Boson Mass with the D0 Detector [Phys. Rev. Lett. 108, 151804] – Analysis on **Cigri and Irods**
- Glassdef-poly (LiPHY):
  - Mechanical properties of semi crystalline polymers
  - 3,5 M-hours in 2014 on **Froggy**
- Scales (LEGI):
  - Validation of a CFD library
  - 1,5 M-hours in 2015 on **Froggy**

CIMENT

Next...

UMS “GRICAD”

« Grenoble Alpes Recherche - Infrastructure de CAlcul intensif et de Données »



CIMENT

Questions ?

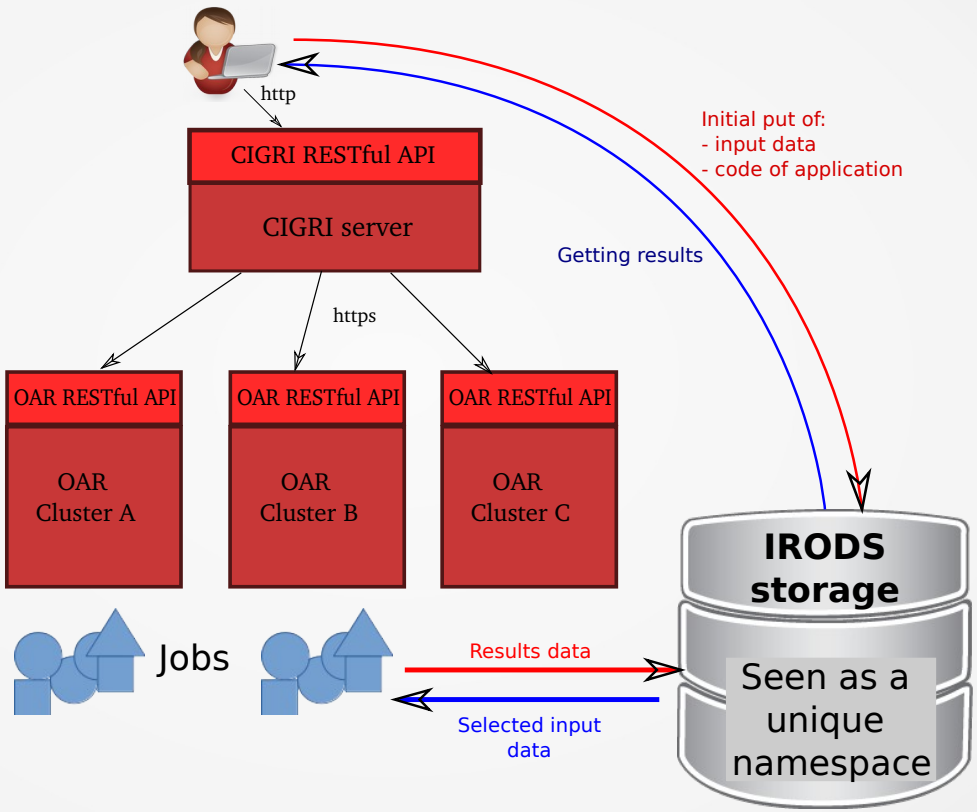
CIMENT

Backup slides

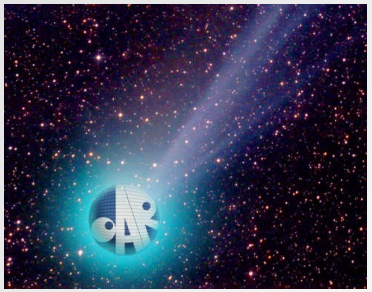
- Fonctionnement CIGRI
- Colmet
- Cigri/irods examples

CIMENT

Cigri and Irods



OAR/Colmet

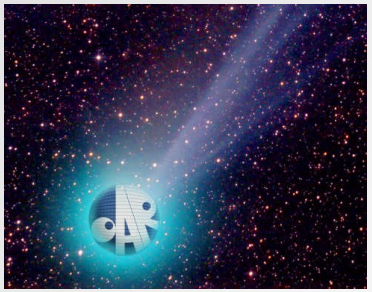


Le projet:

- *En cours depuis 2013 (co-financement d'un développeur sur quelques mois par CIMENT et INRIA)*
- *Expérimenté depuis début 2015 sur Froggy*
- ***P. Le Brouster, B. Bzeznik, S. Harrache, O. Richard***

