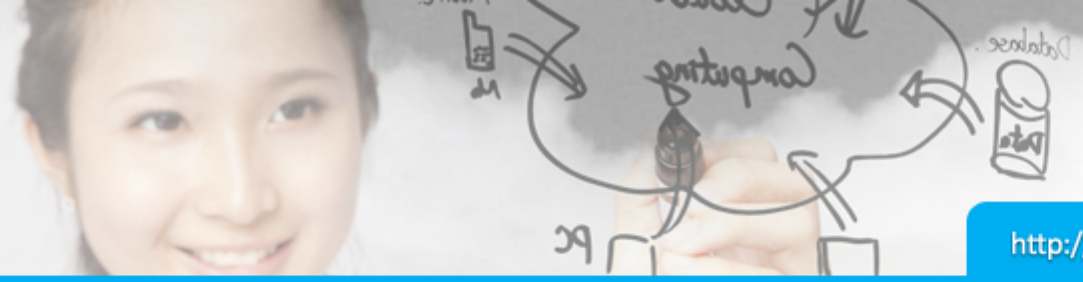
Systèmes de stockage dans les infrastructures de calcul distribuées : quelles opportunités pour le HPC ?

Jérôme Pansanel <jerome.pansanel@iphc.cnrs.fr>

ANF UST4HPC - Fréjus – 17 mai 2018

Hommage

Christophe Caron



À propos

Au sommaire

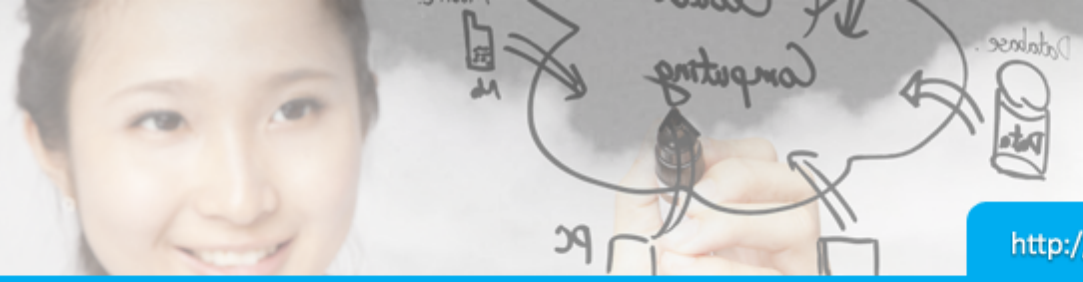
- Les infrastructures de calcul distribuées
- Les systèmes de stockage historiques
- Évolution des solutions
- Questions / discussions

Cette présentation n'est pas

- Une présentation de France Grilles
- Un panorama complet des solutions de stockage
- Un guide pour choisir sa solution de stockage

Qui suis-je ?

- Responsable de la plateforme SCIGNE de l'IPHC
- Directeur technique de France Grilles



Infrastructures de calcul distribuées

Contexte européen

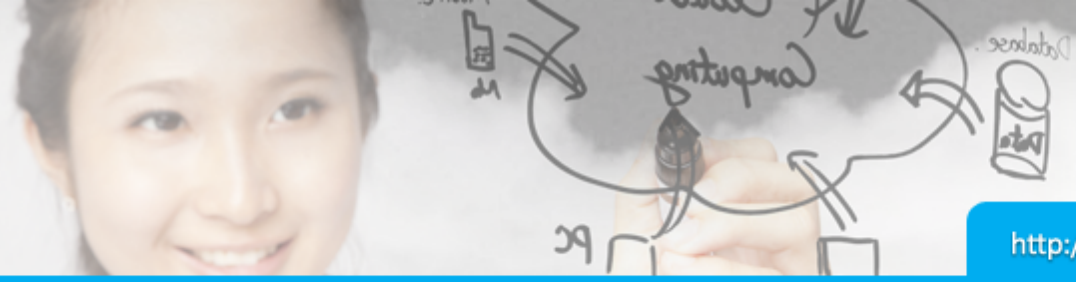
European Open Science Cloud

- Projet ambitieux supporté par le programme d'investissement H2020
- L'utilisation du Big Data en science, dans l'industrie et les services publics
- Accès pour les chercheurs, les PMEs, l'industrie et l'administration publique à des infrastructures (stockage, calcul, services, …) ayant une visibilité mondiale
- Sécuriser le stockage et l'analyse des données
- Un réseau rapide et disponible
- Développement et déploiement à grande échelle d'infrastructures numérique (HPC, Cloud de données) et réseaux à haut-débit européens

