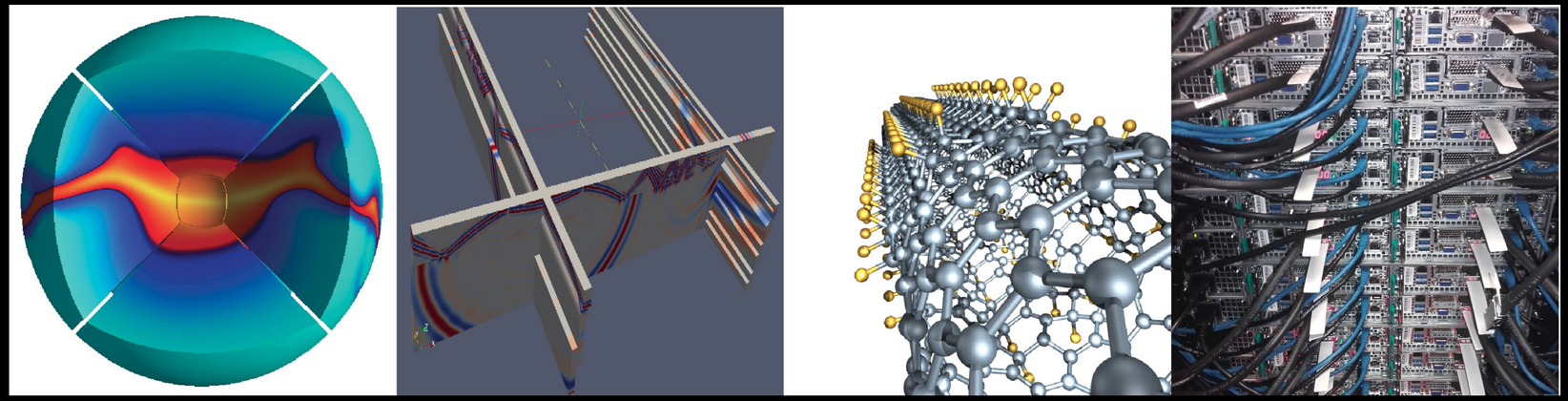
GLiCID



Séminaire SFR : Gestion des données de la Recherche

Le 23 Janvier 2024 – Audrey Bihouée

université angers
CENTRALE NANTES
Nantes Université
Le Mans Université
GLiCID

Ordre du jour

- Généralités sur le calcul scientifique
- Le projet GLiCID
- Le stockage sur une infrastructure de calcul
- En pratique
- Modèle économique

Quelques banalités sur le calcul scientifique

Le calcul scientifique concerne :

- Historiquement, principalement les domaines comme la Physique, la Mécanique, les Géosciences, l'Astrophysique, la Chimie…
- Depuis une dizaine d'années, on constate de nouveaux acteurs comme la Biologie, l'Économie, les Sciences Sociales, les Langues… toutes les communautés scientifiques sont concernées

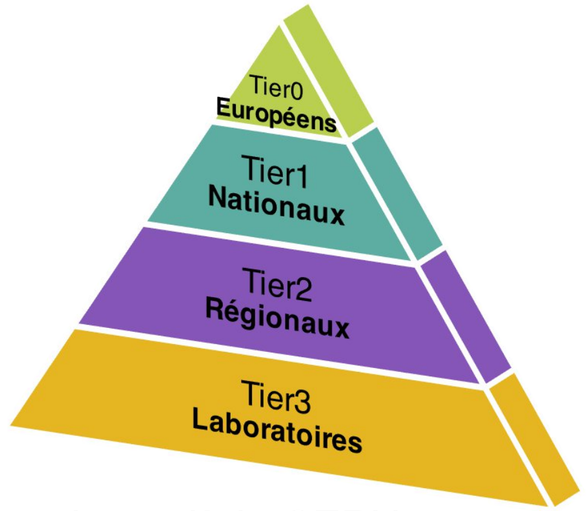
Les besoins explosent :

- en calcul distribué classique (CPU, cœurs)
- en calcul avec accélérateurs (GPU)
- en stockage de données associées

→ Il y a eu beaucoup de changements dans le domaine du HPC, en particulier sur le public concerné

Le calcul scientifique en Europe

- Organisé en Tier
- Quelques exemples (en 2019)
  - Tier0 : SuperMUC-NG (Allemagne) ~ 300 000 cœurs
  - Tier1 : IDRIS (Jean Zay) ~85 000 cœurs ; ~ 3000 GPU
  - Tier2 : GLiCID
    - 16300 cœurs, 100To de RAM, 80 GPU, 1.5Po de stockage rapide et 4Po de stockage tiède CEPH



