Formation Cluster pour les utilisateurs du CTIG

Création de la formation
à partir des supports

Pascal Croiseau (GABI)
Olivier Filangi (PEGASE)
Sylvie Nugier (CTIG)
François Laperruque (SAGA)
Martin Souchal (CTIG)

CTIG – CATI IPBI
Formation Linux pour les utilisateurs du CTIG

Présentation

- **Objectifs**
  - Connaître le Cluster du CTIG
  - Savoir lancer des traitements
  - Suivre le déroulement des traitements

Planning

- Partie I
  - Généralités et Infrastructure du CTIG
  - SGE
  - Lancement par lot de traitement

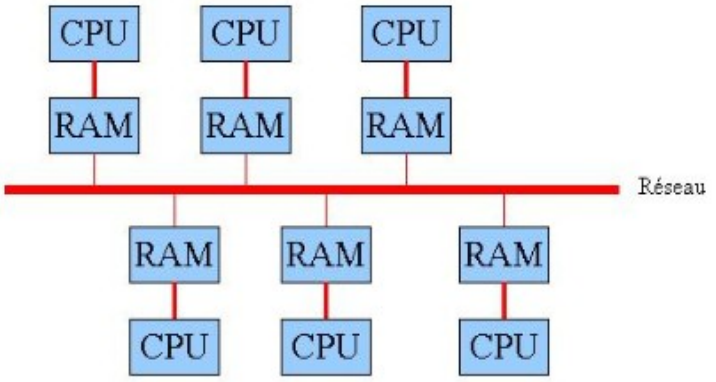
Généralités (1)

Le cluster (*grappe*) de calcul du CTIG est une machine parallèle dédié au calcul intensif (**HPC** ou **High Performance Computing**).

Il existe également des clusters dits à **haute disponibilité** mais qui ne nous concernent pas à ce jour.

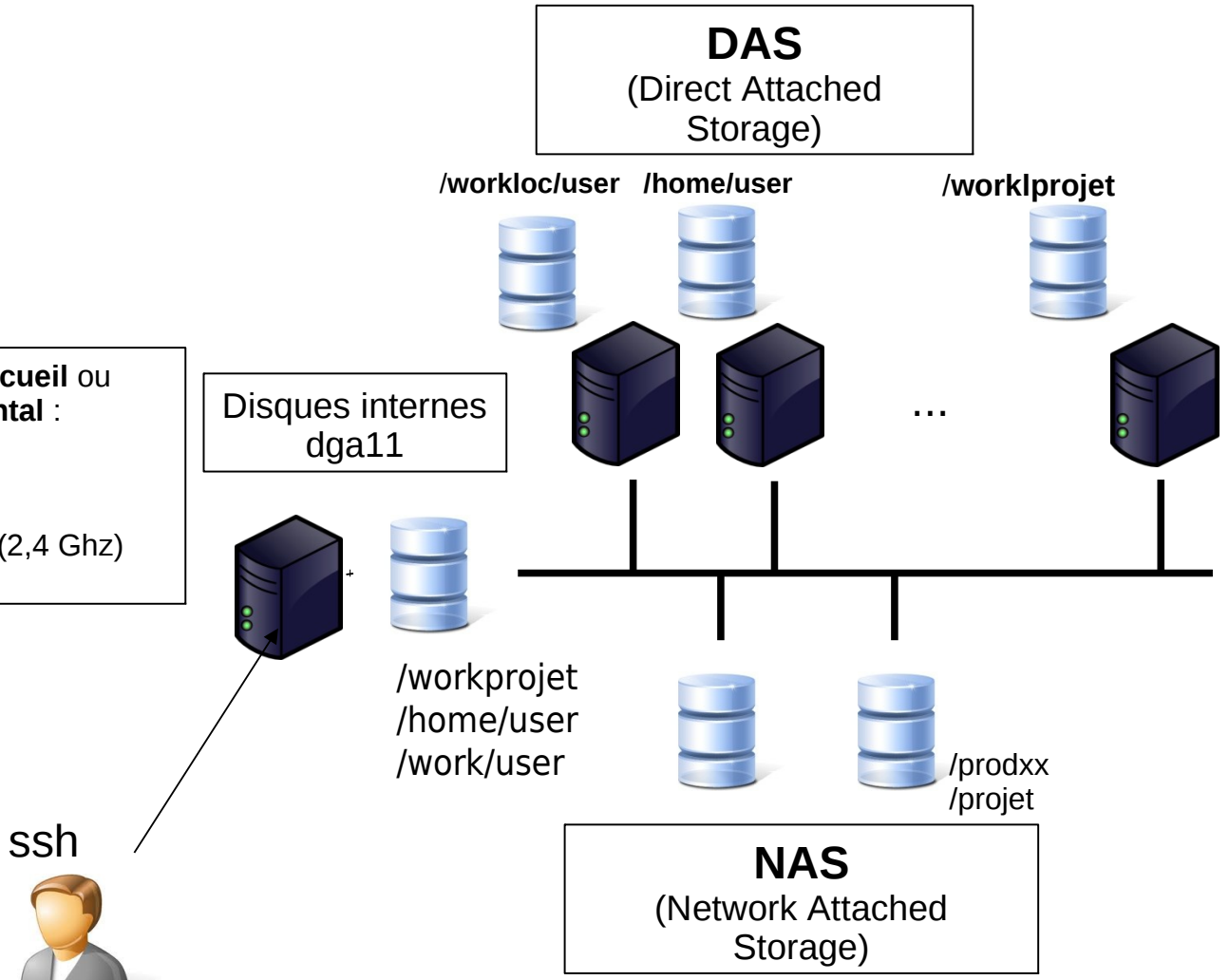
L'idée du cluster de calcul est d'agréger des machines de puissance raisonnable (en général des quadriprocesseurs x86 offrant le meilleur rapport coût/performance) afin d'obtenir une "pseudo" machine virtuelle.

