CENTRE
RÉGIONAL DE
RESSOURCES
INFORMATIQUES

UNIVERSITÉ
Clermont
Auvergne
& associés

Infrastructures du mésocentre

Antoine MAHUL, CRRI

CRRI

Centre Régional de Ressources Informatiques

- Lieu de mutualisation de ressources informatiques
- Point de présence RENATER en Auvergne
- Opérateur du réseau métropolitain CRATERE (17 établissements)
- Gestion du réseau universitaire
- Opérateur du datacenter universitaire
- Appui au SI universitaire
- Appui au mésocentre

Calcul scientifique au CRRI

Mutualisation de ressources de calcul pour la communauté de recherche du site auvergnat

- 1963 : centre de calcul...
- 2005 : grille de calcul Auvergrid (noeud EGI)
- 2006 : PRAI Lifegrid
- 2008 : calculateur SMP (CPER Environnement)
- 2010-2011 : cluster de calcul (CPER Innovapôle)
- 2014 : extension cluster de calcul & stockage (CPER)
- 2014 : structuration mésocentre
- 2015-2020 : projet CPER AUDACE

Mésocentre

- Un mésocentre est :
  - un ensemble de moyens humains
  - de ressources matérielles et logicielles
  - au services de la communautés scientifique
  - fédérateur (EPST, Universités, Industriels) en région
  - doté de sources de financement propres
- Pour fournir un environnement scientifique et technique propice au calcul haute performance
- Faisant l’objet d’évaluations scientifiques régulières
- Pour s'inscrire dans une démarche nationale

Gouvernance du mésocentre

- Une gouvernance pour permettre une utilisation raisonnée des ressources du CRRI :
  - Moyens mutualisés de calcul
  - Services autour du calcul scientifique
  - Stockage de données utiles à la recherche
- Deux comités :
  - Orientation
  - Utilisateurs
- Un seul but : permettre un accès simple à des moyens important

Ressources du mésocentre

- Ressources de calcul pour la Grille EGI
  - En production sur la grille depuis 2006
  - 192 coeurs de calcul, 48 To de stockage
- Cluster de calcul (HPC1)
  - En production depuis 2009
  - 120 coeurs de calcul
- Cluster de calcul (HPC2)
  - En production depuis 2015
  - > 700 coeurs de calcul, 10 To de stockage
- Stockage distribué en mode Objet
  - Production printemps 2016
  - 520 To bruts
- Architecture SMP (multicoeurs, grande quantité de RAM)
- Cloud IaaS

Cluster de calcul HPC2

- Seconde génération du cluster de calcul local
- Mise en production au **printemps 2015**
- Nœuds de calcul :

Noeuds	Qtté	CPU	RAM	HT	Cœurs	Inter	Financement
hpcnode[01-04]	4	2 x Sandy Bridge	64 Go	Non	16	GbE	CNRS / LMGE
hpcnode[05-08]	4	2 x Ivy Bridge	96 Go	Oui	32	GbE	CPER 2011
hpcnode[09-24]	16	2 x Ivy Bridge	128 Go	Oui	32	GbE	CPER 2014
hpcnode[25-28]	4	2 x Ivy Bridge	256 Go	Oui	32	GbE	CPER 2014
hpcnode[29-36]	8	2 x Ivy Bridge	128 Go	Oui	32	QDR	CPER 2014

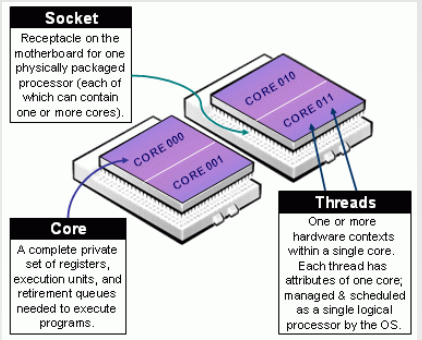
- Installation (OS/Applications) homogène sur tous les nœuds

HPC2 : utilisations

- Utilisateurs
  - 54 utilisateurs enregistrés, 14 laboratoires
  - 32 utilisateurs actifs (au moins 1 job)
- Jobs (depuis janvier 2015)
  - ~ 540 000 jobs soumis
  - 2,925 millions d'heures CPU allouées
- Applications préinstallées
  - 50 modules d'environnements
  - GCC, ICC, OpenMPI, Python, Numpy, Scipy, R, Bioconductor, Perl, Bioperl...