Élargir l'accès aux services et construire la confiance

- PMEs, administration publique, standards



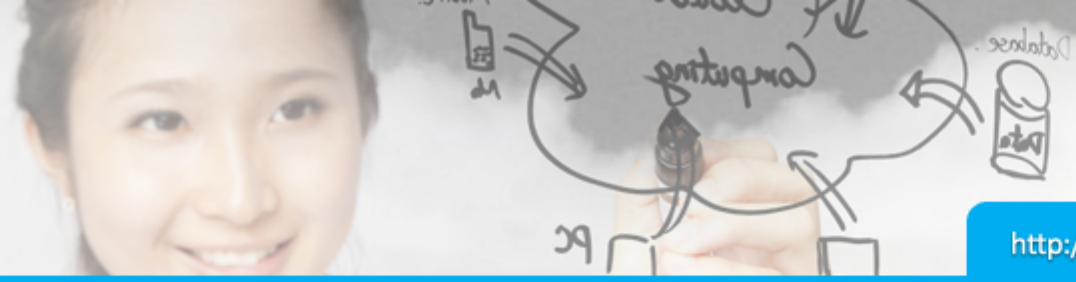
Infrastructures de calcul distribuées

Au niveau européen

- EGI – Infrastructure pour le calcul HTC (grille, Cloud computing) et le stockage géographiquement distribué (OneData, DPM, iRODS, ...)
  https://www.egi.eu
- PRACE – Infrastructure pour le calcul HPC (Tier 0 – projets en millions d'heures de calcul)
  http://www.prace-ri.eu/

Le terme infrastructure de calcul distribuée désigne une association collaborative

- De technologies numériques (matériel, logiciel)
- De ressources (données, services, ...)
- De systèmes de communication (protocoles, droits d'accès et réseaux)
- De structures organisationnelles (France Grilles pour EGI)
- Et des hommes !



Historique EGI

Au début le LHC …

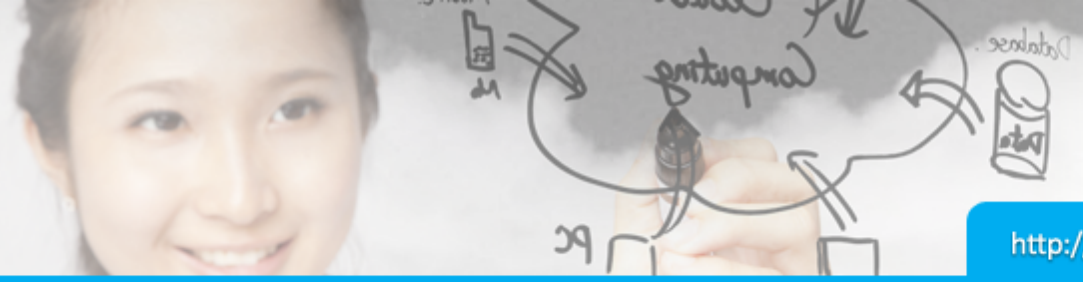
- En 1999, pour répondre aux besoins du LHC, les bases de la grille de calcul sont posées
- Début du premier projet européen sur les grilles de calcul en 2001 : DataGrid
- En 2004, le projet EGEE (Enabling Grid for E-sciencE) ouvre la grille de calcul à d'autres disciplines (géosciences, bioinformatique)

2010, une date clé

- Création de la fondation EGI.eu
- Création du GIS France Grilles
- Les composants logiciels sont (enfin) stables
- Ouvertures à toutes les communautés scientifiques
- Participation à plusieurs projets européens (community-driven)

