MINES
Saint-Étienne

1816
2016

Cluster de calcul CENTAURE

N. Moulin, Th. Louvancourt, J. Mancuso

INSPIRING
INNOVATION
SINCE 1816

MINES Saint-Étienne | INSPIRING INNOVATION SINCE 1816

Un cluster ?

Définition

- Le terme de cluster (ou grappe, en français) désigne un ensemble d'ordinateurs indépendants, appelés nœuds, tous interconnectés par un réseau dédié.
- On dispose ainsi d'une machine capable de traiter des problèmes de très grande taille, en utilisant la puissance cumulée de ses nœuds.
- Liste des plus « gros » clusters : https://www.top500.org/

Objectifs

- Effectuer de très gros calculs comportant typiquement plusieurs millions (milliards) d'inconnues.
- Effectuer de nombreux calculs séquentiels simultanés.

Clusters et EMSE

Petit historique...

- 1973 : La rotonde...*un centre de calcul*
- 2005 : Ferme de PC
- 2009 : Cluster Pegase (centre SMS) 30k€
- 2016 : Cluster Centaure (investissement EMSE : 80k€)

Clusters et EMSE

Petit historique...

- 1973 : La rotonde...*un centre de calcul*
- 2005 : Ferme de PC
- 2009 : Cluster Pegase (centre SMS) 30k€
- 2016 : Cluster Centaure (investissement EMSE

Clusters et EMSE

Petit historique...

- 1973 : La rotonde...*un centre de calcul*
- 2005 : Ferme de PC
- 2009 : Cluster Pegase (centre SMS) 30k€
- 2016 : Cluster Centaure (investissement EMSE

Clusters et EMSE

Petit historique...

- 1973 :
- 2005 :
- 2009 :
- 2016 :

Descriptif technique

Architecture globale

- 1 nœud maître ou frontale,
- 27 nœuds de calcul (404 cœurs de calcul),
- 1 réseau Ethernet Gigabit (administration)
- 1 réseau Infiniband (calcul)
- 250 Go de disque pour le système (RAID1)
- 10 To de disque dur pour les données (NFS),
- 200 Go ~ 1 To de disque (scratch) sur les nœuds.

Descriptif technique

Configuration des nœuds (rack1 – rack2 - rack3)

- Bi-processeurs Intel Xeon E5-2660 v3 (2,6GHz, 10C/20T, 25Mo de mémoire cache, Turbo)
- 64Go de Ram, 700 Go de scratch
- 1 réseau Infiniband / 1 réseau Ethernet Gigabit
- Bi-processeurs Intel Xeon X-5650 (2,6GHz, 6C, 12Mo de mémoire cache)
- 24Go de Ram, 250 Go de scratch
- 2 réseaux Ethernet Gigabit
- Bi-processeurs Intel Xeon E-5530 (2,4GHz, 4C, 8Mo de mémoire cache)
- 32Go de Ram, 150 Go de scratch
- 2 réseaux Ethernet Gigabit

→ http://services-numeriques.emse.fr/pole-modelisation-et-calcul-numerique/cluster-centaure

Charge du cluster

Ganglia : http://centaure/ganglia/

Descriptif technique

Architecture logicielle

- Distribution CentOS 7 64bits (équivalent RedHat, durée des dépôts…),
- Logiciels de monitoring du cluster : ganglia (http://centaure/ganglia/),
- Compilateurs C/C++/fortran (GNU et Intel),
- Debugger (gdb),
- Python,
- Librairies spécifiques pour le calcul parallèle : openmpi, PETSC, BLAS, Lapack...
- Logiciels de gestion de queue de calcul : SLURM

Descriptif technique

Logiciels scientifiques

- Zset/ZeBuLoN : calcul parallèle par éléments finis (Mines-ParisTech – CdM, Onera, Safran, Mines-St-Etienne),
- CimLib : calcul parallèle par éléments finis (Mines-ParisTech – Cemef),
- Comsol / Abaqus /Ansys : codes commerciaux de calcul par éléments finis,
- GMSH : logiciels de maillage,
- Matlab : calcul scientifique,
- CPLEX : optimisation,
- Outils de visualisation : Paraview, Visit…
- …

→ Base de données : référencement des méthodes numériques / logiciels / codes de calculs https://portailmetier.emse.fr/ApplisWeb/modelisation/index.php

Utilisation du Cluster

Comment accéder à Centaure

→ http://services-numeriques.emse.fr/pole-modelisation-et-calcul-numerique/cluster-centaure/acces-au-cluster-centaure

- Rappel : fonctionne sous **linux**.
- Nom de la machine sur le réseau : **centaure**
- Création de compte via **admin-centaure.emse.fr**
  (N. Moulin, Th. Louvancourt, J. Mancuso)
- Protocole de communication distant : **ssh** (pour Windows, utiliser putty ou Xming)
  ssh -X login centaure.emse.fr
- Transfert des données via les commandes **scp** ou rsync ou avec les logiciels filezilla ou **winscp**.
- Chaque utilisateur dispose d'un espace personnel (/export/home/login) mais **le cluster n'est pas un espace de stockage de données**

Utilisation du Cluster

Modules et environnements

→ http://services-numeriques.emse.fr/pole-modelisation-et-calcul-numerique/cluster-centaure/modules-et-environnements

- La commande module permet de lister et d'utiliser simplement les logiciels, librairies ou utilitaires installés sur le cluster et ainsi configurer l'environnement des utilisateurs. Pour lister l'ensemble des modules existant, il faut utiliser la commande :

  `module avail`

- Si un module vous semble manquant, n'hésitez pas à nous le faire savoir (admin-centaure@emse.fr).

