ÉCOLE
NORMALE
SUPÉRIEURE
DE LYON

15 parvis René-Descartes
BP 7000, 69342 Lyon cedex 07
Tél. +33 (0)4 37 37 60 00
www.ens-lyon.fr

Portail Galaxy au Centre Blaise Pascal

requête des laboratoires de biologie de l'ENS-Lyon

Emmanuel Quémener

UNIVERSITÉ DE LYON

ENS DE LYON

1. Contexte

1.1 Où en sommes-nous ?

Dans un courrier daté du 20 janvier 2015, les directeurs des laboratoires IGFL[1], LBMC[2] et RDP[3] de l'ENS-Lyon s'exprimaient sur leur utilisation massive de séquenceurs haut-débit et la conséquente nécessité de disposer d'une infrastructure spécifique pour le traitement de leurs données. Pour se faire, un « standard de fait » a émergé au sein de leur communauté, le portail Galaxy, développé par la *Penn State University*. Début 2014, l'IGFL avait mis en place un premier portail pour ses besoins propres : son exploitation avait notamment permis l'organisation de trois sessions de formation continue sur l'utilisation de Galaxy pour 35 personnes issues de la recherche. De plus, cette plate-forme avait aussi servi de socle pour plusieurs séances de travaux pratiques de L3 et M1 de la formation initiale de l'établissement.

1.2 Où voulons-nous aller ?

Malheureusement, la version de portail Galaxy installée à l'IGFL pour ses besoins propres ne pouvait, ni évoluer, ni accueillir de nouveaux utilisateurs venant de tous les autres laboratoires de l'établissement.

Les laboratoires de biologie IGFL, LBMC et RDP ont donc sollicité Emmanuel Quémener, responsable technique du Centre Blaise Pascal, pour la mise en place et le test d'un portail Galaxy destiné à tous les laboratoires de biologie de l'établissement, y ajoutant par la même occasion le laboratoire CIRI[4], cité en épilogue de la lettre du 20 janvier 2015.

1.3 Comment y allons-nous ?

Emmanuel Quémener a répondu favorablement à cette requête de mise en place d'un portail Galaxy, mais dans le cadre des missions d'« hôtel à projets »du Centre Blaise Pascal. Ainsi convient-il de rappeler le mode de fonctionnement maintenant largement éprouvé des « paillasses numériques » du CBP : ce sont des ressources informatiques attribuées dans le cadre d'un projet de sciences numériques sur un laps de temps limité. Associées aux plateaux techniques du CBP, ces « paillasses » ont donc pour objectif l'expérimentation (la frontière entre science & technologie), le développement (de nouveaux codes de calcul), l'intégration (de codes sur des infrastructures diverses), la qualification (matérielle, système ou logicielle). Le CBP n'a donc pas vocation à fournir une infrastructure de « production » : l'ENS-Lyon dispose dans ses murs du méso-centre Pôle Scientifique de Modélisation Numérique pour ces missions. Cependant, le CBP peut établir finement quelles

1 Institut de Génomique Fonctionnelle de Lyon

2 Laboratoire de Biologie Moléculaire de la Cellule

3 Laboratoire de Reproduction et Développement des Plantes

4 Centre International de Recherche en Infectiologie

spécifications fonctionnelles et techniques sont nécessaires pour la mise en production au PSMN d'une telle plate-forme de production.

Dans ce projet de « paillasse numérique » de portail Galaxy pour les laboratoires de biologie et les départements d'enseignement, les points suivants seront nécessaires :

- une amélioration continue sous forme de déploiement successifs de portail Galaxy permettant, au fur et à mesure des évolutions, de coller le mieux possible aux besoins ;
- un investissement nécessaire et continu des différents demandeurs à savoir les utilisateurs pilotes de tous les laboratoires de biologie ;
- un rétro-planning fixé sur la session de travaux pratiques du 30 avril 2015 et donc la mise à disposition en test de la plate-forme au plus tard le 1er avril 2015 ;
- une perspective de création d'un portail Galaxy exploitant les ressources de calcul du PSMN et donc l'exploitation d'une infrastructure fonctionnellement comparable.

2. Éléments du projet « portail Galaxy à l'ENS-Lyon »

De manière à fixer plus finement les tenants et les aboutissants du projet, nous appliquons la démarche analytique de questionnement basée sur les réponses aux questions CQQCOQP (Comment, Quand, Qui, Combien, Où, Quoi, Pourquoi).

