Informatique Scientifique @UPSaclay

Marco Leoni, marco.leoni@universite-paris-saclay.fr

Réunion des utilisateurs

27 mai 2020

Quelques dates clé depuis dernière rencontre (juillet 2019)

- *Septembre 2019*: Soumission Projet ERM
  - Un grand merci à tous ceux qui ont contribué!
- Fin *Décembre 2019*: Résultats projet ERM
- *Janvier 2020*: plusieurs établissements ont fusionné:
  - UPSud => UPSaclay
  - LAL => IJCLab
  - Ces processus ont contribué à ralentir la dynamique de l'I.S.
- Mars 2020: prévision d'un rencontre => émergence covid-19…

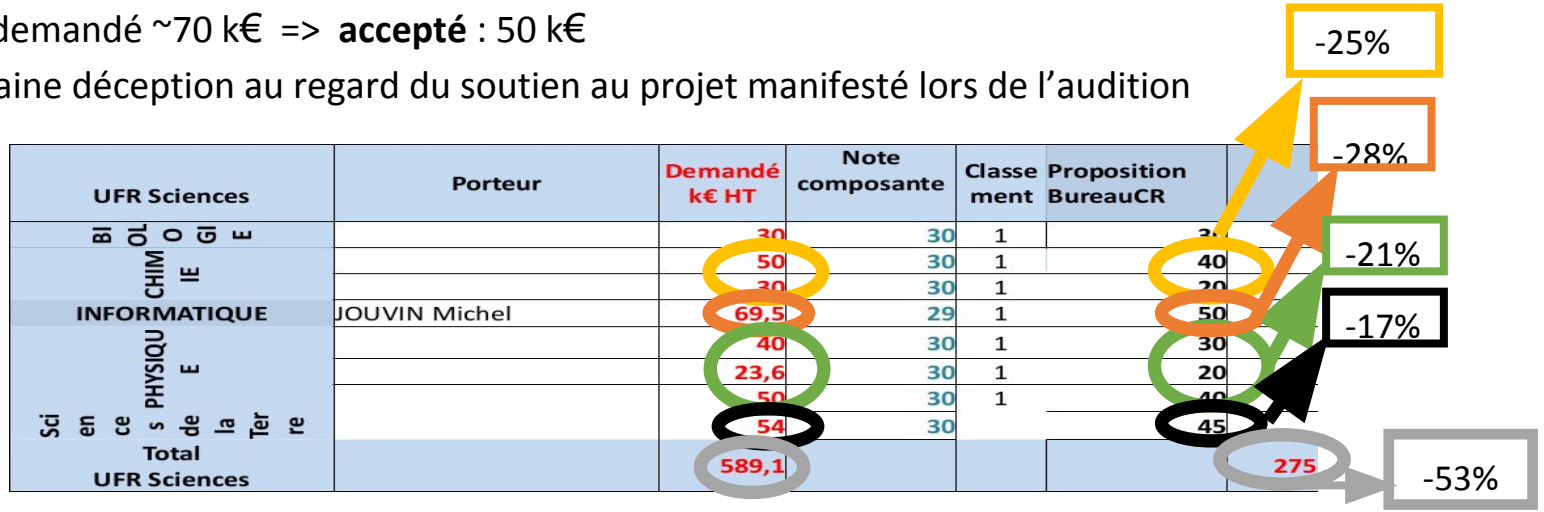
Sommaire

- Update résultats ERM
- Update activités
  - Salle machines/Cloud@VD
  - JupyterHub
  - CLAC (Cloud for Astronomy and Cosmology)
  - Spark
  - Gitlab Université
  - Open Science / Open Data
- Paris-Saclay : projet de mésocentre
- Actions 2020 prévues
  - Formation bonnes pratiques du cloud
  - Questionnaire sur les usages et besoins des utilisateurs : probablement dans le cadre du mésocentre UPSay

