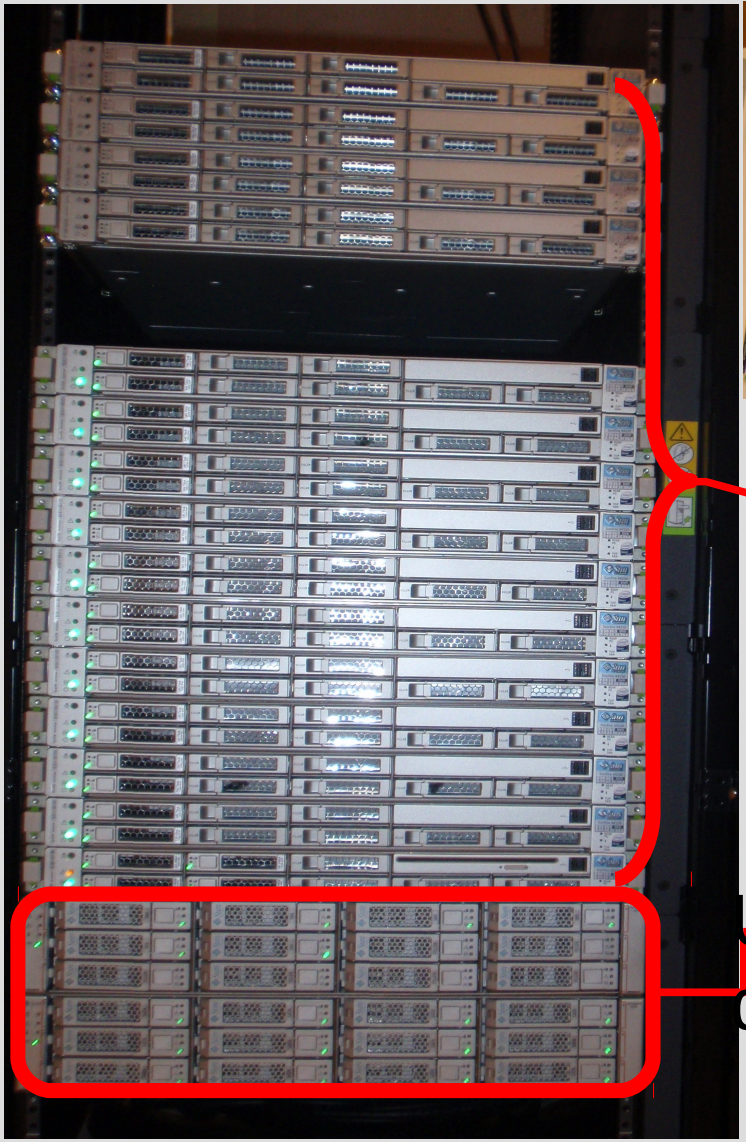
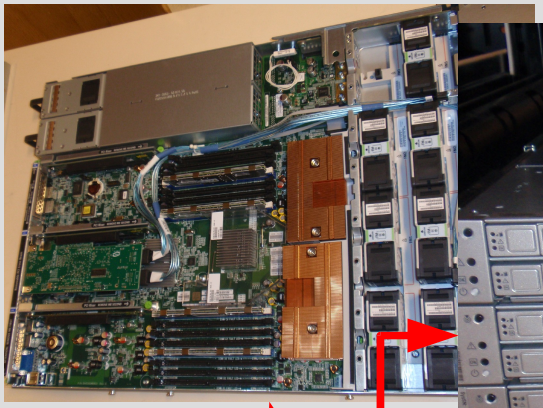
# Cluster URBM-Sys. Dyn.