2.1 Pourquoi ?

Offrir aux laboratoires de biologie de l'ENS-Lyon un portail permettant de faciliter l'accès au traitement des données biologiques.

2.2 Quoi ?

Mise en place d'un portail présenté comme un standard dans la communauté des biologistes, le portail Galaxy. Ce portail devra :

- être le plus proche possible de la version standard fournie par l'université porteuse ;
- supporter la charge de plusieurs dizaines de personnes exploitant simultanément le portail ;
- offrir un accès immédiat pour les personnes de l'établissement ;
- offrir un accès simplifié pour des personnes extérieures à l'établissement ;
- disposer de méthodes permettant le chargement et le déchargement de gros volumes ;
- être scalable pour permettre une intégration simplifiée au centre de calcul de l'école.

2.3 Pour qui ? Par qui ?

2.3.1 Pour qui ?

Le public clairement identifié pour l'exploitation le portail Galaxy est très hétérogène :

- des chercheurs en biologie mais n'ayant pas de pratique particulière des centres de calculs ;
- des ingénieurs d'exploitation de plates-formes expérimentales notamment des séquenceurs ;
- des étudiants en stage de quelques semaines voire quelques mois ;
- des étudiants de formation initiale.

Cette forte hétérogénéité des utilisateurs impose des contraintes opérationnelles fortes : plusieurs dizaines de personnes exploitant le portail simultanément pour une session de travaux pratiques d'étudiants de L3 à une très forte sollicitation, sur un laps de temps très court, pour le traitement de données issues d'un séquenceur par un seul individu.

2.3.2 Par qui ?

- Pour la partie informatique : réalisation de la « paillasse numérique » par Emmanuel Quémener au CBP avec la perspective d'une intégration simplifiée auprès du méso-centre PSMN ;
- Pour la partie fonctionnelle : désignation de responsables fonctionnels pour l'installation et la vérification de fonctionnement des greffons Galaxy.

La question de « gouvernance » associée à l'exploitation du portail est identifiée comme un point clé : n'étant pas biologiste de formation, le responsable de la partie informatique ne peut pas évaluer le caractère fonctionnel d'un composant Galaxy qu'il a installé !

2.4 Où et quand ?

Le portail Galaxy expérimental ne sera disponible que sur le réseau interne de l'ENS-Lyon.

- Début Mars 2015 : disponibilité pour évaluation par quelques utilisateurs tests ;
- Fin avril 2015 : disponibilité « en charge » pour une formation de 40 étudiants de L3 ;
- Point d'étape en juin 2015 : ouverture plus large du nombre d'utilisateurs ;
- Juillet - octobre 2015 : exploitation par de nouveaux utilisateurs ;
- Point d'étape en octobre 2015 : évaluation de la pertinence de poursuivre.

La disponibilité du portail circonscrit au périmètre de l'ENS-Lyon obéit à deux contraintes :

- contrainte de sécurité sur le portail et ses composants :
  - les composants internes de Galaxy doivent être évalués avant toute ouverture globale sur Internet. La sécurité associée aux différents composants est largement transférée sur les développeurs du portail originel ;
  - de manière à éviter que les tuples identifiant/mot de passe circulent en clair sur Internet, un accès chiffré (protocole HTTPS) doit être mis en place avec la fourniture de certificats ;
- contrainte sur la réactivité de la commission Web de l'ENS-Lyon : toute demande de site diffusé à l'extérieur de l'établissement doit être examiné et respecter certaines contraintes cosmétiques.

2.5 Combien ?

Dans un premier temps, les ressources matérielles exploitées sont celles déjà disponibles au CBP. Aucun achat de matériel n'a été réalisé à l'exception d'un disque dur SSD de 1TB. De fait, les ressources matérielles mobilisées dès l'origine ont été les suivantes :

- un Sunfire x4150 avec 32GB, équipé de 6 disques durs de 1TB et un disque SSD de 1TB ;
- une grappe de nœuds modulables de 8 à 24 Sunfire x4150 pour traiter les requêtes ;
- la frontale d'administration pour la grappe de nœuds ;
- toute l'infrastructure réseau associée.

2.6 Comment ?

2.6.1 Clarification sur les paillasses numériques

Avant de développer plus spécifiquement les différentes étapes dans la réalisation de ce portail Galaxy, revenons sur la nature des « paillasses numériques » proposées par le Centre Blaise Pascal et leur dessein :

- Expérience : environnement dédié à l'exploration de nouvelles approches et leur métrologie ;
- Démonstrateur : intégration sur un même socle de nouvelles approches déjà évaluées de manière indépendante ;
- Prototype : définition du système optimal pour une mise en production.