- Missions des Tier2 : développement, petite production, formation, mutualisation, proximité, flexibilité

Labellisation et DACAS

- **Origine** : volonté du MESRI depuis 2016 de **labelliser un** datacenter par Région pour les équipements ESR
- **Objectifs** : rationalisation (fermer 90 % des salles informatiques ESR)
- **Contrainte** : l'ensemble des projets d'équipements financés ou co-financés par des fonds publiques devront être hébergés dans ces datacenters mutualisés (issus de tout projet - PIA, ANR, FUI, Horizon, …)
- **Réponse** pour les Pays de La Loire : le projet **DACAS**
- Le DataCenter UN **et** le projet DACAS ont été labellisés INFRANUM le 10/11/2020

CPER : Data center, Réseau, Calcul

Projet soumis au CPER 2021-2027

Porteurs : Yann Capdeville, Nicolas Wendling / Stéphane Amiard et Thierry Oger

3 volets :

- Datacenter (DACAS, budget 10M€)
- Calcul scientifique (GLiCID 6M€)
- Réseau (RRTHD, 4M€)

Partenaires:

- Datacenter et Réseau : Universités de Nantes, du Mans et d'Angers ; création du Service Inter Établissement Numérique (SIEN) voté par les 3 CA le 4/11/2021
- GLiCID : Universités de Nantes, du Mans et d'Angers et l'École Centrale

L'ensemble du projet a été financé par le CPER pour un montant total de 20M€

Calcul scientifique : état des lieux en Pays de la Loire en 2020

5 acteurs principaux :

- CCIPL (UN)
- BiRD (INSERM, UN, Biologie)
- ICI (ECN)
- MathStic (UA)
- INFRALAB (LMU)

	Utilisateurs actifs	CPU cœurs	GPU	CPU.h (10^6 h)
CCIPL	155	5500	27	28.4
BiRD	142	450	9	1.5
ICI	130	6300	18	19
UA	30	1000	10	8.4
LMU	~50	1900	90	1

Projet GLiCID

Dans le cadre du CPER 2021-2027

- Objectifs principaux du projet :
  - n'avoir qu'un seul acteur calcul Tier2 en Région
  - avoir plus de ressources et de services pour un coût équivalent
  - Mutualiser les ressources et les RH

- GLiCID est la partie calcul HPC du projet DACAS+réseau
- Établissements : UA, LMU, NU, Centrale Nantes, Inserm
- Porteur : Yann Capdeville
- Responsables par partenaire :
  - Yann Capdeville (CCIPL)
  - Luisa Rocha da Silva (ICI, ECN)
  - Audrey Bihouée (BirD/SFR F.Bonamy, INSERM, UN)
  - Frédéric Saubion (UA)
  - Sylvain Meigner (LMU).
- Budget et équipements demandés : 6M€

