

Utilisation et bonnes pratiques d'un centre de calculs

Bastien DI PIERRO

Université Claude Bernard Lyon 1

25 octobre 2018

- 1 Philosophie
- 2 Architecture
- 3 Gestionnaire de batch
- 4 Bonnes pratiques

Première approche d'un calculateur



main.C



kernel.F90

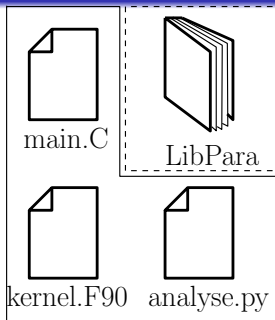


analyse.py

Démarche

- Code de calcul : noyau + analyse + bibliothèques ...

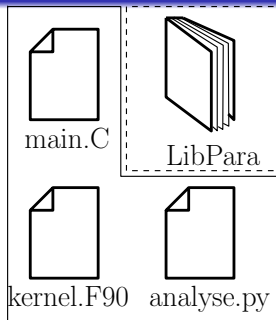
Première approche d'un calculateur



Démarche

- Code de calcul : noyau + analyse + bibliothèques ...
- Parallélisation : + **bibliothèque parallèle** (OpenMP, MPI)

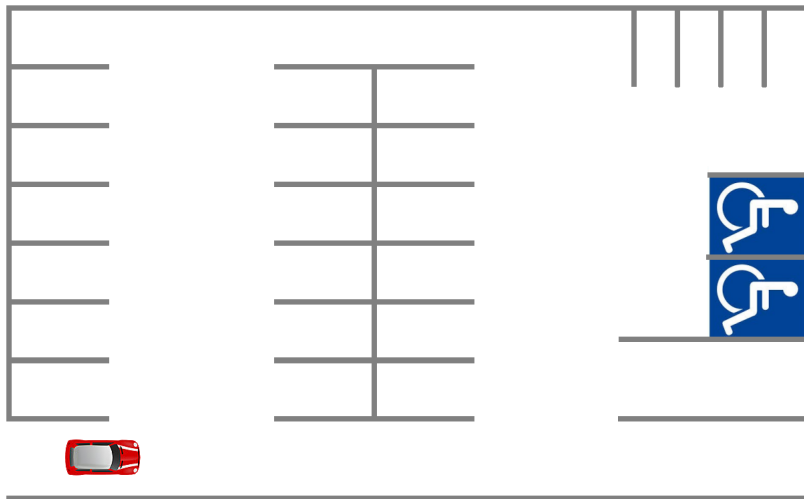
Première approche d'un calculateur



Démarche

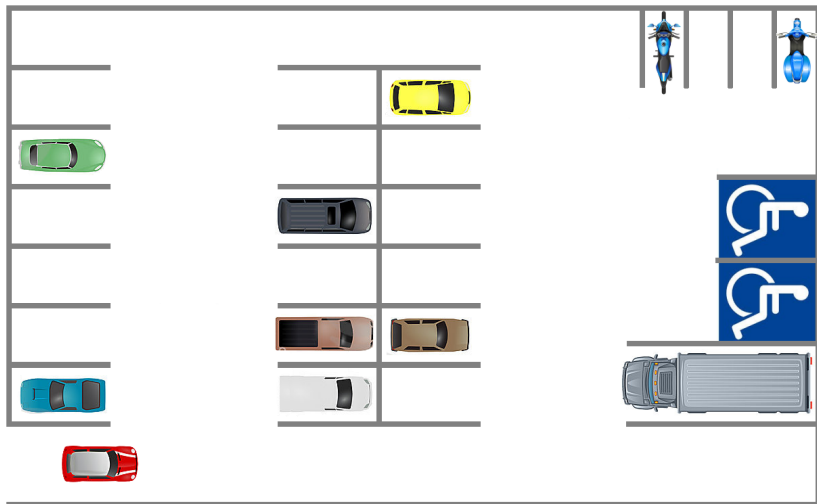
- Code de calcul : noyau + analyse + bibliothèques ...
- Parallélisation : + **bibliothèque parallèle** (OpenMP, MPI)
- Portabilité sur un calculateur :
 - $\approx 10^2 - 10^6$ coeurs de calculs
 - $\approx 100.Go - 100.To$ RAM
 - $\approx 1.To - 1Po$ ROM