Une « paillasse numérique » de « production » du Centre Blaise Pascal n'a pas vocation à exister, sinon pour ses besoins propres : c'est le rôle de structures dédiées, comme le Pôle Scientifique de Modélisation Numérique ou la Direction des Systèmes d'Information, d'offrir le niveau de service associé.

Le portail Galaxy réalisé par le CBP pour cette requête des laboratoires de biologie (et du département d'enseignement) se situe donc, dans sa première mouture, dans les catégories de l'expérience, voire du démonstrateur. La phase de prototypage sera définie sur l'expérience acquise par les 6 premiers mois d'exploitation de la « paillasse numérique » construite.

2.6.2 Réponse aux spécifications fonctionnelles

Si nous revenons au projet originel de portail Galaxy, pour chacune des spécifications fonctionnelles, les solutions suivantes sont mises en œuvre :

- supporter plusieurs dizaines de personnes : distribuer les requêtes du portail sur une grappe de calcul en arrière-plan pour soulager la plate-forme. Ceci exige une interaction avec le gestionnaire de tâches de la grappe ;
- simplifier l'accès aux personnes de l'établissement : exploiter l'authentification de la DSI de l'ENS-Lyon au travers du service LDAP ;
- offrir un accès temporaire aux personnes extérieures à l'établissement : exploiter les

comptes Wifi invités de la DSI au travers de l'annuaire LDAP ;

- disposer de méthodes permettant les chargement/déchargement de gros volumes : exploiter le protocole de transfert de gros fichiers par FTP utilisant l'authentification de la DSI.

2.6.3 Partage des responsabilités

- Emmanuel Quémener : installation, maintenance, évolution de la plate-forme matérielle et logicielle ;
- Bio-informaticiens des unités : vérification, installation composants, contact utilisateurs.

3. Déploiement & Expérimentations

Si installer un portail Galaxy sur son équipement personnel semble « facile » (il existe même des machines virtuelles complètes[5]), l'intégration d'un portail Galaxy modulaire interconnecté à une infrastructure existante et destiné à plusieurs dizaines d'utilisateurs n'est pas aussi immédiate que cela.

Avant de parvenir à la version mise à disposition des utilisateurs de test, plusieurs explorations préliminaires ont été nécessaires, notamment pour appréhender comment le portail Galaxy, utilisé en frontale, pouvait s'interfacer avec une grappe de calcul existante. Le serveur devant héberger le portail Galaxy a été installé dans une machine virtuelle sous KVM[6]. Les spécifications techniques des serveurs hôte & Galaxy étaient les suivantes :

- serveur hôte : Sunfire x4150 avec 2 disques systèmes, 5 disques en RAID5 de 1TB, 1 disque SSD de 1TB, 32GB de mémoire vive et 2 processeurs Intel E5440 équipés de 4 cœurs chacun. Le système d'exploitation est une Debian/Linux Jessie AMD64, la couche de virtualisation KVM et le système de fichiers partagé avec les machines virtuelles du ZFSonLinux ;
- serveur virtuel Galaxy : 6 cœurs sur 8 dédiés, 16GB de mémoire vive, une partition système de 100GB et une partition du disque SSD de 500GB sur le 1TB total. Le système d'exploitation est une Debian/Linux Jessie AMD64.

L'exploitation d'une machine virtuelle plutôt qu'une machine physique se justifie par sa flexibilité :

- pour l'allocation de ressources dynamiques en mémoire et en ressources de calcul (et donc évaluer ce qu'il sera strictement nécessaire sur un portail de production) ;
- pour la sauvegarde du portail avant toute modification trop risquée (et donc offrir un retour en arrière rapide en cas de souci majeur) ;
- pour le déplacement du portail sur d'autres hôtes physiques plus musclés en cas d'insuffisance de ressources ;
- pour un clonage facilité du système complet (ouverture de frontales multiples).

3.1 Déploiements successifs du portail Galaxy

5 https://wiki.galaxyproject.org/VirtualAppliances

6 Kernel Virtual Machine : méthode de virtualisation matérielle du noyau Linux

3.1.1 Premier déploiement

Le premier déploiement a été réalisé sur la machine virtuelle dans un dossier dédié au logiciel du portail Galaxy proprement dit, afin d'évaluer la nature de requêtes entrées et sorties.

