Introduction à SLURM

raphael.pasquier@univ-antilles.fr

Université des Antilles

17 juin 2022

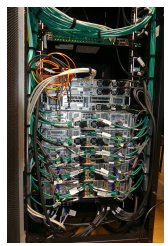
Simple **L**inux **U**tility for **R**esource **M**anagement

Le cluster Exocet se compose :

- Deux serveurs frontaux **exocet1** et **exocet2**
- vingt cinq nœuds "Calcul" (**node01** à **node25**).
- Un nœud "grosse mémoire RAM" (**mem01**) ayant 1 536 Go de mémoire RAM.
- Cinq nœuds graphiques.

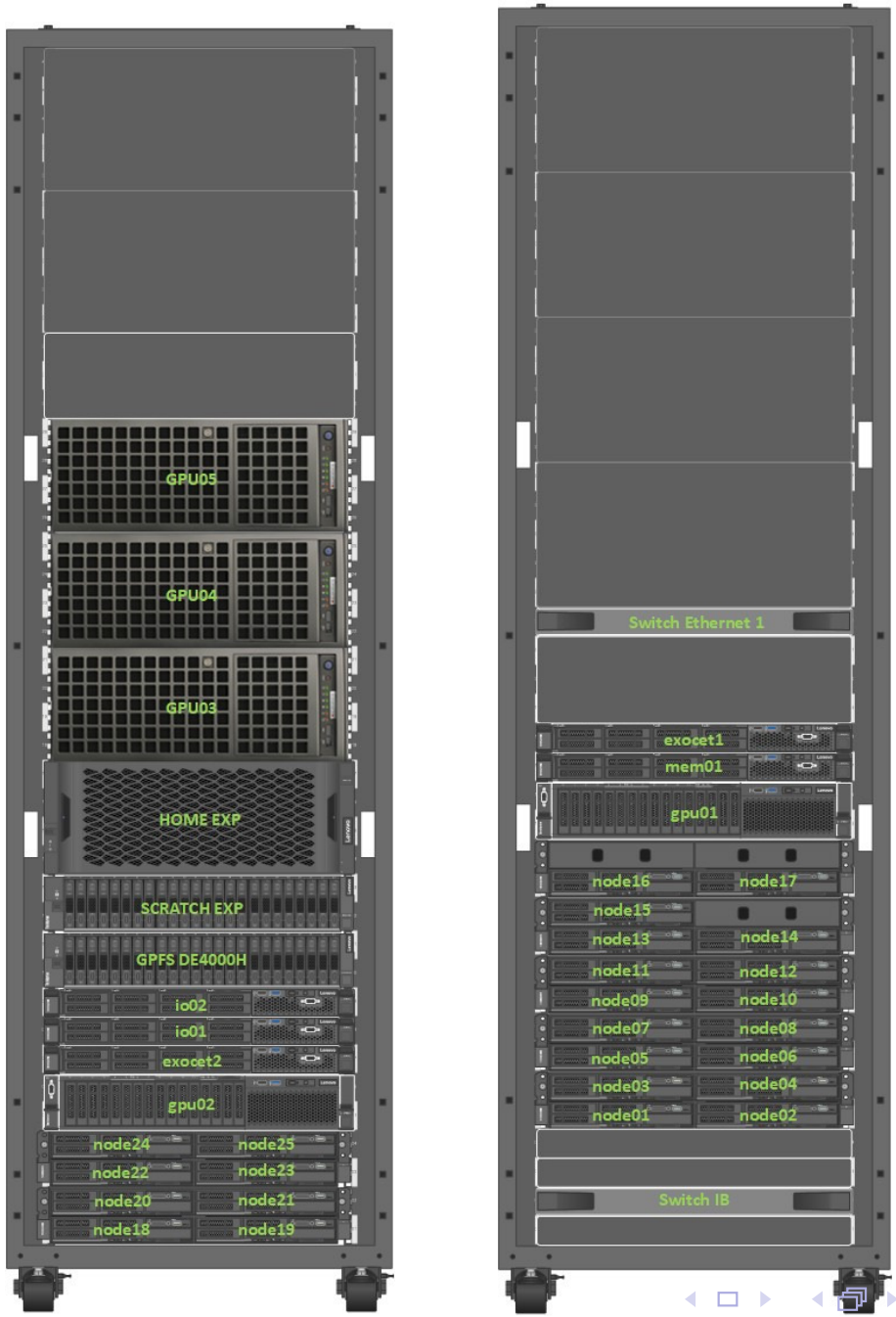
La première baie

la deuxième



Les faces arrières

Le cluster Exocet



Principes

La règle d'or pour utiliser les ressources d'Exocet :

- Pour utiliser un ou plusieurs nœuds (pour un calcul sur CPU ou GPU), il faut d'abord faire une demande auprès de SLURM (faire une soumission ou une requête).

Vocabulaire SLURM :

- Un calcul s'appelle aussi un "Job". On dira qu'on soumet un job. Un calcul demande des ressources (mémoire RAM, cœurs de CPU, GPU).

La raison d'être de SLURM est donc

- éviter les disputes entre utilisateurs,
- utiliser au mieux les ressources du cluster en fonction des demandes des utilisateurs.