- FUNDP le 05/01/2010
- Unité de Recherche en Biologie Moléculaire
- Unité de Systèmes Dynamiques (département de math)
- Nicolas Delsate & André Füzfa

# C'est quoi  un cluster ?





Des noeuds de calculs = 14*8=112 processeurs

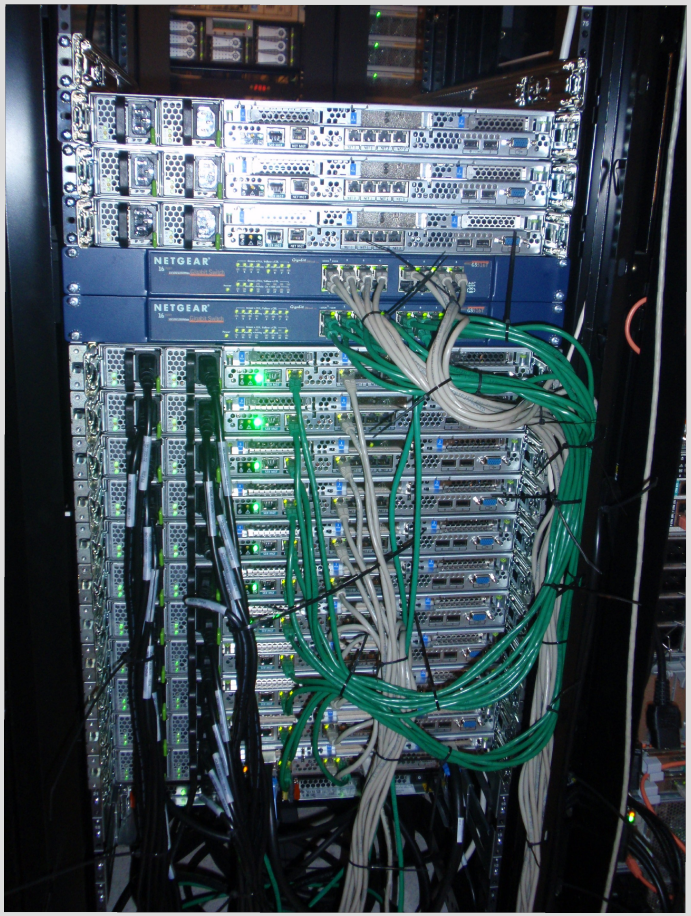
Une baie de disques 7 To

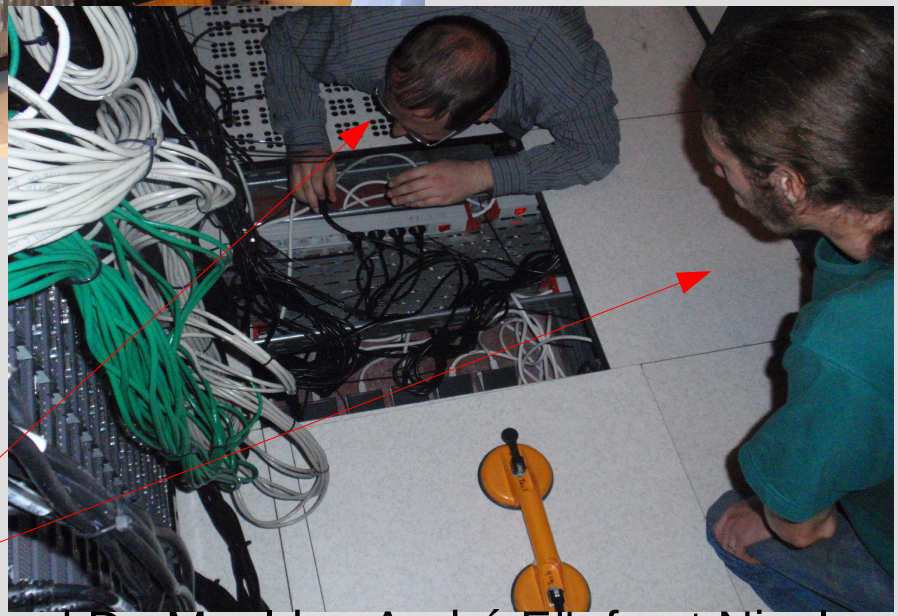
# Mais encore...

Des heures d'installation et de résolution de bugs et problèmes en tout genre

D'autres clusters (nous sommes dans les armoires de l'ISCF).

Des cables et des noeuds ...



Des gens pour la maintenance et l'installation.

Didier Belhomme, Frédéric Wautelet, Fabrice Berger, Eric Bareke, Bertrand De Meulder, André Füzfa et Nicolas Delsate

# Quelques chiffres

- 14 noeuds => 14*8=112 processeurs
- 16 Go RAM/noeud et 4 To de Disque
- 1 machine frontale (machine de connexion sur laquelle vous soumettez vos jobs, le cluster fait le reste grâce à SGE pour les envoyer sur les autres noeuds...)
- Plus de 40 000€ investis par l'URBM et 13 000€ investis par le département et/ou l'unité de systèmes dynamiques.