Analogie : Parking



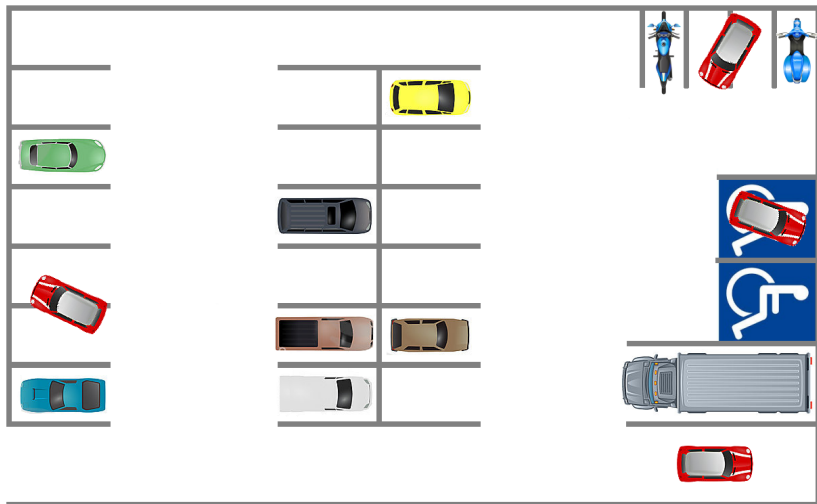
Parking : structure à grande échelle

Analogie : Parking



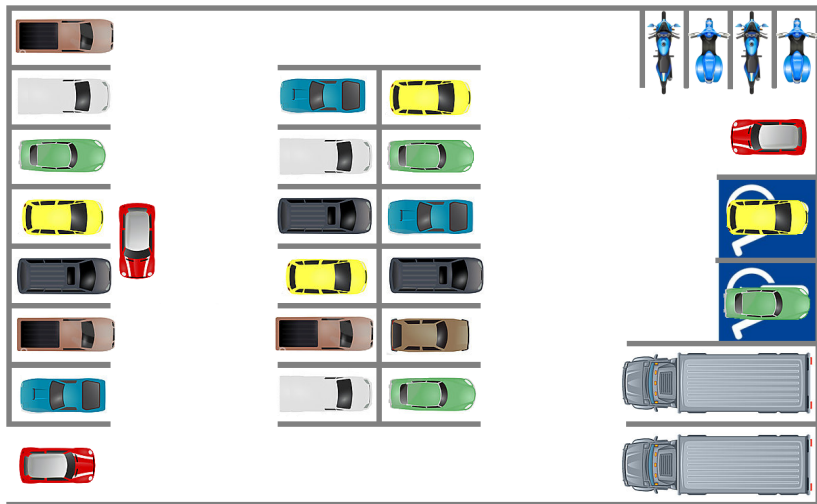
Parking : structure communautaire

Analogie : Parking



Parking : structure communautaire $\Rightarrow$ Règles

Analogie : Parking



Parking : structure communautaire $\Rightarrow$ Règles

Balance

Avantages

- Accès à des ressources numériques performantes
- Mise à disposition gratuite

Balance

Avantages

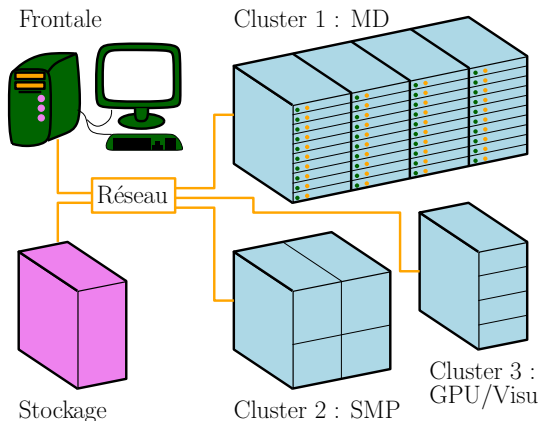
- Accès à des ressources numériques performantes
- Mise à disposition gratuite

Règles

- Utilisation responsable des ressources
- Respect des autres utilisateurs/responsables techniques
- Allocation limitée et souvent sur demande/dossier

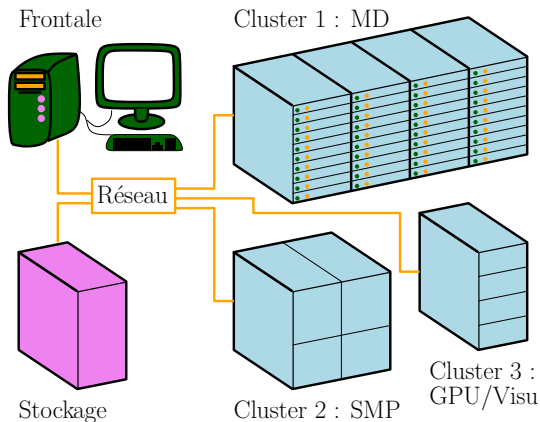
- 1 Philosophie
- 2 Architecture
- 3 Gestionnaire de batch
- 4 Bonnes pratiques

Architecture d'un centre de calcul



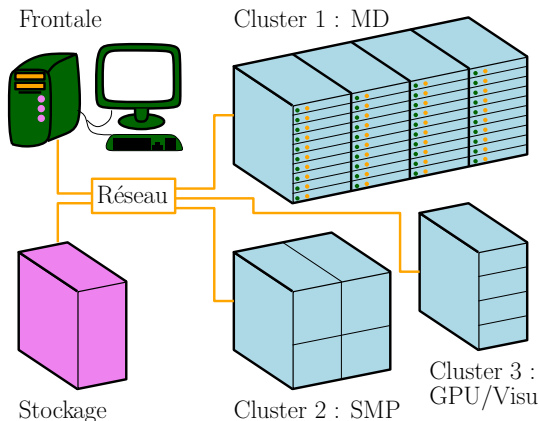
- Frontale
- Calculateurs (Mémoire distribuée, partagée, visualisation)
- Baie de stockage

Architecture d'un centre de calcul



Frontale : Organiser les tâches (réserver des ressources, lancer les calculs, copier/déplacer les fichiers ...) entre les différents organes.
Ne doit surtout pas effectuer de calculs !

