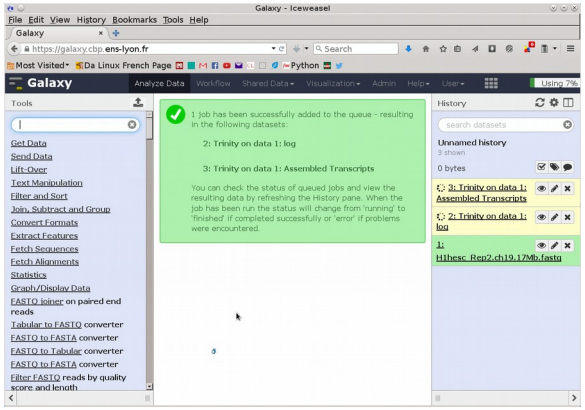
CBP
CENTRE BLAISE PASCAL

Déploiement d'un Portail Galaxy : du démonstrateur... à une production tacite...



Emmanuel Quémener

UNIVERSITÉ DE LYON

ENS

Requête des laboratoires

- Contexte :
  - Explosion des volumes de données en sortie des séquenceurs
  - Analyse de volumes de données massifs
  - Utilisateurs avec peu de formation en calcul scientifique
  - Émergence d'un portail standard dans la communauté de biologie
  - Premier portail Galaxy à l'IGFL, mort aujourd'hui...
- Expressions de besoins :
  - Disposer d'un portail Galaxy hébergé par le CBP accessible à l'ENS
  - Former un comité chargé de suivre les évolutions

Réponse du Centre Blaise Pascal

Oui, mais à certaines conditions

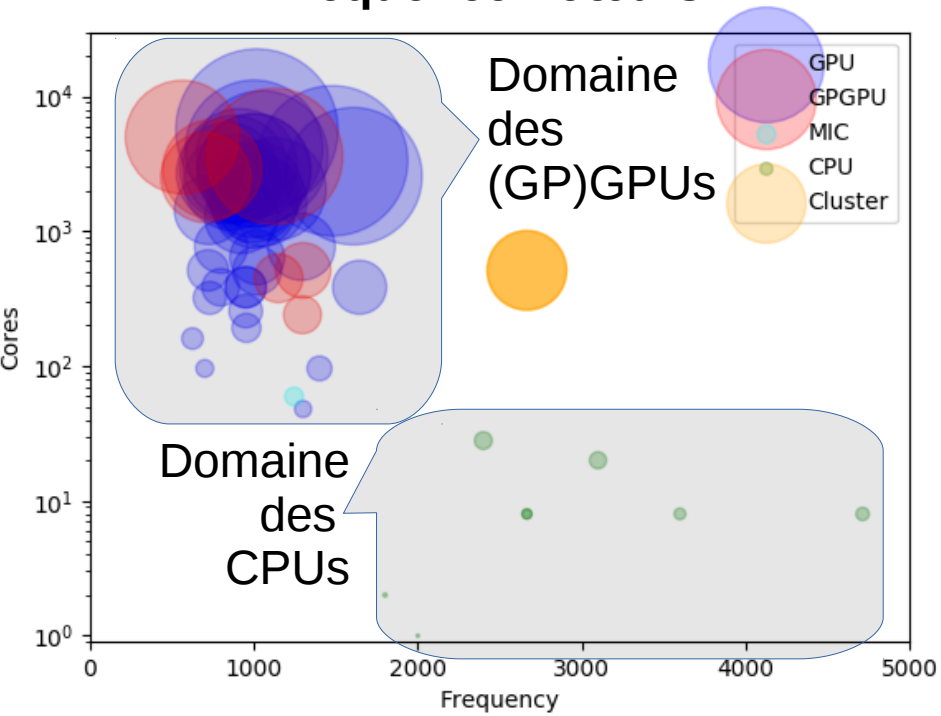
- Respect des « phases » du CBP
  - Expérience : environnement dédié à l'installation et la métrologie
  - Démonstrateur : exploration de nouvelles approches
  - Prototype : définition du système optimal
  - Production : pas la vocation initiale du CBP
- Partage des responsabilités
  - EQ : installation, maintenance, évolution de la plate-forme hard & soft
  - Bio-informaticiens des unités : vérification, installation composants, gestion premier niveau des utilisateurs

D'abord, c'est quoi le CBP ?

