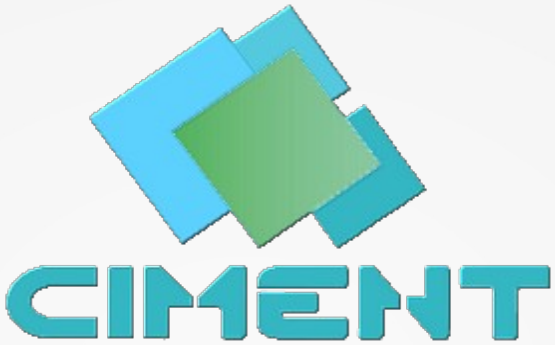


F. Roch
Séminaire RDATADEV
06/2015

La pyramide du HPC

- 2 piliers fondamentaux structurants au niveau national et international
  - GENCI (Grand Equipement National de Calcul Intensif) : PRACE (réseau européen), CC nat. (IDRIS, CINES, CCRT), coordination au niveau national & européen
  - Le CSCI (Comité stratégique pour le Calcul Intensif) : expertise, veille, analyse scientifique, conseil, orientations
- Une base forte
  - Mésocentres
  - Groupes calcul, les maisons de la modélisation (ex. MaiMoSine), de la simulation, les équipes de recherches
  - Des initiatives TER@TEC, ORAP, CERFACS, labo INTEL UVSQ CEA GENCI,
- Bull CEA « extreme computing »...
- Des liens forts
  - Exemple : equip@meso

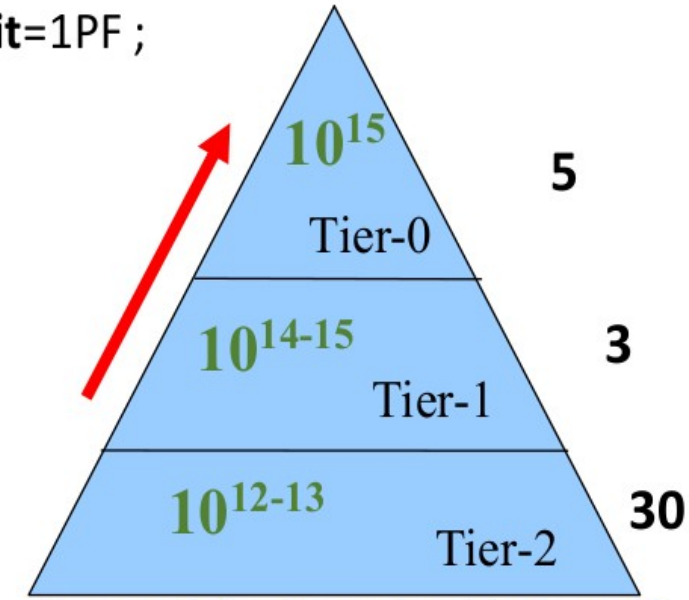
CIMENT

CIMENT mésocentre grenoblois de calcul intensif

~1 Pflops/s sur une plateforme
Jugene=1PF ; **Marenostrum**=1PF; **Hermit**=1PF ;
Juqueen=1.6PF; **Curie**=2PF ; **Fermi**=2PF;
SuperMuc=3PF

~100 Tflops/s sur une plateforme
Turing= 836, **Ada**=230 ; **Occigen**=2100 ;
Curie=~500

126 Tflops/s répartis sur **13 plateformes** de 96 à 3216 cœurs
Accessibles en mode grille (CiGRI)



Pyramide du calcul intensif

Opérations à vf par seconde

Le mésocentre CIMENT

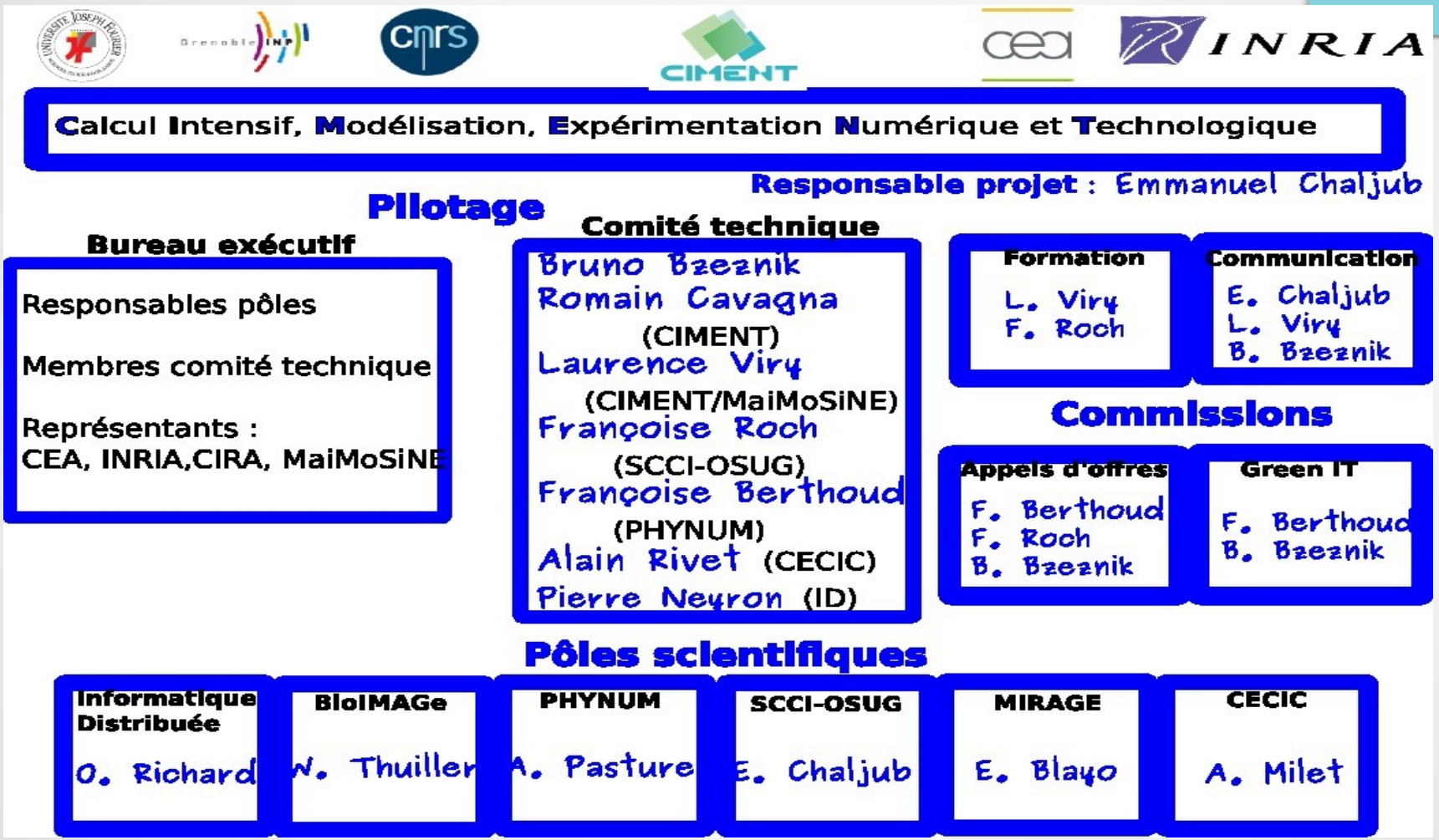
- Création en 1998 (AAP du ministère : SSCI-OSUG, Grappe de PC, MIRAGE)
- 2000-6 CPER CIMENT, séminaires, formations doctorales, expertises partagées, vers l'hébergement à la DSI, collaborations
- 2004 CIMENT GRID
- 2007-13, CIRA (CIMENT+FLMSN(Fédération Lyonaise de +Savoie), plate-forme du RTRA , naissance de MaiMoSiNE,