pour >12000 cœurs, >100 GPU type A100, >8Peta octets de stockage

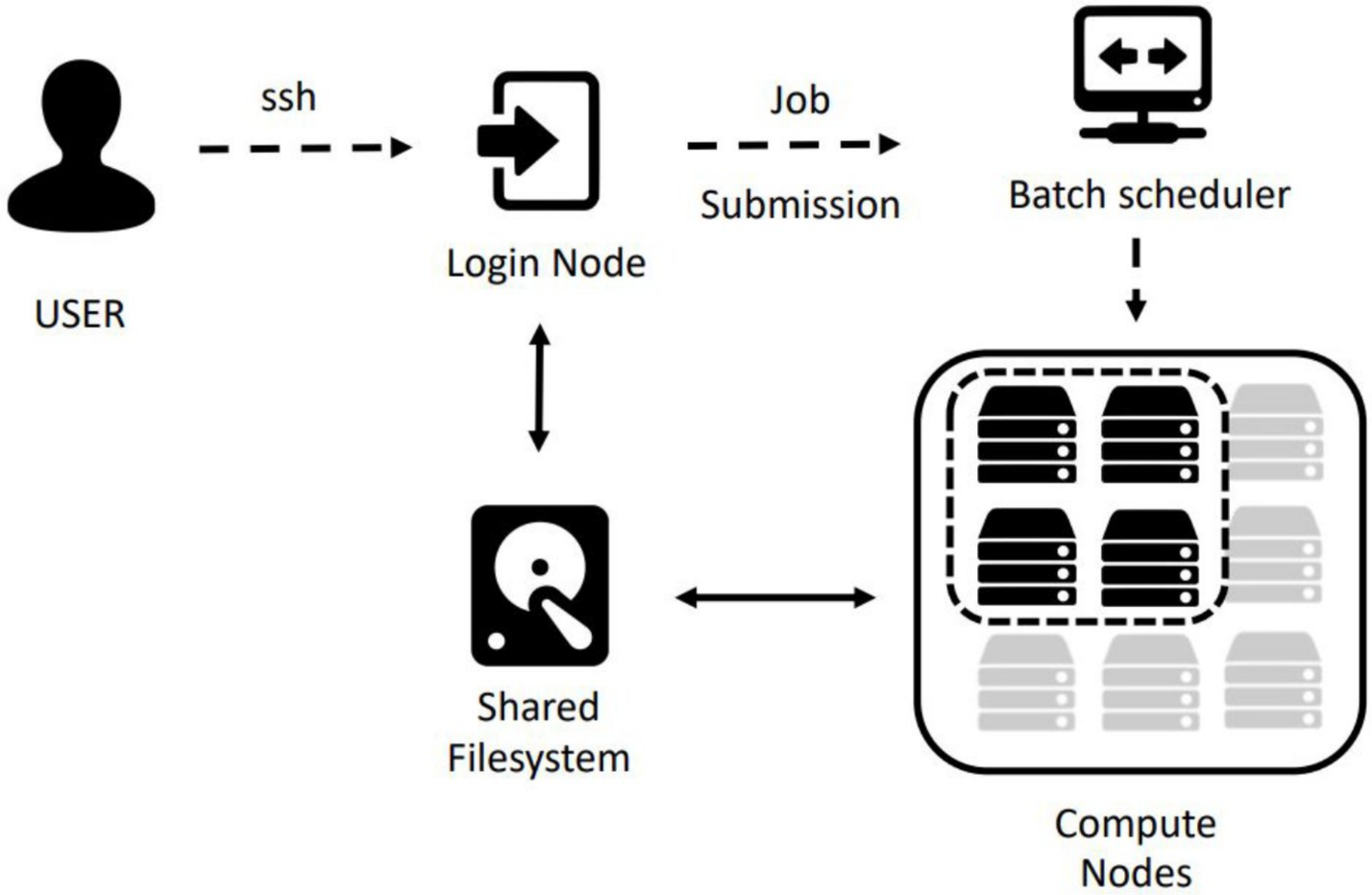
Site web : www.glicid.fr

Unité de service GLiCID

Missions de GLiCID :

- Fournir des moyens de calcul aux chercheurs de la Région et leurs partenaires sans se limiter aux « gros » calculs
- Proposer un volume de stockage de données associées au calcul suffisant
- Apporter un support au développement et une veille technologique
- Proposer des formations au HPC et aux outils s'y rapportant
- Garder une forte proximité avec les utilisateurs
- Participer aux projets nationaux (EQUIPEX+ : MUDIS4LS, MESONET)

Principe d’un cluster de calcul



Source : Junaid MIR (Ecole Centrale de Nantes), Pierre-Emmanuel GUÉRIN (Ecole Centrale de Nantes)

Stockage sur un cluster de calcul

Performance vs Sécurité

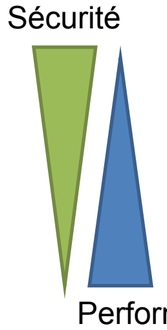
- Une infrastructure de calcul nécessite une solution de stockage **performante** :
  - accès massivement parallèle aux données
  - disques rapides
- Pour gagner en performance, on désactive les mécanismes de sécurité :
  - Moins voire pas de snapshots
  - Pas de réplication
  - Pas de sauvegarde
- Pour gagner en sécurité, on réduit la performance

- A capacité identique, le coût d'une infrastructure performante et d'une infrastructure sécurisé est le même

23

Source : https://moodle.france-bioinformatique.fr/pluginfile.php/360/course/section/57/Module_2.pdf

Stockage sur un cluster de calcul



Usage	Espace de travail	Quota	Sauvegarde
Espace perso	/home	3To	Oui
Espace projet	/LAB-DATA/GLiCID/projects	A la demande (€)	Possible (€)
Espace temporaire de travail	/scratch	Pas de quota, nettoyé régulièrement	Non

- Espace projet : Stockage volumétrique – 1Po (2023)-> 4Po (2028)