L'interconnexion avec le système d'information de l'école a été validée : pour accéder au portail, il suffit de demander à Emmanuel Quemener de rajouter l'identifiant de l'utilisateur. L'utilisateur dispose alors d'un accès à toutes les ressources du CBP et donc du portail Galaxy. Le protocole utilisé pour valider l'authentification est Ldap : ce service offert par la DSI s'appuie sur le système d'information de l'établissement. De plus, pour les utilisateurs ne disposant pas de compte ENS-Lyon, il suffit de créer un parrainage Wifi (service offert par la DSI) et de fournir l'identifiant provisoire préfixé par z. La validité de ce compte temporaire est ainsi gérée au niveau de l'établissement.

Lors de la tentative d'interfaçage avec la grappe de calcul, il est apparu que le dossier complet du portail devait être accessible par tous les nœuds, et tous les travaux soumis au portail Galaxy étaient à exécuter par le même utilisateur.

3.1.2 Second déploiement

Ainsi, toutes les requêtes sont exécutées par un utilisateur « galaxy » à l'intérieur d'un dossier unique « galaxy », partagé par toute l'infrastructure : la solution naturelle est donc de confier à la frontale de la grappe, également serveur de dossiers des utilisateurs, la gestion de cet espace. Cette solution était fonctionnelle et permettait instantanément aux nœuds de la grappe d'accéder à ce dossier partagé. Cependant, la frontale de la grappe supportait toute la charge entrée/sortie du portail Galaxy d'un côté et toute la charge entrée/sortie des nœuds esclaves du portail.

Ce second déploiement a néanmoins permis de valider quel pouvait être le gestionnaire de tâches exploité par le portail afin de soumettre les travaux aux nœuds de la grappe. En effet, Le CBP utilisait comme gestionnaire de tâches OAR depuis 4 ans. Le PSMN utilise pour des raisons historiques le gestionnaire GridEngine. Interfacer le portail Galaxy avec OAR a été essayé, mais cela se révéla impossible : le liant indispensable étant largement imparfait. Seul le même gestionnaire de tâches que le PSMN a permis au portail de distribuer efficacement les tâches auprès des nœuds : GridEngine a donc été déployé sur l'infrastructure du CBP. Toutefois, si ce portail était fonctionnel techniquement, très vite sont apparues des limites dans son exploitation opérationnelle : en effet, la frontale de la grappe de calcul, complètement submergée par la sollicitation, d'un côté permanente de la frontale Galaxy, de l'autre par les nœuds, se retrouvait saturée au point de devenir inaccessible pour toute autre sollicitation. Il fallait donc trouver une autre solution.

3.1.3 Troisième déploiement

Face à la surcharge de la frontale de la grappe, la seule solution était, d'une part, de disposer d'un accès direct du portail Galaxy au réseau d'interconnexion des nœuds et d'autre part, de proposer le montage direct des nœuds du dossier nécessaire du portail Galaxy. Ainsi, à chaque requête, le nœud accède au dossier galaxy pour ses besoins de lecture/écriture.

Ce partage direct entre le portail Galaxy et les nœuds a permis le libérer la frontale de tout trafic inutile et ainsi permettre de mener des tests plus aboutis. Au-delà de l'authentification validée dans le premier déploiement, de la gestion des tâches par GridEngine validée dans le second, de la distribution du montage avec les nœuds dans le troisième, un service d'importation de gros fichiers a été validé, lui aussi exploitant l'authentification de l'établissement. C'est cette version qui a été exploitée par les utilisateurs tests, mais aussi par les étudiants de L3 lors de la séance du 30 avril. A ce déploiement était associé un ensemble de 8 machines disposant de 8 cœurs pour le traitement des requêtes de la frontale.

3.1.4 Quatrième déploiement

A la suite de la session de travaux pratiques du 30 avril, les utilisateurs test ont sollicité la frontale en stockant des données au-delà de ses capacités nominales : le volume de 500GB s'est retrouvé saturé sans possibilité de le purger sans risque pour son intégrité totale. Il a donc été nécessaire de déplacer la machine virtuelle hébergeant le portail Galaxy sur un autre hôte en offrant un espace de stockage plus important. C'est sur cet équipement qu'est actuellement en service le portail Galaxy. Cette migration de socle technique a également permis d'étendre la mémoire vive de la machine virtuelle de 16GB à 24GB (+50%), son nombre de cœurs dédiés de 4 à 6 (+50%) et l'espace de stockage dédié au portail de 500GB à 4TB (+700%).