  Equip@meso : plateforme froggy,

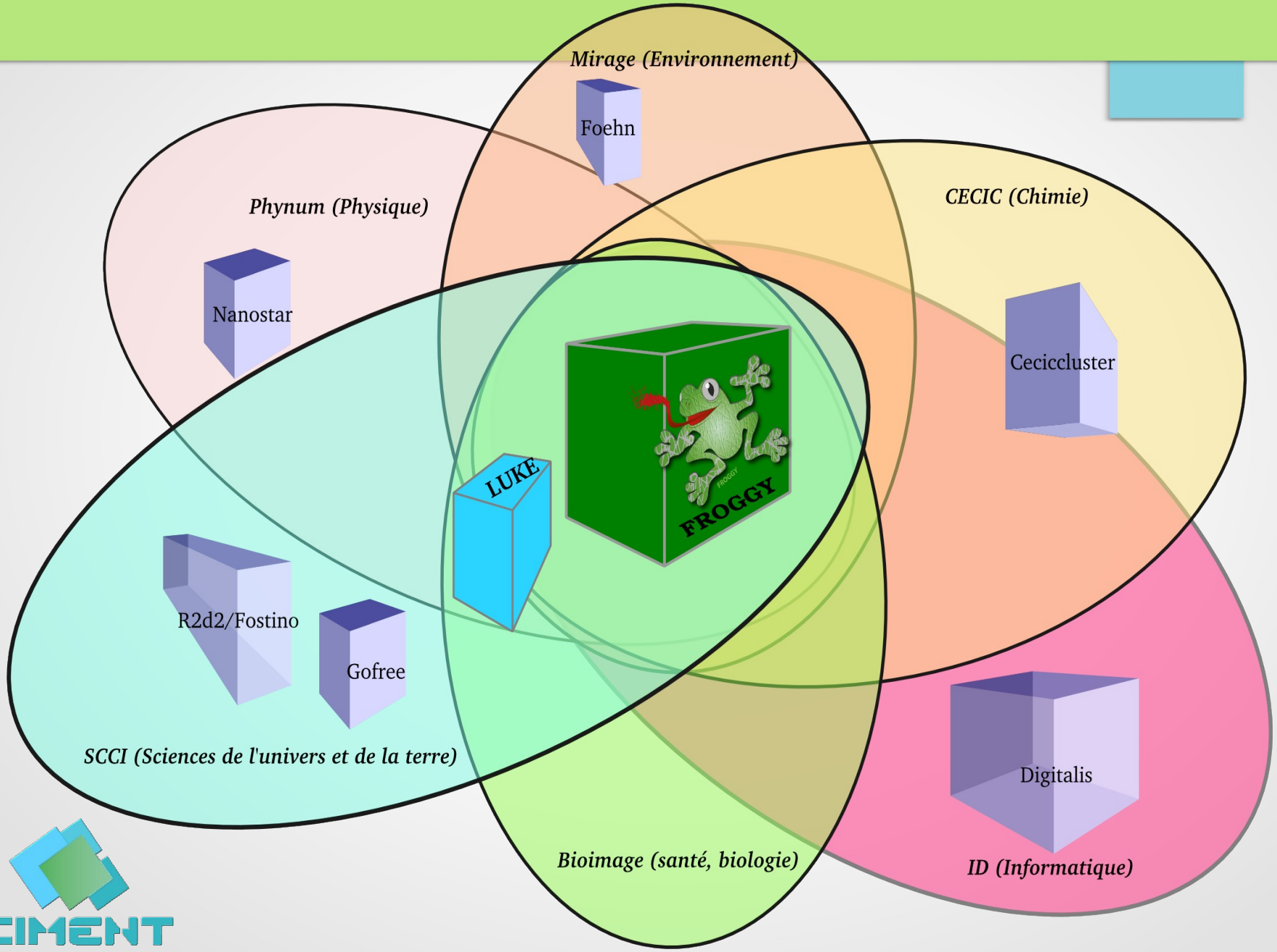
  projet d'infra de salles partagées écoresponsables, etc.
- 2014 CIDRA : CPER Rhône Alpes

CIMENT

L'Organisation CIMENT

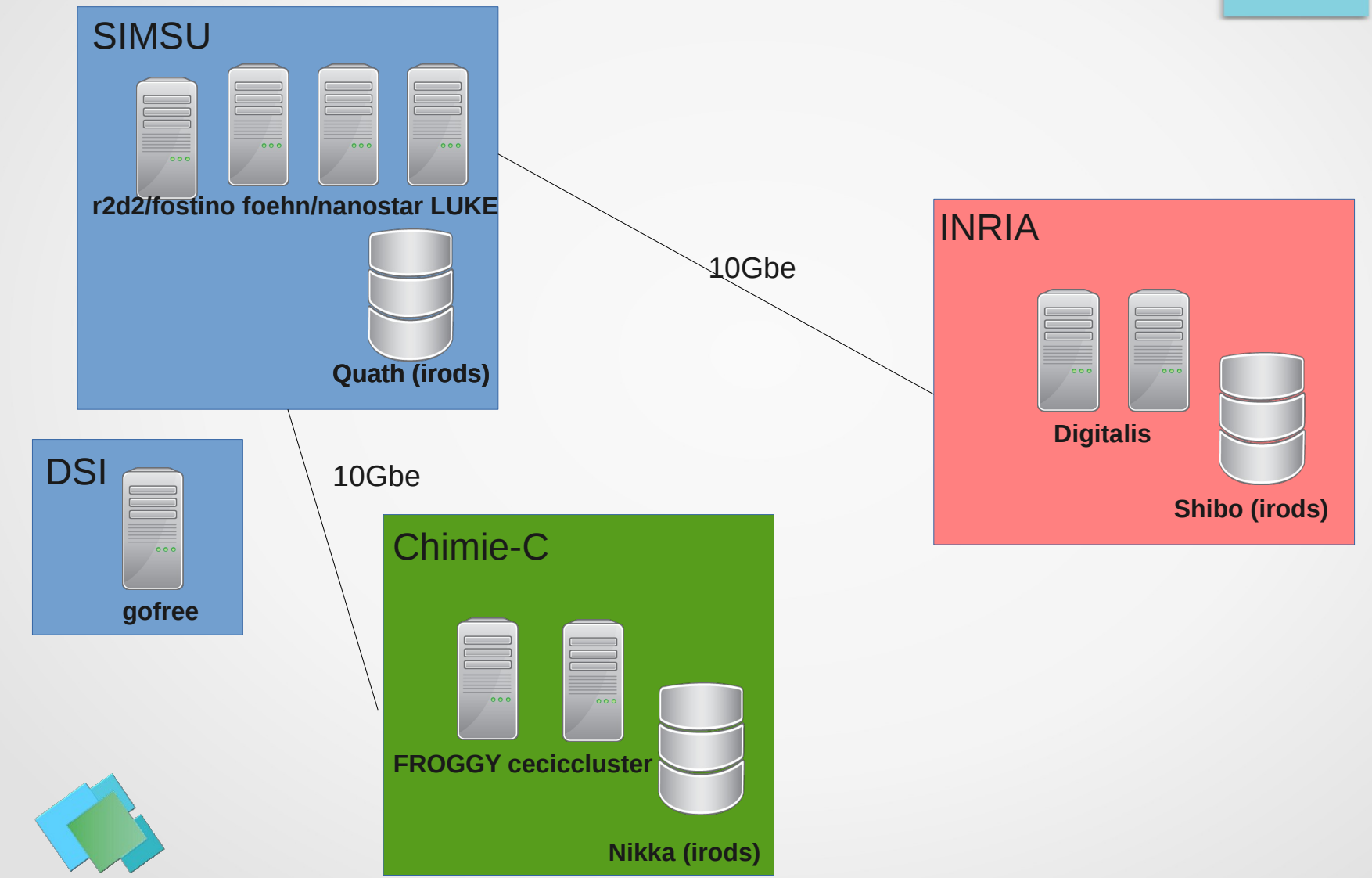


CIMENT: plateformes de calcul et pôles



CIMENT

Les sites


SIMSU
r2d2/fostino foehn/nanostar LUKE
Quath (irods)
10Gbe
INRIA
Digitalis
Shibo (irods)
DSI
gofree
10Gbe
Chimie-C
FROGGY cecicluster
Nikka (irods)

