Journée Calcul 14 juin

Ordre du jour :

- Présentation de la journée
- Stockage capacitif/performant et  distancedonnées/calcul?
- Quels besoins, quels usages pour les cluster de calcul par rapport aux clouds orientés calcul ou lambda compute
- Infrastructure de données
- Développement, optimisation, parallélisation de code
- Logiciels et licences
- Formation/support
- 15:10 - 15:40  Pause café
- Conclusion participative - Etape suivante ?
- 16:55 - 17:00 Fin de la journée

==========

# Présentation de la journée Marc LELIEVRE (IRSTEA)/ Alain FRANC (INRA) / Eric MALDONADO (IRSTEA)

Faciliter le dialogue entre chercheurs et informaticiens infrastructures
Ajouter le lien présenté : https://www.mathinfo.inra.fr/calcul
https://www.mathinfo.inra.fr/groupeCalcul

Accessible depuis IRSTEA uniquement : Wiki scientifique IRSTEA : http://is.irstea.fr/

Ajouter quelques liens

# Stockage capacitif/performant et  distance  données/calcul?  Albert SHIH (Observatoire)

10h22 --> 10h49

Au vue des volumétries, comment peut-on s'en sortir ? Solutions ?

Bande magnétique à considérer si la olumétrie des données est très élevé et les données sont dites ""froides""
On réduit la consommation électrique ainsi que le poids si on compare une baie de stockage classique et des bande
Stockage cloud type Amazon:attention au coût de récupération des données souvent bien plus élevé que le coût de dépôt.

*"Ian Bird and Tim Bell from **CERN**'s IT department were up next. They offered an inside look into how they handle huge streams of data coming in from various experiments (ATLAS, ALICE, etc.) — **around 90 petabytes per year!** — and pinpoint the notable parts of the data quickly so that they can **write that to tape** (yes, it's 2019, and they're writing to tape.)"*
src: https://superuser.openstack.org/articles/takeaways-kendall-openstack-days-cern/

Stockage objet vs posix
Posix accessible facilement mais difficile à étendre
Inverse pour l'objet (courbe d'apprentissage)

Pas de méthode ou de type de stockage "parfait", il faut définir ses besoins sur plusieurs facteurs :
type de donnée
vitesse
volume
temps de conservation
taille des fichiers
utilisation

migration
niveau d'importance

DAS = Direct Attach Storage = On récupère les données sur le disque local de la machine avant de lancer le calcul (a renseigner dans le batch)

Optimiser la partie purement traitement scientifique quand c'est très distribué. Si 1M d'expé, le code est répliqué autant de fois.
Identifier les données, METADONNEES riche = données de valeur

- # Quels besoins, quels usages pour les cluster de calcul par rapport aux clouds orientés calcul  Joelle AMSELEM (INRA)

10h50 --> 11h05 -->11h20

Ajouter les liens :

URGI *Unité de Recherche Génomique Info* *https://urgi.versailles.inra.fr/*
Cloud privée produisant 3 catégorie de services
SaaS Software as a Service URGI Platform
PaaS Platform as a Service URGI Platform
IaaS Infrastructure as a Service, hébergement de l'ensemble des services de base informataique dans un datacenter

REPET: Pipeline et virtualisation pour la reproductibilité des analyses
RepetDB : http://urgi.versailles.inra.fr/repetdb/begin.do : base de données des résultats d'analyse sortis de REPET
05 Automatiser la préparation d'un environnement adapté à une expé

Diffusion :
  Repo Gitlab
  Repo DockerHub => Image Docker Repet "All-in-one"
  IFB à partir des images docker, ils peuvent aller sur les cluster IFB (https://www.france-bioinformatique.fr/fr/cluster)

Offre de services cloud DSI en cours de construction.

Cloud vs cluster : cela converge, vers le clousder !
 gestion d'environnement gix et al

Le cloud peut aussi permettre la création de cluster physique par le moyen de API :
  https://www.openstack.org/software/releases/stein/components/ironic
  On peut y ajouter des cluster kubernetes et/ou des containers

  Plus qu'un concurrent il faut peut être voir le cloud comme un "facilitateur" pour le calcul se raprochant plus de l'utilisateur/trice.