Architecture d'un centre de calcul



```
bastien@uranus:~$ ssh bdiptierro@p2chpd
~~~ Bienvenue au P2CHPD de la FLMSN ~~~

~~~~ Computing ressources ~~~~~

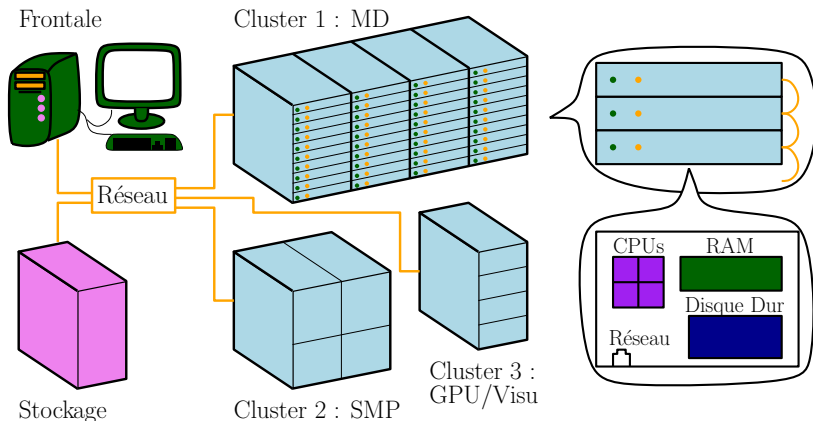
Nodes online = 320
Nodes free on parallel1 = 17/36
Nodes free on parallel2 = 14/36
Nodes free on parallel = 152/220
Jobs currently running = 74

[bdiptierro@p2chpd-login3 ~]$ ./monCode
```



Frontale : Organiser les tâches (réserver des ressources, lancer les calculs, copier/déplacer les fichiers ...) entre les différents organes.
Ne doit surtout pas effectuer de calculs !

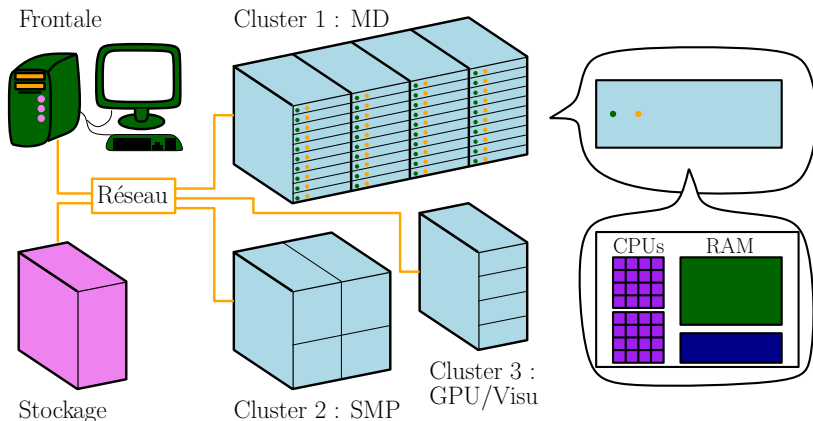
Architecture d'un centre de calcul



Calculateur à mémoire distribuée : Armoire à "rack" reliés entre eux par un réseau dédié.

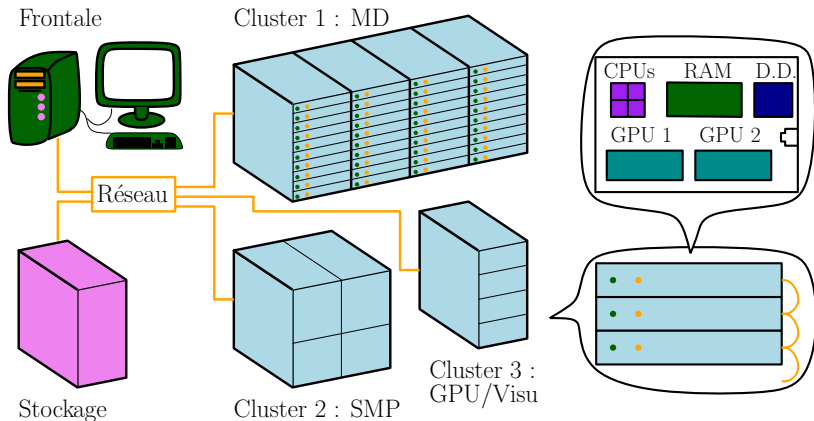
Rack = 1-4 processeurs (4-32 coeurs) + RAM dédiée (4-32Go) + 1 disque dur (mémoire temporaire "scratch") + carte réseau

Architecture d'un centre de calcul



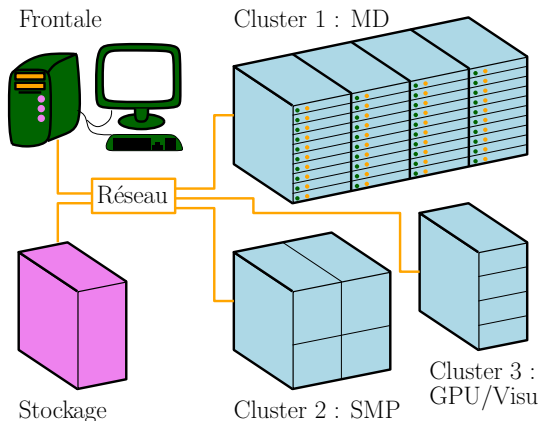
Calculateur à mémoire partagée : connexion de "gros ordinateurs".
"Gros ordinateur" : 4-32 processeurs (32-256 coeurs) + mémoire partagée (O(100 Go)) + disque dur.

