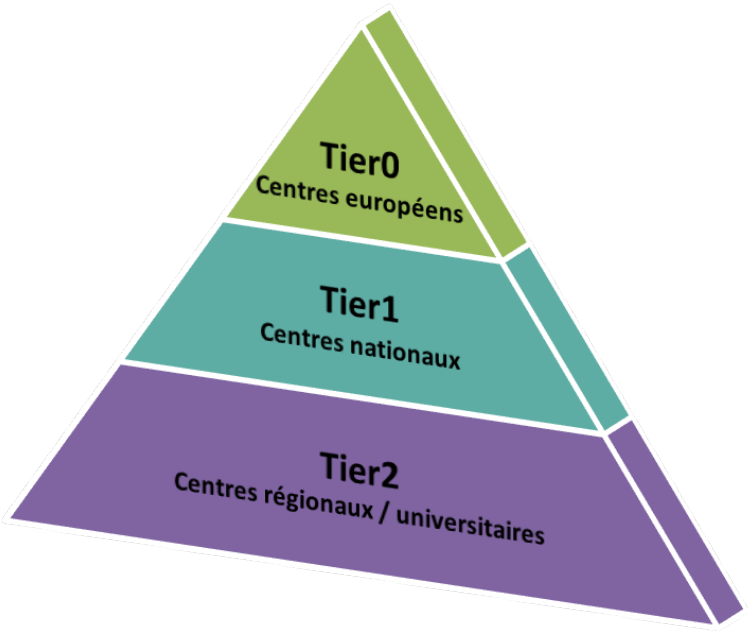
Université de Lille

Mésocentre : Plateformes HPC

Journées Mathrice : 28 septembre 2017

Structuration européenne :



- Tier 0 : heures de calcul sur les centres nationaux (PRACE).
- Tier 1 : 3 centres en France : IDRIS, CINES, TGCC (GENCI).
- Tier 2 : mésocentres régionaux.

Mésocentre de Lille

- **Historique :**
  - 1973-1990 :
    - Les *monstres* de l'époque : CII10070, IRIS80, DPS8, CDC 962, ...
  - 1990-2000 :
    - Plus une activité prioritaire (stations de travail).
  - 2000-maintenant :
    - Création du Mésocentre (projet MNERT national).
    - Financements : MNERT, Région, FEDER, Lille 1.
    - Renouvellement des calculateurs tous les 3-4 ans.
    - Dernier financement Région/FEDER/Lille 1 en 2012.
    - 2 ETP actuellement, 3 après la fusion Université de Lille.
- **Processus de mutualisation :**
  - Premiers financements laboratoires en 2011.
  - Plus de financement direct depuis 2012.
  - Part des labos aujourd'hui : 87% du cluster.

Plateformes :

- **Cluster HPC :**
  - Système d'exploitation Linux (Centos 7).
  - Support sur les applications.
- **Cloud HPC :**
  - Système d'exploitation libre.
  - Les utilisateurs sont administrateurs sur leurs machines virtuelles (VM).
  - Pas de support pour l'administration des VM.
- **Stockage HPC :**
  - Stockage financé par les labos/équipes.
  - Accès uniquement pour les contributeurs.
  - Accès direct aux données depuis le cluster.
- Site web : http://hpc.univ-lille1.fr/

Accès “administratif” aux plateformes de calcul :

- **Pour qui :**
  - Ouvert à tous les établissements de recherche de la “région” NPDC.
  - Chercheurs, ingénieurs, doctorants ....
  - Pour les étudiants : uniquement les stagiaires sur projet de recherche.

- **Comment :**
  - Demande par mail : `http://hpc.univ-lille1.fr/acces`
  - Nécessite la validation du directeur du laboratoire/unité.
  - Directeur : personne qui a délégation de signature des tutelles.

Cluster HPC

Cluster HPC : Architecture

- **Frontale : `zeus.univ-lille1.fr`**
  - Serveur sur lequel se connectent les utilisateurs.
  - Transfert et gestion des données.
  - Compilation des codes.
  - Préparation et lancement des jobs.
- **Nœuds de calcul :**
  - Serveurs sur lequel tournent les jobs.
  - Pas d'accès pour les utilisateurs.
- **Serveurs de services :**
  - Gestionnaire de jobs : Slurm.
  - Serveurs de stockage.
  - Monitoring : `http://hpc-status.univ-lille1.fr/`
  - LDAP : authentification Cluster.
  - DHCP (réseaux privés), TFTP (installation réseau), Nagios/Incinga.