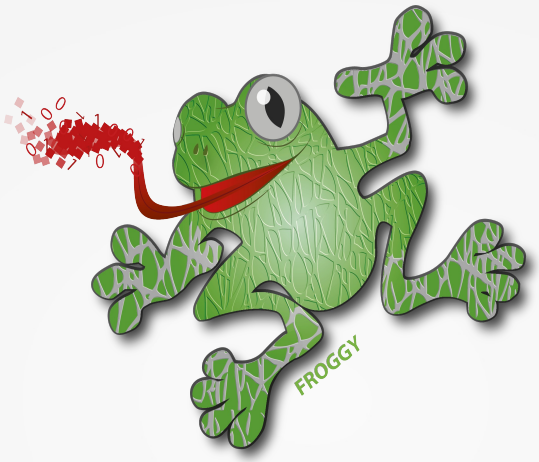
Liste des plateformes (Dec 2014)

Name	Vendor	Number of cpu cores	Total memory	Max memory/node	Total storage (net)	Computing network	Total Gflop/s	Accel Gflop/s *	Purchase date
Luke	Dell	120	800 GB	256 GB	350.28 TB	Gigabit ethernet	2358	0	2014-05-01
Froggy	Bull	3184	15328 GB	512 GB	118.55 TB	Infiniband FDR	82505	16380	2013-05-31
Cecicluster	Dell	216	432 GB	24 GB	12.5 TB	Infiniband QDR	1961	0	2011-12-01
Gofree	Dell	336	2016 GB	72 GB	30 TB	Infiniband QDR	3178	0	2011-01-01
Fontaine	Dell	144	288 GB	24 GB	12 TB	Infiniband QDR	1308	0	2010-11-01
Global_storage	Dell		264 GB	24 GB	520 TB	10Gb/s ethernet	0	0	2010-09-01
Foehn	SGI	128	480 GB	48 GB	7 TB	Infiniband DDR	1367	0	2010-03-01
Adonis	Bull	96	288 GB	24 GB	0 TB	Infiniband DDR	4676	3634	2010-01-01
Nanostar	SGI	256	512 GB	16 GB	7 TB	Infiniband DDR	2560	0	2009-01-01
Edel	Bull	576	1728 GB	24 GB	0 TB	Infiniband DDR	5230	0	2009-01-01
Genepi	Bull	272	272 GB	8 GB	5.44 TB	Infiniband DDR	2720	0	2008-10-10
Fostino	IBM	464	464 GB	8 GB	27.5 TB	Gigabit ethernet	5197	0	2008-09-01
R2d2	IBM	512	1088 GB	32 GB	19.24 TB	Infiniband DDR	5120	0	2008-09-01
Airelle	Dell	276	676 GB	128 GB	9.054 TB	Gigabit ethernet	2563	0	2008-01-01

CIMENT

2 machines communes

- FROGGY

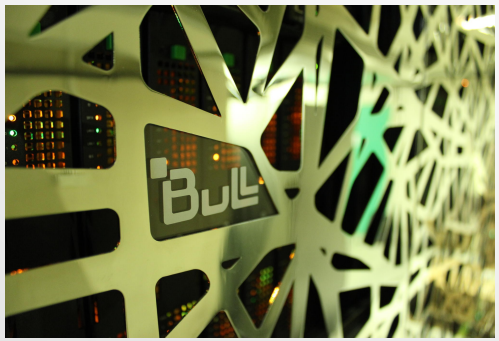


- LUKE

CIMENT

FROGGY

- Plate-forme phare de CIMENT (projet Equip@meso)
- Dédiée au **calcul parallèle** (jusqu'à ~3000 cpu-cores, 18 GPUs)
- Réseau faible latence non bloquant (Infiniband FDR)
- Scratch distribué sur système de fichiers partagé (Lustre), 3Go/s
- Administration centralisée
- Évolutive par ajout de nœuds
- Refroidissement à eau tiède



FROGGY

- **190 noeuds** à 2 CPUs Intel SB 2.6 GHz/20M
- **64Go** RAM par noeud (20% avec **128Go**)
- 1 noeud de **visu**
- Un "**fat node**" (4 cpus, 512Go RAM)
- **9 noeuds bi- GPUs** Nvidia K20
- **90TB** de scratch distribué, **3Go/s**
- Refroidissement DLC (**Direct Liquid Cooling**)

LUKE

- Machine hétérogène construite au fil de l'eau
- Dédiée aux **calculs séquentiels** ou aux traitements de données massives
- Réseau 10Gbe
- Gros scratchs locaux aux noeuds
- Partitions dédiées à des projets
- Certains noeuds dédiés à la grille CiGri
- Possibilité d'intégrer les noeuds aux ressources Irods
- Administration collaborative

CIMENT

LUKE mai 2015

Node	Team	cpus	cores	GB/core	Mem GB	/var/tmp GB
luke0	visu	1	4	2	8	None (None)
luke1	ciment	2	20	3	64	10 (1hd)
luke2	ciment	2	24	5	128	13 (1hd)
luke3	seiscope	2	16	16	256	37252 (12hd6)
luke4	whisper	2	20	6	128	11697 (12hd6)
luke5	gipsalab	2	16	8	128	37252 (12hd6)
luke6	consert	2	12	2	32	11811 (12hd6)
luke7	seiscope	2	12	4	48	40676 (12hd6)
luke8	i2mars	2	16	4	64	93 (2ssd1)
luke9	emabio	2	16	16	256	18349 (12hd6)
luke10	gin	2	20	6	128	756 (1hd)
luke11	gin	2	20	6	128	756 (1hd)
luke12	gin	2	20	6	128	756 (1hd)
luke13	gin	2	20	6	128	756 (1hd)
luke14	cycle	2	16	4	64	18 (2hd0)
luke15	cycle	2	16	4	64	18 (2hd0)
luke16	cycle	2	16	4	64	18 (2hd0)
luke17	cycle	2	16	4	64	18 (2hd0)
luke18	gipsalab	2	8	8	64	5825 (9hd5)
luke19	ciment	2	10	3.2	64	137 (12hd6)
luke20	ciment	2	10	3.2	64	139 (4hd5)

Espaces disques spécifiques

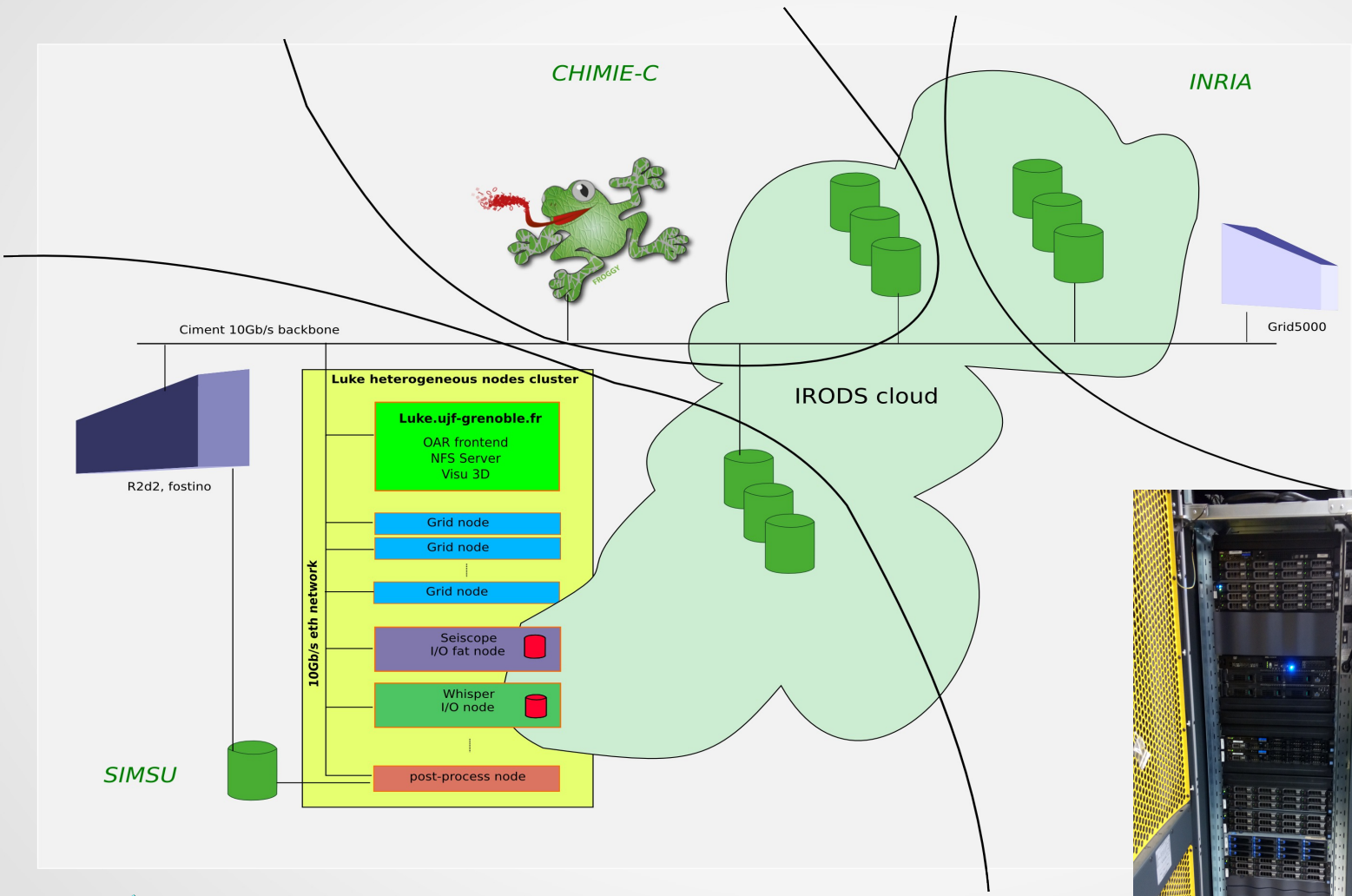
Node	Team	montage	TB
luke0	visu	/home	15
luke-nfs	ciment	/nfs_scratch	37
	ciment	/nfs_scratch2	37
	ciment	/beegfs_scratch	74
	ciment	/applis/site	37
luke3	seiscope	/scratch_r720	37
luke4	whisper	/scratch_r720	37
gin-nfs	gin	/gin/data	48
luke14-17	cycle	/cycle_scratch	20
luke19	ciment	/scratch_r730	11
luke20	ciment	/scratch_r730	11

