Lettre d'information

Avril 2023

Cette lettre d'information est destinée aux membres des équipes de recherche utilisant la plate-forme bio-informatique GenoToul. Elle a pour but de répondre aux questions et commentaires que vous nous avez fait remonter via le questionnaire de satisfaction annuel et de vous annoncer la prochaine mise en production du nouveau cluster.

Logiciels

Pour vous aider, nous avons mis en place des exemples d'utilisation des logiciels dans les répertoires d'installation (/usr/local/bioinfo/src/) et sur notre *site web*. De plus, nous mettons à votre disposition une FAQ dédiée. Par exemple la commande `module avail -t 2>&1 | grep -i soft_name` est utile lorsqu'on cherche un logiciel spécifique dans la liste des modules.

Les logiciels et les workflows sont installés sur demande via un formulaire *Ask / Software installation*.

Il peut arriver que le logiciel ne soit pas installable sur notre infrastructure à cause de conflits de dépendance. N'hésitez pas à aller voir à l'IFB-core et sur le forum de la communauté france bioinformatique : *https://community.france-bioinformatique.fr*

Banques

Les banques sont installées dans /bank. Seules la version courante et la précédente sont gardées, donc attention à l'utilisation des liens symboliques. Certains index bwa, star, bowtie et blast+ sont également générés dans les répertoires suivants /bank/bwadb/, /bank/STARdb/, /bank/bowtie2db/, /bank/blastdb/. Les banques mises à jour via biomaj (et uniquement elles) sont listées sur *cette page* (login nécessaire). Ce lien est accessible en cliquant sur l'image de *cette page* de notre site web.

Retours sur l'enquête de satisfaction

Vous avez été nombreux (123 soit près de 12%) à répondre à notre questionnaire de satisfaction annuel et nous vous en remercions chaleureusement. Suite à vos retours nous avons rajouté des items à notre FAQ, par exemple, concernant les options de `srun`.

Si vous rencontrez des problèmes pour lancer un job ou utiliser un logiciel, n'hésitez pas à nous poser vos questions via le *formulaire* ou le mail du support : *support.bioinfo.genotoul(at)inrae.fr*.

Le principal facteur limitant à l'utilisation du cluster est la quantité de RAM (mémoire vive). Afin de réduire les temps d'attente en queue, il est utile d'ajuster au mieux les réservations de CPU et de mémoire. Pour se faire, la commande `seff` sur un job similaire qui s'est bien terminé vous permettra de connaître la mémoire utilisée. Nous avons tenu compte de ce besoin en RAM dans l'acquisition du nouveau cluster. Le cluster actuel est équipé de 256Go de RAM par nœuds, le nouveau cluster, qui sera mis en production dans les prochaines semaines, sera, quant à lui, équipé de 2048Go par nœud. Nous espérons que ce choix va permettre de répondre aux principaux besoins en RAM.

Si vous souhaitez vous loguer en interactif pour un temps très court, il est possible d'utiliser le paramètre `-t` ou `-time` (HH:MM:SS), par exemple `srun -pty bash -t=00:05:00` (5mns) pour passer en priorité.

Nous vous rappelons qu'il est possible de louer de l'espace disque work supplémentaire via une *demande de ressources* sur notre site web. Le tarif s'explique par le fait que le disque a été optimisé pour la capacité et les performances spécifiquement sur notre cluster de calcul. Il est composé d'une agrégation de matériels (6 serveurs, 300TiB disques SSD, 2PiB disques NL-SAS, switch à haut débit et faible latence). De plus il y a des licences logicielles et une garantie matérielle pendant toute la durée d'exploitation de la solution. Enfin, nous payons des frais d'hébergement et d'électricité de l'ordre de plusieurs centaines de milliers d'euros annuels pour le nouveau cluster et son stockage.

Quand utiliser un cluster plutôt qu'un PC ?

Un cluster est un ensemble de serveurs de calcul appelés nœuds qui sont interconnectés. Les nœuds peuvent être différents les uns des autres, par exemple il y a un nœud de login, des nœuds de calcul standards, des nœuds à forte mémoire, des nœuds GPU…. Il est possible de se connecter aux nœuds de login via ssh. Notre cluster est décrit sur notre site web sur cette *page* (pour avoir plus de détail, cliquer sur chaque composant de l'image). Il s'agit d'une ressource partagée entre plus d'un millier d'utilisateurs.
Il est utile d'utiliser un cluster de calcul plutôt qu'un ordinateur personnel dans les cas suivants :
- besoin de plus de ressources que disponible sur son PC (RAM, CPU, espace disque)
- le même job doit être lancé plusieurs fois sur différentes données (possibilité de le lancer en parallèle sur plusieurs nœud)
- le programme que vous souhaitez lancer gagnerait à être « multi-threadé » via MPI ou OpenMP.

Cependant, l'utilisation d'un cluster ne rend pas toujours le calcul plus rapide. Il peut arriver que votre job sur le nœud du cluster partage les ressources avec d'autres jobs : CPU, mémoire, réseau d'accès aux espaces de stockage... Nous observons par exemple des ralentissements possibles sur les I/O (lecture/écriture). Il faut, par conséquent, faire attention à cette contrainte lorsque vous dimensionnez vos jobs, par exemple éviter de lire des milliers de fois le même fichier en parallèle.
Lorsque vous lancez vos jobs, ils seront tout d'abord placés en file d'attente et leur priorité de passage sera calculée par l'ordonnanceur et leur lancement effectif dépendra également des ressources disponibles.