Nouvel horizon, nouveaux challenges

- H2020 et l'European Open Science Cloud
- Fédération avec EUDAT → EOSC-Hub

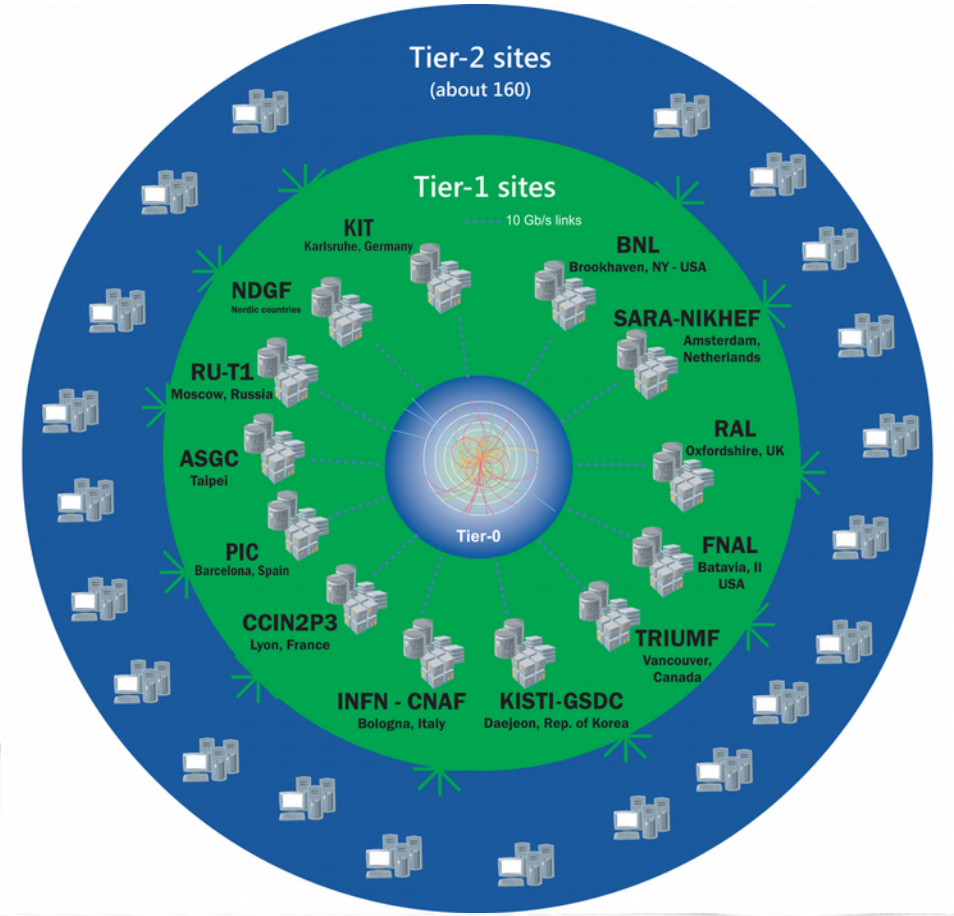


Systèmes de stockage *historiques*

Organisation des données

WLCG

- Organisation par Tier :
  - Tier-0 (CERN, 20 % des ressources)
  - 14 Tier-1 (~ 1/10 des données, 40 % ressources)
  - ~ 160 Tier-2 (40 % des ressources)
  - Tier-3 (pas de pledge, ni d'engagement formel)
- Pas de stockage sur bande dans les Tier-2 (et les Tier-3)
- 670k CPUs, 440 Po de stockage disque, 390 Po de stockage bande
- ~ 4 milliards d'heures de calcul / an
- Gestion des données par expérience
- Authentification unifiée
- Basé sur File Transfer Service (FTS3)



Authentification / Autorisation

Authentification

- Basée sur l'utilisation des certificats X509
- Pilotage des politiques de gestion des certificats et de sécurité par l'International Grid Trust Federation (IGTF)
- En France, les certificats sont délivrés par l'AC GRID2-FR, gérée par RENATER
- Certificats pour les personnes, les serveurs et les services

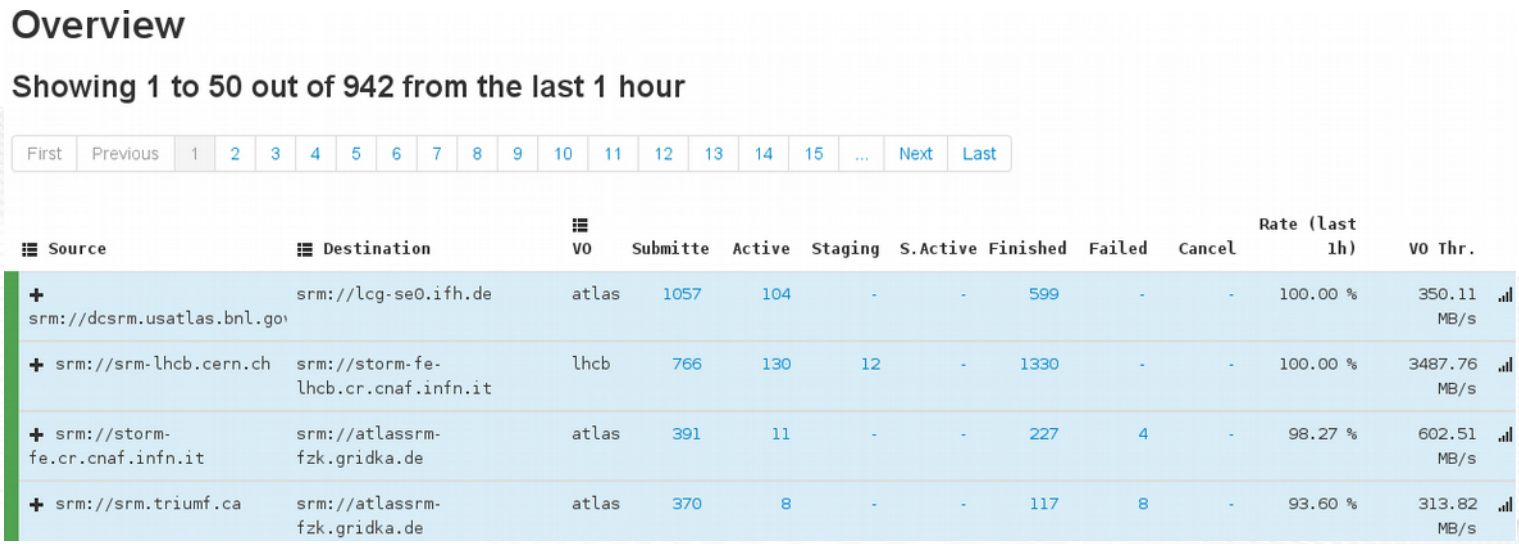
Autorisation

- Basée sur l'utilisation de groupes et de rôles au sein de communautés d'utilisateurs (VO)
- Mécanique LCAS (Local Centre Authorization Service) : DN, FQANs
- Utilisation de serveurs VOMS (ou analogue)
- Ex. : /cms/Role=priorityuser