De plus, ce portail Galaxy reste toujours seulement accessible de l'intérieur de l'ENS-Lyon : il est donc nécessaire d'être connecté dans l'établissement pour y avoir accès. Toutefois, pour les utilisateurs souhaitant l'exploiter de l'extérieur de l'établissement, le CBP offre (et donc de tous ceux du portail) un accès via l'application x2go[7] : ils ont alors, quel que soit leur système (GNU/Linux, Windows ou MacOSX), la possibilité d'ouvrir un bureau virtuel sur les machines de la salle en libre service du CBP[8] pour ensuite y ouvrir un navigateur. De manière à augmenter la sécurité par la confidentialité de l'identifiant/mot de passe, un service HTTPS a été ouvert sur le portail Galaxy. Ce service exploite des certificats certifiés par l'autorité du CNRS. En effet, la DSI de l'ENS-Lyon ne veut pas diffuser de certificats pour les sous-domaines de l'ENS-Lyon. En définitive, l'accès actuel à la frontale se fait donc par :

- http://galaxy.cbp.ens-lyon.fr : accès non sécurisé, redirigé vers le service sécurisé ;
- https://galaxy.cbp.ens-lyon.fr : accès sécurisé.

Pour ne pas avoir de « vociférations » de son navigateur trop « zélé » lors de la première connexion (quand il ne refuse pas catégoriquement d'accéder au service), il est nécessaire d'intégrer l'autorité de certification du CNRS[9].

3.2 Expérimentations

3.2.1 Test grandeur nature : formation du 30 avril 2015

7 http://wiki.x2go.org/doku.php

8 http://www.cbp.ens-lyon.fr/doku.php?id=ressources:ressources

9 https://igc.services.cnrs.fr/load_all_certificate/?CA=CNRS2-Standard&lang=fr&ca=CNRS2-Standard

Cette formation de L3 en biologie a été très instructive puisqu'elle a relevé nombre de limitations qui n'avaient pas été relevées lors de la première phase d'évaluation (période de mars et avril). Ainsi a-t-il fallu, en « *live* », modifier de nombreux réglages pour permettre aux 40 étudiants de compléter leur séance de travaux pratiques sur cette durée de deux heures. Pour mémoire, cette formation s'est déroulée sur le troisième déploiement du portail et les ressources associées.

Tout d'abord est apparu un problème d'authentification, lequel qui a exigé le redémarrage du service local associé à ce service de l'établissement : malheureux désagrément impossible à anticiper. Puis a surgi un problème sur le proxy Web, interface entre le portail Galaxy et un accès Web standard : un passage de 16 à 64 sessions simultanées s'est avérée indispensable. Ensuite, sur les réglages des services du portail Galaxy proprement dit, un utilisateur mobilisait à lui seul toutes les ressources : restreindre le nombre de tâches concurrentes qu'un utilisateur pouvait lancer simultanément fut nécessaire. Enfin, côté matériel, pour faire face à une charge processeur très importante et un début d'utilisation du cache disque, la machine virtuelle fut portée de 16GB à 24GB pour sa mémoire vive et de 4 à 6 cœurs dédiés. La grappe de nœuds disponibles fut étendue de 8 à 16 unités dédiées.

Au terme des deux heures de travaux pratiques, 295 tâches Galaxy ont été exécutées par 22 utilisateurs différents et totalisant de 2 à 48 jobs par utilisateur (un des utilisateurs ayant réussi à se connecter ouvrant alors une session pour ses camarades sur d'autres machines).

Côté métrologie réseau, vers l'extérieur, le portail Galaxy a totalisé 2.5GB de données transmises et 1.4GB de données reçues. Vers les nœuds de la grappe, 57.5GB ont été reçus pour 112.1GB transmises.

La partie de la séance la plus invasive sur les ressources était le transfert de HG19 du portail Galaxy vers chaque nœud (3GB soit 27 secondes minimum de transfert par utilisateur). Ainsi, le lien interne du réseau était saturé durant toute cette période de la séance.

En conclusion, le principal goulot d'étranglement révélé ne résidait pas, comme supposé, dans les accès disques mais sur l'interconnexion entre les nœuds et le portail Galaxy. C'est donc sur ce verrou qu'il faudra prioritairement travailler pour le prochain prototype.