Les images singularity, quant à elles, doivent être construites sur son ordinateur personnel puis utilisées sur le cluster car il est nécessaire d'être root pour faire un singularity build.

Cycles d'apprentissage

Des places disponibles sur nos formations 2023

En partenariat avec Sigenae et SaAB (MIAT), nous vous proposons des cycles d'apprentissage récurrents comme Linux et Cluster (2 fois dans l'année). Il reste actuellement des places pour les sessions suivantes :

Une session *alignement et détection de variants* du 12 au 14 novembre 2023.

Une session sur les « *one line perl* » d'une journée est organisée le 28 novembre 2023.

Les *tarifs* et *le formulaire d'inscription* sont disponibles sur notre site internet.

La plateforme *GenoToul Biostat* met aussi en place différentes formations.
Si vous ne trouvez pas la formation que vous souhaitez, *l'IFB (l'Institut Français de Bioinformatique)* référence une grande diversité de cycles d'apprentissage en bioinformatique, comme des sessions sur la phylogénie ou l'assemblage et l'annotation. Pensez également aux formations permanentes de vos différents instituts. Elles proposent de nombreuses formations en informatique.

Le coin des débutants

Nous essayons de centraliser au maximum les informations importantes pour nos utilisateurs sur notre site web. Vous y retrouverez, par exemple, toutes les newsletters dans le menu about us / *Newsletters*), notre FAQ (menu *FAQ*), les formulaires de contact à propos de nombreux sujets tels que la demande de compte, de ressources supplémentaires ou d'accompagnement de projet (*Ask for*). Nous mettons également à votre disposition les supports de formation de l'ensemble des *cycles d'apprentissage* que nous proposons. Vous trouverez en particulier un tutoriel pour vous aider à lancer des *workflows nextflow nf-core*.
Sur cette *page* vous trouverez une liste de petits scripts qui peuvent être utiles. Ils sont dans le répertoire /usr/local/bioinfo/Scripts/bin du cluster. La *charte* d'utilisation de notre plateforme est accessible depuis notre FAQ (User access / « conditions of access and use of the infrastructure »).

Nous rappelons que les serveurs d'accès sont réservés exclusivement à la connexion, au transfert de données, à la compilation, au test rapide de lignes de commandes et à la soumission de jobs sur le cluster de calcul. Nous vous invitons à lancer R, nextflow et snakemake en batch via un script. Les sessions tmux, screen, vscode ne doivent pas rester ouvertes lorsqu'elles ne sont plus utilisées, car elles surchargent les nœuds de login ce qui peut empêcher les utilisateurs de se connecter. Par conséquent, merci par avance de fermer vos sessions et de killer vos process inactifs.
Dans le cas contraire, sans preuve d'utilisation effective, ils seront killés par les administrateurs.

Trucs et astuces sur le cluster

A la fin du contrat des personnels temporaires, leur compte est supprimé. Pensez à récupérer les données et les metadonnées correspondantes avant le départ de vos collaborateurs. Une extension de 2 mois maximum est possible. Ce délai correspond au délai accordé également par INRAE pour l'accès à l'intranet et aux e-mails, nous ne pouvons pas, par conséquent, aller au-delà. Nous sommes désolés de la gène occasionnée.
Chaque compte est personnel, ne communiquez pas vos identifiants de connexion. Pour permettre l'accès à certains de vos répertoires à un collaborateur, il est nécessaire d'utiliser les commandes expliquées dans le FAQ rubrique Linux / question : « How to change permission on file or folder ? ». Sur le save, il faut utiliser la commande nfs4_setfacl, et sur le work setfacl. Des exemples d'utilisation sont mis à votre disposition dans la FAQ.

Spécial nouveau cluster

Brève description de la nouvelle infrastructure de calcul

La plateforme Bioinfo GenoToul vient de faire l'acquisition d'une nouvelle infrastructure de calcul pour remplacer ses équipements vieillissants. Il s'agit d'un cluster de calcul fonctionnant à travers SLURM (Simple Linux Utility for Resource Management) composé de :

- 1 serveur d'administration
- 2 serveurs de login (accès depuis l'extérieur, transferts, compilation, test et soumission sur le cluster de calcul)
- 1 serveur de visualisation (bureau à distance, rendu 3D)
- 1 serveur GP/GPU : pour faire de l'IA
- 39 nœuds de calcul à forte mémoire RAM (2TB)
- +2PiB de stockage parallèle (dont 223TiB rapide)
- Réseau d'interconnexion Infiniband HDR à 100Gbs

Séminaire d'information sur le nouveau cluster

C'est avec un plaisir non dissimulé que nous vous annonçons que nous organisons un séminaire d'information sur le nouveau cluster le 15 mai 2023 de 9h00 à 11h00. Il se déroulera en salle de conférence MIAT (bâtiment C8 du centre INRAE d'Auzeville-Tolosane). Il sera également accessible en visio. Nous vous présenterons la solution choisie et ses principales caractéristiques, les changements par rapport au cluster actuel et les modalités ainsi que le calendrier de migration. Nous répondrons ensuite à vos questions concernant cette nouvelle infrastructure de calcul et de stockage. Nous espérons que vous serez nombreux à ce rendez-vous.

L'EUROPE S'ENGAGE L'OCCITANIE AGIT
UNION EUROPÉENNE
La Région Occitanie Pyrénées · Méditerranée

Projet cofinancé par le Fonds Européen de Développement Régional
Financement dans le cadre de la réponse de l'Union à la pandémie de COVID-19