Cluster HPC : nœuds de calcul

- **Ressources CPU :**
  - 100 nœuds à 2*12 cœurs (Haswell) et 128 Go de Ram
  - 104 nœuds à 2*8 cœurs (Ivy Bridge) et 64 Go de Ram
  - 8 nœuds à 2*6 cœurs (Westmere) et 48 Go de Ram
  - Total : 4208 cœurs de calcul et 20 To de Ram
  - Puissance : 126 Tflops
- **Ressources Accélérateurs :**
  - 14 GPU Nvidia Tesla 2090 et 2 GPU Nvidia K80
  - 2 coprocesseurs Intel Xeon Phi 5110p
  - 8 coprocesseurs Intel Xeon Phi 7120p
  - Puissance : 27 Tflops
- **Réseau Infiniband Intel/Qlogic :**
  - Réseau rapide : 40 Gbit/s.
  - Communication intra-nœuds pour les calculs parallèles.
  - Accès au stockage.

Cluster HPC : systèmes de stockage réseau

- **/home :**
  - Stockage NFS .
  - 190 To.
  - Vitesse écriture max : 1.6 Go/s
- **/workdir :**
  - Stockage BeeGFS .
  - 290 To.
  - Vitesse écriture max : 6.5 Go/s
  - Stockage à favoriser (notamment pour les codes parallèles).
- **Attention :**
  - Stockage temporaire : les données doivent être rapatriés dans les labos.

Cluster HPC : utilisation

- **Accès technique :**
  - Système linux (Centos).
  - Accès :
    - linux : ssh
    - windows : putty
  - Transferts de données :
    - linux : rsync, scp, ...
    - windows : winscp, filezilla, ...
  - Accès depuis l'extérieur par le VPN de l'Université.
    - Actuellement : accès Lille 2 et Lille 3 par le VPN.
    - Accès direct après la mise en place du réseau UDL.
- **Utilisation :**
  - Uniquement par mode batch : SLURM.
  - 2 noeuds interactifs (GPU et PHI) pour tests rapides.

Cluster HPC : soumission de jobs

- Exemple de script : fichier job.slurm

```
#!/bin/bash
#SBATCH --nodes=1                  # nombre de noeuds
#SBATCH --ntasks-per-node=1        # nombre de coeurs par noeud
#SBATCH --time=24:00:00            # temps demandé (walltime)
#SBATCH --job-name=test            # nom du job dans la queue
#SBATCH --mem=2G                   # mémoire max par noeud

# commandes utilisateurs :
./mon_executable
```

- Commandes SLURM :
  - `sbatch job.slurm` : soumets le job et renvoie son numéro ID
  - `squeue` : état de mes jobs
  - `scancel ID` : arrête un job (uniquement les siens)

Cluster HPC : limite du walltime

Nombre de cœurs	Walltime en heures
48	384
64	288
96	192
128	144
192	96
256	72
384	48
512	36
768	24

- Le walltime dépend du nombre de cœurs.
- En dessous de 48 cœurs, la limite est toujours de 384h.

Environnement par modules

```
[mmarquil@zeus]$ module avail
------------------------------------------------------------ /share/modules/compilers -------------------------------------------
gcc/4.8.5/openmpi/1.6.5   gcc/6.3.0/compilers       intel/2013/intel-mpi      intel/2017/compilers      pgi/16.10/openmpi/2.0.1
gcc/4.8.5/openmpi/1.8.4   intel/2011/compilers      intel/2013/openmpi/1.6.5  intel/2017/intel-mpi
gcc/5.4.0/compilers       intel/2011/openmpi/1.6.2  intel/2015/compilers      intel/2017/openmpi/2.0.1
gcc/6.1.0/compilers       intel/2011/openmpi/1.6.5  intel/2015/intel-mpi      nvidia/cuda/7.5/compilers
gcc/6.1.0/openmpi/2.0.0   intel/2013/compilers      intel/2015/openmpi/1.8.4  pgi/16.10/compilers
------------------------------------------------------------ /share/modules/libraries -------------------------------------------
anaconda2/4.0.0            boost/1.62                 libraw/0.17.2             tensorflow/0.10/python2
anaconda3/4.1.0            hdf5/1.8.18/intel/2017/mpi protobuf/3.0.0            tensorflow/0.10/python3
bazel/0.3.0                hdf5/1.8.18/intel/2017/seq swig/3.0.10               theano/0.8.2/python2
------------------------------------------------------------ /share/modules/applications ----------------------------------------
abaqus/6.13                    cosmologic/turbomole/7.1/seq   gerris/131206/seq              molpro/2012.1.18/mpp
altair/13.0                    cosmologic/turbomole/7.1/smp   gromacs/4.5.5                  molpro/2015.1.11/mpp
altair/14.0                    cosmologic/turbomole/7.1.1/mpi gromacs/4.6.7                  polyrate/2016-2A
ansys/16.2                     cosmologic/turbomole/7.1.1/seq gulp/3.4                       R/3.3.1
caffe/0.9999                   cosmologic/turbomole/7.1.1/smp gulp/4.3                       vasp/5.4.1/vasp
code_saturne/4.0.5             dlpoly/4.07                    lammps/may16                   vasp/5.4.1/vaspsol
code_saturne/4.2.1             dlpoly/classic                 matlab/R2014b                  vasp/5.4.1/vtst
cosmologic/cosmotherm/16       gaussian/g03                   mesmer/4.1                     yadics/04.14a
cosmologic/cosmotherm/17       gaussian/g09.D.01              molcas/8.0/mpi
cosmologic/turbomole/7.1/mpi   gerris/131206/mpi              molcas/8.0/smp
------------------------------------------------------------ /share/modules/utils -----------------------------------------------
benchmark/ior/2.10.3   java/1.7.0            octave/4.0.3          python/2.7.12          python/3.5.1
benchmark/iozone/3.444 java/1.8.0            perl/5.24.1           python/2.7.8
```