3.2.2 Exploitation du portail sur 6 mois

Le portail Galaxy a été ouvert aux utilisateurs de test à partir du début de mois de mars puis à tous les étudiants fin avril. Quelques utilisateurs ont été rajoutés de manière très sporadique entre mai et septembre. Sur cette période, 2567 requêtes Galaxy ont été traitées (requêtes de test du staff technique exclues).

L'illustration 1 page 11 montre l'exploitation du portail sur cette période. Trois types de comportements sont observés : le premier est très dense sur un laps de temps très court, c'est le cas de formations où beaucoup d'utilisateurs différents vont lancer, le même jour un grand nombre de requêtes (le 30 avril 2015 par exemple). Le second est dense mais se répartit sur quelques jours : c'est le cas de traitements massifs de données issues de séquenceurs (première semaine de juillet). Le troisième constitue un usage résiduel avec cependant quelques pics d'activité.

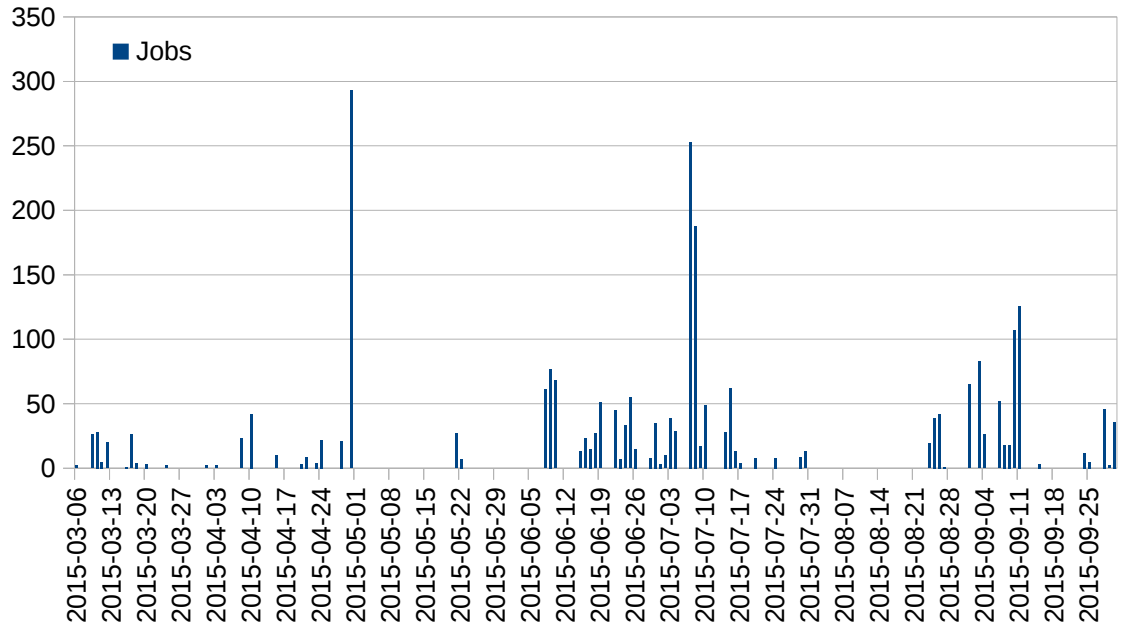

Illustration 1: Exploitation du portail Galaxy sur 6 mois

Il est aussi intéressant d'associer le nombre de requêtes à l'utilisateur. C'est ce que représente le camembert 2 page 11. Ainsi, trois utilisateurs se répartissent plus de 80 % des tâches soumises au portail. Qui plus est, ces trois utilisateurs sont issus de l'IGFL, ce qui illustre une réelle asymétrie entre les laboratoires de biologie pourtant demandeurs de la plate-forme et leur exploitation durant cette période de six mois.