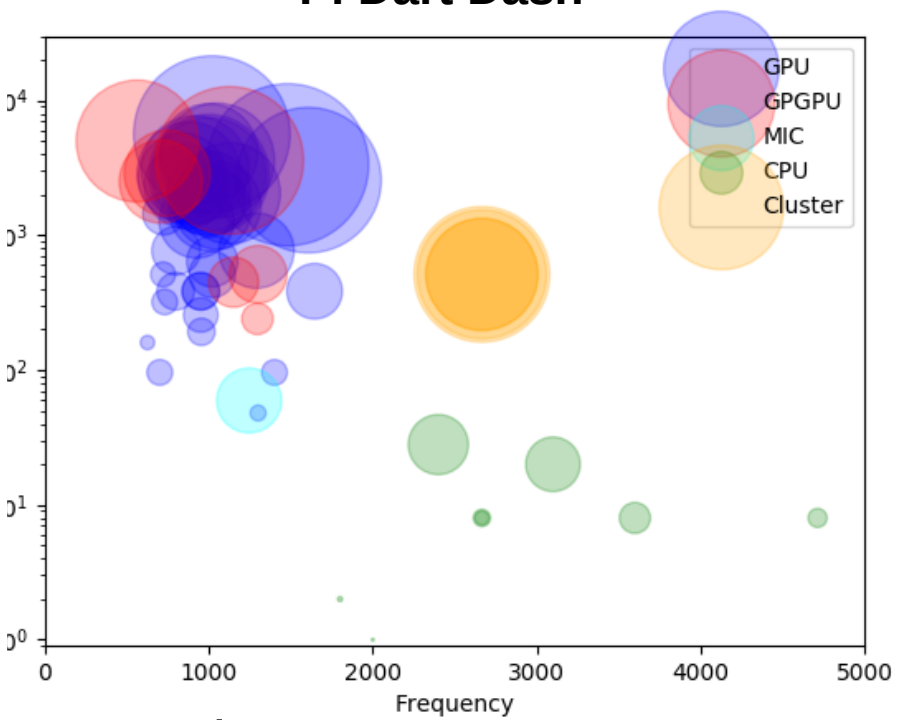
- Hôtel à projets, à conférences, à formations...
- Maison de la modélisation
- Plate-forme expérimentale avec 10 plateaux :
  - Plateaux techniques multi-nœuds, multi-cœurs, multi-shaders, ...
  - Plateaux techniques d'intégration : Debian, Ubuntu, ...
  - Plateau technique « architectures exotiques »
  - « Paillasses numériques » : biologie & SHS
    - Machines virtuelles pour les « Humanités Numériques »
    - Machines de visualisation : 3D avec lunettes, MorphoGraphX, ...
    - Machines de traitement expérimentales : repeat*

Et à quoi ça sert tout ça ?
A comparer.

Performance Théorique
Fréquence * cœurs

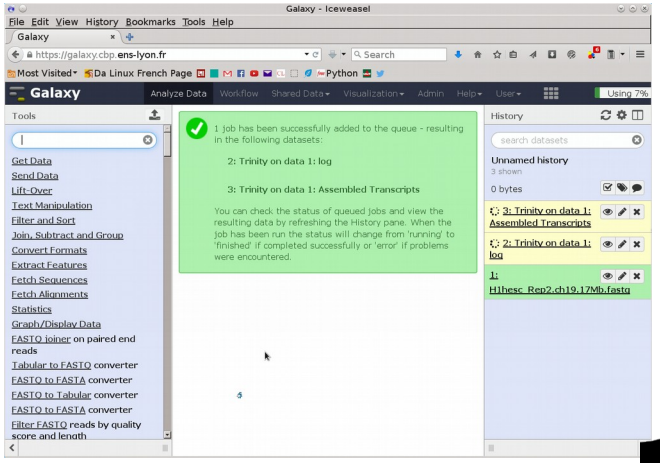


Performance en 32 bits
“Pi Dart Dash”



- Sur un GPU, cohérence entre performances théorique & pratique
- Sur un CPU, performance relative meilleure

Le CBP : un Iceberg
Sur et sous la ligne de flottaison



- Une trentaine de serveurs physiques
- Une centaine de machines virtuelles
- Une centaine de nœuds
- Une cinquantaine d'équipements réseaux
- Une authentification DSI

Plate-forme Galaxy : le Graal ?

