FICHE TECHNIQUE CTIG

Cluster CTIG

Diffusée le :

Destinataires :

OBJET – CHAMP D'APPLICATION

Mode d'emploi destiné aux utilisateurs du Cluster mis à disposition par le CTIG afin d'effectuer des tâches qui nécessitent de faire tourner x jobs en simultané.

CONDITIONS REQUISES

Avoir un compte sur le cluster et un client ssh/nx.

DEFINITIONS ET ABBREVIATIONS

Cluster : On parle de grappe de serveurs ou de ferme de calcul (computer cluster en anglais) pour désigner des techniques consistant à regrouper plusieurs ordinateurs indépendants (appelés nœuds, node en anglais), afin de permettre une gestion globale et de dépasser les limitations d'un ordinateur pour :

- augmenter la disponibilité ;
- faciliter la montée en charge ;
- permettre une répartition de la charge ;
- faciliter la gestion des ressources (processeur, mémoire vive, disques dur, bande passante réseau).

Les grappes de serveurs sont un procédé peu coûteux, résidant dans la mise en place de plusieurs ordinateurs en réseau qui vont apparaître comme un seul ordinateur ayant plus de capacités (plus puissant, etc.), très utilisé pour les calculs parallèles. Cet usage optimisé des ressources permet la répartition des traitements sur les différents nœuds.

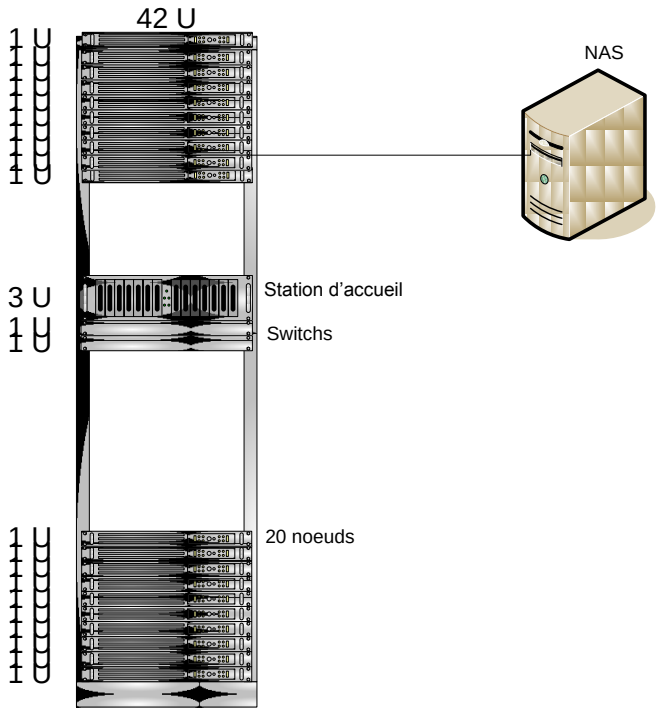
SOMMAIRE

1 PRÉSENTATION DU CLUSTER DU CTIG 3
1.1.1 Caractéristiques du cluster 3
1.1.2 Station d'accueil 3
1.1.3 Nœuds de calculs 3
2 CONNEXION AU CLUSTER 3
2.1.1 Demande de compte 3
2.1.2 Ligne de commande (SSH) 3
2.1.3 Connexion graphique (NX) 3
3 ORGANISATION ESPACE DISQUE 4
4 L'ORDONNANCEUR SGE 5
4.1.1 Présentation de SGE 5
**Durée maximum d'utilisation du processeur 5*
4.1.2 Soumettre un job 5
4.1.3 Suivre l'exécution d'un job 6
4.1.4 Obtenir de l'information sur un job terminé 7
4.1.5 Tuer un job 7
5 LOGICIELS DISPONIBLES 7
6 CONTACTS 7

1 Présentation du cluster du CTIG

1.1.1 Caractéristiques du cluster

- Cluster SGI Altix XE340
- 40 nœuds de calcul et un nœud frontal
- Interconnexion gigabit ethernet
- 2 switchs 24 ports
- 2 switchs 48 ports



1.1.2 Station d'accueil

- Altix XE 270 avec alimentation redondante
- 2 Processeurs Intel Xeon 4 core 5620 2.4 Ghz
- 12 Go de mémoire DDR3
- 2 disques internes SATA II 250 Go, 7200 trs/min
- 4 disques internes SAS 300Go
- 6 emplacements disques

1.1.3 Nœuds de calculs

- 40 nœuds de calcul Altix XE 340
- 20 nœuds 48Go de RAM
- 20 nœuds 96Go de RAM
- 480 cœurs de 2.66 GHz équipés de 4 Go de RAM chacun
- Processeur Intel Xeon 6 core 5650
- 2 Disques SAS 300 Go et 600 Go, 15000 trs/min
- Interconnexion gigabit ethernet

2 Connexion au cluster

2.1.1 Demande de compte

Pour se connecter sur le cluster et soumettre des jobs il faut au préalable faire une demande d'ouverture de compte au CTIG. Pour cela, il faut remplir le formulaire disponible sur le wiki du CTIG (https://ctigwiki.jouy.inra.fr/) et l'envoyer à ctig.systeme@dga.jouy.inra.fr.