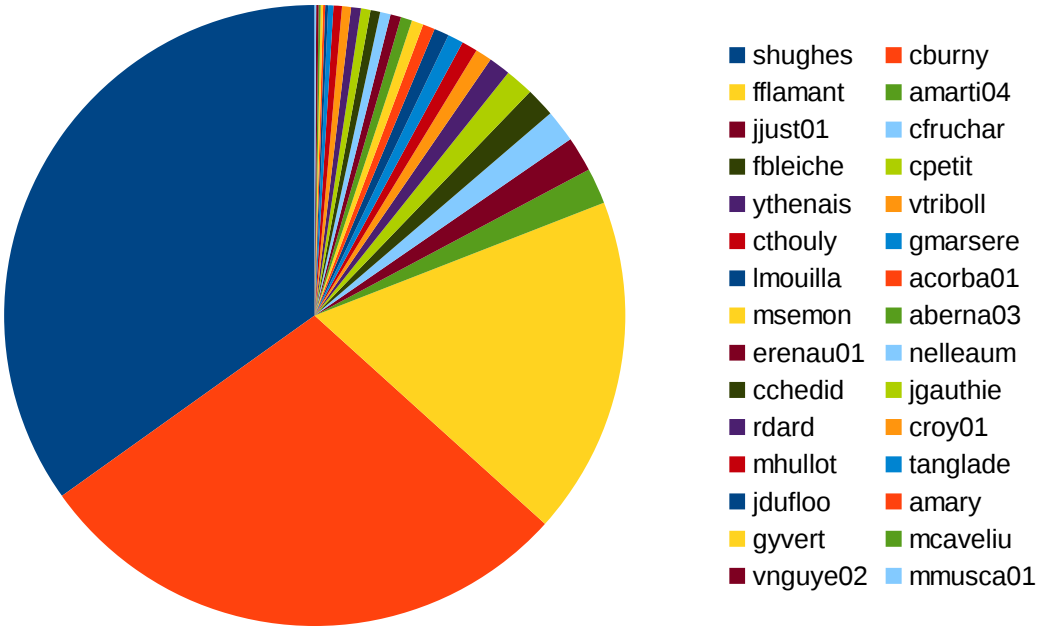

Illustration 2: Exploitation du portail Galaxy par utilisateur

Pour finir, il est aussi intéressant de regarder quelle tâches ont été exécutées sur le portail Galaxy. C'est ce qu'illustre le camembert 3 page 12. Ainsi, les premiers traitements sont des chargements et des opérations de recherche ou de transformation élémentaire.

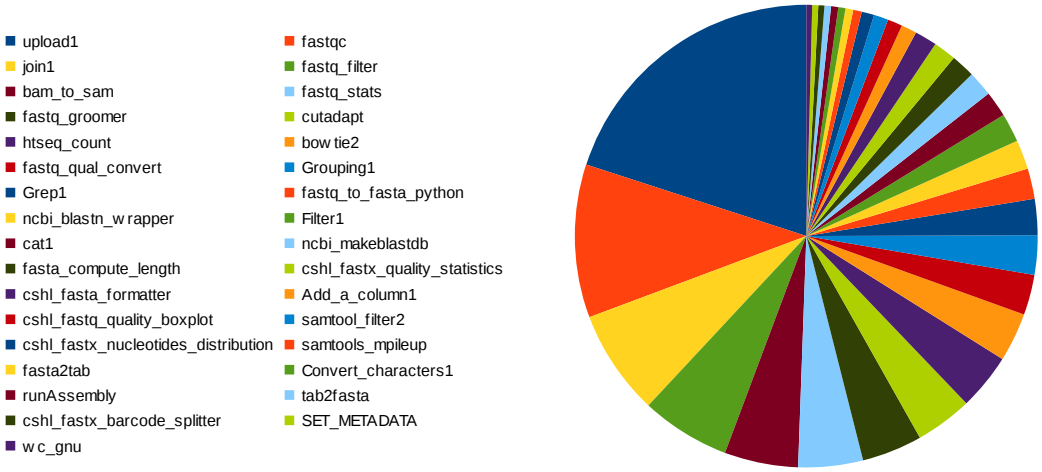

Illustration 3: Distribution des tâches par traitement

4. Conclusion

4.1 Sur 6 mois d'exploitation, quels usages ?

Deux usages sont clairement identifiés : la recherche d'un côté et l'enseignement de l'autre. Ainsi, les usages sollicitent le portail Galaxy et l'infrastructure associée de manière très différente :

- Recherche :
  - Activité « creuse » : quelques utilisateurs simultanés, souvent un seul,
  - Volumes entrée/sortie difficiles à anticiper mais nécessitant un service dédié,
  - Calculs de durée difficile à anticiper ;
- Enseignement :
  - Activité « dense » : plusieurs dizaines d'utilisateurs simultanés,
  - Volumes entrée/sortie raisonnable et prévisible,
  - Calculs de durée raisonnable et prévisible : quelques minutes.

Ce grand écart entre ces usages va exiger de l'infrastructure une polyvalence difficile à appréhender sans une métrologie plus fine, notamment pour les traitements liés à la recherche. Cette métrologie, pour être pertinente, exige que TOUS les acteurs demandeurs exploitent le portail. Force est de constater que l'IGFL s'est investi largement plus que les autres laboratoires ces six derniers mois.

