SLURM : Simple Linux Utility for Resource Management

Alexandre Ancel

ancel@math.unistra.fr

Institut de Recherche Mathématique Avancée (IRMA),
Centre de Modélisation et de Simulation de Strasbourg (Cemosis),
Université de Strasbourg

15 mars 2016

Plan de la présentation

Plan

Slurm : Introduction

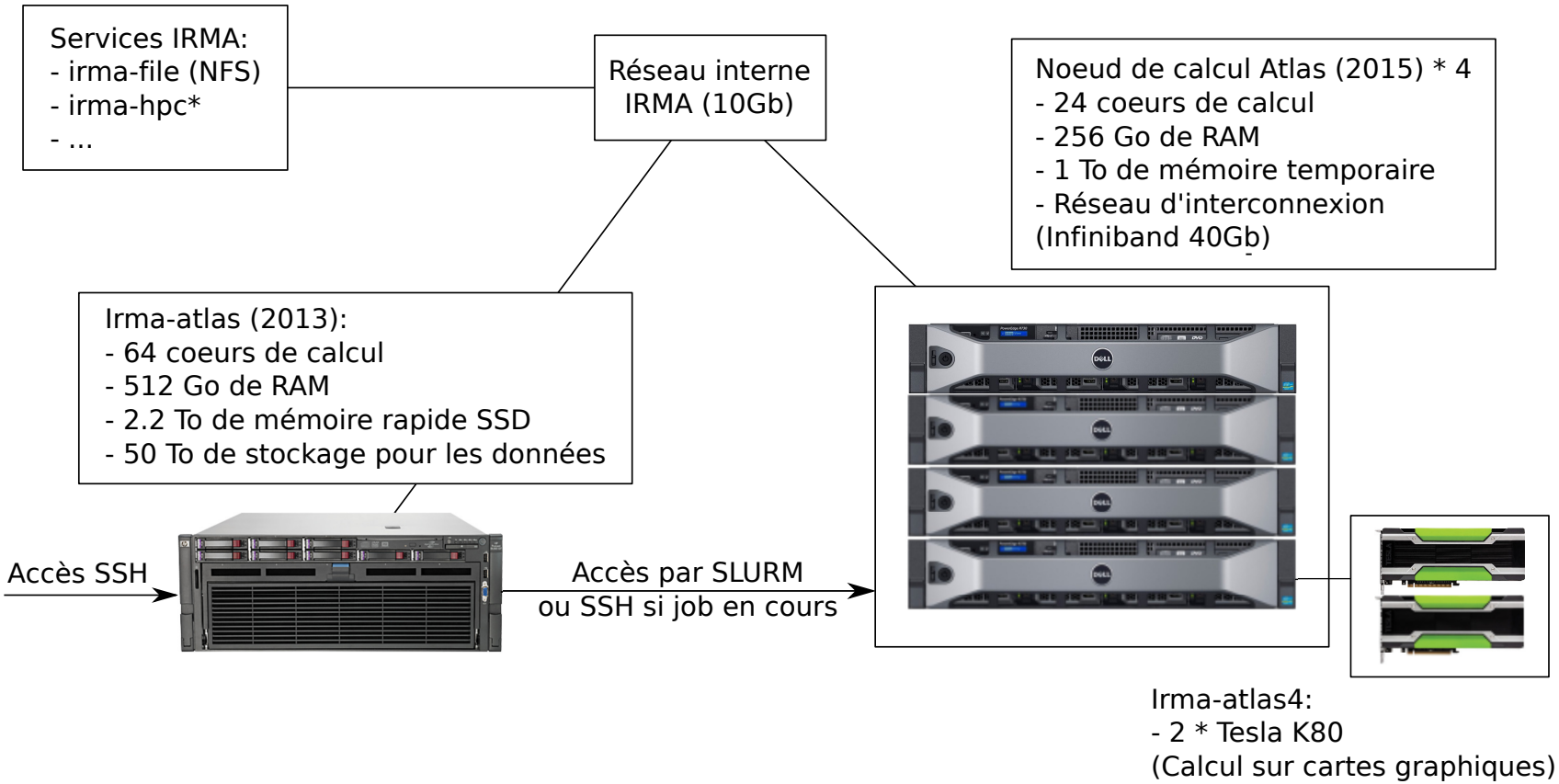
- Slurm :
  - Gestionnaire de ressources et d'ordonnancement de tâches
  - Permet l'allocation de ressources de manière exclusive ou non sur une période de temps définie
  - Permet la gestion de conflits avec l'utilisation de files d'attentes
  - Extensible à l'aide de plugins en C
  - Portable (Linux, OS X), Gratuit et Open-Source
  - Créé au début des années 2000 :
    Slurm: Simple Linux Utility for Resource Management,
    A. Yoo, M. Jette, and M. Grondona,
    Job Scheduling Strategies for Parallel Processing, volume 2862 of Lecture Notes in Computer Science, pages 44-60, Springer-Verlag, 2003.
  - 2010 : Création de SchedMD : Services autour de slurm + maintenance
- Alternatives :
  - OAR (Gratuit/Open-Source, Inria) - ex : Grid'5000
  - Torque (Open-Source, Adaptive Computing)
  - LoadLeveler (IBM) - ex : Machines SGI
  - PBS (Altair)
  - ...