- 4 catégories : compilateurs, libraries, applications, outils.
- Versions multiples de logiciels.
- Possibilité de restreindre l'accès des logiciels commerciaux.

Cluster HPC : mutualisation

- **Calcul du pourcentage d'investissement du cluster :**
  - Investissements pris en compte :
    - Matériels : serveurs, stockage, réseaux, ...
    - Logiciels si mutualisés (ex : compilateur).
  - Index de performances :
    - Matériel récent plus performant à investissement égal.
    - Amortissement "comptable" du matériel.
    - Pourcentage dégressif : 100%-95%-85%-70%-50%-0% sur 5 ans.
- **Garantie de temps de calcul :**
  - Calcul de l'utilisation sur deux mois glisants.
  - On compare l'utilisation au pourcentage d'investissement :
    - Utilisation $<$ pourcentage : priorités dans la queue.
    - Utilisation $>$ pourcentage : pas de priorités dans la queue.

Cloud HPC

Cloud HPC :

- **Principe du cloud :**
  - Permet de faire fonctionner plusieurs systèmes d'exploitation sur une même machine physique en partageant les ressources.
- **Architecture du cloud HPC :**
  - **Frontale :** `ouranos.univ-lille1.fr`
    - Serveur sur lequel les utilisateurs gèrent leurs machines virtuelles (VM).
    - Gestion par interface graphique ou par ligne de commandes.
  - **Nœuds de calcul :**
    - Serveurs sur lequel tournent les VM.
    - Les utilisateurs ont accès à leurs VM.
  - **Serveurs de services :**
    - Serveur de stockage.

Cloud HPC : ressources

- **Nœuds de calcul :**
  - 5 nœuds à 2*12 cœurs (Haswell) et 128/256/512 Go de Ram
  - 6 nœuds à 2*4 cœurs (Westmere) et 96 Go de Ram
  - 16 nœuds à 2*6 cœurs (Westmere) et 48 Go de Ram
  - 22 nœuds à 2*4 cœurs (Westmere) et 12 Go de Ram
  - Total : 536 cœurs de calcul et 3 To de Ram
  - Puissance : 8.4 Tflops
- **Infrastructure :**
  - Réseau 10 Gbit/s
  - 215 To de stockage (non garanti)

Cloud HPC : usage et mutualisation

- **Usage :**
  - Permet d'avoir des environnements personnalisés (linux et "windows").
  - L'utilisateur est administrateur sur ses VM.
  - Une VM peut être partagé par plusieurs utilisateurs.
  - Accessible depuis l'extérieur de l'Université.
- **Mutualisation :**
  - Pas de mécanisme de priorité équivalent au Cluster pour l'instant.
  - Ressources dédiées aux financeurs.
- **A noter :**
  - Utilisé principalement par la bioinformatique (Bilille).
  - Intégré à la fédération nationale de Cloud France-Grilles.
  - Intégration future à EGI (European Grid Infrastructure).

Stockage HPC

Stockage HPC :

- **Principes :**
  - Stockage financé par les laboratoires.
  - Liaison 40 Gpbs avec le cluster HPC.
  - Baies MD3860i, iscsi, 60 disques.
  - peut-être mutualisé par plusieurs labos/équipes.
- **Premiers utilisateurs :**
  - Pasteur/Lille 2 : 780 To
  - Inserm : 290 To
  - LML : 390 To
  - Besoin : nécessité d'avoir un stockage important proche du calcul.

Services :

- **Support :**
  - Utilisation des plateformes.
  - Portage et installation des codes (cluster).
- **Expertise en calcul scientifique :**
  - Analyse de performances.
  - Optimisation.
  - Parallélisation (OpenMP et MPI).
- **Formations :**
  - Initiation à Linux pour nos plateformes : 1 jour.
  - Cluster HPC : 1/2 journée.
  - Cloud HPC : 1/2 journée.
  - Parallélisation OpenMP et MPI.
  - Plusieurs sessions de formations possibles par an.
- Aide au passage sur les centres de calculs nationaux.
- Pour toute demande : admin-calcul@univ-lille1.fr

Questions ?