File Transfer Service (FTS)

Un outil multi-protocoles

- Utilise la librairie GFAL2 (Grid File Access Library)
- Organisation automatique en fonction de la disponibilité des sites et de la charge du réseau
- GridFTP, XRootD, SRM or HTTP
- Accessible via CLI et API REST
- Outil de monitoring Web



GFAL2

Le couteau suisse pour accéder au données sur les stockages distribués

- Librairie en C avec une API POSIX
- Basé sur un système de plugins
- LFC, RFIO, **SRM**, **GridFTP**, *HTTP* (Davix), XROOTD, local, ...
- Plusieurs modules / projets :
  - GFAL2 utility tool (CLI)
  - GFAL2 PYTHON (python bindings)
  - gfalFS : module FUSE
- Licence Apache 2

→ https://dmc.web.cern.ch/projects/gfal-2

Plusieurs solutions

- DPM
- dCache
- StoRM
- XRootD
- EOS

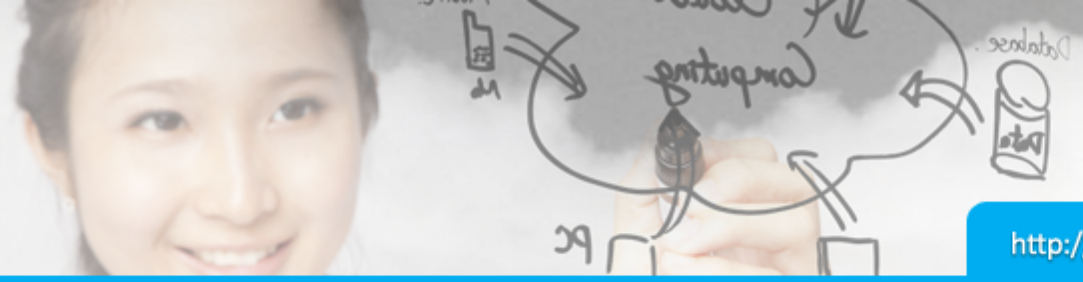
DPM – Disk Pool Manager

Une solution Scale Out

- Disk Pool Manager (DPM) est une solution simple pour créer un service de stockage grille
- Produit stable et mature (développé depuis les années 2000)
- Composé d'un serveur de tête et de plusieurs serveurs disques
- Un seul espace de noms
- Nombreux protocoles supportés (HTTP, WebDAV, XRootD, SRM, gridFTP, RFIO)
- Facile à installer et à maintenir
- Utilisation de Memcached en cache de la bases de données
- Disponible dans les dépôts EPEL (CentOS 6 et 7) et Ubuntu
- Licence Apache 2.0

Clients

- CLI et streaming (GFAL2)
- XrootD, RFIO
- Webdav / CURL



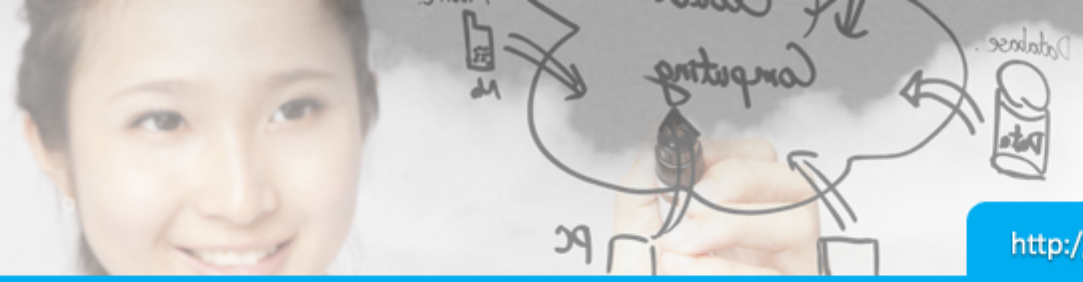
Conclusion

Une solution fonctionnelle

- Passage à l'échelle
- Maintenance du système
- Gestion aisée de plusieurs Po de données
- Possibilité de fédération entre sites

Mais limitée

- Nœud de tête : SpoF
- Limitation des I/O parallèles à cause de l'authentification
- Pas de support de protocoles d'authentification tel que OAuth 2.0 / OIDC
- Problème de performance sur les petits fichiers
- Évolution du mode de lectures des fichiers ROOT
- Difficile à intégrer dans les nouvelles infrastructures de Cloud
- Évolution du cadre des appels à projets européens (DMP)