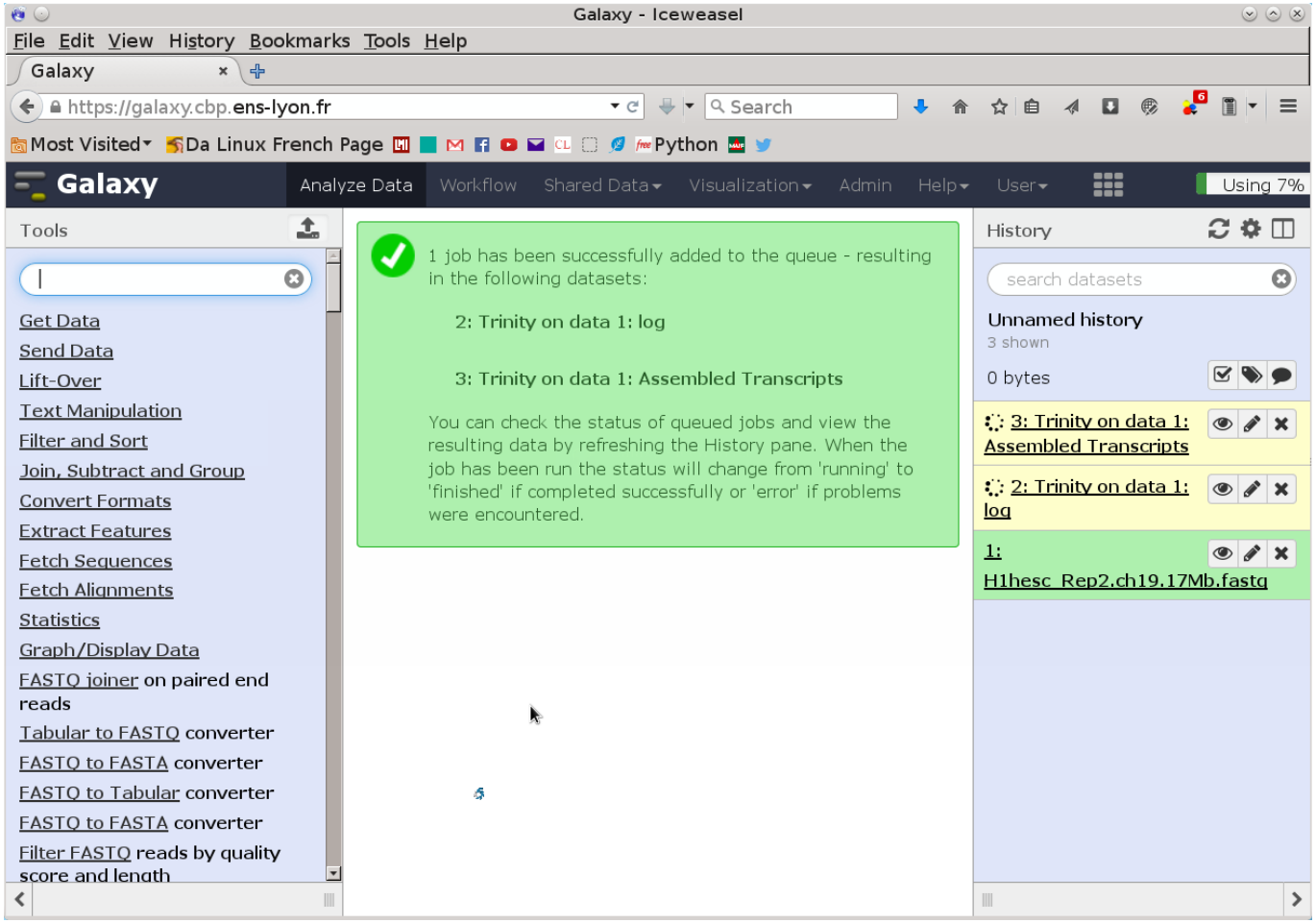
- « Galaxy, je l'ai installé sur mon portable, ça marche ! »
- « Pas de souci, ça s'installe tout seul ! »

Galaxy « personnel »

Galaxy « universel » ?

Pour un portail Galaxy universel
Quels outils ?

A quoi ça ressemble, un portail Galaxy ?