luke : frontale du cluster
luke-nfs : serveur nfs dédié

2 filesystems // BeeGFS performants:
/cycle_scratch
/beegfs_scratch

CIMENT

LUKE


CHIMIE-C
INRIA
Ciment 10Gb/s backbone
Grid5000
IRODS cloud
R2d2, fostino
Luke heterogeneous nodes cluster
Luke.ujf-grenoble.fr
OAR frontend
NFS Server
Visu 3D
Grid node
Grid node
Grid node
Seiscope
I/O fat node
Whisper
I/O node
post-process node
10Gb/s eth network
SIMSU

LUKE

```
roch@luke:~$ chandler

luke0 (visu)       AAAA                          luke1 (ciment)     BBBBBBBBBBBBBBBBBBB
luke2 (ciment)     JJJJJJJJJJJJJJJJJJJJJJ        luke3 (seiscope)   BBBBBBBBBBBBBBB
luke4 (whisper)    JJJJJJJJ                      luke5 (gipsalab)   BBBBBBBBBBBBBBB
luke6 (consert)    BBBBBBBBBBB                   luke7 (seiscope)   BBBBBBBBBBBBBBBBBBBBBBB
luke8 (i2mars)     JJJJJJJJJJJJJ                 luke9 (emabio)     BBBBBBBBBBBBBBB
luke10 (gin)       BBBBBBBBBBBBBBBBBBB           luke11 (gin)       BBBBBBBBBBBBBBBBBBB
luke12 (gin)       BBBBBBBBBBBBBBBBBBB           luke13 (gin)       BBBBBBBBBBBBBBBBBBB
luke14 (cycle)     BBBBBBBBBBBBBBBB              luke15 (cycle)     BBBBBBBBBBBBBBBB
luke16 (cycle)     BBBBBBBBBBBBBBBB              luke17 (cycle)     JJJJJJJJJJ
luke18 (gipsalab)  BBBBBBBB                      luke19 (ciment)    BBBBBBBBBBBBBBBBBBB
luke20 (ciment)    JJJJJJJJJJJJJJJJJJJ

 =Free B=Besteffort  =Standby J=Job S=Suspected A=Absent D=Dead

34 jobs, 360 resources, 4 down, 336 used

                Jobs       Jobs
User           running    waiting   Resources    Nodes
=================================================
lecointre        8          0          8          1
cotere           1          0          1          1
vieillam         1          0          10         1
volat            1          0          44         2
rogezy           15         6          260        15
doutes           1          0          12         1
escorneg         1          0          1          1
```

20 noeuds de calcul
Totalisant :
- 344 cores
- 2To RAM
- 450 To stockage
- 7.4 Tflops

La grille CIMENT

- Constituée de l'ensemble des clusters de CIMENT
- Dédiée aux applications de type multi-paramétriques

Cigri est le middleware de gestion des jobs grilles

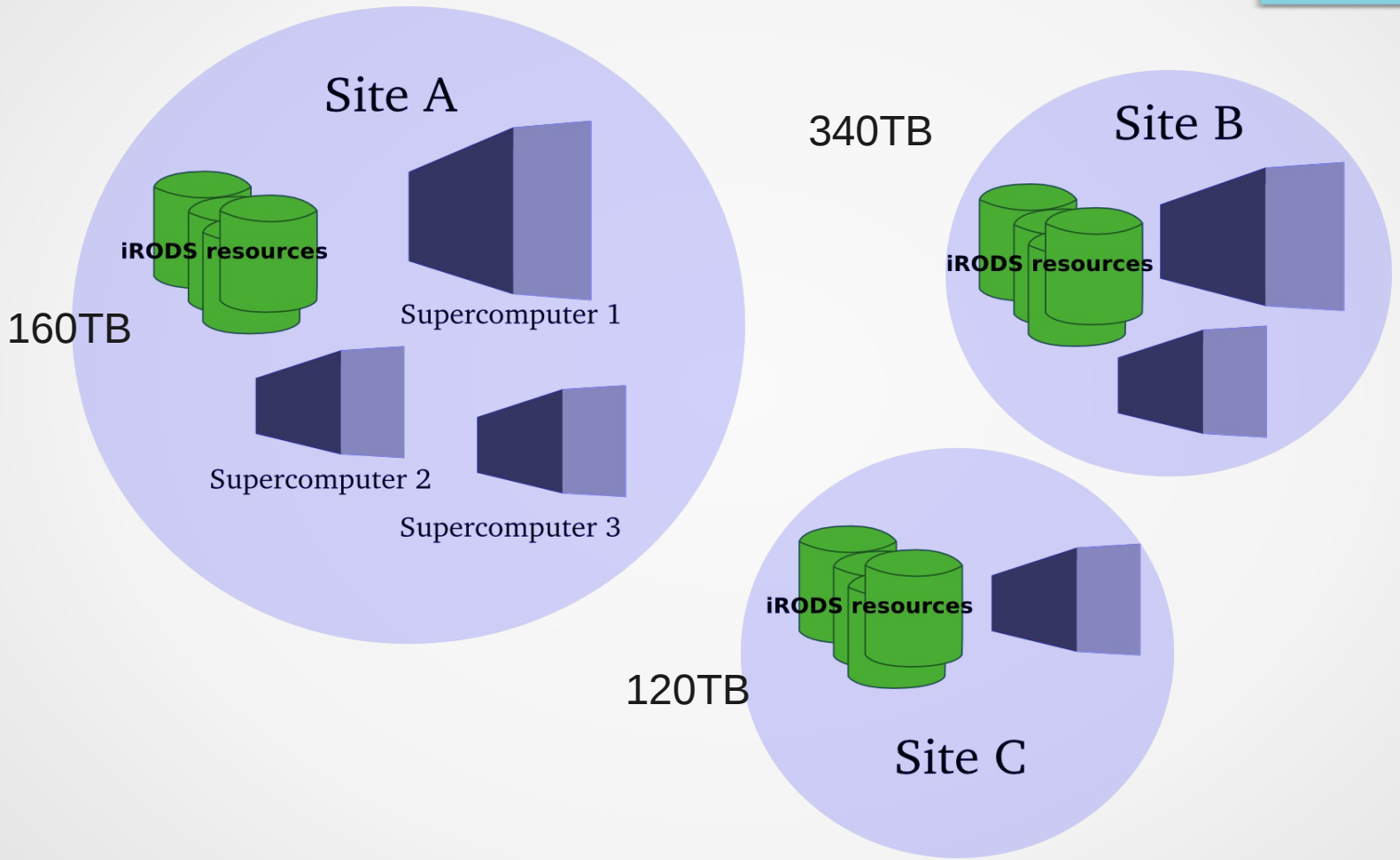
Développement LIG-INRIA-CIMENT

http://cigri.ujf-grenoble.fr

Irods est le middleware qui permet de gérer les données distribuées sur la grille, il permet l'accès à un gros espace de stockage