- # Infrastructure de données - François Laperruque (INRA)

11h20 --> 11h--> 11h50

INRA : Phénotypage animal à haut débit
Phénotype : Ensemble des caractères apparents d'un individu

autres compétences nosql / big data : cati bbric codex crig

retour : difficultés à trouver des infras pour tester une infra nosql
Interet d'une infra de test, proposée par l'Institut, pour ces nouvelles technos?

A-t-on la connaissnce du nombre de déploiement de spark dans l'Institut?
Grande difficultés pour recruter des compétences "BigData". Impossible de concurrencer les offres du privé...

- # Développement, optimisation, parallélisation de code Sylvain JASSON

11h45 --> 12h15

Ajouter les liens :
https://www6.inra.fr/cahier_des_techniques/content/download/5249/53482/version/1/file/CTh2018bis_Art9_FRA.pdf
https://www.toptal.com/developers/sorting-algorithms

Implémentation informatique, les différentes couches :
FLOPS Floating-Points operations per seconds (silicium)
OS
Réseau
Communs
Base de données & Bibliothéque de codes
Gestionnaire de workflow
Environnements spécifique dont données et logiciels maisons
Sauvegarde des résultats
Moyens de visualisation des résultats

Optimiser c'est bien avoir un résultat c'est mieux

Les différents tiers :
Machine perso
Tier 3 (labo)
Tier 2 mésocentre
Tier 1 national
Tier 0 européen

Prendre le temps et faire des efforts pour optimiser son code
Language interpreté vs compilé peu ou pas être rentable
Ne pas avoir de préjugé !

relation mémoire / CPU et même bande passante

- **12:30 - 14:00 Buffet et café**

- # Logiciels et licences Didier Laborie (MIA)

14h05 --> 14h35

Ajouter des liens :
https://mia.toulouse.inra.fr/Accueil

Genotoul BioInfo
http://bioinfo.genotoul.fr/

Logiciels (>600/an)

comptes actifs retombés en dessous de 1000, donc réouverture contrôlée de nouveaux comptes dans peu de temps.

Installation de logiciels/librairies sur la plateforme de calcul par les admin sys
mise à jour par le gestionnaire de paquets, problème, vérifier les dépendances, certaines librairies peuvent être update et fait planter un workflow spécifique

guix : https://fr.wikipedia.org/wiki/GNU_Guix
autre présentation de Guix : https://linuxfr.org/news/gnu-guix-version-un-point-zero

Est ce que les solutions docker sont pertinentes pour les problématiques de portabilités et reproductibilité ?
Redhat 8/CentOS 8 n'utilisent plus docker : https://access.redhat.com/solutions/3696691, même si cela reste compatible, jusqu'à quand ?
https://www.opencontainers.org/ une solution ?

IFB :https://www.france-bioinformatique.fr/

- 

# • Formation/support

14h45 -->

## Developpement de pipelines au GAFL (Jacques Lagnel, INRA)

> 40 conteneurs crées sous singularity
Gestion de workflow avec Snakemake

Quels moyens pour former et sous quelles formes ?

- **15:10 - 15:40 Pause café**

# • Conclusion participative - Etape suivante ?

Présentation Alain Franc

Vos remarques à chaud sur le déroulement de cette journée (densité, planning...)
l'espace à la discusison était très bein
Présence chercheurs et infra
Qu'attendez-vous / quelles suites à donner ?

Quels besoins pour un scientifique ?

- IA & eGPU
- 

Il faut cataloguer l'ensemble des moyens de calculs

- 

Quelques expérimentations à Irstea mais sans usage concret (à cette heure)

Où calculer et à quel prix ?

Si vous aviez un service idéal (quelle est votre idéal) ?

Répondre aux besoins des chercheurs n'est pas chose facile
Une personne d'appui doit pouvoir orienter et accompagner les chercheurs vers les ressources.

Le domaine/thématiue scientifique peuvent avoir des réponses différentes en terme de moyens de calcul.
==> identifier les besoins par thématiques scientifiques pour obtenir le cluster le plus adapté

Exemple : offre riche pour la thématique BIOINFORMATIQUE

Accompagnement / besoin

- **Besoin de service pour le calcul. 15 ingé pour une machine (idris).**
- 
- 

16:30 Fin de la journéé