2.1.2 Ligne de commande (SSH)

La station d'accueil du cluster se nomme dga11 : elle est accessible par l'adresse dga11.jouy.inra.fr ou via son adresse IP 193.54.97.152

Pour se connecter utiliser un client SSH tel que PuTTY sous Windows, ou SSH sous linux.

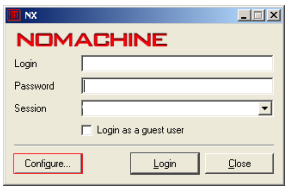
```
ssh user@dga11.jouy.inra.fr
```

Votre mot de passe vous est demandé. Vous arrivez ensuite dans votre répertoire personnel.

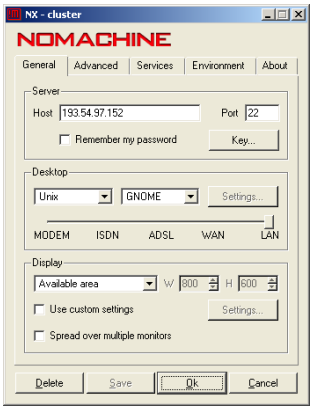
2.1.3 Connexion graphique (NX)

Pour se connecter au bureau graphique GNOME il faut utiliser un client NX (disponible pour Windows Linux et MacOS) : vous pouvez récupérer le client NX Nomachine à l'adresse : http://www.nomachine.com/download.php

Lors de la première connexion vous devez configurer votre accès :



Cliquer sur configure et remplir les champs comme ci dessous :



Cliquer sur OK. Pour se connecter entrer son login/mdp.

3 Organisation espace disque

Tous les utilisateurs ont un répertoire personnel sauvegardé situé dans /home qui contient leurs fichiers de configuration. Ils disposent également d'un répertoire de travail appelé /travail. Ce répertoire de travail est exporté sur chaque nœud du cluster depuis une baie NAS dediée. Il est ouvert à tous les utilisateurs du cluster.

Il existe en plus de ces répertoires, un répertoire de travail situé sur les disques internes des nœuds (/workloc) : il permet de travailler directement sur le nœud ce qui permet un gain non négligeable de performance, en évitant des opérations de lecture écriture sur des disques distants. Cependant il faut bien veiller à copier ses données au début et a la fin des travaux.

D'autres parts vos répertoires habituels du NAS sont accessibles en lecture/écriture depuis chaque nœud et sur la station d'accueil.

En fin de calcul ne pas oublier de recopier les fichiers de sortie sur vos répertoires habituels et de faire le ménage dans les répertoires de travail du cluster.

Un script est disponible pour effectuer du ménage sur le cluster après l'exécution d'un job : qls pour lister vos fichiers, et qclean pour les supprimer.

4 L'ordonnanceur SGE

4.1.1 Présentation de SGE

L'ordonnanceur utilisé pour soumettre des jobs de calcul sur le cluster du CTIG est SGE (Sun Grid Engine), c'est un logiciel open source dont la documentation complète est disponible librement à l'adresse http://docs.oracle.com/cd/E24901_01/index.htm

Il a été mis en place plusieurs files d'attentes (queues) suivant la durée maximum d'exécution du job : ainsi, plus le job sera court et plus la priorité sera élevée. Actuellement le cluster dispose de 480 cœurs qui permettent de faire tourner 480 jobs simultanément. Il n'est possible dans la configuration actuelle que d'utiliser un cœur par job (sauf en lançant des jobs en parallèles).

L'objectif des files est de répondre à la majeure partie des besoins tout en optimisant l'utilisation du cluster.

Attention : tout job lancé hors SGE est supprimé sans préavis.

Voici les files d'attentes disponibles, le nombre de cœurs disponible indiqué est pour une utilisation de 4G de mémoire :

Nom de la file	Durée maximum du job*	Nombre de cœurs disponibles (si les autres files ne sont pas utilisées)
longq	24h	120 ou 240-N avec N>120
unlimitq	illimité	30
workq	2h	480 –N / limité à 440 par util.
bigmem	illimité	240

*Durée maximum d'utilisation du processeur

N : Nombre de jobs actifs

Les jobs en trop sont mis en attente.