un cluster peut être représenté comme une seule machine multi-processeurs où chaque processeur possède sa propre mémoire vive.
Ce modèle s'appelle modèle **NUMA**
(Non Uniform Access Memory)



Source : http://igm.univ-mlv.fr/~dr/XPOSE2006/BACHIMONT_BRUNET_PIASZCZYNKSI/systeme.htm

Infrastructure



Station d'accueil ou **Serveur frontal** :
DGA11

12 Go RAM
2 x 4 cœurs (2,4 Ghz)

40 Nœuds de calcul

20 Nœuds x 48 Go
20 Nœuds x 96 Go RAM
2 x 6 cœurs (2,66 Ghz)
300 Go + 600 Go

Gigabit ethernet (1Gbit/s)

Puissance crête théorique

2,66*40*2*6*4=5,107TFLOP

4 opérations simultanées

CTIG – CATI IPBI
Formation Linux pour les utilisateurs du CTIG

Connexion

Pour se connecter sur le cluster et soumettre des jobs il faut au préalable faire une demande d'ouverture de compte au **CTIG**. Adresser votre demande à ctig.systeme@jouy.inra.fr
Formulaire disponible sur le wiki du ctig :
https://ctigwiki.jouy.inra.fr/dokuwiki/doku.php?id=docctig:accesctig:demandecompte

Connexion au serveur frontal dga11 accessible via le protocole SSH à l'adresse dga11.jouy.inra.fr

- Client SSH (mode ligne de commande) :
  - A partir d'Unix : ssh jdupont@dga11.jouy.inra.fr
  - A partir de Windows : utilisation de putty

- Client NX
  - voir installation et paramétrage dans https://ctigwiki.jouy.inra.fr/dokuwiki/doku.php?id=docctig:accesctig:accesmoyens

Le serveur frontal/station d'accueil DGA11

- Les missions du serveur frontal :
  - Mettre à disposition un environnement de développement
  - Exécuter les programmes sur le système de calcul
  - Contrôler le déroulement de l'exécution des programmes
  - Récupérer le résultat de l'exécution des programmes
- Ce qu'il ne faut pas faire avec le serveur frontal:
  - S'en servir de système de calcul !!!
  - Cela impacterait l'ensemble des utilisateurs
    - Les jobs seront tués systématiquement et sans avertissement.