Expression de Besoins d'un Galaxy pour Recherche & Enseignement Mariage de la carpe & du lapin

- Recherche
  - Activité « creuse » : quelques utilisateurs simultanés
  - Grand nombre de jobs sur durée courte (quelques jours)
  - Volumes entrée/sortie difficiles à anticiper
  - Traitements de durée impossible à anticiper
- Enseignement
  - Activité « dense » : plusieurs dizaines d'utilisateurs simultanés
  - Grand nombre de jobs sur durée très courte (2 heures)
  - Volumes entrée/sortie raisonnable et prévisible
  - Traitements de durée raisonnable et prévisible : quelques minutes

Des expressions de besoins aux Spécifications fonctionnelles & techniques

- Spécifications fonctionnelles
  - Accessible via navigateur directement
  - Chargement de gros volumes par méthode tierce
  - Authentification établissement
- Spécifications techniques
  - Machine virtuelle sous KVM
    - Pour permettre une meilleure portabilité d'une configuration à l'autre
  - Délégation des traitements à des ressources tierces
    - Pour bénéficier de la métrologie du gestionnaire de batchs
    - Pour libérer la frontale de taille de traitements

Autour du portail Galaxy : les autres services nécessaires

- Services périphériques :
  - Portail Galaxy : serveur d'application en Python sur port particulier
  - Redirecteur (Proxy) Web : NGINX pour accès Web facilité
  - Serveur FTP : chargement de gros volumes de données
  - Serveur NFS : partage dossier Galaxy avec nœuds
- Services externes :
  - Authentification avec LDAP par ENS-Lyon & filtrage par login par CBP
  - Serveur de Batch avec GridEngine
  - Serveur de $HOME pour nœuds
  - Serveur SIDUS des nœuds

Expérience & Démonstrateur Quelques éléments techniques

- Socle :
  - Sunfire X4150, 32 GB RAM, 8 cœurs E5440 à 2.93 GHz, 4TB nets
  - Debian Jessie, noyau rétroporté
  - « Charge » de VM sur 6 disques en ZFS/raidz ou 5 disques HD + SSD 1TB
- Machine virtuelle avec portail Galaxy :
  - 4 cœurs, 16GB de RAM
  - espaces root de 100GB & data de 1TB
  - Debian Jessie
- Nœuds :
  - 8 Sunfire X4150, 32GB RAM
  - Système SIDUS : « Single Instance Distributing Universal System »
- Interconnexion réseau GigaBit Ethernet

Construction d'un portail Galaxy

- Objectif : intégrer un portail Galaxy à un mésocentre
  - Ça veut dire :
    - Interfacer la passerelle à des ressources distribuées de type « cluster »
    - Distribuer les requêtes des utilisateurs sur les nœuds des clusters
    - Examiner le « comportement » en fonction de la charge
- Expériences : 4 déploiements successifs pour le PoC
  - Déploiement en local dans un dossier dédié
  - Déploiement dans le dossier de l'utilisateur « galaxy »
  - Déploiement dans un dossier partagé avec les nœuds
  - Déploiement (migration) vers une machine « confortable »

Lors de la phase d'exploration Application des « perturbations »

- Choix du gestionnaire de batch : nécessité respect DRMAA
  - OAR : actuel CBP, pressenti PSMN : API non opérante
  - Slurm : très utilisé, API DRMAA inopérante (malgré la doc)
  - GridEngine : jamais utilisé au CBP, aide à la configuration par PSMN
- Dossier partagé « galaxy » contre le portail :
  - Mauvais choix du $HOME de l'utilisateur Galaxy : passage dans dossier
    - Déni de service sur la frontale des clusters CBP
  - Mauvais choix d'un volume disque « standard » :
    - Réactivité des disques insuffisante à la sollicitation
  - Mauvais choix d'un réseau uniquement GE pour le serveur
- Base de données en SQLite3 :
  - Soin particulier sur l'analyse des logs Galaxy

Période de « chauffe »

- Sur déploiement #3 :
  - Déploiement du portail Galaxy sur dossier partagé avec nœuds
    - Dossier Galaxy avec données génomiques sur disque SSD rapide de 1To
    - Interconnexion clusters via NFS sur réseau Gigabit Ethernet (125Mo/s)
    - Immobilisation 8 nœuds x41z avec 8 cœurs
    - Tous services activés : NFS, FTP, NGINX
  - Déploiement de plus de 80 modules
  - Intégration de bases de données externes
  - Vérification FTP
    - Attention ! Le fichier une fois intégré au portail Galaxy est supprimé !
  - Vérification composants