HPC2 : accès

- Accès par ssh depuis le campus
- Gestionnaire de job : SLURM
  - Allocation de ressources (CPU, RAM, GPU)
  - Gestion de l'architecture des processeurs
  - Confinement et placement des processus
  - Gestion de l'énergie (pas encore en place)
- Gestion des priorités
  - Fair Queueing : plus on utilise, moins on est prioritaire
  - De nombreux critères possibles (projets, QoS, par taille...)



HPC2 : jobs

- Exécution asynchrone des calculs
  - Permet le partage, la réservation des ressources et l'ordonnancement
  - Uniquement des codes de calcul sous linux en ligne de commande (pas d'interface graphique)
- Séquence typique:
  - L'utilisateur se connecte par ssh au frontal
  - L'utilisateur décrit son job dans un script shell (CPU, RAM, durée, ligne de commande)
  - L'utilisateur soumet son job (sbatch)
  - Le Job manager planifie l'exécution, met le job en attente
  - Le job est exécuté quand les ressources sont disponibles
  - L'utilisateur peut récupérer ses résultats à la fin de l'exécution

HPC2 : applications

- Applications utilisateurs
  - Déployer ses applications dans /home
  - Quota actuel : 100 Go
- Tool chain disponibles:
  - GCC + OpenMPI (C / C++ / F90)
  - Intel + OpenMPI (C / C++)
- Applications partagées
  - Déployées par les administrateurs
  - Pour éviter la duplication des codes
  - Gérées via modules d'environnements

HPC2 : modules d'environnements

- Solution de gestion des variables d'environnement
  - Classique dans les infras HPC (mésocentre, centres nationaux)
  - Cloisonnement des applications
  - Permet de référencer les applications installées
  - Permet de gérer différentes versions
  - Permet de gérer les modules de façon hiérarchique

```
[anmahul@hpc2 ~]$ module load gcc/4.8.4 openmpi
[anmahul@hpc2 ~]$ module list
Currently Loaded Modules:
  1) crri   2) binutils/2.25   3) gcc/4.8.4   4) openmpi/1.8.4

[anmahul@hpc2 ~]$ module avail
------------------- /etc/modulefiles/compiler/gcc/4.8.4 -------------
   openblas/0.2.14    openmpi/1.8.4

------------------- /etc/modulefiles/apps ---------------------------
   bioperl/1.6.924    blast-legacy/2.2.26    ncbi-blast/2.2.30+
```

Stockage objet CEPH

- CPER 2014 :

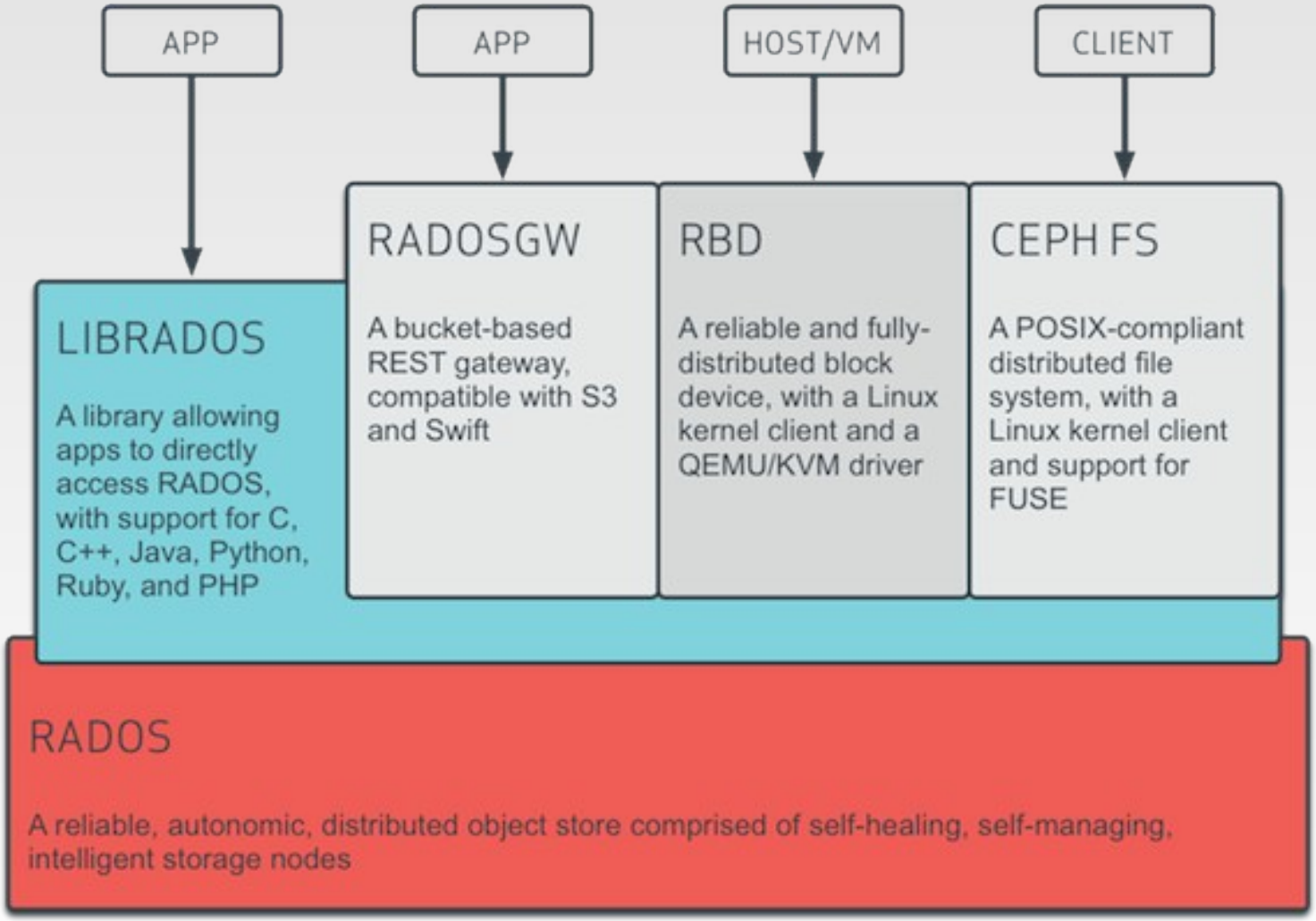
  “acquisition d’outils mutualisés pour le stockage, l’analyse de grandes masses de données, la modélisation et le calcul intensif”