Le serveur frontal/station d'accueil DGA11

- Pour faire des tests
  - Il faut se loguer en direct sur un nœud du serveur et pas sur la station d'accueil (qlogin)
  - Une autre alternative est de travailler sur DGA12 qui sera configuré à l'identique du cluster (mais il faut que la structure du programme puisse s'adapter)

Le stockage des données

- Répertoire personnel sauvegardé : */home/username*
- Répertoire de travail commun : */travail/*
- Petit projet ou tâches ponctuelles : */work/username*
- Accès aux disques locaux : */workloc/username (50Go par projet de recherche)*
  - qls : etat des */workloc/username*
  - qclean : nettoyage des */workloc/username*

Utilisation de SGE

CTIG – CATI IPBI
Formation Linux pour les utilisateurs du CTIG

SGE (Sun Grid Engine)

- Ordonnanceur
  - *« L'ordonnancement de tâches informatiques concerne exclusivement la manière de lancer des traitements (batchs) sur un ou plusieurs composants de son système d'information au moyen de progiciels spécifiques. ... »* (Wikipedia)
  - Application de règles d'exécution
  - Soumission dans une file d'attente = queue
  - Enchaînement successif des traitements
  - Pas de garantie de date d'exécution

Contraintes/limitations de soumission

- Chaque job est executé par défaut:
  - Sur 1 cœur de calcul
  - Avec 4 Go de RAM
  - Sur une queue (workq) du cluster
- Les queues disponibles :
  - Workq: 480 jobs ; 440 max par utilisateur avec 4G de mémoire ; limité à 2h
  - Longq: 120 jobs ; 100 max par utilisateur avec 4G de mémoire ; limité à 24h
  - Unlimitq: 30 jobs ; pas de limite de temps
  - Bigmem: 240 jobs mais accès possible à 96Go de mémoire par jobs au lieu de 48Go pour les autres queues ; pas de limite de temps

Ne pas (trop) surestimer les ressources nécessaires à l'exécution de ses travaux, afin de tirer le meilleur parti du système

Soumission de jobs : qsub

Le job ID

```
[user@dga11 ~]$ qsub mon_script.sh
Your job 22967 («mon_script.sh») has been submitted

[user@dga11 ~]$ ls
mon_script.sh mon_script.sh.e22967 mon_script.sh.o22967
```

2 fichiers crées : sortie standard et sortie erreur

Soumission de jobs : qsub (options)

- -N job_name : pour donner un nom à son job
- -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)

qsub -q formationq mon_script.sh

Soumission de jobs : qsub (options)

Le classique:

```
[user@dga11 ~]$ qsub -q workq mon_script.sh
Your job 22967 («mon_script.sh») has been submitted
```

- **4 Go de RAM**
- **1 coeur**

Soumission de jobs : qsub (options)

Besoin plus de mémoires

```
[user@dga11 ~]$ qsub -q workq -l h_vmem=16G mon_script.sh
Your job 22967 («mon_script.sh») has been submitted
```

- **16 Go de RAM**
- **1 coeur**

h_vmem : Upper limit for virtual memory

Soumission de jobs : qsub (options)

- Besoin plus de cœurs : les environnements parallèles
  - openmp : 1=<X<=12 cœurs sur 1 nœud

```
[user@dga11 ~]$ qsub -q workq -pe openmp 4 mon_script.sh
Your job 22967 («mon_script.sh») has been submitted
```

Queues	Max Slots
workq	12
longq	6
unlimitq	2
bigmem	12

qstat -f : pour savoir le nombre de slots disponibles

- **4 Go de RAM**
- **4 cœurs**

```
#!/bin/bash

export OMP_NUM_THREADS=$NSLOTS
./calcf_omp $*
```

Soumission de jobs : qsub (options)