4.2 Quelles plates-formes aujourd'hui & demain ?

Le portail Galaxy *hic et nunc* dispose des ressources dédiées suivantes :

- Un hôte de machines virtuelles Dell R410 disposant de 4 disques de 2TB, 32GB de RAM, 8 cœurs physiques (16 logiques), une interconnexion 10G partagée entre le réseau interne à la grappe et le réseau externe. Il héberge le seul portail Galaxy ;
- Une machine virtuelle disposant de 6 cœurs, de 24 GB de RAM, d'un espace de stockage pour les données Galaxy de 4TB dont 2.9TB sont libres début novembre ;
- 8 nœuds Sunfire x4150 avec 32GB de RAM, 8 nœuds Dell R410 avec 24 GB de RAM, 5 nœuds HP Moonshot de capacités différentes soit un total de 232 cœurs logiques ;
- un réseau Gigabit Ethernet avec les nœuds et 10G avec le portail Galaxy.

Sur ce portail ont été intégrées toutes les applications (disponibles sous formes de greffons) demandées par les laboratoires de l'IGFL et du LBMC. Au 5 novembre étaient disponibles et opérationnels pas moins de 173 greffons différents.

Les premières expérimentations et leur métrologie ont permis de définir les spécifications matérielles du futur portail et de préparer cet équipement :

- Un hôte de machines virtuelles Dell R510 disposant de 12 disques SSHD de 4TB, 64GB de RAM, 12 cœurs physiques (24 logiques), une interconnexion 10G avec le réseau externe, 10G avec le réseau interne, une interconnexion InfiniBand QDR à 40GB/s ;
- Une machine virtuelle avec 8 cœurs physiques dédiés, 48GB de RAM, 32TB de stockage utile pour les données Galaxy et une interface InfiniBand très haut débit dédiée ;
- Les mêmes 22 nœuds de grappe pour 232 cœurs exploitables avec la possibilité d'y associer en fonction de la demande d'autres nœuds issus des clusters d'expérimentation ;
- Une interconnexion InfiniBand entre le portail Galaxy et les nœuds.

Ce futur portail, plutôt à classer dans la catégorie du démonstrateur, voire du prototype, intégrera :

- une évolution sur la base de données du portail Galaxy pour une meilleure gestion des accès concurrentiels : lorsque plusieurs utilisateurs se connectaient simultanément, la base de données actuelle et locale montrait d'étranges comportements qui empêchaient ensuite un nettoyage des données temporaires ;
- sur le gestionnaire de tâches : le gestionnaire GridEngine, exploité par le portail Galaxy (et aussi le PSMN) n'a plus de support dans la distribution Debian, signe de son archaïsme. Il est envisagé de migrer sur un gestionnaire de tâches largement exploité par la communauté HPC et réputé pour sa scalabilité : Slurm.

4.3 Quel bilan pour quelle suite ?

Les laboratoires de biologie de l'ENS-Lyon IGFL, LBMC et RDP ont sollicité le Centre Blaise Pascal pour la fourniture d'un portail Galaxy fin janvier 2015.

Le CBP a répondu à cette requête en offrant, dès début mars 2015, un portail Galaxy de test, lequel a bénéficié d'une amélioration continue depuis cette date.

Ce portail Galaxy de test a traité près de 3000 tâches sur cette période mars-octobre 2015. Il a été exploité par le département d'enseignement en avril dernier, le sera aussi en décembre prochain. Bien qu'ayant bénéficié à plusieurs dizaines d'utilisateurs, seuls trois du même laboratoire l'ont véritablement exploité de manière dense et récurrente.

La réalisation de ce premier portail et son exploitation ont mobilisé le responsable technique du CBP, Emmanuel Quémener, pour une période consolidée d'au moins trois mois sur l'année 2015. De plus, le portail Galaxy réquisitionne également un ensemble de ressources matérielles non négligeables. En outre, un second portail, aux spécifications techniques héritées de l'expérience acquise depuis mars, est en préparation mais son installation reste suspendue…

Cependant, le CBP a pointé du doigt le manque patent d'utilisateurs de tests dans l'exploitation de cet outil mis à disposition, lourd à installer et à administrer.

Il revient donc aux directeurs des unités de biologie de maintenant de s'exprimer sur la poursuite ou non du travail entrepris au Centre Blaise Pascal et sur la pérennisation de ce nouveau plateau technique numérique.