CIMENT

Grille de stockage (IRODS)



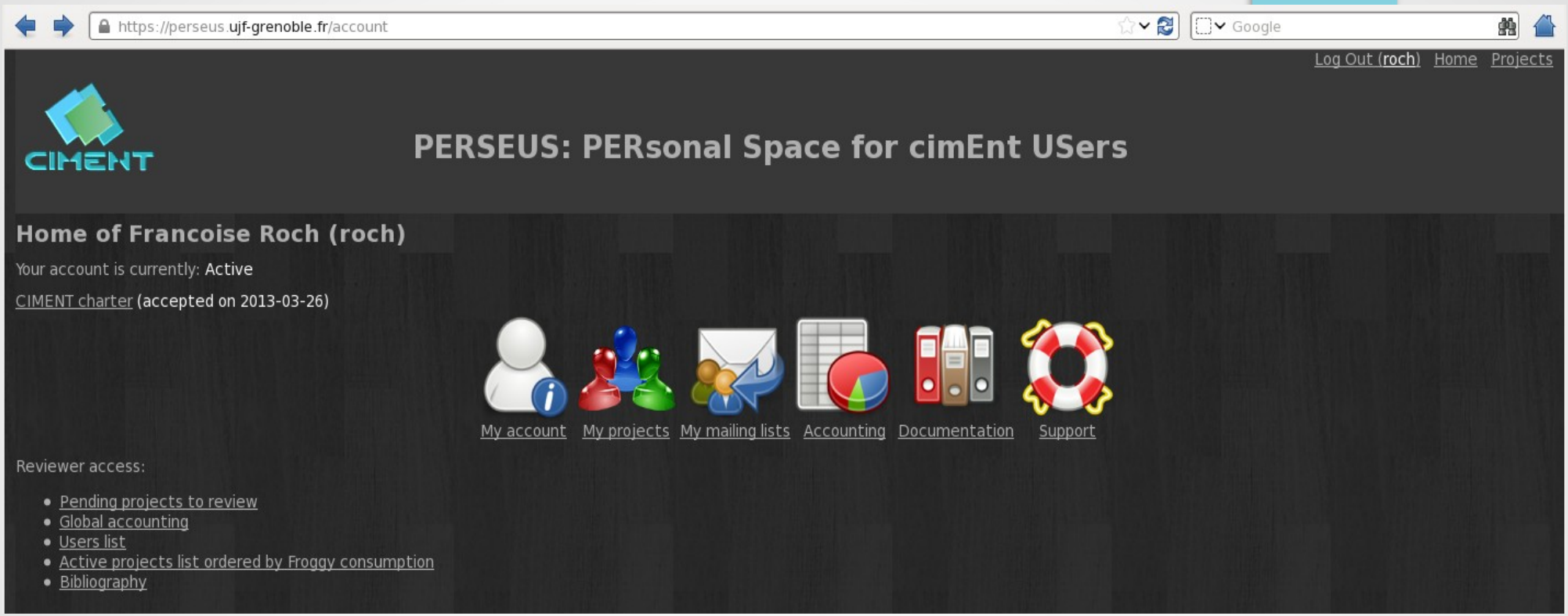
+Luke: 45OTB

Au total, en 2015...

- 14 HPC platforms
- **6972** cpu-cores
- 25 Tera Bytes of RAM
- **1,2 Peta Bytes of disk storage**
- **125 Tflop/s** (20 Tflops of accelerators)

CIMENT

Perseus : PERsonal Space for cimEnt USers

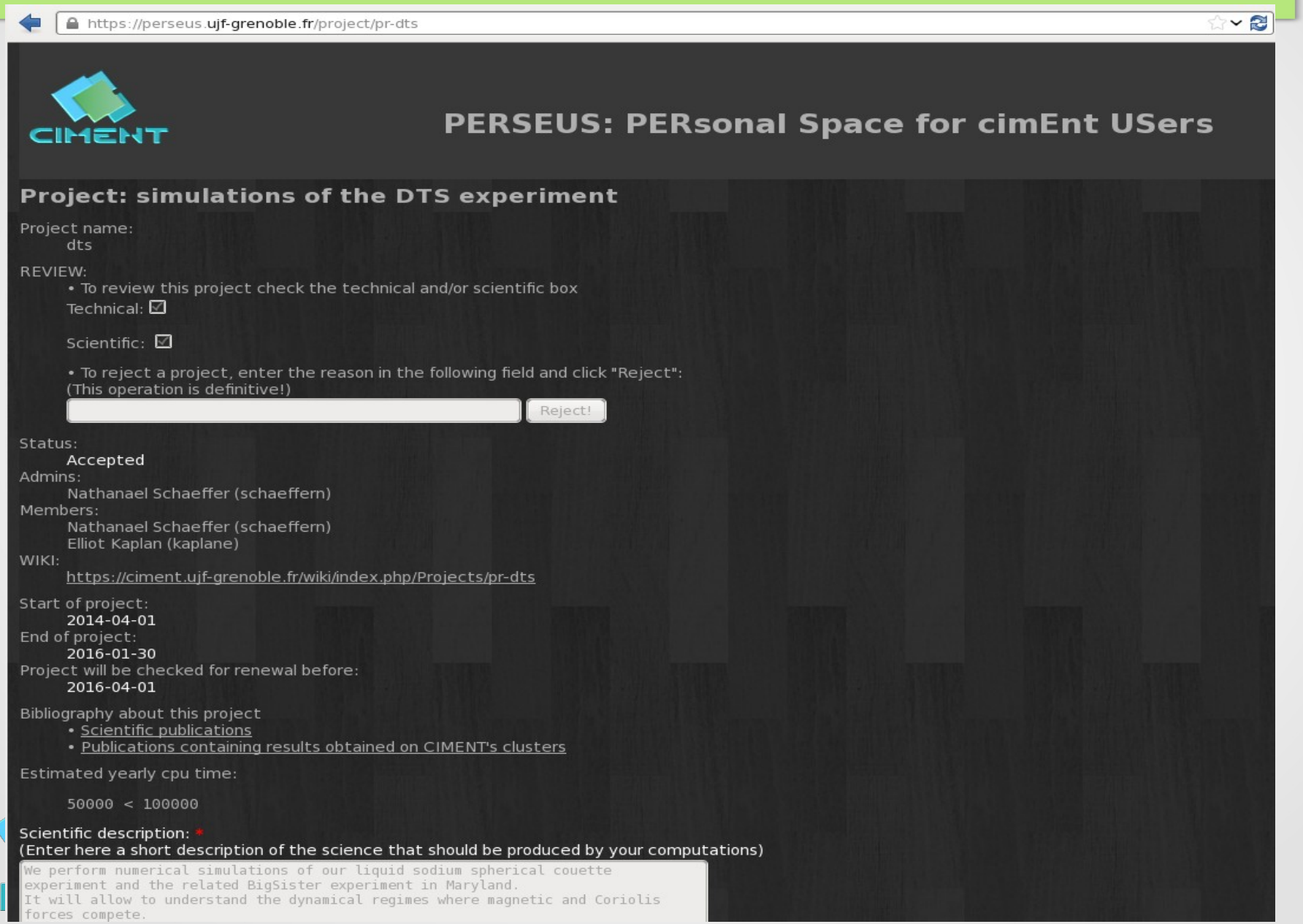


Développement : R. Cavagna & B. Bzeznik

- Point d'entrée unique au service
- Gestion des projets, espace wiki par projet
- Gestion des publications (bibtex et DOI)

CIMENT

Perseus / projects



Perseus / accounting

Ciment Accounting for roch

Values are given in hours or cpucore-time

You can also check the Global Accounting page

- **Froggy:**

Project	Your hours	Total hours
formation-hpc-130614:	1.2	143.8
admin:	2678.9	57853.7
formation-ced-mdm-2015:	9.0	531.8