Toutes les options précédemment décrites peuvent être intégrées directement dans le script soumis à SGE avec qsub

```
[user@dga11 ~]$ cat mon_script.sh
#!/bin/sh
#$ -M user@maville.inra.fr
#$ -m a
#$ -q workq
#Mon programme commence ici
ls
#Fin du programme
```

Sans passer par un script : l'option -b

```
[user@dga11 ~]$ qsub -b y -cwd ls
```

Soumission de jobs : qlogin (mode interactif)

L'utilisateur est connecté sur le nœud 231

```
[user@dga11 ~]$ qlogin
Your job 22972 ("QLOGIN") has been submitted
waiting for interactive job to be scheduled ...
Your interactive job 22972 has been successfully scheduled.
Establishing builtin session to host node231 ...
[user@node231 ~]$ exit
Logout
[user@dga11 ~]$
```

L'utilisateur est déconnecté

Contrôle des jobs : qstat

```
job-ID  prior   name       user         state submit/start at     queue                          slots ja-task-ID
-----------------------------------------------------------------------------------------------------------------
  22993 560.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node238                  1 3
  22993 560.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node237                  1 14
  22993 560.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node243                  1 17
  22993 560.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node240                  1 21
  22993 560.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node240                  1 22
  22993 560.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node202                  1 25
  22795 560.00000 VerifParBT mnfouilloux  Eqw   01/02/2012 16:25:11                                1
```

job-ID : identifiant unique du job assigné par l'ordonnanceur
prior : la priorité du job déterminant sa position dans la queue. Sa valeur est déterminée dynamiquement.
state : l'état du job en abréviation :

- d(eletion) : le job est en cours de suppression (initier par un qdel).
- E(rror) : le job n'a pu être exécuter. La raison de cette erreur peut être visualiser à l'aide de l'option -j (qstat -j).
- qh(old) : le job est en attente d'exécution, il n'y a pas de ressource disponible pour l'exécuter.
- r(unning) : l'exécution du job est relancée.
- R(estarted) : c'est un Job Array en cours d'exécution
- s(uspended) : l'exécution du job est suspendue (initier par un qmod).
- S(uspended) : l'exécution du job est suspendue (initier par une suspension de toute la queue contenant ce job).
- t(ransfering) : le job est en cours de transfert afin d'être executer.
- T(hreshold) : l'exécution du job est suspendue (initier par l'atteinte du seuil d'exécution de la queue).
- qw(aiting) : le job est en attente d'exécution, il n'y a pas de ressource disponible pour l'exécuter.

slots : nombre de slots utilisés par le job.
ja-task-ID : identifiant du job array (vide dans le cas ou le job n'est pas un job array).

Contrôle des jobs : qstat

- qstat -u <*nom_utilisateur*> : donne uniquement les informations sur l'utilisateur
- qstat -s r : donne uniquement les jobs avec le status r(unning)
- qstat -f : pour afficher le résultat par noeud
- man qstat

qstat -s r

```
job-ID  prior   name       user         state submit/start at     queue                          slots ja-task-ID
-----------------------------------------------------------------------------------------------------------------
  22993 610.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node238                  1 3
  22993 610.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node237                  1 14
  22993 610.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node243                  1 17
  22993 610.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node240                  1 21
  22993 610.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node240                  1 22
  22993 610.00000 R66        pcroiseau    r     01/05/2012 14:29:47 longq@node202                  1 25
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node227                  1 11
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node227                  1 12
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node228                  1 13
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node228                  1 14
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node228                  1 15
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node228                  1 16
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node228                  1 17
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node228                  1 18
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node231                  1 19
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node231                  1 20
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node231                  1 21
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node231                  1 22
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node231                  1 23
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node231                  1 24
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node232                  1 25
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node232                  1 26
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node232                  1 27
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node232                  1 28
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node232                  1 29
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node232                  1 30
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node233                  1 31
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node233                  1 32
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node233                  1 33
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node233                  1 34
  22994 502.77364 qarray     root         r     01/05/2012 14:41:17 workq@node233                  1 35
```