Test « grandeur nature » du 30/4

- Jeudi 30 avril 2015
- 40 étudiants sur deux salles
- Problème de login
  - Accès au serveur de la DSI très « lent » : redémarrage cache
- Problème pendant la séance de TP
  - Charge de la machine Galaxy : passage de 4 à 6 cœurs, de 16 à 24 Go de RAM
    - Processeur, entrées/sorties sur disque, entrées/sorties réseau
  - Charge du proxy Web : passage à 64 sessions simultanés
  - Charge du portail Galaxy :
    - Un utilisateur connectait trustait toutes les ressources
  - Charge des nœuds : passage de 8 à 16 nœuds disponibles

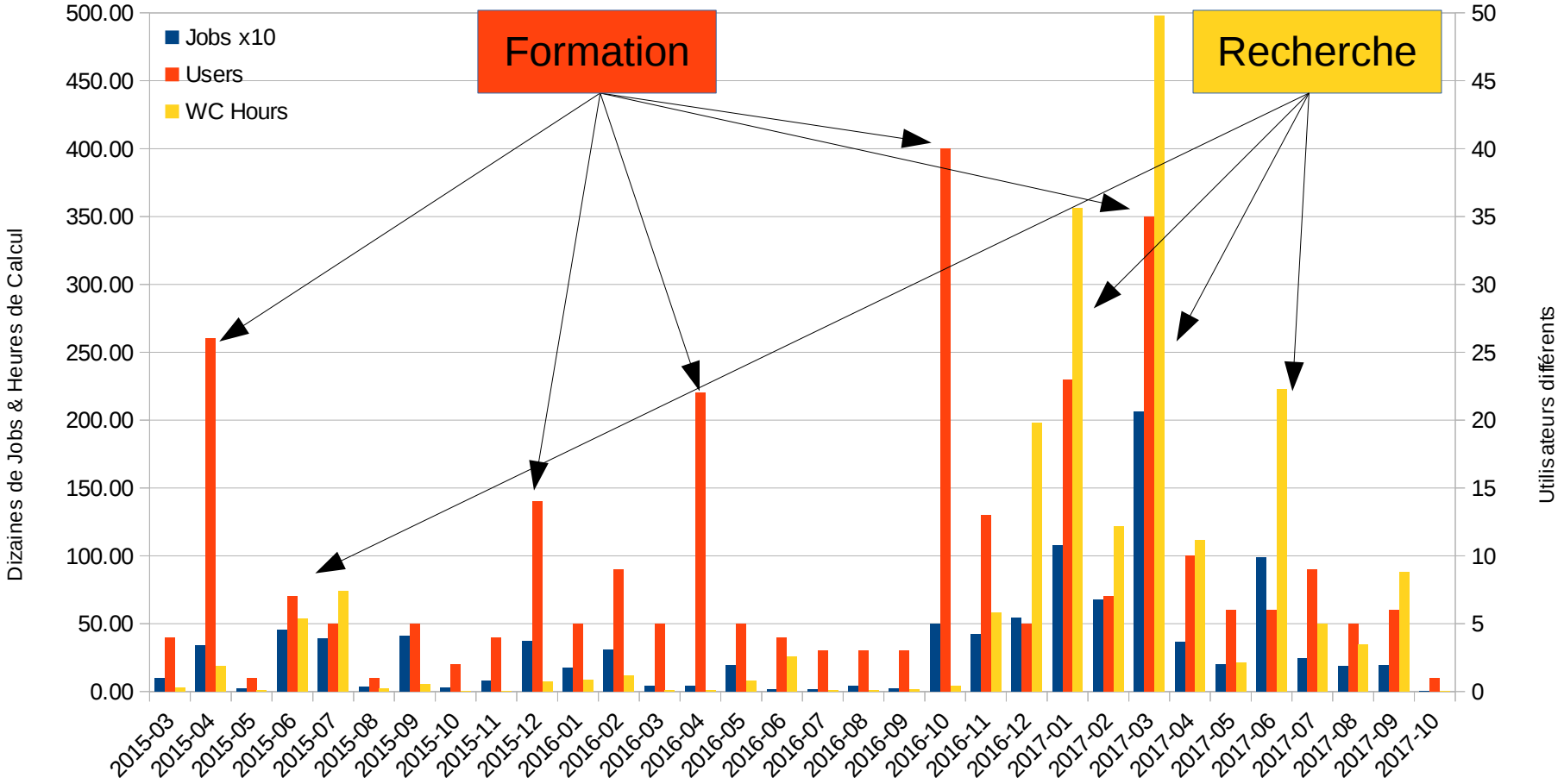
Test « grandeur nature » du 30/4

- Métrologie pendant le TP
  - 295 jobs exécutés