4.1.2 Soumettre un job

En ligne de commande :

```
qsub - submit a batch job to Sun Grid Engine.
qsh - submit an interactive X-windows session to Sun Grid Engine.
qlogin - submit an interactive login session to Sun Grid Engine.
qrsh - submit an interactive rsh session to Sun Grid Engine.
qalter - modify a pending batch job of Sun Grid Engine.
qresub - submit a copy of an existing Sun Grid Engine job.
```

Exemple pour un job simple :

1 - Il faut préparer un fichier (script) contenant la (ou les) ligne(s) de commande
2 - Vos fichiers de sortie doivent être impérativement dirigés vers l'espace disque /work
3- Il faut **impérativement** préciser la quantité de mémoire à utiliser avec –l h_vmem= sinon il vous sera alloué 4Go maximum par job.
4 - Soumettre le job avec la commande de soumission (qsub)

Voici un exemple de script de soumission (il est aussi possible de lancer des jobs en interactif):

`monscript.sh`

```
#!/bin/sh
```

```
#$ -o /work/.../output.txt
#$ -e /work/.../error.txt
#$ -q longq
#$ -M mon_email@inra.fr
#$ -m bea
#$ -l h_vmem=3G
# Mon programme commence ici :
blastall -d swissprot -p blastx -i /work/.../z72882.fa
```

Toute ligne commençant par #$ indique une option à exécuter par sge.

-l h_vmem=8G : pour spécifier à l'ordonnanceur l'allocation de 8Go de mémoire pour l'exécution

de ce job.

-q queue_name : spécifier le nom de la queue

-o output_filename : redirection de la sortie standard

-e error_filename : redirection de la sortie d'erreur

-M mon_adresse@mail : si un problème survient pendant l'exécution, un mail est envoyé à cette adresse.

-m bae : quand ce mail doit être envoyé (b : begin, a : abort, e : end)

-N job_name : pour donner un nom à son job

Pour soumettre :
qsub monscript.sh

Alternativement, vous pouvez ne faire apparaître que les lignes de commandes dans le script et indiquer les options à l'appel de qsub.

Soit qsub –l h_vmem=3G –q longq –M mon_email@inra.fr –m bea Monscript.sh

Avec l'interface graphique ce sont les mêmes commandes. Pour lancer l'interface taper « qmon ».

4.1.3 Suivre l'exécution d'un job

Utiliser la commande qstat dont voici quelques options :

`qstat` : liste les jobs de tous les utilisateurs en cours
`qstat -u user` : donne les informations uniquement sur l'utilisateur
`qstat -j job_id` : détail sur un job en particulier (numéro id attribué par SGE)
`qstat -s r` : donne uniquement les jobs avec le status r(unning)
`qstat -f` : visualise les jobs en cours par file et par nœud

La commande `qmon` permet de donner le même type d'information via une interface graphique.

Suivre l'exécution avec Ganglia :

En se connectant (depuis dga11) sur http://dga11.jouy.inra.fr/ganglia/ à l'aide d'un navigateur, vous pouvez obtenir une vue synthétique de l'état du cluster. Les informations essentielles apparaissant sur cette interface sont le nombre nœuds et le nombre de jobs potentiels présents sur le cluster. Le nombre de jobs en cours d'exécution sont également représenté à l'aide d'une ligne bleu, tandis qu'une aire grisée indique le load average de la plateforme. Note importante lors d'une utilisation normale des ressources la courbe du nombre de processus en cours d'exécution et l'aire tracée par le load average doivent concorder. En cas contraire, le script en cours d'exécution devra être repensé.

4.1.4 Obtenir de l'information sur un job terminé

Utiliser la commande qacct :

#`qacct -j job_id` : donne les informations sur un job en particulier.

4.1.5 Tuer un job

`qdel -j job_id` : tue un job en particulier
`qdel -u user` : tue l'ensemble de mes jobs

On ne peut pas tuer les jobs d'un autre utilisateur.

5 Logiciels disponibles

Les logiciels disponibles sur le cluster sont les suivants :

Compilateur et outils de debug	IntelFortran	compilateur
	gcc	GNU C
	gdb	GNU Debugger
	ddd	Interface graphique pour debug
	gprof	Profilage de code
	valgrind	Profilage de code
éditeurs	nedit	
	emacs	
Bureau	gnome	Bureau par défaut de RedHat
shell	ksh	Installé en standard avec RedHat
outils	gawk	Installé en standard avec RedHat
langages	Java 1.6	
	Latex	Suite Texlive
	R	

6 Contacts

Pour toute demande, question ou problème, contacter l'équipe système du CTIG à l'adresse ctig.systeme@jouy.inra.fr