Contrôle des jobs : qstat -j

```
[ofilangi@dga11 ~]$ qstat -j 22795
==============================================================
job_number:                 22795
exec_file:                  job_scripts/22795
submission_time:            Mon Jan  2 16:25:11 2012
owner:                      mnfouilloux
uid:                        243
group:                      gembal
gid:                        303
sge_o_home:                 /home/mnfouilloux
sge_o_log_name:             mnfouilloux
sge_o_path:                 /opt/intel/composerxe-2011.2.137/bin/intel64:/opt/intel/composerxe-2011.2.137/mpirt/bin/intel64:/usr/lo
-3.3/bin:/usr/local/bin:/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/sbin:/usr/java/latest/bin:/home/mnfouilloux/bin
sge_o_shell:                /bin/bash
sge_o_workdir:              /gembal/mnfouilloux/CONVERT
sge_o_host:                 dga11
account:                    sge
cwd:                        /gembal/mnfouilloux/CONVERT
stderr_path_list:           NONE:NONE:/gembal/mnfouilloux/CONVERT/LOG/
mail_list:                  mnfouilloux@dga11.jouy.inra.fr
notify:                     FALSE
job_name:                   VerifParBTA.sh
stdout_path_list:           NONE:NONE:/gembal/mnfouilloux/CONVERT/LOG/
jobshare:                   0
hard_queue_list:            workq
env_list:                   LD_LIBRARY_PATH=/opt/intel/composerxe-2011.2.137/compiler/lib/intel64:/opt/intel/composerxe-2011.2.137/
/intel/composerxe-2011.2.137/mkl/lib/intel64:/lib64,R_LIBS_USER=/logiciels/lib_R
job_args:                   29,r57
script_file:                VerifParBTA.sh
error reason    1:          01/02/2012 16:29:57 [243:4721]: error: can't open output file "/gembal/mnfouilloux/CONVERT/LOG/": Is
scheduling info:            Job is in error state
```