Architecture d'un centre de calcul



Calculateur de GPU/Visualisation : calculateur disposant de plusieurs cartes graphiques. Dédié aux calculs OpenACC/Cuda (GPGPU) ou pour la visualisation de résultats.

Architecture d'un centre de calcul



Baie de stockage : Serveur de sauvegarde ("store") des données de façon durable. **On ne doit pas lancer de caculs depuis ce point de montage !**

- 1 Philosophie
- 2 Architecture
- 3 Gestionnaire de batch
- 4 Bonnes pratiques

Gestionnaire de batch (Job Scheduler)

Qu'est-ce ?

Programme de gestion de la file d'exécution des différents calculateurs.

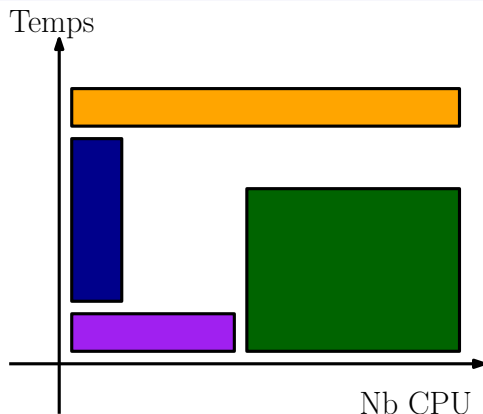
Accessible depuis la frontale.

Rôle

Organiser et agencer l'exécution des tâches (calculs des utilisateurs) en respectant des règles précises

“TETRIS”

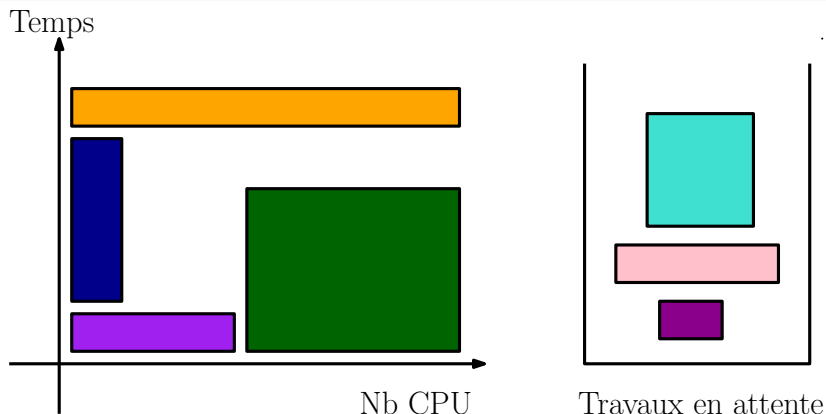
Gestionnaire de batch (Job Scheduler)



Rôle

Organiser et agencer l'exécution des tâches (calculs des utilisateurs) en respectant des règles précises

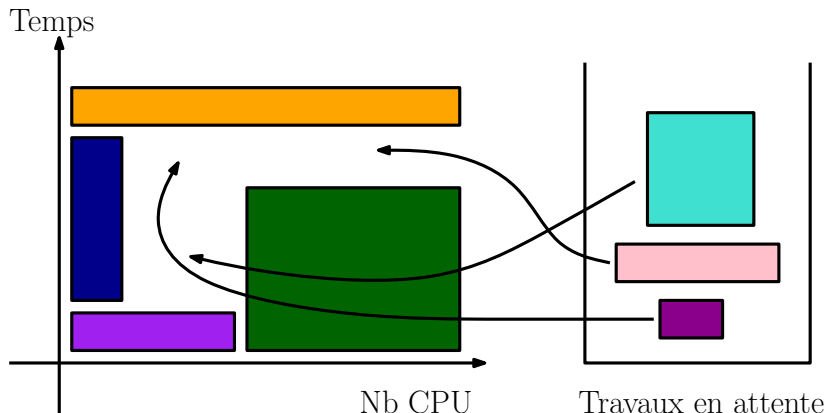
Gestionnaire de batch (Job Scheduler)



Rôle

Organiser et agencer l'exécution des tâches (calculs des utilisateurs) en respectant des règles précises

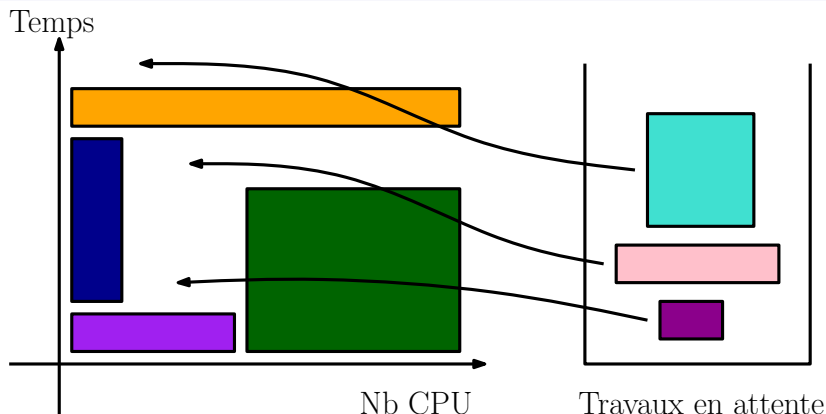
Gestionnaire de batch (Job Scheduler)