Slurm

- Utilisateurs de slurm ? (Source : http://slurm.schedmd.com/)
  - Tianhe-2, National Supercomputing Center, Chine (1/500) (Linpack Max : 33.863 PFLOPS)
  - Sequoia, Oak Ridge National Laboratory, Etats-Unis (3/500) (Linpack Max : 20.133 PFLOPS)
  - Piz Daint, Swiss National Supercomputing Centre, Suisse (7/500) (Linpack Max : 7.779 PFLOPS)
  - TGCC, GENCI/CEA, France
  - ...
  - Plus proche de nous, mésocentre de Strasbourg

Plan

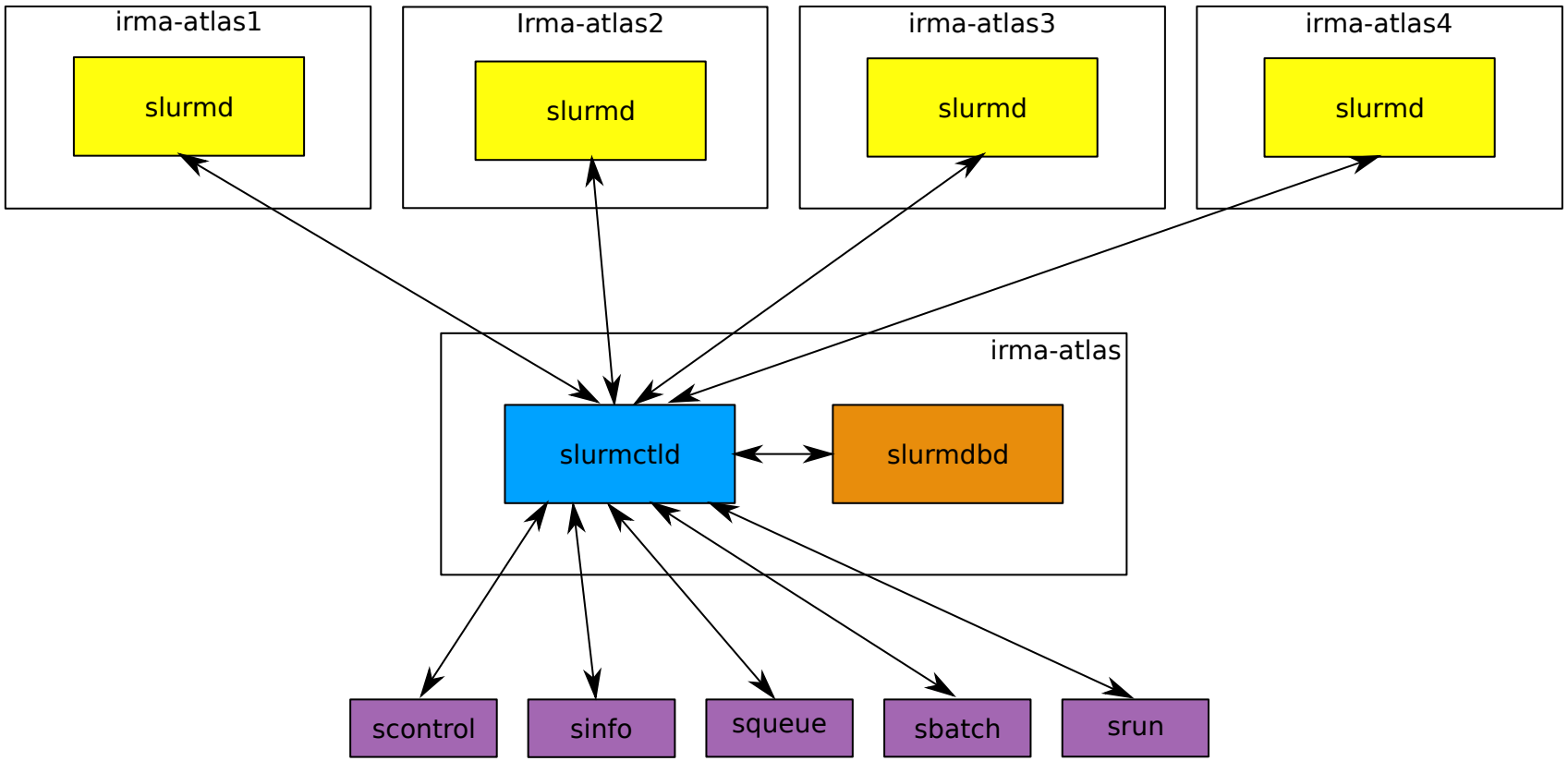
Configuration actuelle : Irma-atlas



- Ubuntu 14.04 : choix un peu à contre courant
  Flexibilité, mises à jour, support (landscape)

Slurm : Services

- Slurm : slurmctld, slurmd, slurmdbd (optionnel)
- Munge : Service d'authentification, créé pour le HPC



Slurm

- Installation : slurm-llnl
- Un seul fichier de configuration permet de configurer slurm :
  - /etc/slurm-llnl/slurm.conf
  - Ce fichier doit être identique sur l'ensemble des machines du cluster
  - Quickstart : https://computing.llnl.gov/linux/slurm/quickstart_admin.html
- Quelques options générales :
  - ClusterName : Nom du cluster
  - ControlMachine, ControlAddr : Machine où est hébergé slurmctld
  - AuthType : Type d'authentification, Munge par défaut
  - SlurmLogFile, SlurmctldLogFile : Fichiers de log
  - Prolog, Epilog : Fichier executé en début ou fin de job
  - UsePam : Utilisation de PAM pour l'accès aux noeuds
  - PriorityType : Configuration des priorités