# Quels programmes sont disponibles ?

- Pour compiler :
  - Tout compilateur "instalable" par paquet gfortran, gcc, …
  - Ifort (compilateur fortran + debugger + openMP, lapack)
- Pour développer :
  - Emacs, gedit, nano, vi, …
- Ce qu'il n'y a pas, parce que l'on ne permet pas de post-traitement :
  - Matlab, gnuplot

# Sur la frontale :

## ssh -X monLogin@urbm-cluster.urbm.fundp.ac.be

- Ce que vous pouvez faire
  - Développer et soumettre vos jobs (SGE)
  - Tester : exécuter votre programme (même en parallèle) mais pas plus de 2 minutes,
  - Un peu de pré-post-traitement type cat, tail, head, ...
- Ce que vous NE pouvez pas faire
  - Exécuter des programmes long (>2min)
- Vous ne pouvez JAMAIS vous connecter aux autres noeuds !

# SGE (1/6) : Sun Grid Engine

- SGE est un logiciel de gestion de queues distribuées. Chaque utilisateur peut envoyer des travaux sous forme de script.
- L’avantage principal consiste en la répartition de la charge de calcul sur les machines du cluster.
- Après la lecture de ce document l’utilisateur "saura" soumettre des jobs, suivre les exécutions, et les arrêter éventuellement. Il doit être uniquement considéré comme une première aide.
- Pour toute information complémentaire veuillez consulter les man pages ou la documentation de SGE.

# SGE (2/6) : Sun Grid Engine

- qstat pour voir l'état des jobs
- qdel pour supprimer un job
- qsub pour soumettre un job (ou un job paramétrique ou un job parallèle)

pour avoir de l'aide : man qsub par exemple

Mais aussi

- qmon pour lancer l'interface graphique
- qhost pour voir l'état des machines

# SGE (3/6) : qstat informations

- qstat donne l'état de SES propres jobs (pas ceux des autres)
- qstat -f donne l'état de tous les noeuds
- qstat -u UserBidon
  donne l'état des jobs de "UserBidon"

```
qstat -u bdemeulder
job-ID  prior   name       user         state submit/start at     queue                          slots ja-task-ID
------------------------------------------------------------------------------------------------------------------
   2839 0.75031 Network_fi bdemeulder   r     11/26/2009 12:00:20 all.q@n10.bioinfo.urbm         1
```

- qstat -j 2839
  donne tout plein d'informations sur le job portant le job-ID numéro 2839

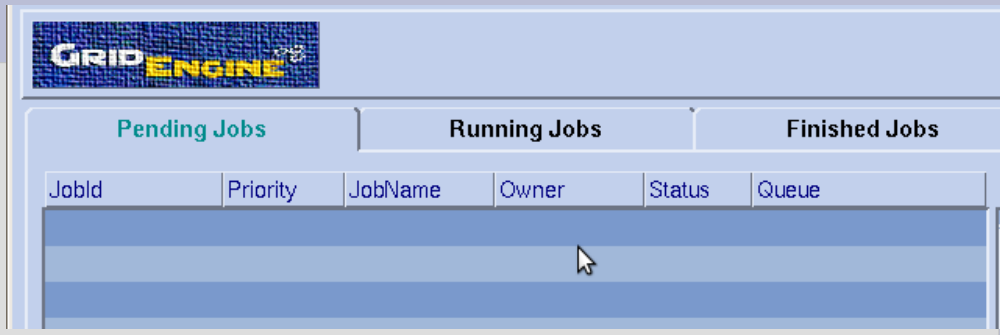
# SGE (4/6) : qdel supprimer

- qdel 2839 supprimer le job ayant la job-ID 2839 (seulement si c'est le vôtre !)
- qdel -f 2839 pour forcer (on ne sait jamais que la première instruction ne fonctionne pas)
- États d'un job