- Accounting des ressources utilisées par les jobs OAR
- Récolte des données “taskstat” du noyau Linux toutes les 5 secondes (impact sur perfs mesuré → négligeable)
- Stockage des données dans fichier HDF5 sur serveur centralisé avec ZeroMQ
- Extraction des données depuis l'API REST de OAR
- L'index est le « job » (et non le processus ou le noeud)

OAR/Colmet

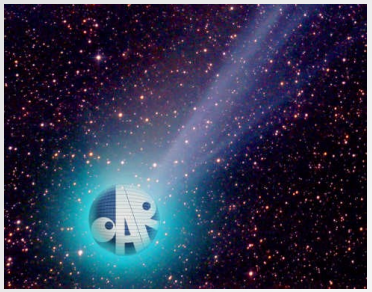


Utilisation: vers une meilleure exploitation des ressources

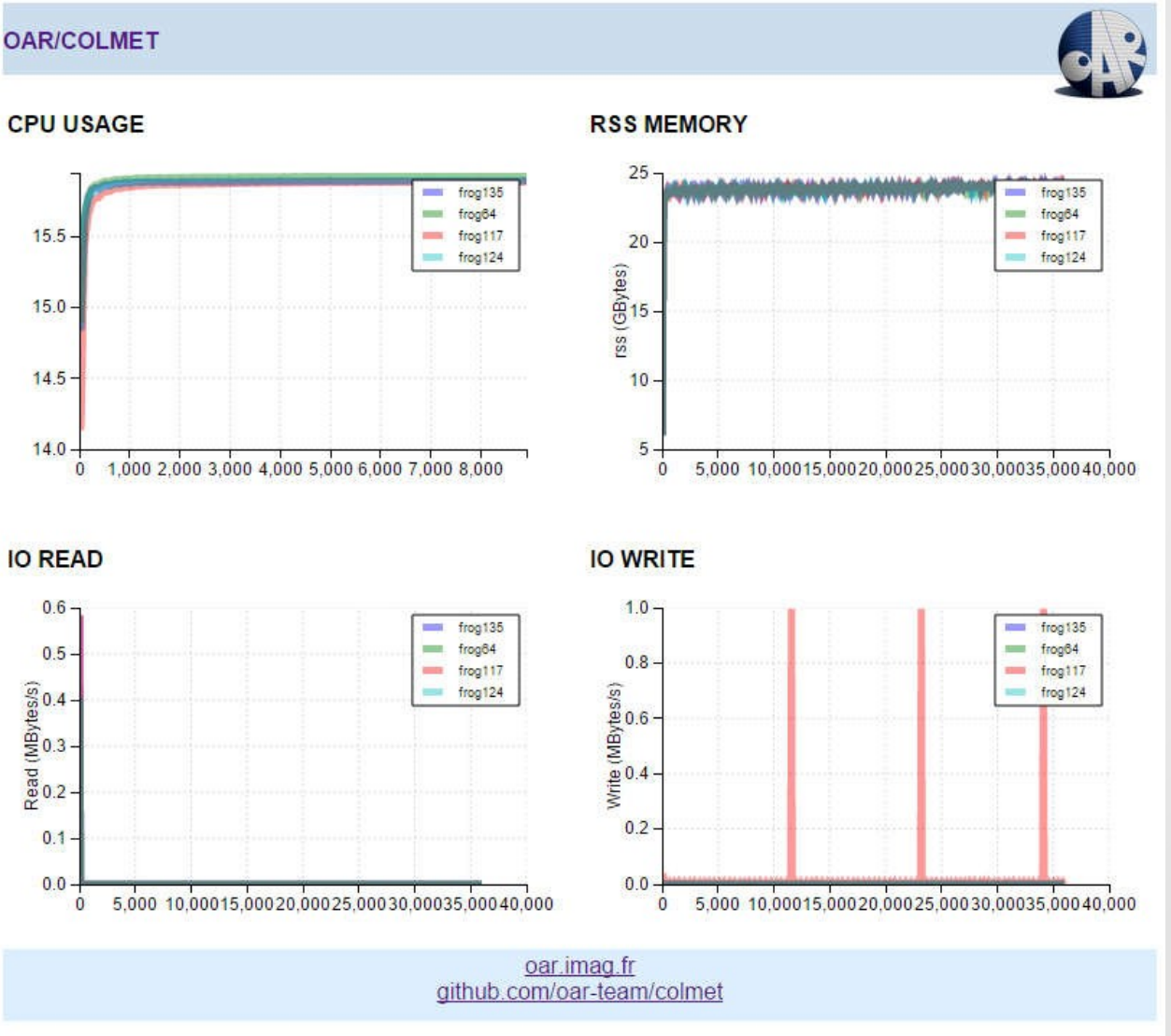
- Interface web en cours de développement
- Orienté vers une utilisation par l'admin pour le moment, mais pour les utilisateurs à terme.
- Permet de dégager des profils de jobs par rapport aux ressources matérielles exploitées (à gros grain; ne pas confondre avec le profiling d'un code)
- Détection de jobs déviants par étude statistique (un job déviant est un job qui n'exploite pas correctement les ressources : par exemple, les processus tournent tous sur un seul nœud)