Évolution des solutions

iRODS : présentation

Un *middleware* pour la gestion des données

- Open Source (licence BSD)
- Supportant plusieurs milliers d'utilisateurs et de groupes
- Permettant d'accéder, de gérer et de partager des données stockées sur différents types de stockage
- Facilitant l'accès à des ressources hétérogènes (Unix, S3, DDN, HPSS, …), à travers un seul espace de nom (zone)
- Permettant de contrôler finement les données grâce à un moteur de règles et un ensemble de micro-services (réplication, vérification des types, …)
- Disposant d'un support officiel
- Accessible via CLI, API et portail Web
- Authentification LDAP, GSI, password / pam
- Possibilité de faire du tiering

iRODS

iRODS : présentation

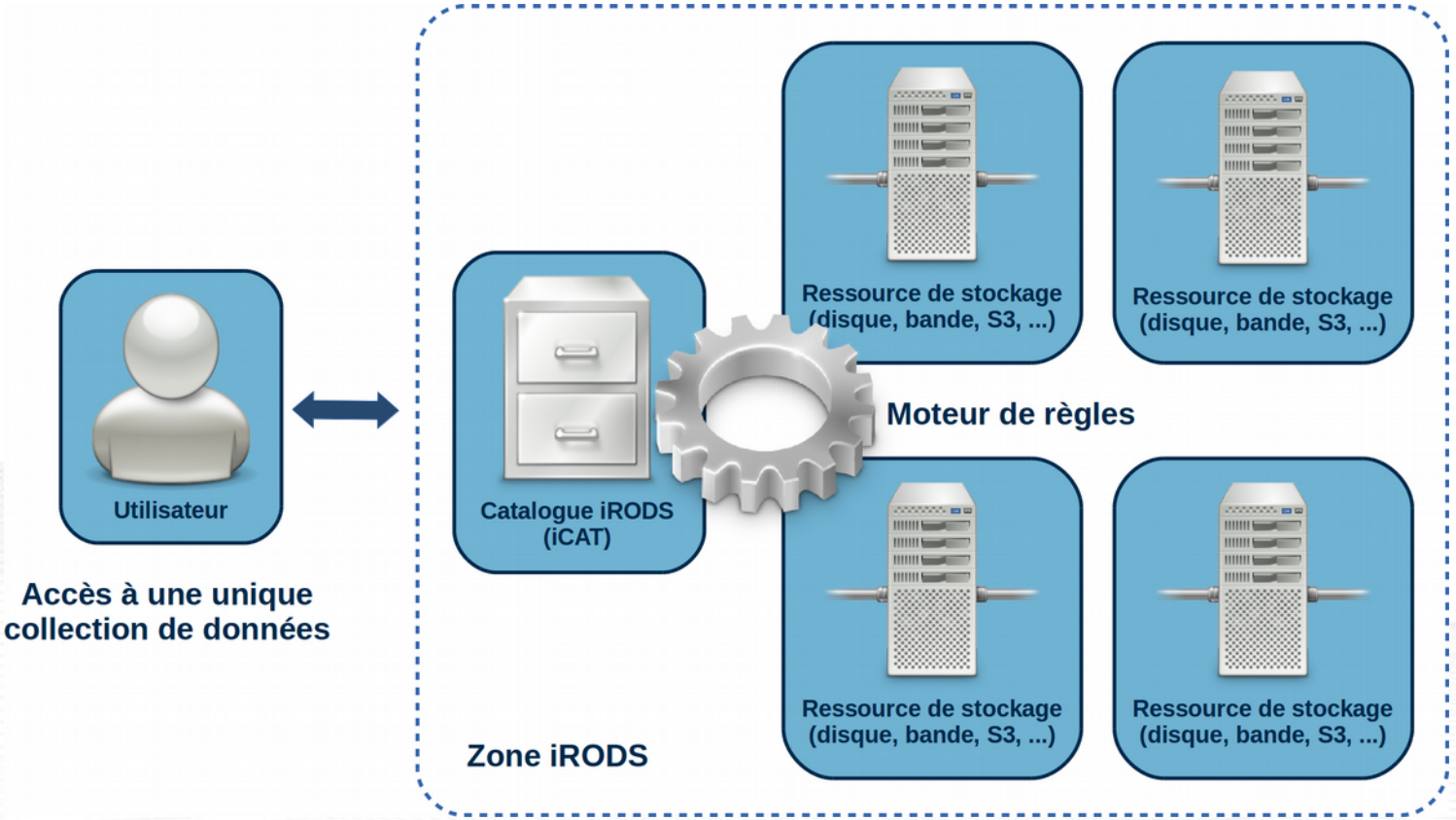
Une solution puissante

- La virtualisation de l'accès au stockage
- La gestion de plusieurs péta-octets de données
- Le transfert parallèle pour les données volumineuses
- La recherche des données (méta-données)
- L'automatisation des processus grâce aux règles et aux micro-services
- La sécurisation des données grâce à la réplication et la gestion des accès
- L'accessibilité des données sur différents types de terminaux

