Contrôler ses jobs

La liste des jobs pris en charge par SGE est accessible via la commande qstat. Afin de filtrer la sortie pour visualiser uniquement ses propres jobs :

qstat -u <user_login>

Un job peut être supprimé à l'aide de la commande qdel :

qdel <job_id>

Toutes ces informations peuvent être visualisées à l'aide de l'interface **qmon**.

L'état du cluster et de l'ensemble des nœuds peut être visualisé à l'aide de l'interface ganglia accessible à l'adresse : http://serviclust.toulouse.inra.fr/ganglia/

Le shell interactif : qlogin

Si vous souhaitez lancer des commandes interactives sans passer par l'écriture d'un script, il suffit de lancer la commande qlogin. Elle exécute un shell sur un des nœuds du cluster.

Les logiciels / banques disponibles sur le cluster

L'ensemble des logiciels mis à votre disposition est disponible dans le répertoire /usr/local/bioinfo/bin avec une description sur le site à l'adresse : http://bioinfo.genotoul.fr/index.php?id=73.

Les banques mises à disposition se trouvent sous /bank. Un test hebdomadaire est réalisé afin de s'assurer de leur cohérence. Vous pourrez trouver le résultat de ce test sur le site à l'adresse suivante : http://bioinfo.genotoul.fr/index.php?id=65.

Afin de rendre l'utilisation du cluster la plus facile possible, l'environnement qui est mis à votre disposition (disques, logiciels, banques) est identique à partir de n'importe quel endroit du cluster.

Quelques logiciels disponibles

Logiciel	Description
Alignement de séquences	
bwa	Alignement via la transformation de Burrows-Wheeler
ncbi-blast	Recherche de similarité dans des banques de données, NCBI Blast
Annotation	
Apollo	Environnement d'annotation expert
ARNnc	
RNA Vienna Package	Dédié à la prédiction / comparaison des ARN
tRNAscan-SE	Prédiction d'ARN de transfert
Assemblage	
newbler	Assembleur de Roche pour le 454
Amos	A Modular, Open-Source whole genome assembler
velvet	Assemblage de novo de « very short reads »
Carte génétique	
Carthagene	Logiciel d'ordonnancement de marqueurs.
Phylogénie / Metagénomique	
Mothur	The one-stop source for microbial ecology needs
Phylip	Package de 34 programmes dédiés à la reconstruction phylogénétique
Utilitaires	
R	Logiciel pour des calculs statistiques
EMBOSS	Programmes qui permettent de couvrir l'ensemble des besoins dans l'exploitation des séquences biologiques

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

MEMENTO CLUSTER DE CALCUL

bioinfo.genotoul.fr

support.genopole@toulouse.inra.fr

Avril 2011

L'environnement de calcul

La plateforme Bioinformatique Genotoul met à votre disposition :

✔ un cluster de 44 nœuds bi-quad cores, soit un ensemble de 352 cœurs pour exécuter vos jobs (4G de RAM par cœur),

✔ bigmem, une machine entièrement dédiée aux calculs coûteux en mémoire possédant 32 cœurs et 256G de RAM,

✔ n46 et n47 possédant 24 cœurs et 144G de RAM,

✔ hypermem possédant 64 cœurs et 1T de RAM accessible sur demande au support,

✔ un espace de stockage « lent » (/save/user), pour vos données, sauvegardé chaque jour et un espace de stockage « rapide » (/work/user) optimisé en écriture pour l'exécution de vos jobs.

Cet environnement de calcul, représenté sur la figure 1, est accessible via le serveur frontal SNP depuis une simple connexion internet.