- Pour charger un module, la commande est :

  `module load abaqus/6-14.1`

  Cette commande charge l'ensemble de l'environnement nécessaire à l'exécution du code Abaqus en version 6.14 .

- Pour lister les modules chargés dans votre environnement :

  `module list`

- Pour décharger un module chargés dans votre terminal ou un script :

  `module unload abaqus/6-14.1`

- Cette commande va décharger l'ensemble des modules chargés par le module abaqus/6-14.1.

- Pour décharger tous les modules :

  `module purge`

Utilisation du Cluster

Gestionnaire de travaux et soumission

→ http://services-numeriques.emse.fr/pole-modelisation-et-calcul-numerique/cluster-centaure/gestionnaire-de-travaux-et-soumission

- **Les calculs sur le Cluster s'effectuent par l'intermédiaire d'un gestionnaire de travaux** qui s'occupe de gérer la file d'attente et de lancer les calculs lorsque les ressources demandées sont disponibles.
- Le gestionnaire de travaux du Cluster est **SLURM** (Simple Linux Utility for Resource Management).

Soumission des travaux

- La soumission d'un job se fait avec la commande

  `sbatch slurm.job`

  où `slurm.job` est un fichier de script dans lequel sont contenues des instructions pour SLURM ainsi que des instructions pour le lancement de votre programme.
- Cette commande retourne un numéro de job (JOBID)

MINES Saint-Étienne | INSPIRING INNOVATION SINCE 1816

Utilisation du Cluster

Exemple de script SLURM

```
#!/bin/bash
#SBATCH --job-name=job-slurm-mpi
#SBATCH --mail-user=you@emse.fr
#SBATCH --mail-type=ALL
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=2
#SBATCH --time=01:00:00

module load mpi/openmpi-x86_64

echo ----------------------------------------------------------
echo SLURM_NNODES: $SLURM_NNODES
echo SLURM_JOB_NODELIST: $SLURM_JOB_NODELIST
echo SLURM_SUBMIT_DIR: $SLURM_SUBMIT_DIR
echo SLURM_SUBMIT_HOST: $SLURM_SUBMIT_HOST
echo SLURM_JOB_ID: $SLURM_JOB_ID
echo SLURM_JOB_NAME: $SLURM_JOB_NAME
echo SLURM_JOB_PARTITION: $SLURM_JOB_PARTITION
echo SLURM_NTASKS: $SLURM_NTASKS
echo SLURM_TASKS_PER_NODE: $SLURM_TASKS_PER_NODE
echo SLURM_NTASKS_PER_NODE: $SLURM_NTASKS_PER_NODE
echo ----------------------------------------------------------

echo Generating hostname list...
COMPUTEHOSTLIST=$( scontrol show hostnames $SLURM_JOB_NODELIST |paste -d, -s )
echo ----------------------------------------------------------

echo Creating SCRATCH directories on nodes $SLURM_JOB_NODELIST...
SCRATCH=/scratch/$USER-$SLURM_JOB_ID
srun -n$SLURM_NNODES mkdir -m 770 -p $SCRATCH  || exit $?
echo ----------------------------------------------------------
echo Transferring files from frontend to compute nodes $SLURM_JOB_NODELIST
srun -n$SLURM_NNODES cp -rvf $SLURM_SUBMIT_DIR/* $SCRATCH  || exit $?
echo ----------------------------------------------------------

echo Run -mpi program...
mpirun -np 4 -npernode 2 --mca btl openib,self -host $COMPUTEHOSTLIST
$SLURM_SUBMIT_DIR/mpi_hello_world-host
echo ----------------------------------------------------------

echo Transferring result files from compute nodes to frontend
srun -n$SLURM_NNODES cp -rvf $SCRATCH  $SLURM_SUBMIT_DIR   || exit $?
echo ----------------------------------------------------------
echo Deleting scratch...
srun -n$SLURM_NNODES rm -rvf $SCRATCH  || exit 0
echo ----------------------------------------------------------
```

Utilisation du Cluster

pour simplifier...

- Un script spécifique a été développé pour générer automatiquement différents fichiers SLURM. Pour utiliser ce script, il faut charger d'abord le module correspondant :

  `module load tools/cluster-bin`

- Le fichier .job est créé en exécutant la commande :

  `cluster-create-slurm-script-01.sh`

  suivie d'une option permettant de spécifier le modèle de fichier .job que vous voulez créer (fichier pour lancer Abaqus, Zset, ...).

- La commande **`cluster-create-slurm-script-01.sh -h`** permet de connaître les différents modèles disponibles.

MINES Saint-Étienne | INSPIRING INNOVATION SINCE 1816

Utilisation du Cluster

Gestion des travaux

- La commande pour voir l'état des jobs est :

  `squeue`

  Cette commande ne montre que vos propres jobs !

- La commande pour arrêter un job est :

  `scancel JOBID`

  avec JOBID le numéro du job.

- La commande pour vérifier l'état des nœuds est :

  `sinfo`

- La version graphique :

  `sview`

Conclusions

Quelques recommandations...

- Ne pas hésiter à utiliser le cluster même pour des calculs modestes, cela décharge vos machines personnelles
-  à vos données : ce n’est pas espace de stockage !
-  Utilisation raisonnée (jetons de licences, walltime...)!

Évolutions pour 2018...

- Augmentation/renouvellement de la puissance de calcul
- Machines dédiées au traitement des données à distance
- Mise en place de nouvelles queues de calcul

Merci de votre attention

INSTITUT
Mines-Télécom

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE L'ÉCONOMIE
DE L'INDUSTRIE ET DU NUMÉRIQUE

INSPIRING
INNOVATION
SINCE 1816