Rôle

Organiser et agencer l'exécution des tâches (calculs des utilisateurs) en respectant des règles précises

Gestionnaire de batch (Job Scheduler)



Rôle

Organiser et agencer l'exécution des tâches (calculs des utilisateurs) en respectant des règles précises

Gestionnaire de batch (Job Scheduler)

Règles

- Fixées par les responsables
- **Identiques pour tout le monde**
- Balance entre “placement idéal” et bien-être des utilisateurs

Gestionnaire de batch (Job Scheduler)

Règles

- Fixées par les responsables
- **Identiques pour tout le monde**
- Balance entre “placement idéal” et bien-être des utilisateurs

Placement idéal (consommation)

- Les tâches volumineuses sont prioritaires
- Les petites tâches “bouchent les trous”

Bien-être utilisateurs

- Les petits travaux sont prioritaires
- Plus on demande de ressources, plus on attend !
i.e. on laisse la place aux autres

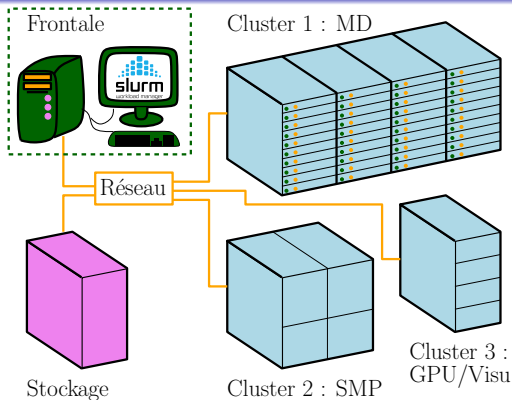
Les différents gestionnaires

- Portable Batch System (PBS)
- TORQUE PBS
- IBM Load Sharing Facility (LSF)
- Open Lava
- Sun Grid Engine (SGE)
- Univa Grid Engine (UGE)
- Argent Job Scheduler (AJS)
- Simple Linux Utility for Ressource Management (SLURM)
- ...

Les différents gestionnaires

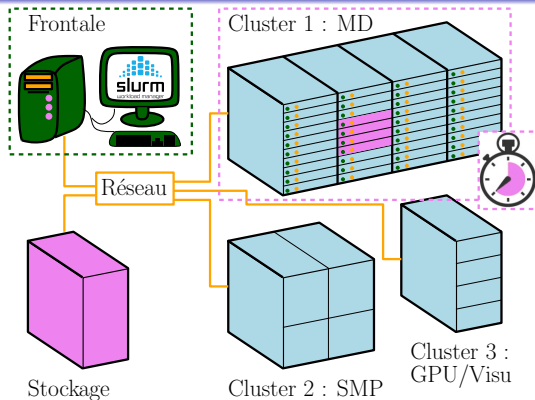
- Simple Linux Utility for Ressource Management (SLURM)

Soumission d'un job



- Depuis la frontale, avec le **gestionnaire de batch**

Soumission d'un job



- Depuis la frontale, avec le **gestionnaire de batch**
- On spécifie :
 - Le calculateur : file d'exécution (partition)
 - Le nombre de noeuds, le nombre de coeurs par noeud
 - La durée du calcul

Gestionnaire de batch : script de lancement (run.sh)

```
#!/bin/bash
```

```
## parametre batch manager slurm
```

```
#SBATCH --partition=parallel1    ## choix du calculateur
#SBATCH --time=24:00:00          ## temps d'exécution
#SBATCH --nodes=8                 ## nombre de noeuds
#SBATCH --ntasks-per-node=8      ## nombre coeurs par noeud
#SBATCH --mem-per-cpu=780        ## RAM par coeur
#SBATCH --mail-type=end          ## envoi d'un email
#SBATCH --mail-user=bastien.di-pierro@univ-lyon1.fr
#SBATCH -o output_%j.txt         ## fichier de sortie standard
#SBATCH -e stderr_%j.txt         ## fichier d'erreur
```

```
echo "Starting at `date`"        ## affichage recapitulatif
echo "Running on hosts: $SLURM_NODELIST"
echo "Running on $SLURM_NNODES nodes."
echo "Running on $SLURM_NPROCS processors."
```

```
## lancement du calcul
```

```
## v----- utilitaire d execution MPI (mpirun)
```

```
    srun ./poussins
```

```
## ^----- mon executable
```


Gestionnaire de batch : commandes utiles

- Exécution du script (sbatch)

```
[bdipierro@p2chpd-login3 poussins]$ sbatch run.sh  
Submitted batch job 121191  
[bdipierro@p2chpd-login3 poussins]$
```

Mon job part en file d'attente ou d'exécution avec l'id 121191

Gestionnaire de batch : commandes utiles

- Exécution du script (sbatch)

```
[bdipierro@p2chpd-login3 poussins]$ sbatch run.sh  
Submitted batch job 121191  
[bdipierro@p2chpd-login3 poussins]$
```

Mon job part en file d'attente ou d'exécution avec l'id 121191

- Etat de mes jobs (squeue)

```
[bdipierro@p2chpd-login3 poussins]$ squeue -u bdipierro
```

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST(Reason)
121095	parallel	CRV.sh	bdipierr	R	03:43:50	4	node[303-307]
121191	parallel	run.sh	bdipierr	PD	0:00	8	(Resources)

R : Run (exécution)

PD : Pending (attente)