Et légère

- Les paquets iRODS occupent moins de 100 Mo (mais les dépendances ont besoin de plus d'espace) et disponibles pour différents OS en v4
- Serveur iCAT avec minimum 2 Go de RAM
- La taille de la base de données dimensionne le choix du serveur
- Base de données clusterisée en environnement de production

iRODS : architecture



iRODS : utilisation

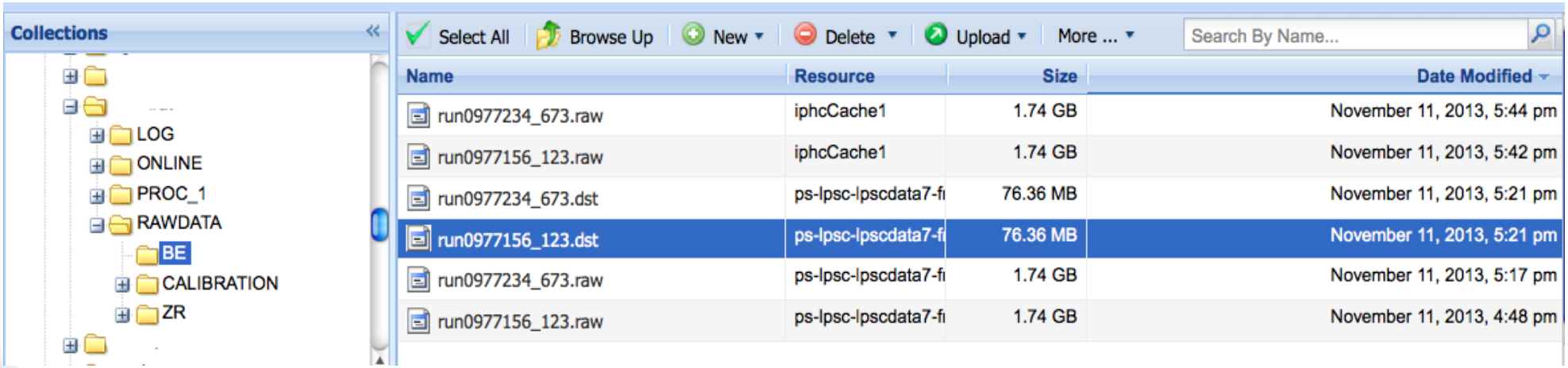
Des commandes de type Unix

```
ipasswd
ils [-l] [-L] [-A]
imkdir
icd
ipwd
imv
icp
irm [-a]
...
```

Et de type FTP

```
iput
iget
```

iRODS : utilisation



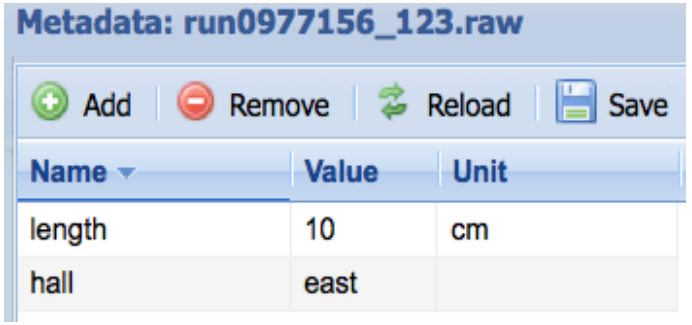

```
[user ~]$ ils
/frgrid/home/UNECOLLAB/RAWDATA:
 C- /frgrid/home/UNECOLLAB/RAWDATA/CALIBRATION
 C- /frgrid/home/UNECOLLAB/RAWDATA/BE
 C- /frgrid/home/UNECOLLAB/RAWDATA/ZR
[user ~]$ ils -l BE/
/frgrid/home/UNECOLLAB/RAWDATA/BE:
  owner    0 ps-lpsc-lpscdata7-fr   80072192 2013-11-11.16:21 & run0977156_123.dst
  owner    0 ps-lpsc-lpscdata7-fr 1748189011 2013-11-11.15:48 & run0977156_123.raw
  owner    1 iphcCache1           1748189011 2013-11-11.16:42 & run0977156_123.raw
  owner    0 ps-lpsc-lpscdata7-fr   80072192 2013-11-11.16:21 & run0977234_673.dst
...
```

iRODS : métadonnées

Les métadonnées

- Triplet (nom, valeur, unité)