Résultats **ERM**

- **Projet transversal**, porté par la Faculté des Sciences
  - Participants: Biologie, Chimie, Physique, Géosciences, Informatique, Écologie, Droit
- Financement: demandé ~70 k€ => **accepté** : 50 k€
  - Une certaine déception au regard du soutien au projet manifesté lors de l'audition

UFR Sciences	Porteur	Demandé k€ HT	Note composante	Classement	Proposition BureauCR	
BIOLOGIE		30	30	1	30	
CHIMIE		50	30	1	40	
		30	30	1	20	
INFORMATIQUE	JOUVIN Michel	69,5	29	1	50	
PHYSIQUE		40	30	1	30	
		23,6	30	1	20	
		50	30	1	40	
Sciences de la Terre		54	30		45	
Total UFR Sciences		589,1			275	



- Questions pour l'avenir
  - Quelles ressources avec les 50 k€ : calcul ou stockage ?
  - Quel type de financement pour une ressource aussi transversale : ERM pas très adapté, encore moins dans le contexte Paris-Saclay

Update des activités

Site Web de l'Informatique Scientifique

- Site officiel depuis le printemps 2018 https://www.mesocentre.u-psud.fr
    - Information est à jour : en particulier les points de contacts des différents services
    - Destiné à être enrichi (réflechir à inclure l'informations contenu sur blog informel http://hebergement.universite-paris-saclay.fr/info.sci.blog/blog)
    - Site hébergé sur GitHub : permet la contribution de tous avec un système de revue des contributions avant approbation (merge request)
    - un mode d'emploi reste à écrire, contacter Michel Jouvin en attendant...

Salle Machine

- Extension en cours, financée par le labex P2IO (CPER, 2 M€), (presque) terminée
  - Du retard à cause de problème administratifs + 1 problème sur la commande des PDU : nouvelle capacité utilisable à partir de l'été
  - 21 racks supplémentaires : possibilité de 40 racks de plus après l'extension (max = 90)
  - Puissance électrique/cooling IT : 600 kW
  - Sécurisation électrique par double alimentation HT/BT pour 300 kW (non localisé) : appelée « niveau argent » par opposition au « niveau bronze » actuel (simple alimentation non secourue)
- Extension en cours, financée par UPSud/Paris-Saclay, pour ajouter un « niveau or »
  - Secours de 10 racks/80 kW par onduleur + groupe électrogène (localisé, rangée basse densité)
  - Permettre l'hébergement de services des DSI des composantes Paris-Saclay, en particulier UPSud
  - Sécurisation électrique de l'accès réseau
- Une ressource unique dans le contexte Paris-Saclay : intérêt grandissant

Cloud : Situation Actuelle

- Croissance des ressources du cloud : 4500 cœurs, 500 TB disque (non permanent)
  - Divers contributeurs, y compris extérieurs à UPSud
  - Du fait de la mutualisation, bénéficie à tout le monde
- Gestion d'une notion de « durée de vie » des VMs pour récupérer les ressources des VMs oubliées (https://lease-it.lal.in2p3.fr)
  - Rappel/mail après 3 mois invitant à confirmer qu'il a toujours besoin de la VM
  - Important de ne pas préempter des ressources inutilement
- Mise en production du service Magnum qui permet de créer des clusters de containers
  - Supporte les orchestrateurs Swarm, Mesos, Kubernetes
  - Kubernetes : le plus prometteur mais la version actuellement disponible est ancienne. Update en cours pour avoir la version 1.16 ou plus.