- Projet de stockage capacitif
  - Stockage de grands volumes de données (>100 To)
  - Allocation d'espace de stockage à la demande
  - Extensibilité à moindre coût
  - Maîtrise des frais de fonctionnements
- Première brique d'un cloud IaaS

CEPH

- Plateforme de stockage distribué
  - Stockage en mode objet
  - Architecture *Scale-out* adaptée au passage à l'échelle
  - S'appuie sur du matériel ordinaire (*commodity hardware*)
  - Sans SPOF
- Logiciel Libre
  - Technologie issue de l'université de Californie (Santa Cruz)
  - Thèse de Sage A. Weil (2007)
  - Développement et support commercial assuré par Inktank
  - Inktank racheté par Redhat le 30 avril 2014

CEPH : principes



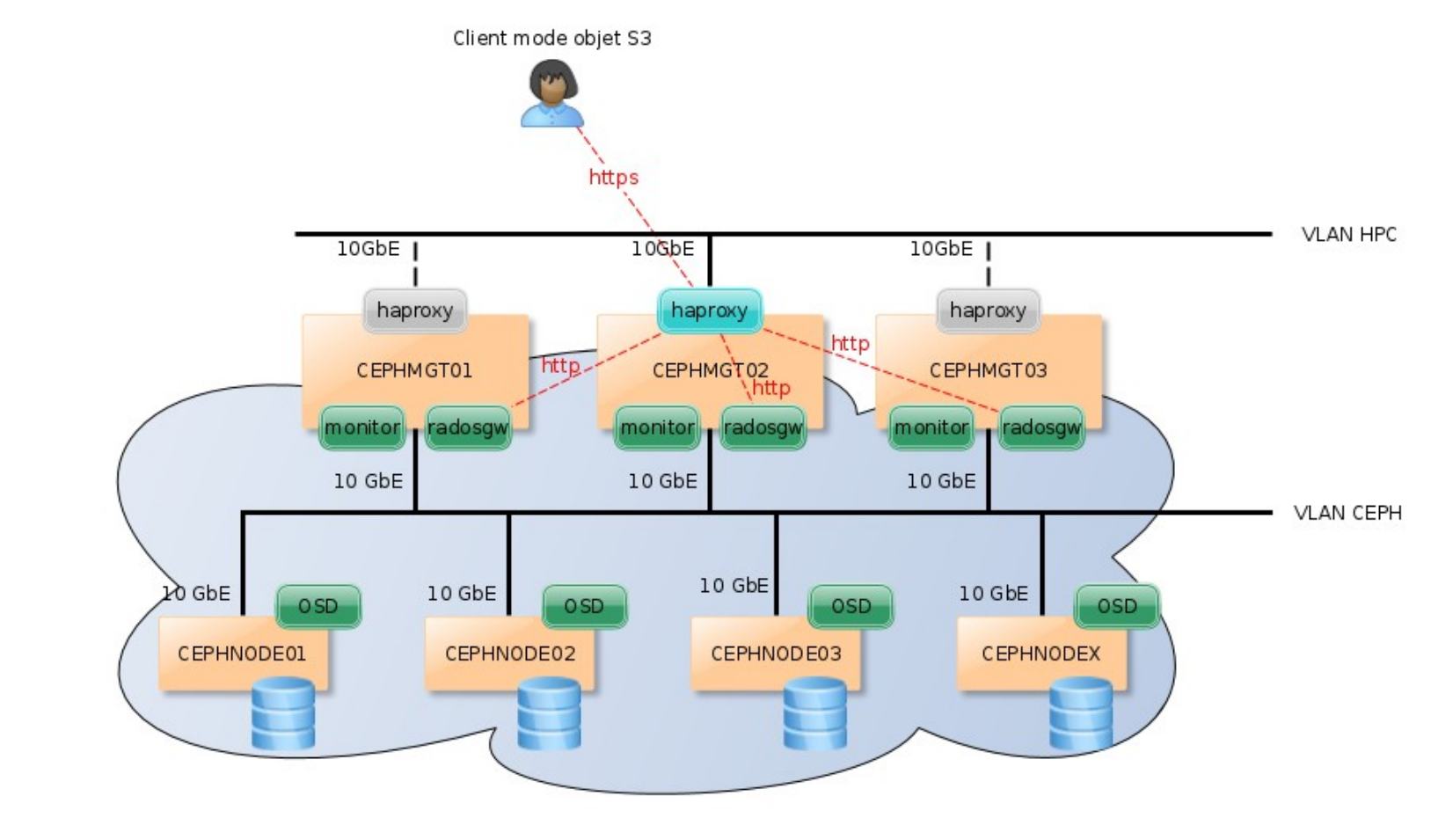
CEPH : fonctionalités

- Cluster de stockage RADOS
  - Réplication des données chaudes (surcout : x2, x3...)
  - Correction d'erreur pour les données froides (x1.5)
  - Equilibrage automatique
  - Algo de placement (CRUSH) assez évolué et modulable
- Client Ceph : interface vers les utilsateurs
  - RADOSGW : API compatible Swift et S3
  - RBD : mode bloc (intégré au kernel Linux)
  - CephFS : mode fichier (stable depuis peu)

CEPH : infrastructure

- Infrastructure (CPER 2014)
  - 13 nœuds de 40 To (6 actuellement en production)
  - 3 nœuds « contrôleurs »
  - Interconnection 10GbE
- API S3
  - RadosGW déployé sur les 3 contrôleurs
  - Load balancing (avec haproxy)
  - Haute disponibilité (avec keepalived / VRRP)

CEPH : architecture



Stockage en mode objet (API S3)