  - 22 utilisateurs connectés : de 2 à 48 jobs par utilisateur
  - Transfert réseau : grosse asymétrie (compréhensible)
    - Vers les utilisateurs : 1.4 GB reçus et 2.5 GB transmis
    - Vers les nœuds : 57.5 GB reçus et 112.1 GB transmis
  - Durant les « bowtie » : transfert de HG19 pour chaque job
    - HG19 fasta ~ 3GB, soit 27 secondes minimum pour chaque chargement
    - Plus d'accès disque (cache)
- Conclusions :
  - Communication portail/nœuds : goulet d'étranglement
  - Accès concurrentiel à la base SQLite limitant
  - Très mauvaise gestion des sockets réseau par Galaxy (tuning noyau)

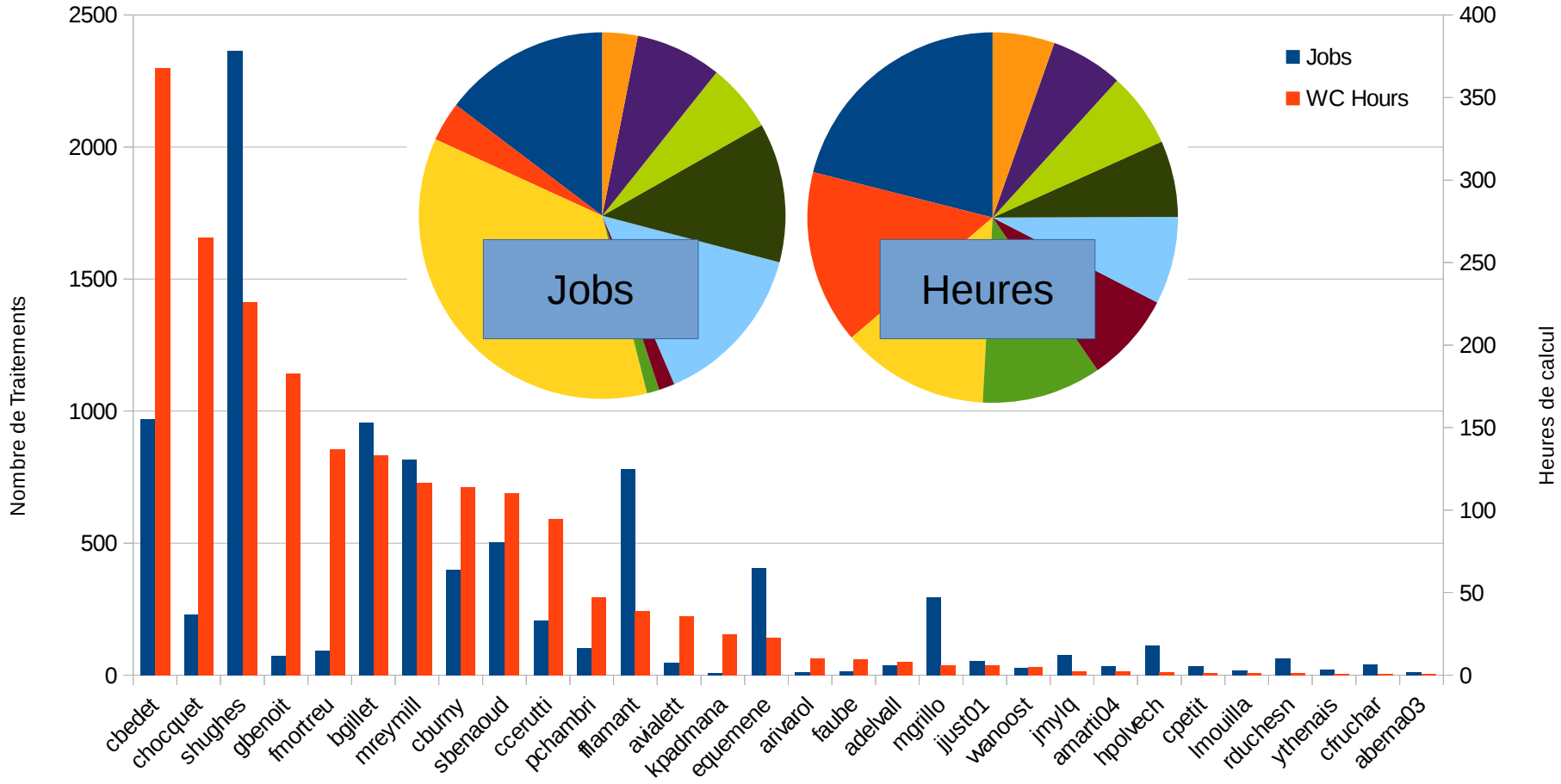
Galaxy version 2
Évolution sans révolution

