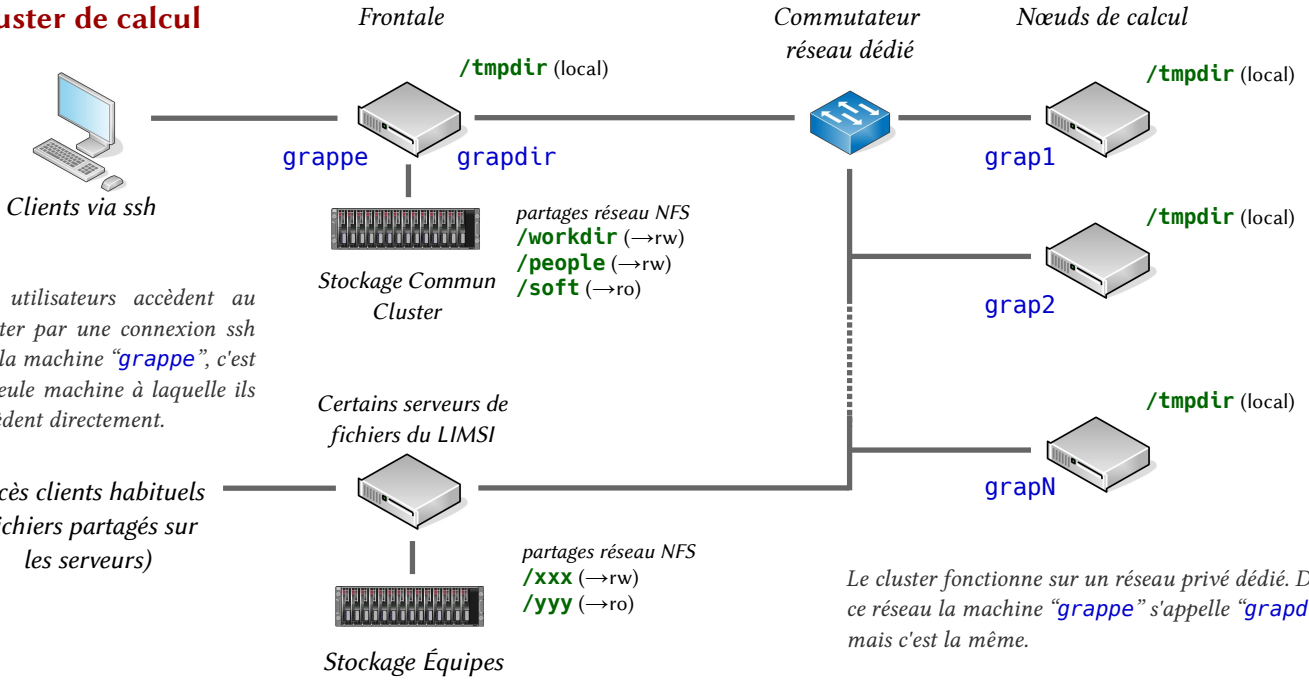
L.Pointal

Mémento GRAPPE
Cluster de calcul mutualisé du LIMSI

http://p2i.limsi.fr/grappe

version du 21/10/2016

Cluster de calcul



Pour les caractéristiques des nœuds, les volumes des espaces de stockage, le réseau, les serveurs de fichiers et les chemins de partage, les files d'attentes... voir la page dédiée : http://p2i.limsi.fr/grappe

Accès au Cluster

pour commencer à l'utiliser...

1) demande d'accès, envoi d'une demande par email à p2i@limsi.fr en précisant son login labo (ici *me*)
   ...attendre l'email de réponse validant l'accès sur grappe
   Lors de cette procédure, vos répertoires personnel et de travail sont créés (avec des quotas limite de stockage), un fichier ~/.forward relaie les emails locaux vers votre compte labo, et un jeu de clés ssh spécifiques au cluster est mis en place.
2) connection sur la machine grappe via ssh : `me@host:~$ ssh me@grappe` *(mot de passe labo)*
3) création et soumission des scripts de jobs sur le cluster... (doc ci-après et en ligne)

Bonnes pratiques stockage

*pour que le cluster soit **bien** utilisé...*

$HOME
/people/*me*
Partagé entre les nœuds, peu d'espace.
Pour stocker vos exécutables personnels, scripts de jobs et autres fichiers de paramètres.

$WORKDIR
/workdir/*me*
Partagé entre les nœuds, utilisé pour stocker des données à traiter volumineuses, non accessibles autrement par les nœuds de calcul, et pour rapatrier les résultats volumineux.
⚠ Un **nettoyage périodique** des anciennes données est réalisé automatiquement sur $WORKDIR.
👉 Une fois vos jobs terminés, pensez à récupérer vos résultats et à les stocker ailleurs que sur le cluster - celui-ci n'est pas un système de stockage pérenne..

$TMPDIR
/tempdir/*idjob*
Local sur chaque machine. Le gestionnaire de jobs y crée un répertoire spécifique temporaire pour la durée de vie de chaque job (répertoire automatiquement purgé à la fin du job).
👉 **Utilisez de préférence ce disque local pour vos données** (copie de données à traiter au début du job, stockage pendant le job, récupération des résultats à la fin du job).