CIMENT

OAR/Colmet



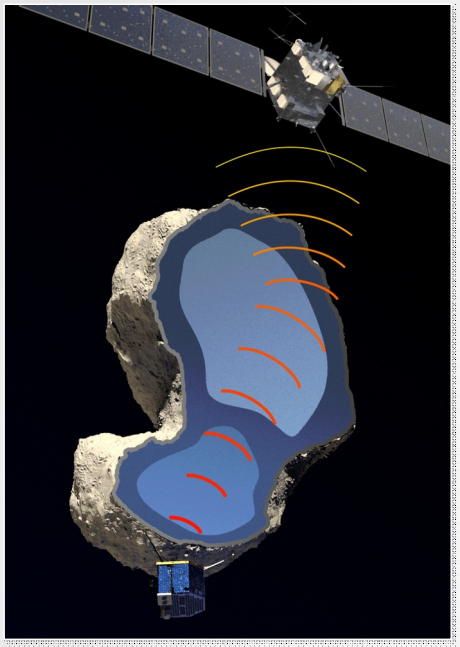
DEMO



Job “standard”:

- Charge ~16 sur tous les
- noeuds du job
- Emprunte mémoire importante
- Quelques écritures disque à interval régulier

Rosetta / CONSERT

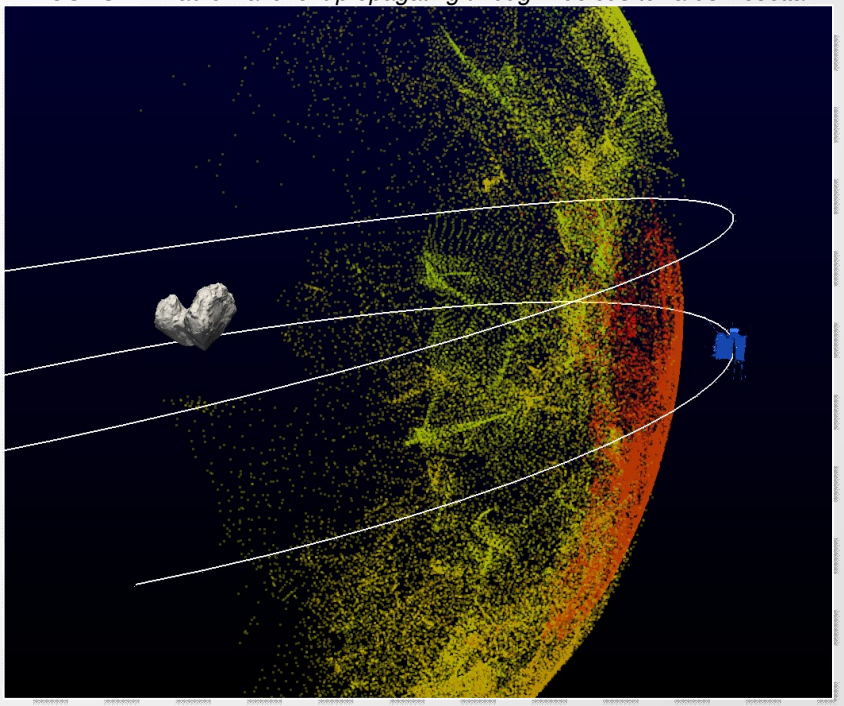


COmet Nucleus Sounding by Radiowave Transmission

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

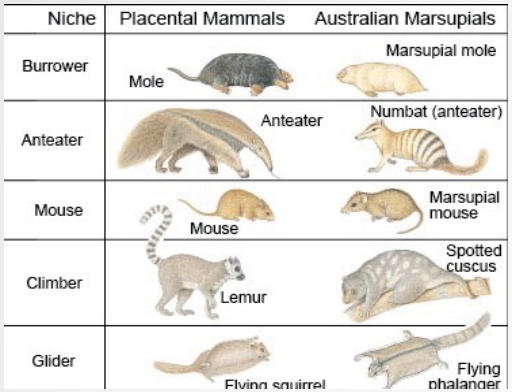


CIMENT with iRods were used for:

- 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

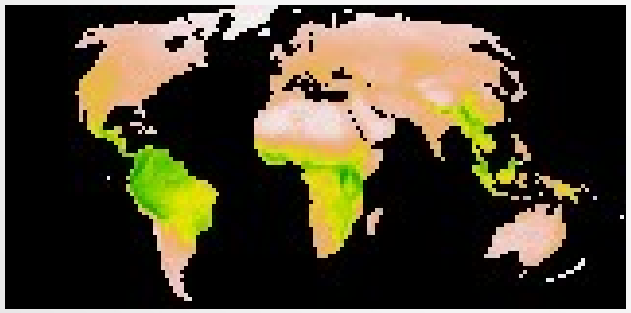


Computations

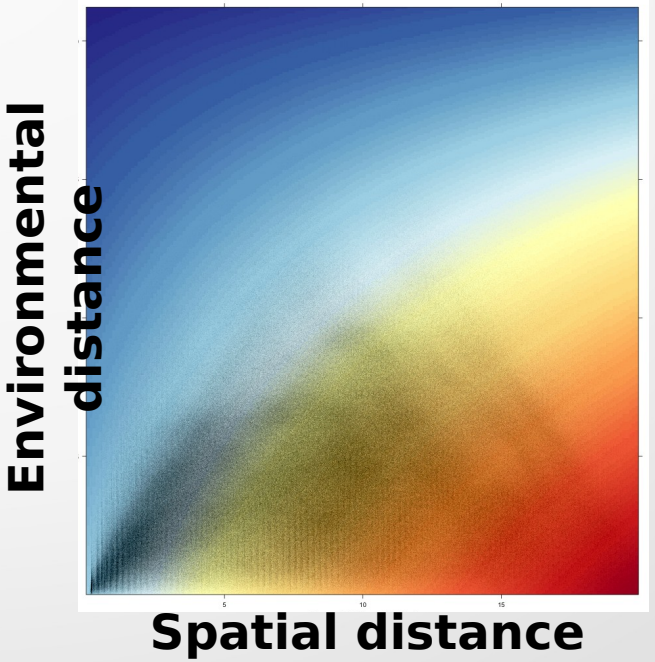
- Measure of morphological and species similarities between sites
  ⇒ 6 000 000 values

- 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

	CIMENT	Grid'5000
Primary Usage	HPC All Grenoble academic community Production workloads	Experimentation Distributed Comp. Research/National Production workloads **« allowed »**
Platform specificities	Optimized HPC software stack, many HPC software via modules Grid middleware for « bag of tasks »	Hardware-aaS / Reconfiguration Basic functionalities for HPC Users deploy their own software stack
Machines	HPC optimized hardware ~7000 cores Froggy : 3200 cores cluster GPUs, Xeon PHIs	11 sites, ~1000 nodes, ~6000 core Various hardware, some HPC, GPUs, Xeon PHIs
Network	IinfiniBand up to FDR (Froggy)	Ethernet 1GE, more and more 10GE, some InfiniBand DDR /QDR
Storage	NFS homedir 30GB Lustre scratch 90TB Irods global storage 700TB	NFS homedir 25GB Storage5K Dfs5k (ceph, ...)
Accounts	CIMENT LDAP / Perseus	Grid'5000 LDAP / UMS
Access	SSH + 1 visualization node www from frontends only	SSH/RestAPI whitelisted www from any node
OS	Linux / RedHat or Debian /appli with modules for HPC software	Default env is Linux Debian Users deploy any OS/software
Resources management	OAR (1/cluster) CIGRI for grid campaigns Walltime: < 4days (froggy)	OAR, kadeploy,... (1/site) Adv. Reservations / Interactive jobs Walltime: <2h or night+we
Support	sos-ciment@ujf-grenoble.fr → helpdesk (GLPI) + **MaiMoSiNE**	mailing list + bugzilla users@lists.grid5000.fr

CIMENT