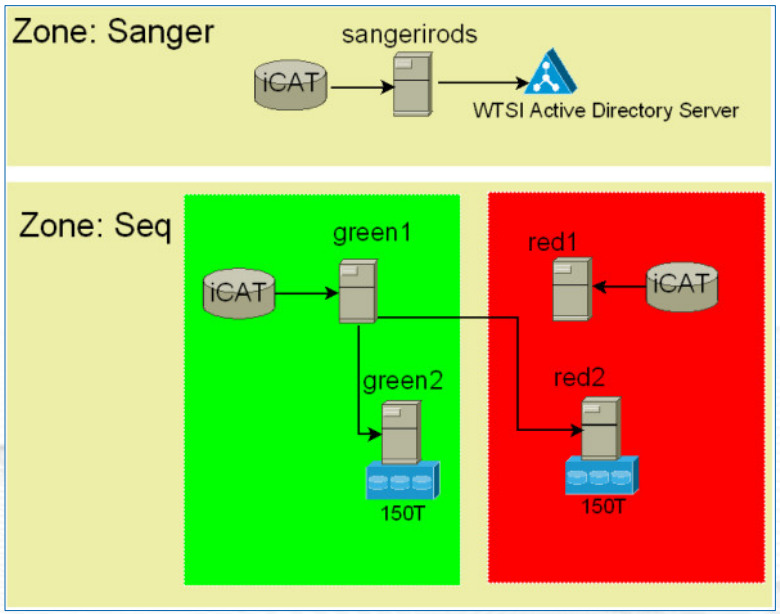
```
[user ~]$ imeta add -d run0977156_123.raw length 10 cm
[user ~]$ imeta add -d run0977156_123.raw hall east
[user ~]$ imeta ls -d run0977156_123.raw
AVUs defined for dataObj run0977156_123.raw:
attribute: length
value: 10
units: cm
----
attribute: hall
value: east
units:
[user ~]$ imeta -d qu hall east
collection: /frgrid/home/UNECOLLAB/RAWDATA/ZR
dataObj: run0977156_123.raw
----
collection: /frgrid/home/UNECOLLAB/RAWDATA/ZR
dataObj: run0817773_556.raw
```



iRODS : cas d'usage du WTSI

Utilisation pour les données de NGS (Next Generation Sequencing)

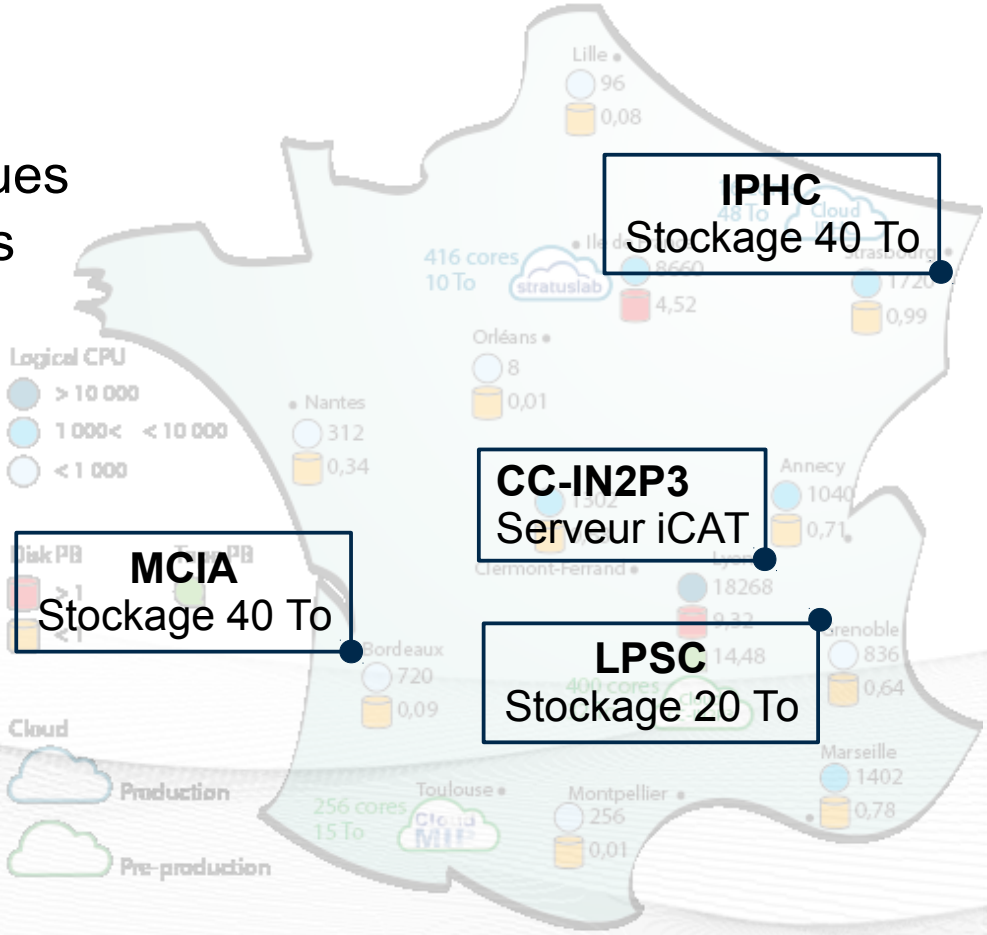
- Plusieurs GB / génome
- Besoin de stockage sur une longue durée (re-exploitation des données)
- Données médicales (cancer, maladie rare, pathogène) → confidentialité
- Volume de 2 PB → besoin d'évolutivité dans le temps
- Disponibilité / réplication
- Extraction automatique de métadonnées