    - priority/basic : First In First Out
    - priority/multifactor : Age, taille du job, partition, qos, fair-share (slurmdbd)

Slurm

- Définition des noeuds et leurs propriétés :

```
NodeName=irma-atlas[1-3] feature=intel,xeon,haswell \
    Sockets=2 CoresPerSocket=12 ThreadsPerCore=2 \
    State=UNKNOWN
NodeName=irma-atlas4 feature=intel,xeon,haswell,gpu, \
    nvidia,K80 Sockets=2 CoresPerSocket=12 \
    ThreadsPerCore=2 State=UNKNOWN
```

- Configuration des partitions (files d'attentes) :

```
PartitionName=public Nodes=irma-atlas[1-4] \
    Default=YES MaxTime=INFINITE State=UP
PartitionName=K80 Nodes=irma-atlas4 Default=NO \
    MaxTime=INFINITE State=UP
```

Utilisation de PAM

- Que se passe-t-il si l'utilisateur veut vérifier en temps réel l'exécution de son job ?
  - Solution : SSH
  - Mais limiter la connexions uniquement aux noeuds où un job est lancé !
  - Eviter les connexions intempestives sur les noeuds et les perturbations d'autres jobs
- Utilisation de PAM
  - Installer une librairie supplémentaire : libpam-slurm
  - Editer la configuration d'authentification de PAM (/etc/pam.d/common-auth) et ajouter :

```
account required /lib/security/pam_slurm.so
```

  - Vérifier que l'accès soit désactivé par SSH (AllowUsers)
  - Modifier le script d'Epilog de Slurm pour exclure des utilisateurs n'ayant pas de job lancé
    Référence :
    https://github.com/SchedMD/slurm/blob/master/etc/slurm.epilog.clean