- **Gofree:**

Project	Your hours	Total hours
default:	10519.2	5067415.3

- **Fostino_r2d2:**

Project	Your hours	Total hours
default:	2913.2	20412325.6

- **Fontaine:**

Project	Your hours	Total hours

- **Luke:**

Project	Your hours	Total hours
admin:	151.6	731.6

Ciment Global Accounting

From: 2014 - 1 - 1

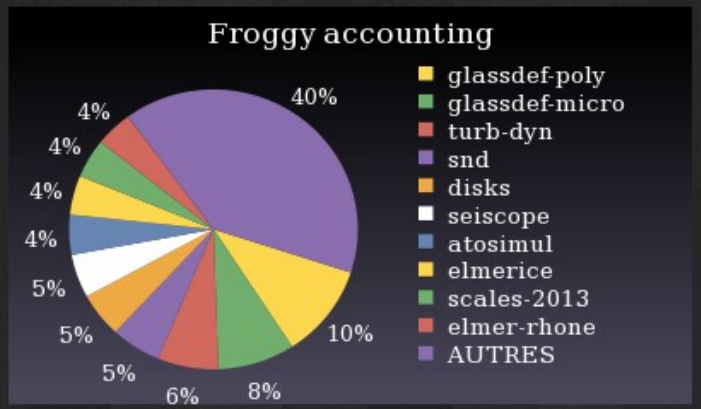
TO: 2015 - 12 - 31

Include users details: ☐

update

Values are given in hours of cpucore-time

- **Froggy:**



TOTAL: 40603780

glassdef-poly	4393218
glassdef-micro	3517183
turb-dyn	2755902
snd	2288042
disks	2093273
seiscope	2031322
atosimul	1882386
elmerice	1833509
scales-2013	1832017
elmer-rhone	1696100
simul_battery	1279454
glassdef-meso	1244912
teembio	1068771
cecipex	1030023
liqsim	983986
tbsim-freeride	765662
braess-dynam	750613

OAR

- Gestionnaire de jobs et de ressources, développé au LIG
- Fonctionalités standard :
  - Soumission/destruction de jobs, demande de ressources, ..
- + :
  - Advance Reservation
  - Expression des hiérarchies dans les requêtes
  - Support de ressources de types différents (licences, stockage, réseau, ...)
  - Container (récursivité, soumettre dans une tâche)
  - Tâche besteffort (tâche à priorité nulle, utilisé par CiGri)
  - Type multiple de tâches (besteffort, deploy, timesharing, ...) (personnalisable)
  - Economie d’énergie

CIMENT

OAR

- Soumission pour tâche interactive :

  [toto@froggy1 ~]$ oarsub -l nodes=4 -I --project myproj

  [ADMISSION RULE] Set default walltime to 1800.

  [ADMISSION RULE] Set default walltime to 1800.

  [ADMISSION RULE] You requested 64 cores

  OAR_JOB_ID=6556533

- Soumission en batch :

  oarsub -l walltime=2 :00,nodes=10 --project myproj /home/toto/script

- Soumission d'une réservation :

  oarsub -r "2015-06-27 11 :00:00" -l nodes=12

  OAR_JOB_ID=6647984

- Connexion à une réservation (utilise le numéro de tâche) :

  oarsub -C 6647984

CIMENT

OAR/sélection des ressources

- Soumission sur 4 cores répartis sur 2 cpus de 2 noeuds distincts (2 par noeud):

  [toto@froggy1 ~]$ oarsub -l nodes=2/cpu=1/core=2 --project myproj

- Soumission dans le “bac à sable”, walltime limité, orienté développement

  [toto@froggy1 ~]$ oarsub -t devel -l --project myproj

- Soumission sur les noeuds de froggy possédant 128 Go de RAM (8go/core) :

  [toto@froggy1 ~]$ oarsub -p "memcore >= 8" -l walltime=2 :00,nodes=10 --project myproj /home/toto/script

  => Avoir une connaissance minimum de l'achitecture de la machine

  Cf . Document : les 10 choses à connaître sur froggy

  http://ciment.ujf-grenoble.fr/wiki/index.php/10_choses_froggy

CIMENT

OAR/ordonnancement

- L’ordonnancement est l'étape où le système choisit les ressources attribuées aux tâches et les dates de lancement.

- L’ordonnancement est défini suivant une politique basée sur l’utilisation d’algorithmes d’ordonnancement
    - FIFO (First-In First-Out)
    - First-Fit (Backfilling) -> *on remplit les trous*
    - FairSharing -> *on tient compte de la consommation passée*
    - Timesharing
    - Advance reservation
    - Récursivité

- De nombreux critères et paramètres sont utilisés pour guider et cadrer les allocations et les priorités.

    Ex: projet prioritaire sur une période donnée, contribution financière, limitation de l'occupation par les jobs séquentiels ...

CIMENT

Bibliothèques de programmation

- Ciment fournit aux utilisateurs des applications et bibliothèques communes :
  - **/applis/site :** répertoire spécifique à la plateforme
  - **/applis/ciment/v2** : répertoire d'applications distribué avec rsync sur tous les clusters (utilisation principale : grille)
    - Embarque sa propre glibc
    - Un seul binaire pour toutes les plateformes
- Licences flottantes du compilateur Intel ouvert à toutes les plateformes, outils pour le tuning et le debugging (sur budget equip@meso)

CIMENT

Outil module

chargement d'un environnement par défaut, accès à la commande module

[froggy1]$ **. /applis/site/env.bash**

Listing de tous les modules disponibles au chargement

[froggy1]$ **module avail**

```
---------------------- /applis/site/modulefiles ------------------------
allinea-ddt/4.0                 isl/0.11.1_gcc-4.4.6
atlas/3.11.17_gcc-4.4.6         isl/0.12.2_gcc-4.4.6
autoconf/2.69_gcc-4.4.6         libffi/3.0.13_gcc-4.4.6
bison/2.7.1_gcc-4.4.6           libtool/2.4.2_gcc-4.4.6
blacs/19.04.2011_gcc-4.4.6      metis/5.0.2_gcc-4.4.6
boost/1.54.0_gcc-4.4.6          metis/5.1.0_gcc-4.4.6
bullxde/2.0                     mpc/1.0.1_gcc-4.4.6
bullxmpi/bullxmpi-1.2.4.1       mpfr/3.1.2_gcc-4.4.6
cloog/0.18.1_gcc-4.4.6          mpich/3.0.4_gcc-4.4.6
cmake/2.8.11.2_gcc-4.4.6        mumps/4.10.0_gcc-4.4.6
doxygen/1.8.3_gcc-4.4.6         ncurses/5.9_gcc-4.4.6
eigen/3.2.0_gcc-4.4.6           netcdf/4.3.0_intel-13.0.1
fftw/3.3.3_gcc-4.4.6            netcdf-fortran/4.2_intel-13.0.1
fftw/3.3.3_gcc-4.8.1            openmpi/1.6.4_gcc-4.4.6
imagemagick/6.8.7-4_gcc-4.4.6   python/3.4.1_gcc-4.4.6
intel-compilers/13.0.1          qt/4.8.3_gcc-4.4.6
intel-compilers/14.0.2          qt/4.8.5_gcc-4.4.6
intel-itac/13.0.1               qt/5.1.1
intel-itac/14.0.2               scalapack/2.0.2_gcc-4.4.6
intel-mkl/13.0.1                scotch/6.0.0_gcc-4.4.6
intel-mkl/14.0.2                slepc/3.4.4_gcc-4.4.6
intel-mpi/13.0.1                suitesparse/4.2.1_gcc-4.4.6
intel-mpi/14.0.2                superlu/4.3_gcc-4.4.6
intel-vtune/2013u17             superlu-dist/3.3_gcc-4.4.6
intel-vtune/2013u8              vmd/1.9.1_other
irods/3.1_gcc-4.4.6             zlib/1.2.8_gcc-4.4.6

---------------------- /applis/site/meta_modulefiles ----------------------
gnu-devel/4.4.6 intel-devel/13  intel-devel/14
```

chargement des outils intel

[froggy1]$ **module load intel-devel**

CIMENT

Lancement de script OAR

- Eviter de charger les modules dans le .bashrc
  Risque d'erreur par manque de visibilité
- Lancement de job via OAR
  Charger l'environnement dans le script de lancement