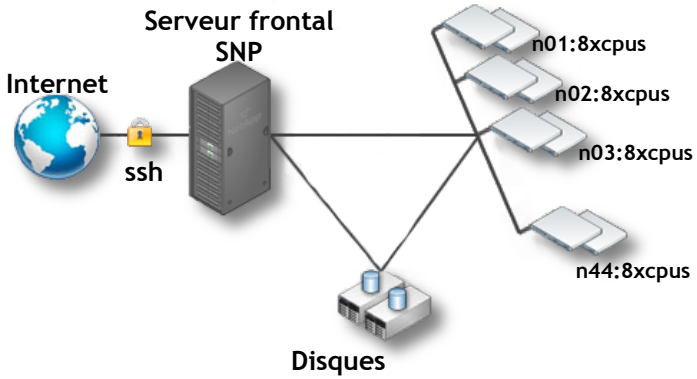

Figure 1 : schéma de l'environnement de calcul.

Connexion au serveur frontal SNP

Depuis **Windows :** Putty et Xming
Depuis **Linux** et **MacOS :** la commande ssh
ssh -X <login>@snp.toulouse.inra.fr

Lancer un calcul

Chaque calcul doit être encapsulé dans un script pour pouvoir être exécuté sur le cluster à partir de SNP. Par défaut tout job s'exécute sur 1 cœur avec 4Go de RAM et sur la queue workq.

Exemple1 : le plus classique
#! /bin/bash
#$ -M <votre-adresse>@mail.fr
#$ -m a
blastall -p blastn -d genbank -i seq.fasta

L'utilisateur disposera d'1 cœur de calcul, de 4G de RAM et de 24h de temps de calcul.

Exemple2 : plus de mémoire et plus de cpu
#! /bin/bash
#$ -M <votre-adresse>@mail.fr
#$ -m a
#$ -q bigmemq
#$ -l h_vmem=50G -l mem=50G -pe parallel_smp 8
blastall -a 8 -p blastn -d genbank -i seq.fasta

L'utilisateur disposera de 8 cœurs de calcul, de 50G de RAM et d'un temps de calcul illimité.

Lancement du script sur le cluster :
[user@snp ~]$ qsub mon_script.sh

Lancer un calcul par lot

Exemple : découper un fichier multifasta et exécuter un blast par fichier découpé.

Découper un fichier multi-fasta :
SplitMultifasta.pl --nb_seq_per_file 200 --output_dir all_fasta

Créer le fichier de blast à l'aide des commandes shell
Comment connaître son shell ? echo $SHELL

Avec le shell tcsh
foreach i (`ls all_fasta/*.fasta`)
foreach? echo "blastall -p blastn -i $i -d refseq_rna -o $i.blast" >> mes_commandes
foreach? End

Avec le shell sh (bash)
$ for i in `ls *.fasta`
> do
> echo "blastall -p blastn -i $i -d refseq_rna -o $i.blast" >> mes_commandes
> done
Lancer les calculs
qarray mes_commandes

Les paramètres de soumission

Paramètres	Utilisation
-M	Adresse email.
-m b|e|a|s	Envoie d'un email au début (b), à la fin (e), à l'arrêt (a) du job ou à la suspension (s).
-N	Nom du script.
-cwd	Lancement du script dans le répertoire courant.
-o	Fichier de sortie du job.
-e	Fichier erreur du job.
-q workq|longq|unlimitq|bigmemq	Spécifie la queue à utiliser pour l'exécution du job.
-l h_vmem *X*G -l mem *Y*G	Spécifie la taille mémoire virtuelle (*X*) et la taille mémoire physique (*Y*) requise par le job.
-pe parallel_*X* Y	Spécifie le nombre de cpu requis par le job (*Y*) dans le mode *X* souhaité (8|smp|fill|rr).

Pour plus d'options sur la soumission : *man qsub*.

Les queues disponibles

Queue	Caractéristiques
workq	Priorité = 300 Temps max = 12h
longq	Priorité = 200 Temps max = 48h
unlimitq	Priorité = 100 Temps max = illimité
bigmemq	Priorité = 0 Temps max = illimité

Pour plus d'information : qstat -q *queue_name*.