- URL : https://s3.calcul.crri.fr
- Stockage grands volumes de données non structurés
- Mode objet :
  - Objet = données + metadonnées + identifiant unique
  - organisation plate de type clé/valeur
- Bucket
  - Ensemble d'objets (conteneur)
  - 1 propriétaire
  - Gestion des ACL
- Clients : CrossFTP, CyberDuck, S3CMD, Owncloud, GALAXY...

Clients s3cmd

- Outil en ligne de commande disponible sur le cluster HPC2
- Un fichier de config :

```
[default]
access_key = XXXXX
secret_key = YYYYY
host_base = s3.calcul.crri.fr
host_bucket = %(bucket)s.s3.calcul.crri.fr
server_side_encryption = False
use_https = True
signature_v2 = True
```

- Quelques opérations simples

`s3cmd ls`	Lister les buckets
`s3cmd mb monbucket`	Créer un bucket
`s3cmd ls s3://monbucket`	Lister les objets d'un bucket
`s3cmd put toto.txt s3://monbucket/toto.txt`	Transférer un fichier local dans un bucket
`s3cmd get s3://monbucket/toto.txt`	Récupérer un objet depuis un bucket
`s3cmd del s3://monbucket/toto.txt`	Supprimer un objet
`s3cmd sync dir s3://monbucket`	Synchroniser un répertoire local dans un bucket

Modalités d'accès

- Actuellement : comptes créés à la demande
  - Via le système de tickets: supportcrri@clermont-universite.fr
- Accès distant
  - Passerelle ssh avec une double authentification (One time Password)
- Portail du CRRI
  - https://portail.crri.clermont-universite.fr (uniquement depuis réseau CRATERE)
  - Accès aux suivi des tickets, au monitoring, au wiki
  - Dashboard mésocentre en cours de développement