FG-iRODS

Une instance de production

- Fédération de ressources humaines et matérielles
- Ouverte à tous les domaines scientifiques
- Taillée pour les petits et moyens projets
- Service de niveau production
- Support utilisateur et formation
- Investissement financier minimal
- Participation à PICO2



iRODS et HPC

Une intégration en pleine évolution

- Des acteurs du stockage HPC dans le consortium iRODS (OpenSFS, DDN, OCF, DELL/EMC, Quantum, NetAPP)
- Des développements autour de Lustre (prototype) : https://github.com/irods-contrib/irods_tools_lustre
- Des plugins pour accéder aux ressources HPC avec des clients natifs (par ex.: DDN)

CEPH

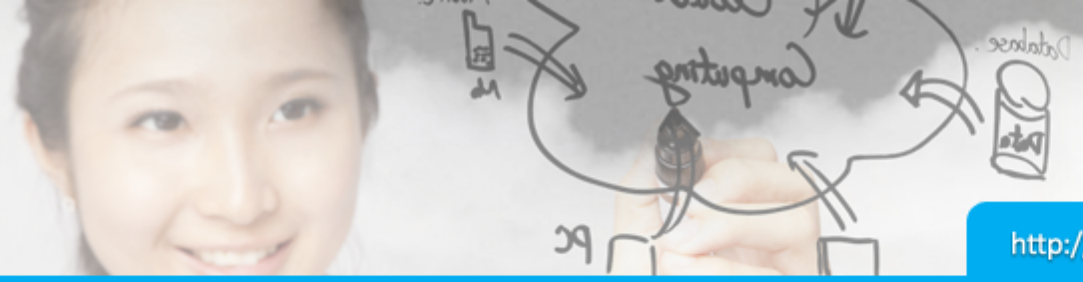
Solution de stockage distribué

- Pas de SPoF
- Passage à l'échelle jusqu'à l'exaoctet
- Auto-réparation
- Scale-out
- Logiciel libre (licence LGPL)
- Protection des données (réplication, erasure-coding)

Système de fichiers unifiant

- Objet : API native
- Bloc : supporté par le noyau linux (krbd) et KVM (librdb)
- Fichier : CephFS

ceph



CEPH : la convergence ?

CEPH et HPC

- Nécessite CephFS pour la gestion des accès concurrents
- CEPH n'est pas adapté pour du matériel haut de gamme
- Livre blanc RedHat CEPH avec disques SAMSUNG NVMe (~700 k IOPS en séquentiel et 90 k IOPS en aléatoire)

CEPH et iRODS

- iRODS peut accéder à CEPH à travers une passerelle RADOS (API S3)
- Déploiement en cours à l'IPHC
- Une solution également pour DPM ?

OpenIO

OpenIO en quelques mots

- Solution de stockage objet
- Léger
- Résilient
- Scale-out
- Tiering et erasure coding
- Intègre une *conscience* pour la gestion du système
- API S3
- Test en cours entre l'IPHC et l'IHES / école polytechnique

→ présentation de Guillaume Delaporte

Annonces

Workshop Opérations France Grilles

- 27 au 29 juin à Montpellier
- Interopérabilité Cloud / HPC / HTC
- iRODS
- Cloud

Formation CEPH

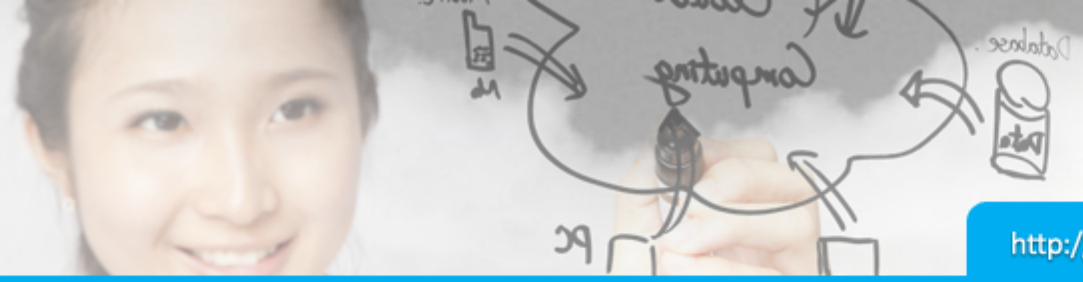
- Deuxième semestre 2018
- Programme en cours – n'hésitez pas à participer !

Formation Administrateur iRODS

- Première édition en décembre 2017
- Refaite (en mieux) si suffisamment de demande

Journées JCAD

- Fusion des journées SUCCES et journées EQUIP@MESO
- Du 24 au 26 octobre, ENS de Lyon



Question / Discussion