Cloud : Évolutions Prévues

- Interface job (batch) au cloud : éviter à l'utilisateur de gérer des VMs pour exécuter des applications de type batch
  - Basé sur HTCondor
  - Potentiellement exécution du job dans une VM ou un container, au choix
  - Probablement pas avant l'automne...
- Support d'une fonctionnalité de type « spot instance »
  - Des VMs au-delà du quota du projet qui peuvent être arrêtées à tout moment si besoin pour un autre projet plus prioritaire
  - D'ici la fin de l'année (2020)
- Support de file systems partagés entre VM

JupyterHub http://jupytercloud.lal.in2p3.fr/

- Instance JupyterHub multi-tenant exécutée le cloud@VD
  - Permet l'exécution de Jupyter notebooks avec plusieurs langages (kernels) : Python, C++, R
  - Intégré avec Adonis (courriel UPSud): même crédentiels
  - Projet commun avec le CEMAP (maths appliquées) à Polytechnique
  - Utilisé avec succès par des enseignants de mathématiques, informatiques, physique
- Évolutions récentes : mise à jour de l'image avec NBGrader (notation), kernel R (stats.), et Binder (système pour construire les containers à partir d'une description de haut niveau)
  - JupyterHub configuration/image: https://gitlab.in2p3.fr/jupyterhub-paris-saclay
  - Support: https://gitlab.in2p3.fr/jupyterhub-paris-saclay/support-requests
  - Référents @UPSaclay : Nicolas Thiery (LRI); Marco Leoni (DSI)
- Limitation actuelle : 1 machine (VM) pour le front-end et l'exécution des notebooks
  - Premières expérimentations avec cluster Kubernetes en cours, en attente de la mise à jour du cloud pour le passage en production
- Discussions pour fournir un kernel orienté Machine Learning

CLAC

- Cloud pour l'Astrophysique et la Cosmologie
    - Des ressources spécifiques pour les besoins de l'astrophysique et de la cosmologie
    - Intégré dans le cloud VirtualData
    - Financé par le DIM ACAV+ (région IdF)
- Phase 1 (2019) : 4 serveurs avec une grande quantité de mémoire
    - 1,5 TB par serveur, 40 GB/coeur
    - Cible prioritaire : simulation des objets complexes du ciel
- Phase 2 (2020, demande en cours) : 9 serveurs 256 thread/512 GB mémoire
    - Réponse attendue prochainement
    - Cible principale : applications avec un fort parallélisme interne type HPC ou Spark (HPC in a box)

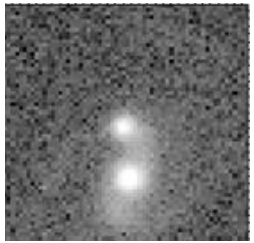
Spark

- R&D Spark démarré conjointement à IJCLab (ex-LAL) et pour Génomique2025 il y a plusieurs années
  - Plusieurs actions de formation
  - Animation UPSud par Chris Arnault (Sapphire CNRS)/Julien Peloton (IgR, CNRS)

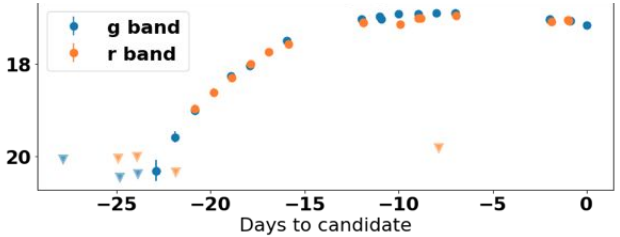
- 2 axes majeurs : infrastructure scalable et expertise applicative
  - Cluster Spark à la demande : provisioning des workers en fonction des besoins dans le cloud. Connexion à venir avec Kubernetes
  - Déploiement des logiciels contrôlé par les utilisateurs via le service CVMFS
  - Projet Astrolab Software : https://astrolabsoftware.github.io

- Développement important de l'activité à IJCLab depuis 2 ans dans le cadre de LSST
  - Télescope LSST : 10 TB/nuit à partir de 2022 avec la génération de millions d'alertes par nuit
  - Projet "*Fink*" : broker d'événements qui analyse et cross-check les alertes pour des besoins physique particuliers. Sélection en cours pour être un des brokers officiels du télescope

Fink Broker - module de Machine learning



au fil du temps



Télescope LSST

images

Courbe de lumière (2 bandes)

Problème:

- *Automatic early detection* (en utilisant uniquement la partie initiale de la courbe de lumière)
- Pour une base de données qui, elle même, évolue dans le temps