| Category | State | SGE Letter Code |
|---|---|---|
| Pending | pending | qw |
| | pending, user hold | qw |
| | pending, system hold | hqw |
| | pending, user and system hold | hqw |
| | pending, user hold, re-queue | hRwq |
| | pending, system hold, re-queue | hRwq |
| | pending, user and system hold, re-queue | hRwq |
| Running | running | r |
| | transferring | t |
| | running, re-submit | Rr |
| | transferring, re-submit | Rt |
| Suspended | job suspended | s, ts |
| | queue suspended | S, tS |
| | queue suspended by alarm | T, tT |
| | all suspended with re-submit | Rs, Rts, RS, RtS, RT, RtT |
| Error | all pending states with error | Eqw, Ehqw, EhRqw |
| Deleted | all running and suspended states with deletion | dr, dt, dRr, dRt, ds, dS, dT, dRs, dRS, dRT |

# SGE (5/6) : qmon

## interface graphique





- Job control : pour voir les jobs
  - Pending job :
    soumis non encore exécuté
  - Running job :
    en exécutions
  - Finished job :
    terminés

    possibilité de suspendre un job (également avec la commande qhold)
- Queue control — pour voir l'état des noeuds
- Le reste ne vous intéresse pas !
- NE PAS UTILISER submit job pour soumettre ses jobs

# SGE (6/6) : qsub soumettre un job

- Remarque : une fois un job soumis, vous pouvez vous déconnecter et il continue.
- qsub script.sh soumettre votre script
- qsub -t 1-10 job.sh soumettre un job 10 fois
- Sans argument
- Avec les options indispensable :
- -cwd : lance le job a partir du répertoire courant. Utile, car les fichiers de sortie iront alors dans ce répertoire.
- -j y : pour concaténer les fichiers de sortie (*.o *.e)
- -m e : pour envoyer un email a la fin du job (quand ça fonctionne ...)
- -S /bin/bash: pour définir le shell par défaut

# Soumettre un job séquentiel

- Créer un script de soumission pgmSoumis.sh

Sera le nom que vous verrez avec qmon

```
#!/bin/sh
#$ -S /bin/bash
#$ -cwd

cd /home/math/nomUser/RECHERCHE/
./monEx.x > monOut.dat && ./unAutreExe
exit 0
```

C'est du bash

Dis à SGE que c'est du bash

Dis à SGE qu'on travaille tjs dans le répertoire dans lequel on se trouve

On se met au bon endroit TJS

Un exécutable

C'est du bash

Et éventuellement un autre

- Soumettre votre script dans une console

```
qsub pgmSoumis.sh
```

# Soumettre un job paramétrique (1/3)

## *Préparer son code*

- En C

```
#include <stdio.h>
#include <iostream>
using namespace std;
int main(int argc, char **argv)
{
  long double a,b,c;
  a=(long double)atof(argv[1]);
  b=(long double)atof(argv[2]);
  c=(long double)atof(argv[3]);

}
```

- En Fortran

```
PROGRAM coucou
   CHARACTER(LEN=60) :: NomFich
   IF(IARGC()==0) THEN
      NomFich='donnee.dat'
   ELSE
      CALL GETARG(1,NomFich)
si vous voulez un nombre
      read(NomFich,*) NomNbre
   END IF
END PROGRAM coucou

équivalent de IARGC()
   - COMMAND_ARGUMENT_COUNT()
   - NARGS()

équivalent de GETARG(1,NF)
   - GET_COMMAND_ARGUMENT(1,NF)
```

# Soumettre un job paramétrique (2/3)

## *Faire un script et soumettre*

- Créer un script de soumission `scriptAr1.sh`

```
#!/bin/bash
#$ -S /bin/bash
cd /home/math/nomUser/RECHERCHE/
#$ -cwd
   #Dit a SGE que c''est un "array job", avec
   #"tasks" allant de 1 a cptL par pas de 1.
#$ -v SGE_TASK_FIRST=1
#$ -v SGE_TASK_LAST=12
#$ -t 1-60
#Quand une commande simple dans un "array job" est envoye a un
#noeud, son "task number" est  stocke dans la variable
#SGE_TASK_ID. Donc nous pouvons utiliser la valeur pour mettre
#les resultats qu''on veut.
    ./monExe $SGE_TASK_ID $SGE_TASK_LAST > a.dat
```