Gestionnaire de batch : commandes utiles

- Exécution du script (sbatch)

```
[bdipierro@p2chpd-login3 poussins]$ sbatch run.sh  
Submitted batch job 121191  
[bdipierro@p2chpd-login3 poussins]$
```

Mon job part en file d'attente ou d'exécution avec l'id 121191

- Etat de mes jobs (squeue)

```
[bdipierro@p2chpd-login3 poussins]$ squeue -u bdipierro
```

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST(Reason)
121095	parallel	CRV.sh	bdipierr	R	03:43:50	4	node[303-307]
121191	parallel	run.sh	bdipierr	PD	0:00	8	(Resources)

R : Run (exécution)

PD : Pending (attente)

- Annulation d'un job

```
[bdipierro@p2chpd-login3 poussins]$ scancel 121191
```

Exécution en mode Interactif

- Réserve des ressources

Depuis la frontale, je réserve 2 unités de calculs :

```
[bdipierro@p2chpd-login3 src_anne]$ salloc -n 2 bash
```

Mêmes options que le script !

Exécution en mode Interactif

- Réserve des ressources

Depuis la frontale, je réserve 2 unités de calculs :

```
[bdipierro@p2chpd-login3 src_anne]$ salloc -n 2 bash
```

Mêmes options que le script !

- Exécution du programme

```
srun --mpi=pmi2 --cpu_bind=Verbose ./poussins
```

```
Parallel OpenFlow Unsteady Spectral Solver
for Inhomogeneous Navier-Stokes
```

```
*****
```

```
Maillage du domaine :      Taille du domaine :
512 x 512 x 512           12.56637 x 12.56637 x 12.56637
```

```
Pas de temps = 1.0E-004      CFL = 0.5
Schema integration : Adams Bashforth 2
Resolution NON LINEAIRE des equations
```

```
Donnees de controle :
Re= 1000.0, Sc= 1.0, q= 0.50, s= 1.0
```

Files d'exécution - Information

```
[bdipierro@p2chpd-login3 ~]$ sinfo
```

PARTITION	AVAIL	TIMELIMIT	NODES	STATE	NODELIST
parallel1	up	31-03:00:0	36	down*	node[001-032,069-072]
parallel2	up	31-03:00:0	36	down*	node[033-068]
parallel	up	2-04:00:00	9	down*	node[312,318,338,403]
parallel	up	2-04:00:00	8	drain	node[308,310,342,347]
parallel	up	2-04:00:00	152	mix	node[303-307,309,311,
parallel	up	2-04:00:00	6	idle	node[432,440,449,456]
test	up	5:00	6	idle	node[432,440,449,456]
debug	up	6:00:00	152	mix	node[303-307,309,311,
debug	up	6:00:00	6	idle	node[432,440,449,456]
dev	up	6:00:00	1	mix	node302
fatjob	up	2-12:00:00	2	down*	node[476-477]
fatjob	up	2-12:00:00	10	mix	node[397-402,472-475]
long	up	5-12:00:00	3	down*	node[312,318,338]
long	up	5-12:00:00	3	drain	node[310,342,347]
long	up	5-12:00:00	37	alloc	node[309,311,313-317]
longRes	up	30-12:00:0	152	mix	node[303-307,309,311,
longRes	up	30-12:00:0	6	idle	node[432,440,449,456]

Files d'exécution

- 1 file = 1 besoin = 1 type de calcul !
- Exemple du p2chpd :

Files d'exécution

- 1 file = 1 besoin = 1 type de calcul !
- Exemple du p2chpd :
 - parallel : jobs "classiques" (jusqu'à 16 noeuds, 2 jours)
 - paralle1 & parallel2 : anciens calculateurs (moins performants)

Files d'exécution

- 1 file = 1 besoin = 1 type de calcul !
 - Exemple du p2chpd :
 - parallel : jobs "classiques" (jusqu'à 16 noeuds, 2 jours)
 - parallel1 & parallel2 : anciens calculateurs (moins performants)
 - test : accès **très** rapide, **limité à 5 min** !
 - debug : accès rapide, limité à 6h ! Réservé au debugage
 - dev : accès rapide, limité à 6h ! Réservé aux développements.
- 1 noeud max !**

Files d'exécution

- 1 file = 1 besoin = 1 type de calcul !
- Exemple du p2chpd :
 - parallel : jobs "classiques" (jusqu'à 16 noeuds, 2 jours)
 - paralle1 & parallel2 : anciens calculateurs (moins performants)
 - test : accès **très** rapide, **limité à 5 min** !
 - debug : accès rapide, limité à 6h ! Réservé au debugage
 - dev : accès rapide, limité à 6h ! Réservé aux développements.
1 noeud max !
 - fatjob : file de noeuds avec beaucoup de mémoire RAM.
Réservé aux calculs très gourmands

Files d'exécution

- 1 file = 1 besoin = 1 type de calcul !
- Exemple du p2chpd :
 - parallel : jobs "classiques" (jusqu'à 16 noeuds, 2 jours)
 - parallel1 & parallel2 : anciens calculateurs (moins performants)
 - test : accès **très** rapide, **limité à 5 min !**
 - debug : accès rapide, limité à 6h ! Réservé au debugage
 - dev : accès rapide, limité à 6h ! Réservé aux développements.
1 noeud max !
 - fatjob : file de noeuds avec beaucoup de mémoire RAM.
Réservé aux calculs très gourmands
 - long : file réservée aux calculs longs (jusqu'à 5 jours)
 - longRes : file réservée aux calculs **très longs**
(jusqu'à 30 jours !). **8 Noeuds max**

- 1 Philosophie
- 2 Architecture
- 3 Gestionnaire de batch
- 4 Bonnes pratiques

Bonnes pratiques
=
respect des autres

Exécution sur la frontale

Règles

- Ne jamais calculer sur la frontale
- Risque de saturer la machine
- Risque de bloquer les autres utilisateurs !