- E/S & Transfert de données : goulets d'étranglement
  - Gestion de cache, socle ZFS avec SSD, passage à l'infiniband
- Spécifications techniques :
  - D'un socle R510 Westmere à R730 E5-2687v3
  - KVM avec passage d'une InfiniBand QDR en passthrough
  - Système de fichiers BtrFS puis XFS
  - Remplacement du serveur FTP de ProFTPd à WS-FTPd
  - Remplacement du SGBD SQLite3 à PostGreSQL
- Implémentation : du TUNING partout !
  - VM, socle, système, cache, NFS, InfiniBand, Nginx, ...

Bilan de 30 mois & 10556 jobs plus tard
Une activité croissante & évolutive…
Les deux utilisations très marquées...

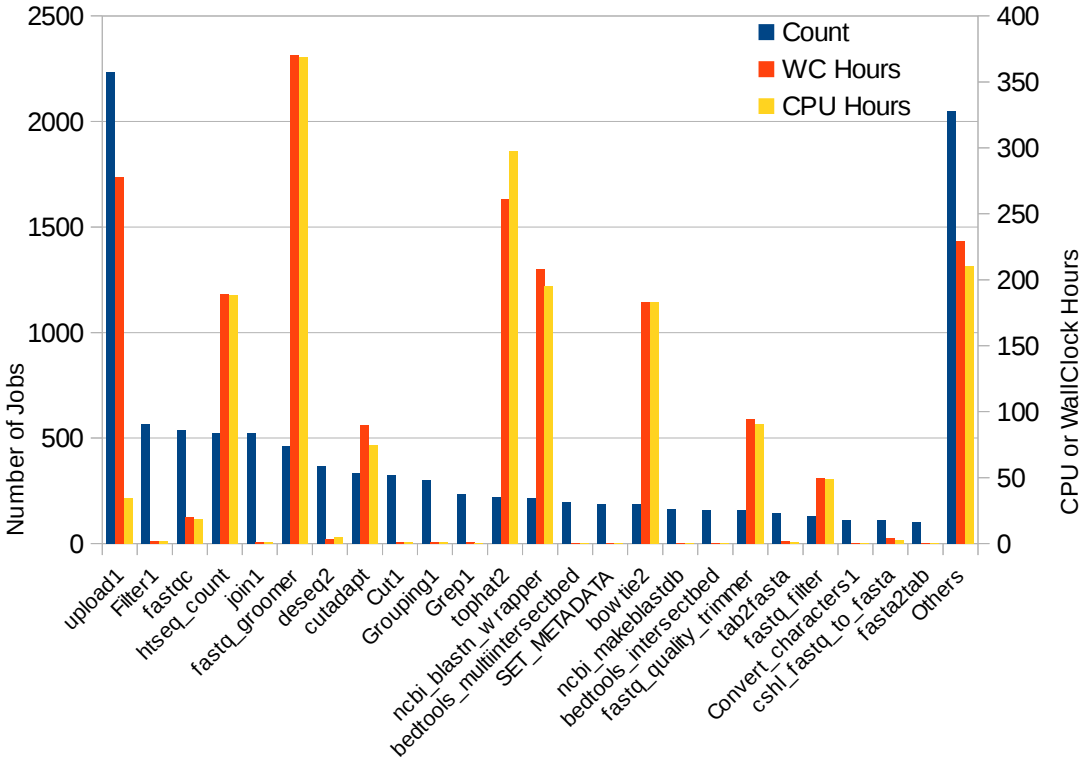


Des comportements utilisateurs très différents !