SLURM se compose :

- de démons communiquant avec des bases de données,
- d'applications (srun, sbatch, sinfo, etc) communiquant avec ces démons.

Les concepts essentiels

- **Partitions**. Le cluster est partagé en plusieurs partitions (un nœud n'appartient qu'à une seule partition).
  - Partition "**cpu**" (partition pa défaut) : les nœuds node01 à node25,
  - Partition "**mem**" : le nœud mem01
  - Partition "**gpu-v100**" : le nœud gpu01,
  - Partition "**gpu-t4**" : le nœud gpu02,
  - Partition "**gpu-p5000**" : les nœuds gpu03, à gpu05.

Pour vos soumissions à SLURM, il faudra préciser la partition que vous souhaitez utiliser (sauf éventuellement "cpu").

- **Modes** d'utilisation du cluster
  - mode "**direct**", "immédiat" ou "intéractif" (MATLAB, R, écrire ou déboguer un script, etc). On utilise la commande **srun**.
  - mode "**batch**" ou "différé" (script R qui dure très longtemps, etc). On utilise la commande **sbatch**.

Visualisation de l'utilisation du cluster

SLURM permet de connaître la disponibilité des ressources via la commande **sinfo**. Quelques exemples d'utilisation :

- **$ sinfo**

PARTITION	AVAIL	TIMELIMIT	NODES	STATE	NODELIST
gpu-v100	up	1-00:00:00	1	idle	gpu01
gpu-t4	up	1-00:00:00	1	idle	gpu02
gpu-p5000	up	1-00:00:00	3	idle	gpu[03-05]
cpu*	up	infinite	1	mix	node01
cpu*	up	infinite	24	idle	node[02-25]
mem	up	infinite	1	idle	mem01

- **$ sinfo -o%C**

CPUS(A/I/O/T)
365/140/539/1044

Visualisation de l'utilisation du cluster

- **$ sinfo -o"%C %R"**

```
CPUS(A/I/O/T) PARTITION
0/0/36/36     gpu-v100
0/0/36/36     gpu-t4
0/36/0/36     gpu-p5000
365/104/431/900cpu
0/0/36/36     mem
```

Pour voir les jobs qui tournent, ou qui sont en attente, on utilise la comamnde **squeue** :

- **$ squeue**

JOBID	PARTITION	NAME	USER	STATE	TIME	NODES	NODELIST(REASON)
66378	cpu	bash	dcouvin	R	42:10	1	node11
66376	cpu	lance_si	dcouvin	R	16:35:25	1	node11
66375	cpu	lance_kl	dcouvin	R	19:22:20	1	node11
20775	cpu	singular	ebarnaci	R	3-17:19:24	1	node01
66373	cpu	afai-sil	ebiabian	R	21:23:20	1	node11
66377	cpu	test_WRF	rcece	R	2:32:24	10	node[02-09,13-14]

Visualisation de l'utilisation du cluster et suppression d'un job

- **$ squeue -o"%C %u"**

  CPUS USER
  1 dcouvin
  1 dcouvin
  1 dcouvin
  1 ebarnacin
  1 ebiabiany
  360 rcece

Pour supprimer le job d'identifiant ID (**squeue** affiche l'ID des jobs), on tape :

- **scancel ID**

Mode direct (srun)

On demande un noeud pour l'utiliser maintenant avec son clavier, son écran (et sa souris si l'application qu'on souhaite utiliser a une interface graphique).
Quelques exemples d'utilisation :

- **srun -p gpu-v100 –x11 –gres=gpu:1 –pty bash**
  On demande un noeud de la partition gpu-v100 avec le renvoi X11, et une seule carte graphique pour exécuter un terminal Bash.
- **srun -p cpu -N 1 -n 36 –pty bash**
  On demande un noeud de la partition cpu, 36 coeurs CPU sur un même noeud (-N 1) pour exécuter un terminal Bash.
- **srun -p cpu hostname**
  Pour s'amuser, on demande un noeud de la partition puis on exécute le programme `hostname` qui affiche le nom de l'ordinateur sur lequel il est exécuté. On revient aussitôt sur exocet1.

Mode différé ou batch (sbatch)

On demande un ou plusieurs noeud pour exécuter un programme.
Cette demande (soumission) est enregistrée dans une file d'attente.
SLURM lancera le programme quand les ressources nécessaires seront disponibles.
On **ne** peut **pas intéragir** avec le **programme** avec son clavier et son écran.
Contrairement à **srun**, il est préférable de mettre dans un fichier les options de **sbatch**, disons `lance_job.sh`, et d'invoquer la commande **sbatch** ainsi :

- **sbatch** `lance_job.sh`

Les programmes **srun** et **sbatch** ont les mêmes options.

Exemple de fichier d'entrée pour sbatch

```
#!/bin/bash

### Nom du job
#SBATCH -J test_openmpi

#SBATCH -p cpu

### Choix du nombre de noeuds
#SBATCH -N 1

### Choix du nombre de processus
#SBATCH -n 25

### Envoi d'un courriel a la fin du job
#SBATCH –mail-type END #SBATCH –mail-user
Bugs.Bunny@gmail.com

### lancement du programme

mpirun -np 1 ./master_mpi : -np 24 ./slave_mpi
```