Limitations

Il nous est impossible de vous en empêcher !

⇒ On compte sur vous !

```
bastien@uranus:~$ ssh bdipierro@p2chpd

~~~ Bienvenue au P2CHPD de la FLMSN ~~~

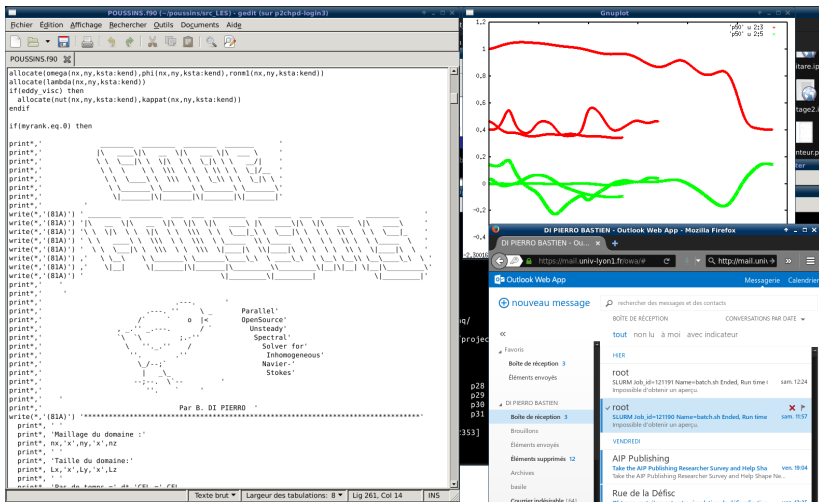
~~~~~ Computing ressources ~~~~~

Nodes online = 320
Nodes free on parallel1 = 17/36
Nodes free on parallel2 = 14/36
Nodes free on parallel = 152/220
Jobs currently running = 74

[bdipierro@p2chpd-login3 ~]$ ./monCode
```



Exécution sur la frontale



Editeur de texte, tracé de courbes, navigateur ...
= pollution de la machine !

Adapter la réservation à ses besoins

```
#SBATCH --partition=parallel1  ## choix du calculateur
#SBATCH --time=24:00:00        ## temps d'exécution
#SBATCH --nodes=8              ## nombre de noeuds
#SBATCH --ntasks-per-node=8    ## nombre coeurs par noeud
#SBATCH --mem-per-cpu=780      ## RAM par coeur
#SBATCH --mail-type=end        ## envoi d'un email
#SBATCH --mail-user=bastien.di-pierro@univ-lyon1.fr
#SBATCH -o output_%j.txt       ## fichier de sortie standard
#SBATCH -e stderr_%j.txt       ## fichier d'erreur
```

Options

- File de calcul

Adapter la réservation à ses besoins

```
#SBATCH --partition=parallel1    ## choix du calculateur
#SBATCH --time=24:00:00          ## temps d'exécution
#SBATCH --nodes=8                 ## nombre de noeuds
#SBATCH --ntasks-per-node=8       ## nombre coeurs par noeud
#SBATCH --mem-per-cpu=780         ## RAM par coeur
#SBATCH --mail-type=end           ## envoi d'un email
#SBATCH --mail-user=bastien.di-pierro@univ-lyon1.fr
#SBATCH -o output_%j.txt         ## fichier de sortie standard
#SBATCH -e stderr_%j.txt         ## fichier d'erreur
```

Options

- File de calcul

Ne pas utiliser la file **“fatjob”** pour un calcul **peu gourmand**

Ne pas utiliser la file **“debug”** pour un calcul de production

Ne pas utiliser la file **“long”** pour un calcul de 12h !

Adapter la réservation à ses besoins

```
#SBATCH --partition=parallel1    ## choix du calculateur
#SBATCH --time=24:00:00          ## temps d'exécution
#SBATCH --nodes=8                ## nombre de noeuds
#SBATCH --ntasks-per-node=8      ## nombre coeurs par noeud
#SBATCH --mem-per-cpu=780        ## RAM par coeur
#SBATCH --mail-type=end          ## envoi d'un email
#SBATCH --mail-user=bastien.di-pierro@univ-lyon1.fr
#SBATCH -o output_%j.txt        ## fichier de sortie standard
#SBATCH -e stderr_%j.txt        ## fichier d'erreur
```

Options

- File de calcul
 - Ne pas utiliser la file **“fatjob”** pour un calcul **peu gourmand**
 - Ne pas utiliser la file **“debug”** pour un calcul de production
 - Ne pas utiliser la file **“long”** pour un calcul de 12h !
- Ajuster au mieux le temps de calcul !
 - ⇒ Permet de placer au mieux vos jobs !
 - ⇒ Dépassement du temps ($\approx 10\%$) : job arrêté !