- Ex de script bash à soumettre via oar :

$ **cat ./oar_script**

```
#!/bin/bash
#OAR -l nodes=1/core=1,walltime=0:10:00
#OAR -t devel
#OAR --project admin
. /applis/ciment/v2/env.bash
module load intel-devel
cd $OAR_WORKDIR
./my_program
```

$ **chmod u+x ./oar_script**

$ **oarsub -S ./oar_script**

CIMENT

CiGri

- Utilise le temps cpu libre des clusters
- Fonctionne en mode besteffort : un job peut être tué si un job *normal* requiert les ressources ; le job grille est resoumis de façon automatique

- Les # étapes :
  - Déployer l'appli sur les clusters
  - Préparer
    un script de lancement,
    un fichier de paramètre,
    un fichier de description (JDL) au format JSON
  - Démarrer et suivre le job

```
~$ gridsub -f povray.jdl
Campaign successfully submitted
CAMPAIGN_ID=189

~$ gridstat
Campaign id Name        User       Submission time   S  Progress
----------- -------- ----------- ------------------- -- --------
178         test3     bzizou     2013-09-13 12-22-17 Re 2002/2000 (100%)
189         povray    daemon       2013-09-23 09-53-55 R  0/721 (0%)
```

  - Récupérer les résultats

povray.jdl

```
{
  "name": "povray_demo",
  "resources": "core=1",
  "exec_file": "\$HOME/povray/start.bash",
  "exec_directory": "\$HOME/povray",
  "param_file": "/home/deamon/povray_params.txt",
  "test_mode": "true",
  "type": "best-effort",
  "prologue": [
    "set -e",
    "source /applis/ciment/v2/env.bash",
    "module load irods",
    "cd \\$HOME",
    "imkdir -p povray_results/\\$CIGRI_CAMPAIGN_ID",
    "iget -r -f povray"
  ],
  "clusters": {
    "gofree": {
      "walltime": "00:10:00"
    },
    "fontaine": {
      "walltime": "00:10:00"
    },
    "froggy": {
      "project": "test",
      "walltime": "00:5:00"
    },
    "fostino": {
      "walltime": "00:10:00"
    }
  }
}
```

Irods

- Système de stockage distribué
- Stockage objet : 1 fichier = 1 objet
- Ce n'est pas un système de fichiers

module load irods

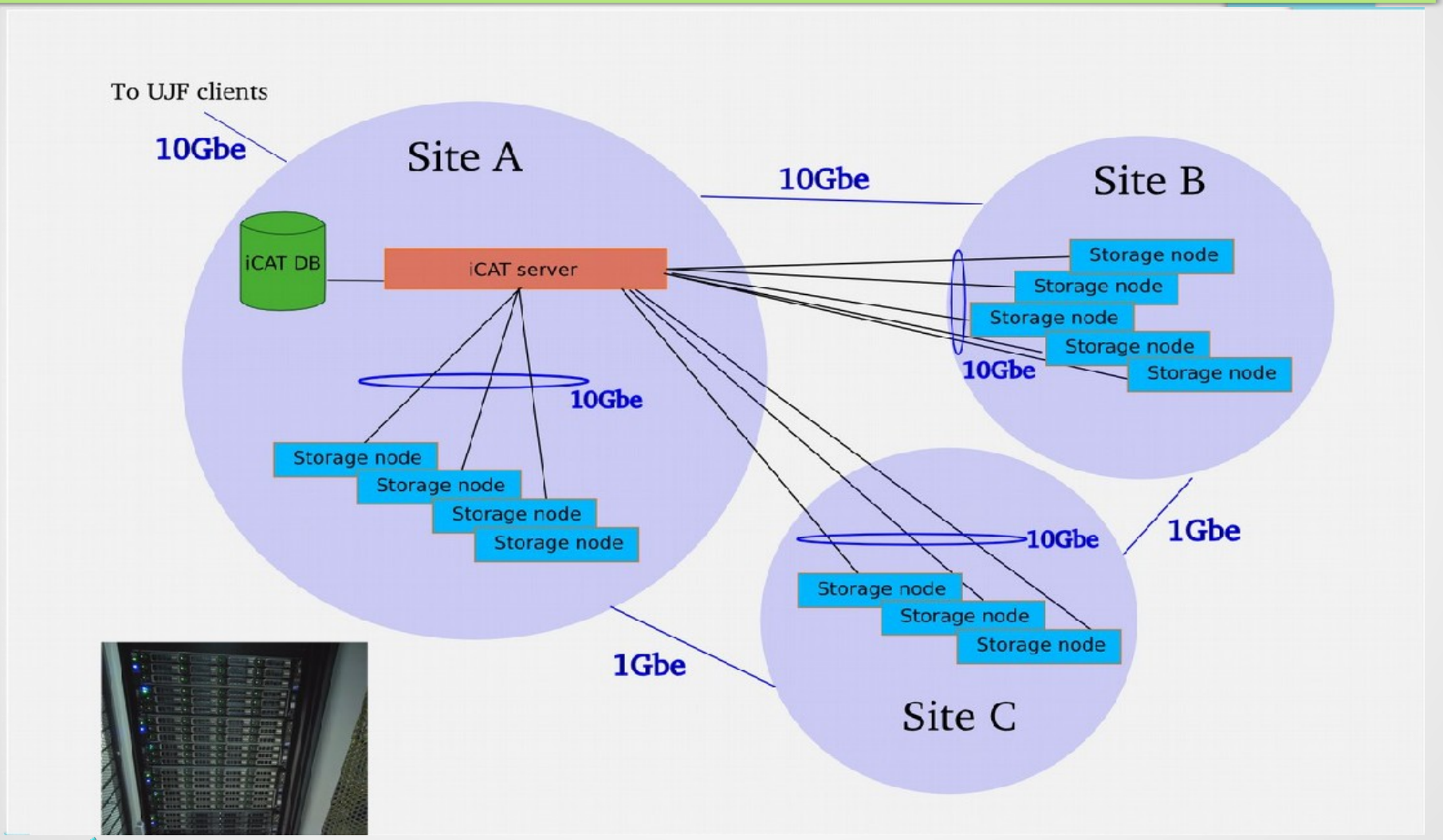
→ accès aux commandes ils, iput, iget, irsync, ...

Pour transférer des fichiers, on fait des iput et des iget

D'autres moyens d'accès : api python, appli web, webdav

CIMENT

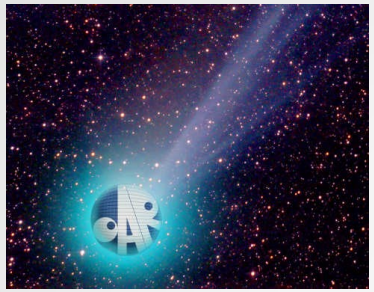
Irods



CIMENT

OAR/Colmet

P. Le Brouster, B. Bzeznik, S. Harrache, O. Richard



Objectif du projet :

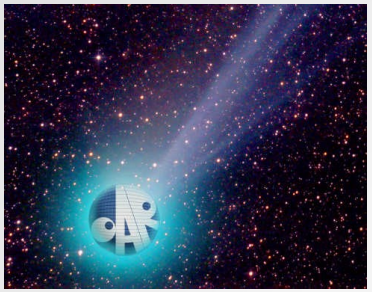
- dégager des profils de jobs par rapport aux ressources matérielles exploitées

- Identifier des jobs n'exploitant pas correctement les ressources

- 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
- L'index est le « job »

CIMENT

OAR/Colmet



OAR/COLMET

Colmet OAR job data extraction

Job ID:	6197366
Begin (seconds):	0
End (seconds):	185375
Resolution (points):	500
Type:	Graph ▼