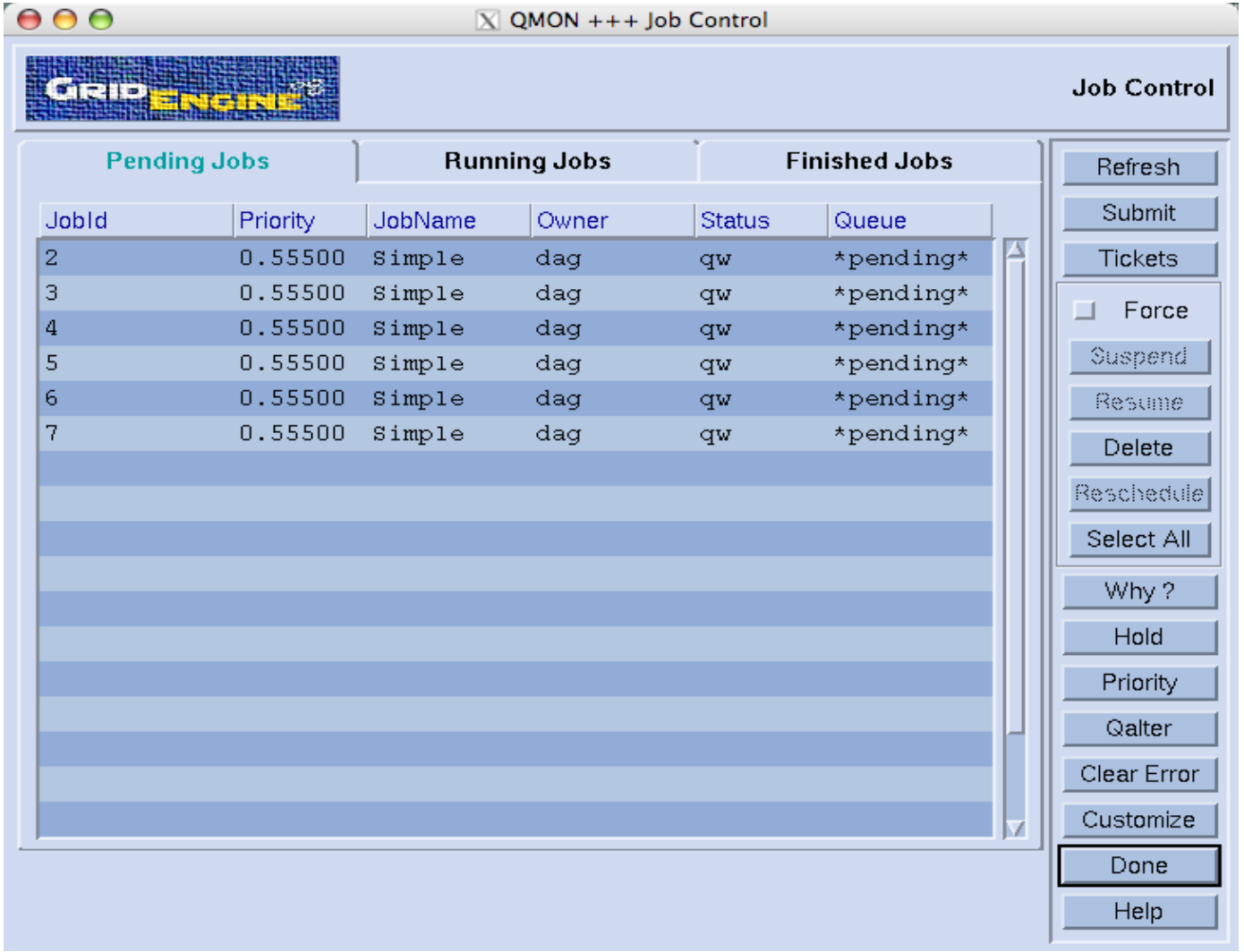
Modifier un job : qalter

Condition : être propriétaire du job

```
[user@dga11 ~]$ qstat -u user
job-ID prior name user state submit/start at   queue slots ja-task-ID
-----------------------------------------------------------------------------
15833 0.56000 mon_script.sh user w 04/07/2009 16:39:41 workq@node236


[user@dga11 ~]$ qalter -l h_vmem=16G
Modified hard queue list of job 15833
```

Contrôle des jobs : qmon



Supprimer un job : qdel

Condition : être propriétaire du job

```
[user@dga11 ~]$ qsub -q workq mon_script.sh
Your job 15833 («mon_script.sh») has been submitted

[user@dga11 ~]$ qstat -u user
job-ID prior name user state submit/start at   queue slots ja-task-ID
-----------------------------------------------------------------------------
15833 0.56000 mon_script.sh user w 04/07/2009 16:39:41 workq@node236


[user@dga11 ~]$ qdel 15833
User has registered the job 15833 for deletion
```

Tracer les jobs

- Redirection de la sortie d'erreur et la sortie standard vers des fichiers de textes :
  - #$ -o standard.txt
  - #$ -e error.txt
- Retour informations par email :
  - #$ -M user@toulouse.inra.fr
  - #$ -m bea
    - b : envoie un mail au départ de l'exécution du job
    - e : envoie un mail à la fin de l'exécution du job
    - a : envoie un mail en cas d'erreur lors de l'exécution du job

Lancement de lots de traitements

Lancer des lots de traitements

- Concept : segmenter un job en plusieurs petits jobs atomiques
- Intérêts :
  - Améliorer le temps de traitement de façon très significative : le calcul est réalisé sur plusieurs CPUs différents
  - Accroître les chances que le job soit réalisé plus rapidement parl'ordonnanceur
- Condition : le job doit être « découpable » en sous job