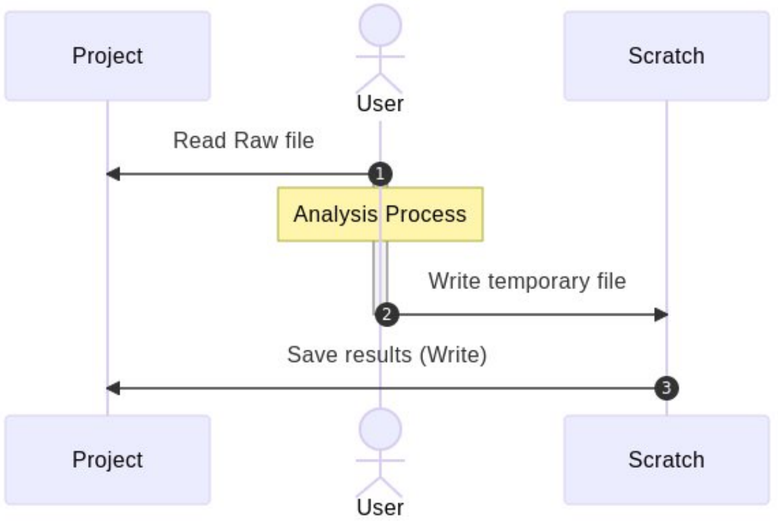
- Stockage sécurisé = Stockage distribué (CEPH)
- Aujourd'hui : 3 points géographiques : Labo maths, Datacenter Lombarderie, ECN
- A terme : 3 salles différentes dans DACAS
- Tolérance élevée aux pannes (disques)

- Pas de sauvegarde automatique. Backup possible sur demande (€)

En (bonne) pratique

- Transfert des données [protocole SSH]
  - (quasi)automatique depuis les équipements producteur de données (Séquenceur, Microscopes…)
  - ou depuis poste de travail

1) Copie des données brutes sur espace projet
2) Calcul/Ecriture sur espace scratch lors de l'analyse
3) Transfert des résultats importants sur espace projet



- Une organisation par projet
  - Avec des méta-données/nommage ad hoc

GliCID : Le modèle économique

Objectifs : garder le modèle accessible à tous tout en finançant au moins les postes GLiCID sur fonds propres.

- Heures de calculs
  - Au moins 60 % des ressources sur GLiCID accessible (sur projet) sans contre partie financière
  - Une partie des ressources achetée par les projets financés de recherche sous forme d'heures prioritaires
  - Une partie des ressources vendue aux établissements externes et au privé (max 20%)
- Stockage de données :
  - Pour les ayants droits : 3To de base sans contre partie financière
  - Payant au To par année au-delà. Tarifs différents entre les ayants droits, le privé etc.

Il est important pour la pérennité de GLiCID que tout le monde prévoit des financements (Prestation) dans ses projets pour les ressources de calcul et de stockage.

Merci de votre attention

LiCID

Directeur : Yann Capdeville (CNRS-NU)

Co-directrices.eurs :
Luisa Rocha da Silva (ICI, ECN)
Audrey Bihouée (UN)
Frédéric Saubion (UA)
Sylvain Meigner (LMU)

Equipe Technique

Pierre-Emmanuel Guérin	IGR, ECN	1 ETP CDI	Ingénieur système et réseau
Pablo Bondia-Luttiau	IGR, ECN	1 ETP CDD	Ingénieur système et réseau
Yann Dupont	IGR, NU	0.8 ETP	Ingénieur système et réseau
Jean-Francois Guillaume	IGE, BiRD-SFR, NU	1 ETP CDI	Ingénieur système et réseau
Guy Moebs	IGR, CNRS-LPG	0.5 ETP	Ingénieur calcul scientifique
Aymeric Blondel	IGE, CNRS-CEISAM	0.4 ETP	Ingénieur calcul scientifique
Jérome Coatanéa	IGE, NU	0.1 ETP	Ingénieur système et réseau
Jean-Christian Feufeu	IGE, NU	0.1 ETP	Ingénieur système et réseau
Damien Fligiel	IGE, CNRS-OSUNA, NU	0.1 ETP	Ingénieur système et réseau
Jumaid Mir	IGR, MesoNET, ECN	1 ETP CDD	Ingénieur calcul scientifique
Hugues Digonnet	MdC, ECN	0.2 ETP	calcul scientifique
à recruter	IGR, UA	0.7 ETP	Ingénieur système et réseau
à recruter	IGR, LMU	0.5 ETP	Ingénieur système et réseau

université angers · CENTRALE NANTES · Nantes Université · Le Mans Université