⚠ ***Les espaces de stockage du cluster ne sont pas sauvegardés.***

❗ ***Les espaces de stockage du cluster ne sont pas faits pour du partage de données.***

Définition d'un job

Un fichier texte script-shell contient des **directives** permettant de spécifier des options du job, et des **commandes** pour lancer les calculs et transférer les données.
Il est soumis à l'ordonnanceur de jobs avec la commande :

`qsub` *script*

Il sera exécuté sur les nœuds de calcul.

```
#!/bin/bash
#PBS -N LongRun3                                  # Directives pour l'exécution du job
#PBS -l walltime=23:00:00,mem=1gb
cd $TMPDIR                                        # Répertoire temporaire pour l'exécution
cp $PBS_O_HOME/* ./                               # Transfert des données/binaires/params
cp $WORKDIR/data1/* ./
./monproc                                         # Traitements
./postproc3
cp data.res $WORKDIR/$PBS_JOBNAME.res             # Transfert des résultats
cp status $PBS_O_HOME/$PBS_JOBNAME.sta
```

voir détails au verso

Variable à l'exécution des jobs

Vous pouvez utiliser ces variables dans vos scripts de job pour les adapter à l'exécution sur les nœuds.

Reprises (en mettant PBS_O_) lors de la soumission par qsub sur grapdir :

PBS_O_HOME	Répertoire perso utilisateur $HOME
PBS_O_LOGNAME	Login utilisateur $LOGNAME

...PBS_O_LANG, PBS_O_PATH, PBS_O_MAIL, PBS_O_SHELL, PBS_O_TZ (...)

Définies lors de la soumission par qsub sur grapdir :

PBS_O_HOST	Nom de la machine (grapdir)
PBS_O_WORKDIR	Répertoire courant sur grapdir *lors de la soumission*
PBS_O_QUEUE	Queue de soumission initiale
PBS_O_INITDIR	Répertoire de travail au démarrage (si spécifié via l'option -d)

L'existence d'un espace WORKDIR partagé entre tous les nœuds, ainsi que la possibilité d'accéder à certains volumes de stockages hors du cluster à partir des nœuds n'est pas possible sur tous les clusters de calcul.

Définies lors de l'exécution du job sur le nœud :

PBS_ENVIRONMENT	PBS_BATCH ou PBS_INTERACTIVE
PBS_JOBID	Identificateur associé au job par le cluster (ex. 56.grapdir ou 56-32.grapdir pour un job-array)
PBS_JOBNAME	Nom donné au job (cf option -N)
PBS_NODEFILE	Fichier contenant la liste des nœuds participants au job
PBS_ARRAYID	ID associé pour un job d'un job-array (cf option -t)
TMPDIR	Répertoire de stockage **local** du job, pour la durée du job, à utiliser prioritairement

Variables du cluster, définies à la soumission et à l'exécution :

WORKDIR	Répertoire utilisateur sur le stockage commun en partage réseau (spécial cluster grappe)

Options d'exécution des jobs

Options #PBS pour les scripts de job (ces options peuvent aussi être fournies en ligne de commande) :

-d *rep*	Répertoire de travail au démarrage (positionne PBS_O_INITDIR), qui est par défaut $HOME
-N *jobname*	Nom donné au job (défaut nom du script soumis)
-o *fich*	Fichier de capture de stdout (défaut *jobname*.o*jobid* dans PBS_O_WORKDIR)
-e *fich*	Fichier de capture de stderr (défaut *jobname*.e*jobid* dans PBS_O_WORKDIR)
-M *me*@limsi.fr	Adresses emails de suivi de la vie du job
-m bea	Choix des événements du job à suivre : b=begin, e=end, a=aborted
-t *arrayids*	Lancement de N jobs via un job-array avec les *arrayids* indiqués (ex. 15-30 ; 1,3,7-12)
-V	Export de toutes les variables d'environnement vers le job
-T *script*	Indication de scripts prologue/épilogue au job (nommés prologue.*script* & epilogue.*script*)
-l *limites*	Limites pour le job

Pour les *limites*, on a entre autres (plusieurs -l ou séparateur virgule) :

pmem=100mb	mémoire physique
walltime=03:45:00	durée maximale
nodes=2	nombre de nœuds sollicités
nodes=2:ppn=3	nombre de processeurs/nœuds (cœur) sollicités
file=150gb	espace de stockage à réserver

Voir la doc et les autres (nombreuses) options disponibles avec man qsub.

Commandes de gestion des jobs & d'état du cluster

Commandes de base et options les plus courantes :

qsub *[options] script*	Soumission d'un script à l'ordonnanceur du cluster
qstat *[options][jobid]*	Statut des jobs, options : -a all, -r running -n nodes, -f infos complètes, -u *user*, -Q *queue*
	-B statut du serveur de jobs
qdel *jobid*	Suppression d'un job
qhold *jobid*	Mise en attente d'un job
qrls *jobid*	Réactivation d'un job mis en attente
pbsnodes	État des nœuds de calcul du cluster