- Soumettre le script en console : `qsub scriptAr1.sh`

# Soumettre un job paramétrique (3/3)

## *Ou faire un autre script et soumettre*

- Créer un script de soumission `scriptArray.sh`

```
#!/bin/bash
#$ -S /bin/bash
#$ -cwd
cd /home/math/nomUser/RECHERCHE/
paramIN=$(cat abc.dat | head -$SGE_TASK_ID | tail -1)

./monExectuable.x $paramIN >> fichierOut.dat
```

- Soumettre votre script dans une console

```
qsub -t 1-7260 scriptArray.sh
```

Pour dire que c'est du "array"

le nombre de fois que votre pgm sera lancé
MAX 75000

# Soumettre un job parallèle

- Paralléliser son code avec MPI ou openMP

- Apprendre à le soumettre via qsub en lui disant que c'est un job parallèle
  Pas encore fait !

# Trucs à savoir dans la console (1/2)

- size monPgm : permet de connaitre (le premier chiffre que la commande renvoit) la taille minimum, en mémoire, en ko de votre programme. Par exemple 1478438 correspond à 147 Mo.
- time -p ./monPgm : permet de connaître le temps d’exécution de votre programme (affiché à la fin de l'éxécution).

# Trucs à savoir dans la console (2/2)

- scp -Crq login@machine:/Repertoire autreLogin@autrePC:/AutreRepertoire/.
  Copie Compressée Récursive (q)silencieuse d'un répertoire d'une machine vers une autre
- tail -9 monFichier.dat
  affiche à l'écran les 9 dernières lignes d'un fichier
- head -3 monFichier.f90
  affiche à l'écran les 3 premières lignes d'un fichier
- head -15 | tail -1 monFichier.txt
  affiche à l'écran la 15ième ligne d'un fichier
- ...

# Ifort option de compilation

- -check all : run-time checks on whether array subscript & substring references are within declared bounds + runtime checks for valid pointers +runtime checks for uninitialized variables.
- -g : Produces symbolic debug information in the object file
- -traceback : the compiler generate information in the object file to allow the display of source file traceback information at run time when a severe error occurs
- -debug all : enable debug information and control output of enhanced debug information.
- -ftrapuv : trap uninitialized variables. Initializes stack local variables to an unusual value to aid error detection
- -prec-div : improve precision of floating-point divides (some speed impact)
- -prec-sqrt: determine if certain square root optimizations are enabled
- -O3 : enable -O2 plus more aggressive optimizations that may not improve performance for all programs

# gcc option de compilation

- -Os : optimisation
- -Wall : warning all
- ...

# Attention

- Vous n'êtes pas seul donc on vous demande une bonne "gestion" de vos jobs
- Il n'y a pas de quota mais ce n'est pas une raison pour exagérer !
- Il est toujours possible (pcq petit groupe) de demander pour faire passer son job devant les autres en cas de besoin URGENT.
- Faites l'effort d'optimiser un peu vos codes quand même. Sur des clusters "plus professionnels", l'efficacité des jobs est contrôlée.
- Un peu d'indulgence vis à vis des administrateurs (comme vous l'avez fait jusque maintenant)

# La Doc

- http://perso.fundp.ac.be/~ndelsate/CLUSTER/
- http://wikis.sun.com/display/gridengine62u2/Submitting+Jobs
- https://www.wiki.ed.ac.uk/display/EaStCHEMresearchwiki/How+to+write+a+SGE+job+submission+script
- http://migale.jouy.inra.fr/faq/calcul/utilisation
- http://www.cis.upenn.edu/acg/sge.html
- SGE dans google
- ...

# A faire dès la première connexion

- À votre première connexion vous changez votre mot de passe :

  passwd
  puis vous suivez les instructions

## Dans vos papiers

- Numerical simulations were made on the local computing ressources at Unité de Systèmes Dynamiques (FUNDP, Belgium).

# Nos problèmes

- Ifort Intel Cluster Toolkit n'est pas installer et pose des problèmes
- Disque quasi plein
- Maintenance en amateur...

# Merci

- Aux gens de l'URBM, Didier Belhomme, Frédéric Wautelet
- Aux "béta-testeur" : André, Benoît et Moi
- Aux gros utilisateurs pour les gros tests : Audrey et moi
- Aux gens que j'ai oublié