qarray

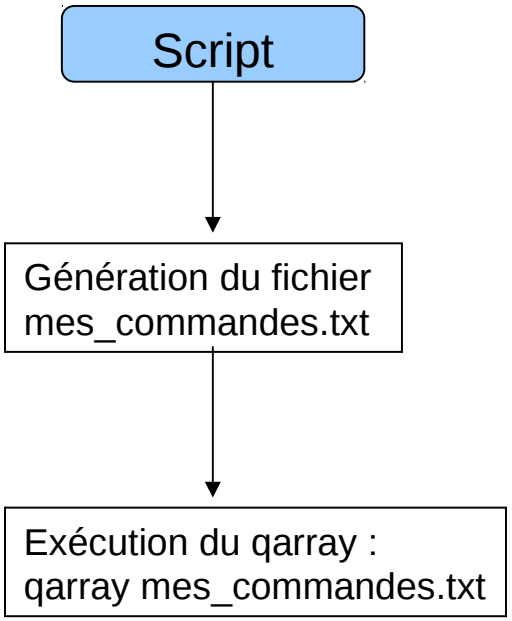
- qarray = qsub pour n lignes
  - Toutes les options du qsub fonctionnent pour le qarray
  - Ex fichier d'entrée :
    - fasta_to_fastq.pl –fasta sequence1.fasta –qual sequence1.qual
    - fasta_to_fastq.pl –fasta sequence2.fasta –qual sequence2.qual
    - fasta_to_fastq.pl –fasta sequence3.fasta –qual sequence3.qual
  - Ex commande :
    - qarray -q workq mon_fichier.txt

Créer mes_commandes.txt

traitements par fichiers

v est un argument chaine de caractère
de la commande <*mon action*>

```
#!/bin/bash
for i in `ls dir/*.txt`
do
  echo "<mon action> $i" >> mes_commandes.txt
done
```



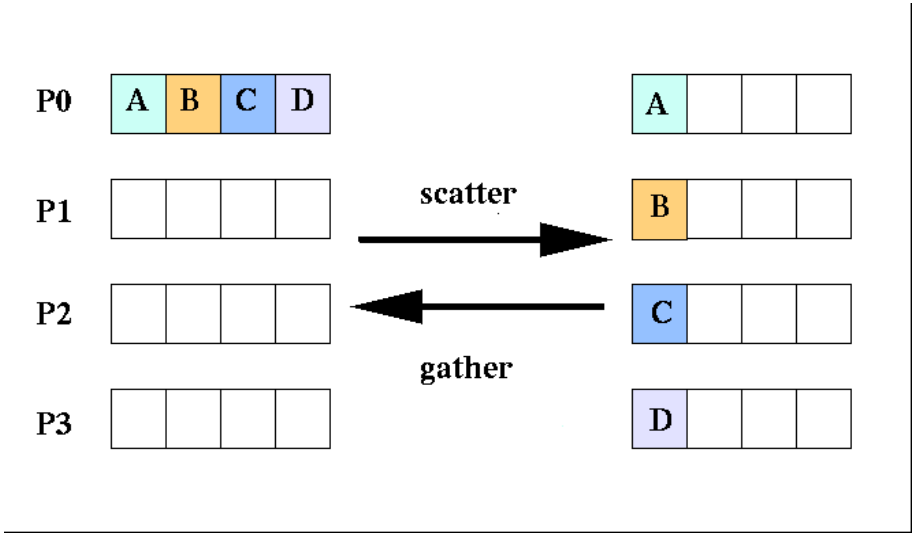
traitements dépendants d'un itérateur

v est un argument entier
de la commande <*mon action*>

```
#!/bin/bash
nsnp=10
for (( v = 1; v <= $nsnp; v++ ))
do
   echo "<mon action> $v" >> mes_commandes.txt
done
```

Remarques

- **qarray** n'est pas une commande SGE (/usr/bin/qarray).
  - Masque un **qsub** ***-t n*** avec utilisation **SGE_TASK_ID**
- **qarray** réalise la plupart du temps un **scatter.** Pour faire un **gather**, il faut utiliser l'option **-sync y** (attente active sur qarray)
  - Applicable dans un modèle **SIMD** *(single instruction multiple data)*
- **-tc max_running_task** : limite le nombre d'exécution à <max_running_task> jobs en parallèle



TP