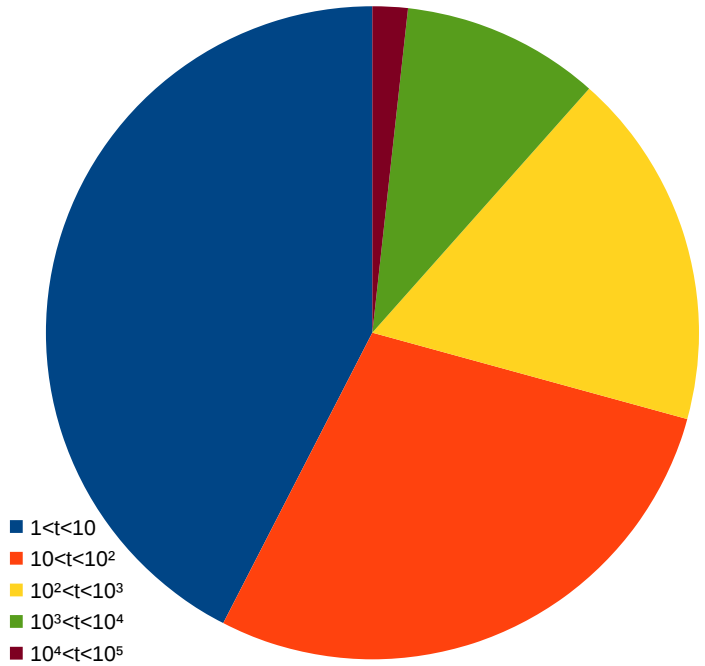


Des usages pour des 137 outils très diversifiés, mais dense pour 1 !

Distribution entre outils



Distribution WallClock



Bilan après 30 mois

- Au départ :
  - Définition par le CBP d'un prototype par cycles courts
  - Investissement des bio(logistes|informaticiens) dans la gestion
- A l'arrivée :
  - Le portail Galaxy est loin d'être parfait :
    - Prévoir le coussin & le sac à proximité lors du déploiement & exploitation d'outils
    - Des contrôles « qualité » de paquets trop sommaires (en comparaison avec Debian)
  - Des usages en recherche très différents brisant toute généralisation
    - Mais nécessitant une utilisation & récurrente
    - Donc un service « expérimental » avec une disponibilité « de production »
  - Une implication des (directions des) laboratoires (à mon avis) trop faible
  - Des « stress » ou « burn out » dans les traitements (E/S maîtresses du jeu!)

Avenir : ouverture, Galaxy3 intégration sidus4labs

- Ouverture à d'autres disciplines
  - Création de greffons Galaxy sur les applications « typiques » de chimie, physique
- Galaxy 3 :
  - Ouverture à l'extérieur de l'établissement
  - Passage du dossier « Galaxy » sur stockage distribué
    - Exploitation GlusterFS (ou autre) sur 8 nœuds SIDUS
  - Intégration à un cluster exploitant slurm
    - Association par allocation de ressouces
  - Distribution en « OpenCluster » avec du « SparseComputing »
    - Exploitation des stations de travail distribuées, mode déconnecté
- Sidus4galaxy :
  - Intégration du portail Galaxy dans Sidus4Labs

Remerciements aux biologistes qui ont aidé !

- Sandrine Hugues de l'IGFL
  - Pour ses tests & son engagement dans actions de formation pour les laboratoires de biologie
- Jeremy Just du RDP
  - Pour ses tests & lle chargement des génomes du RDP dans la base persistante
- Hélène Polvèche du LBMC
  - Pour les retours sur le fonctionnement de certains outils & ses formations
- Frédéric Brunet de l'IGFL
  - Pour la fourniture des éléments permettant le chargement de génomes
- Marie Sémon du LBMC
  - Pour l'exploitation par les départements d'enseignements

Requête pour 3IP
Donnez le vieux matériel !

- 3IP : Prononcez « Trip »
- Introduction Inductive à l'Informatique et au Parallélisme
- Approche portant l'accent sur la manipulation
  - Donc du matériel démonté généralement archaïque
  - Donc du très vieux matériel (tout matériel) montrant l'évolution
- Si vous en avez, ne jetez pas ! Donnez les, SVP !!!