Comptabilité

- Permet d'établir des rapports d'utilisation (statistiques)
- Mise en place de fair-share (en fonction des ressources déjà allouées et déjà consommées)
- Réglage de la comptabilité : (/etc/slurm-llnl/slurm.conf)
  - AccountingStorageType :
    accounting_storage/none,
    accounting_storage/filetxt,
    accounting_storage/slurmdbd
  - Fichiers texte : AccountingStorageLoc
  - SlurmDBD : AccountingStorageHost, AccountingStoragePort
- Pour plus d'informations :
  https://computing.llnl.gov/linux/slurm/accounting.html

SlurmDBD

- /etc/slurm-llnl/slurmdbd.conf :
  - Configuration du type de base de données (MySQL, PostgreSQL)
  - Configuration de l'accès à la base de données
  - Configuration initiale de la base de données à faire
- Configuration de la base : sacctmgr
  - Association = cluster + account + user name + partition name (optionnel)
  - cluster = Nom du cluster actuel
  - account = Groupes d'utilisateurs
  - user = Nom d'utilisateur (même que le nom linux)
  - Activer les associations :
    /etc/slurm-llnl/slurm.conf : AccountingStorageEnforce=associations
- Ajout de limites : AccountingStorageEnforce=associations, limits
  Ordre : user > account > cluster > pas de limites
  - Exemples de limites :
    MaxSubmitJobs, MaxNodesPerJob, MaxCPUsPerJob, ...

Commandes : Administrateur

- scontrol : Gestion de la configuration de slurm

```
sudo scontrol show <nodes|partitions>
sudo scontrol update NodeName="irma-atlas4" State=DOWN
sudo scontrol reboot_nodes irma-atlas4
sudo scontrol suspend <job_id>
sudo scontrol resume <job_id>
```

- sacctmgr : Gestion des différents comptes pour la comptabilité

```
sudo sacctmgr add cluster atlas
sudo sacctmgr add account slurm-users
sudo sacctmgr add user ancel DefaultAccount=slurm-users
sudo sacctmgr list associations user=ancel
sudo sacctmgr modify user ancel account=slurm-users \
set MaxJobs=2
sudo sacctmgr delete user ancel
```

Plan

Commandes utiles : Utilisateur

- sinfo : Informations sur les noeuds et les partitions
- squeue : Afficher la file d'attente actuelle
- salloc : Allouer des ressources et lance une commande (e.g. job interactif)
- sbatch : Lancer un script en batch
- srun : Lancer un job en parallèle (dans une allocation salloc ou sbatch)
  - mpirun : Peut être utilisé à la place de srun. L'infrastructure de slurm est automatiquement utilisée.
- scancel : Annuler un job à partir de son identifiant

Exemples de sessions types : session interactive

- Demande d'allocation de ressources avec la commande salloc

```
salloc -t "02:00:00" -p public -n 24 -w irma-atlas4
```

- La commande se met en attente jusqu'à disponibilité de la ressource
- Chargement des librairies et exécutables dans l'environnement (environnement-modules)

```
module load ...
```

- Lancement du code utilisateur

```
mpirun -np 24 ./feelpp_qs_laplacian \
--config-file ./qs_laplacian_2d.cfg
```

- Déconnexion

Exemples de sessions types : session en batch

- Création d'un fichier batch : test.slurm

```
#!/bin/bash
#SBATCH -p public
# number of cores
#SBATCH -n 96
# If hyperthreading is enabled and you do not want to use it
#SBATCH --ntasks-per-core 1
# max time of exec (will be killed afterwards)
#SBATCH -t 12:00:00
# specify execution constraitns
#SBATCH --constraint \"intel\"
# send a mail at the end of the exec
#SBATCH --mail-type=END
#SBATCH --mail-user=login@server.com

module load ...

mpirun -np 24 ./feelpp_qs_laplacian \
--config-file ./qs_laplacian_2d.cfg
```

Exemples de sessions types : session en batch

- Soumission du job

```
sbatch test.slurm
```

- Placement en file d'attente jusqu'à disponibilité des ressources
- Exécution du code jusqu'à terminaison

Conclusion

- Slurm propose un environnement pour cluster permettant :
  - L'allocation de ressources
  - La soumission de jobs
- Installation simple et rapide pour un environnement de base
- Perspectives :
  - Ajout d'outils de monitoring pour les utilisateurs :
    Prévenir lors de dépassement RAM/Swap ...