Solution:

- classification en utilisant un algorithme d'apprentissage machine dit "actif" ('*active' machine learning),* puis *random forest*

(*M. Leoni* en collaboration avec *J. Peloton*, *E. Ishida*, *A. Moeller*)

Gitlab @ UPSaclay

- Gitlab: service pour gérer et partager du code avec des fonctionnalités de suivi (tickets…)
  - Similaire à GitHub
- Instance hébergée sur les VMs de la DSI https://gitlab.u-psud.fr/
  - Intégré avec Adonis : même crédentiels
  - Déjà utilisé par Enseignants/Chercheurs/Doctorants (> 600 comptes)

Évolutions récentes:

- Mise en place de l'intégration continue (CI), integré dans Gitlab
- Serveur de test: https://gitlabtest1.di.u-psud.fr/
- Support: https://sos.dsi.universite-paris-saclay.fr/ (Ticket DSI, catégorie Informatique Scientifique)
- Référent: Marco Leoni (DSI)

Science Ouverte

- Un enjeu majeur de la science actuelle regroupant plusieurs problématiques
  - Accès aux publications : open access
  - Gestion des données : Data Management Plan (DMP)
  - Accès aux données : open data, science reproductible
  - Partage d'infrastructure de données et de traitement : EOSC (European Open Science Cloud)
- Un chargé de mission à Paris Saclay : Étienne Augé (VP adjoint science ouverte)
  - Stratégie et actions en cours de définition
- Redéfinir la place de l'infrastructure IODS / Linked Open Data dans ce contexte
  - Assurer la maintenance pour les 2 projets qui en dépendent : Gregorius (droit), DAAP (pharmacie)
  - Faire un bilan de l'exploitation de ces 2 services : identifier les points critiques pour leur maintenance
  - S'insérer dans les infrastructures françaises et européennes, en particulier EOSC

Paris-Saclay : un nouveau mésocentre ?

- Depuis 1 mois, relance des discussions sur l'informatique scientifique de Paris Saclay
  - Première itération à l'automne 2019 : projet CPER HPC@UPSaclay
  - Relance mi-avril à l'occasion du projet PIA3 MesoNet porté par GENCI
- Objectif : créer un mésocentre unique regroupant les différentes plateformes d'informatique scientifique
  - FUSION (CentraleSupelec, HPC/IA), LabIA (IA), VirtualData (cloud, stockage distribué)
  - Une offre cohérente et complémentaire mise en oeuvre par des "opérateurs" différents
- Organiser des réseaux d'expertise autour du mésocentre
  - HPC/calcul intensif : proposition "d'objet transverse" avec la MdS (E. Audit, M. Jouvin)
  - IA : DataIA
  - ML : Center for Data Science (CDS)
  - …
- Un réseau d'utilisateurs : élargir Informatique Scientifique@UPSud ?
  - Initiative Paris Sud semble unique parmi les partenaires ?
  - AgroParisTech : motivé et présent aujourd'hui
  - En discussion : un chargé de mission pour l'animation du réseau

Conclusions

- Communication parfois peu visible mais beaucoup d'actions qui ont permis de relancer la dynamique informatique scientifique
  - Importance du rôle de Marco Leoni (maintenant CDI)
- Un lieu d'hébergement des ressources informatiques unique dans le contexte Paris-Saclay et au-delà
  - Extension en cours va permettre de faire face à de nouveaux besoins
- Une plateforme pour héberger des services mutualisées : le cloud
  - Une ressource déjà significative dont l'extension ne pose pas de problèmes
  - CLAC : exemple de communauté apportant des ressources pour des besoins spécifiques
- Plusieurs services avancés déjà disponibles pour le traitement de données, le partage de données et de connaissance et le support à l'enseignement
  - Favoriser des solutions communes aux problèmes communautaires autant que possible
- Paris-Saclay et le projet de mésocentre sont une chance pour le développement de la dynamique mise en place à Paris-Sud