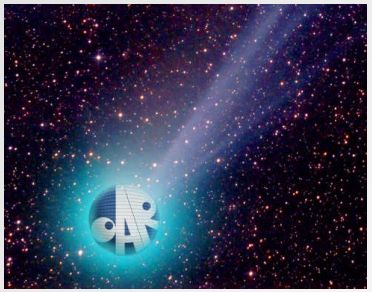
Graph
Extract only

GO

oar.imag.fr
github.com/oar-team/colmet

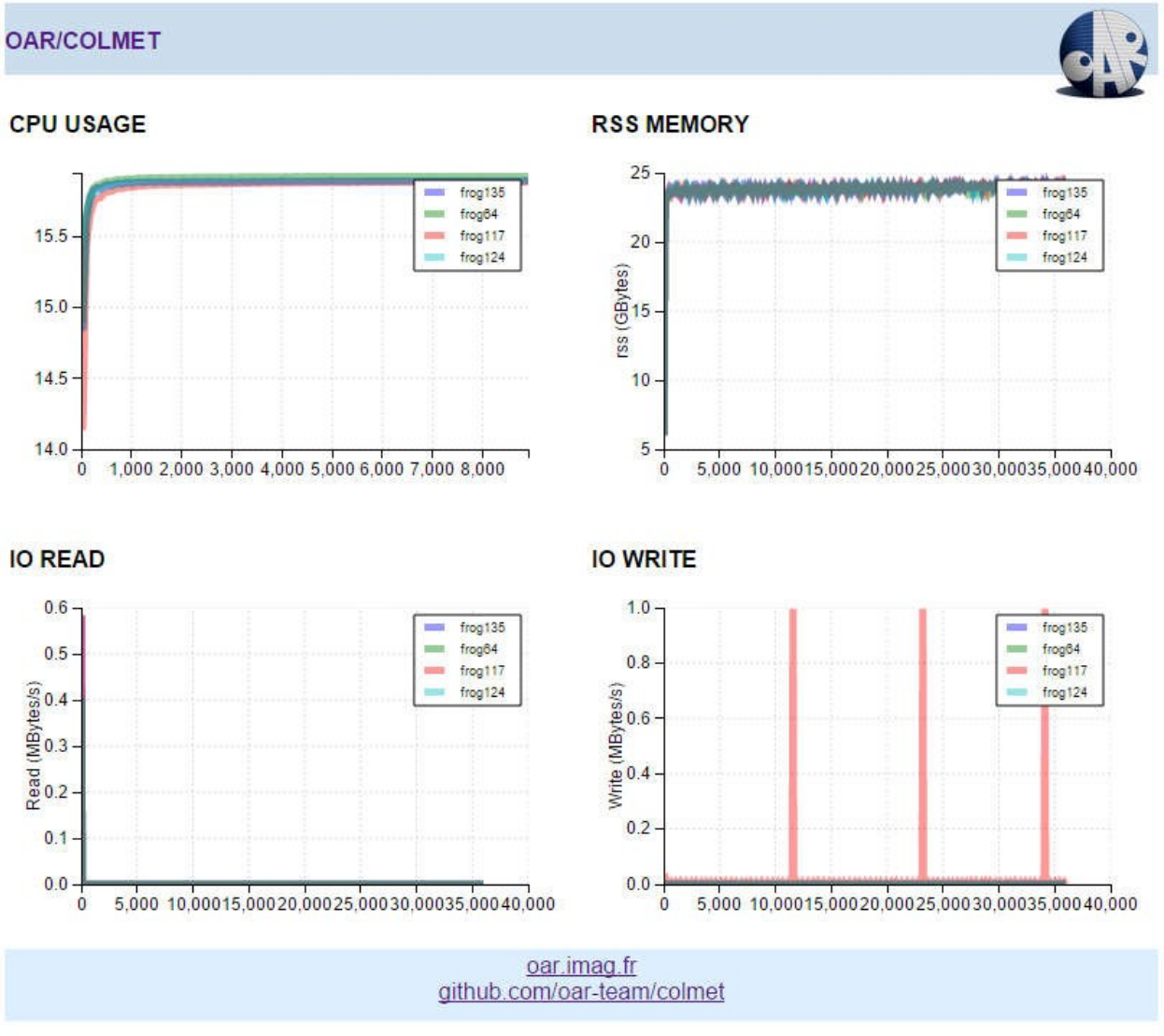
CIMENT

OAR/Colmet

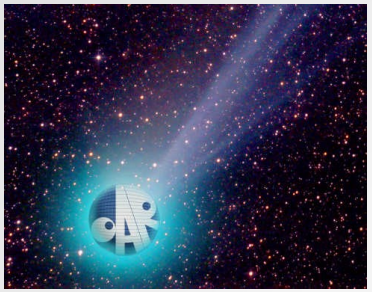


Job “standard”:

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

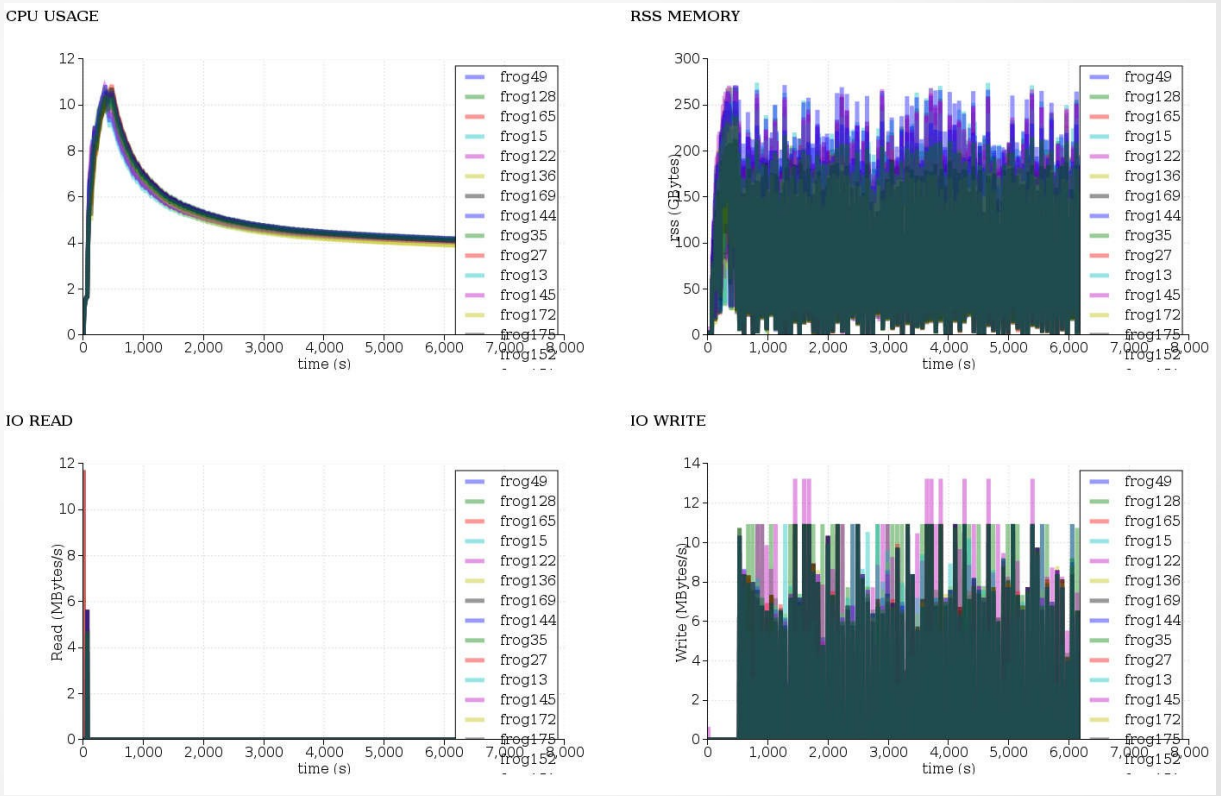


OAR/Colmet



Job “i/o intensif”:

- Charge cpu n'arrive pas à monter
- Emprunte mémoire importante
- Eccritures disques importantes depuis tous les noeuds



Formations

- Formations CED (college doctoral), MaiMoSiNE, CIMENT
- Utilisation des plateformes CIMENT et mise en commun de l'infrastructure de formation du CED avec MaiMoSiNEe (salle CED au LJK)
- Modules autour du calcul scientifique et du HPC

  Environnement de developpement d'application de calcul scientifique (15h)

  Calcul parallele (33h) sous forme de 2 modules

    Introduction au le calcul parallele (architecture, OpenMP,MPI, multicore)

    Vers le massivement parallele (accelerateurs,...).
- Module sur les statistiques de base et le logiciel R (30h)

  ⇒ Modules complementaires a l'etude pour 2015-2016:

    Fouille de donneeś

    Statistiques geospatiales

    Analyse de sensibilité

CIMENT

Merci

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