Calculs séquentiels

120336	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1 (Associati
120337	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1 (Associati
121099	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1 (Associati
121100	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1 (Associati
121109	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1 (Associati
121110	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1 (Associati
121115	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1 (Associati
121118	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1 (Associati
121165	parallel	OP700	jschreib	PD	0:00	8 (Associati
119378	parallel	43_2304t	flaadhar	PD	0:00	8 (JobHeldUs
119380	parallel	48_2304t	flaadhar	PD	0:00	8 (JobHeldUs
120181	parallel	49_2304t	flaadhar	PD	0:00	8 (JobHeldUs
121223	parallel	POIS	falizard	R	1:11:11	2 node[404-4
120264	parallel	ZZZZZZZZ	XXXXXXXX	R	12:25:07	1 node329
120263	parallel	ZZZZZZZZ	XXXXXXXX	R	13:51:20	1 node357

Calculs séquentiels

120336	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1	(Associati
120337	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1	(Associati
121099	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1	(Associati
121100	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1	(Associati
121109	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1	(Associati
121110	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1	(Associati
121115	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1	(Associati
121118	parallel	ZZZZZZZZ	XXXXXXXX	PD	0:00	1	(Associati
121165	parallel	OP700	jschreib	PD	0:00	8	(Associati
119378	parallel	43_2304t	flaadhar	PD	0:00	8	(JobHeldUs
119380	parallel	48_2304t	flaadhar	PD	0:00	8	(JobHeldUs
120181	parallel	49_2304t	flaadhar	PD	0:00	8	(JobHeldUs
121223	parallel	POIS	falizard	R	1:11:11	2	node[404 - 4
120264	parallel	ZZZZZZZZ	XXXXXXXX	R	12:25:07	1	node329
120263	parallel	ZZZZZZZZ	XXXXXXXX	R	13:51:20	1	node357

Calculs séquentiels

- 1 calcul séquentiel = 1 noeud (16 CPU) pollué !
- Utilisez les files adaptées (anciennes : parallel1 & parallel2)

Scalabilité

- Code parallélisé “qui fonctionne”
- Accès à 128 coeurs de calculs

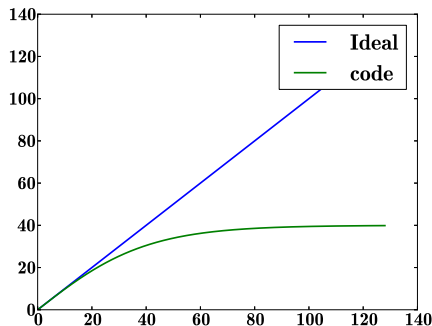
Scalabilité

- Code parallélisé “qui fonctionne”
- Accès à 128 coeurs de calculs
⇒ `mpirun -n 128 ./monCode`

Scalabilité

- Code parallélisé “qui fonctionne”
- Accès à 128 coeurs de calculs
⇒ `mpirun -n 128 ./monCode`

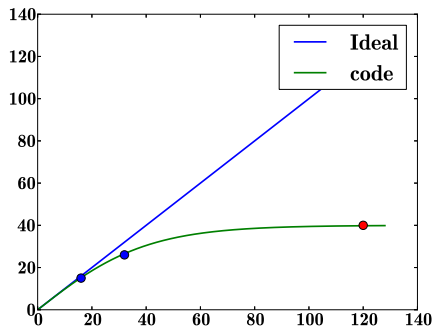
Test d'accélération



Scalabilité

- Code parallélisé “qui fonctionne”
- Accès à 128 coeurs de calculs
⇒ `mpirun -n 128 ./monCode`

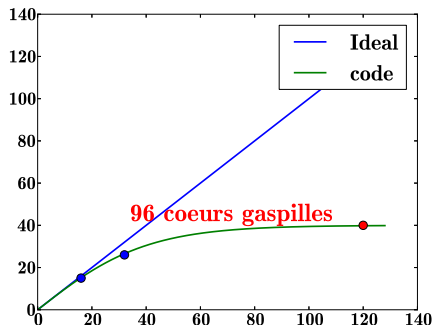
Test d'accélération



Scalabilité

- Code parallélisé “qui fonctionne”
- Accès à 128 coeurs de calculs
⇒ `mpirun -n 128 ./monCode`

Test d'accélération



Gestion des données

Espaces disques

- Espace stockage partagé $\Rightarrow$ limité !

Gestion des données

Espaces disques

- Espace stockage partagé $\Rightarrow$ limité !
- Utilisez au maximum l'espace "store" (s'il existe)
- Faites le ménage dans vos données

Gestion des données

Espaces disques

- Espace stockage partagé $\Rightarrow$ limité !
- Utilisez au maximum l'espace "store" (s'il existe)
- Faites le ménage dans vos données

Nous ne pouvons pas décider à votre place quelles sont les données à préserver !

Gestion des données

Espaces disques

- Espace stockage partagé $\Rightarrow$ limité !
- Utilisez au maximum l'espace "store" (s'il existe)
- Faites le ménage dans vos données

Nous ne pouvons pas décider à votre place quelles sont les données à préserver !

- Mise en place de quotas

```
[bdipierro@p2chpd-login3 ~]$ quota -s
Disk quotas for user bdipierro (uid 17146):
Système fichiers  blocs  quota  limite  sursisfichiers  quota  limite  sursis
p2chpd-storage6.local:/softs
                    144G    400G    410G          17290  50000  50000
[bdipierro@p2chpd-login3 ~]$
```

Conclusions

Conclusions

Soyez responsable !