Déploiement d'applications parallèles et prise en main des outils de gestion de jobs sur le cluster IBNBADIS

Dr. Ahcène BENDJOUDI

Equipe : CAlcul Parallèle et Applications CAPA

DTISI/CERIST

abendjoudi@cerist.dz, ahcene.bendjoudi@gmail.com

http://bendjoudi.ahcene.free.fr

Plan

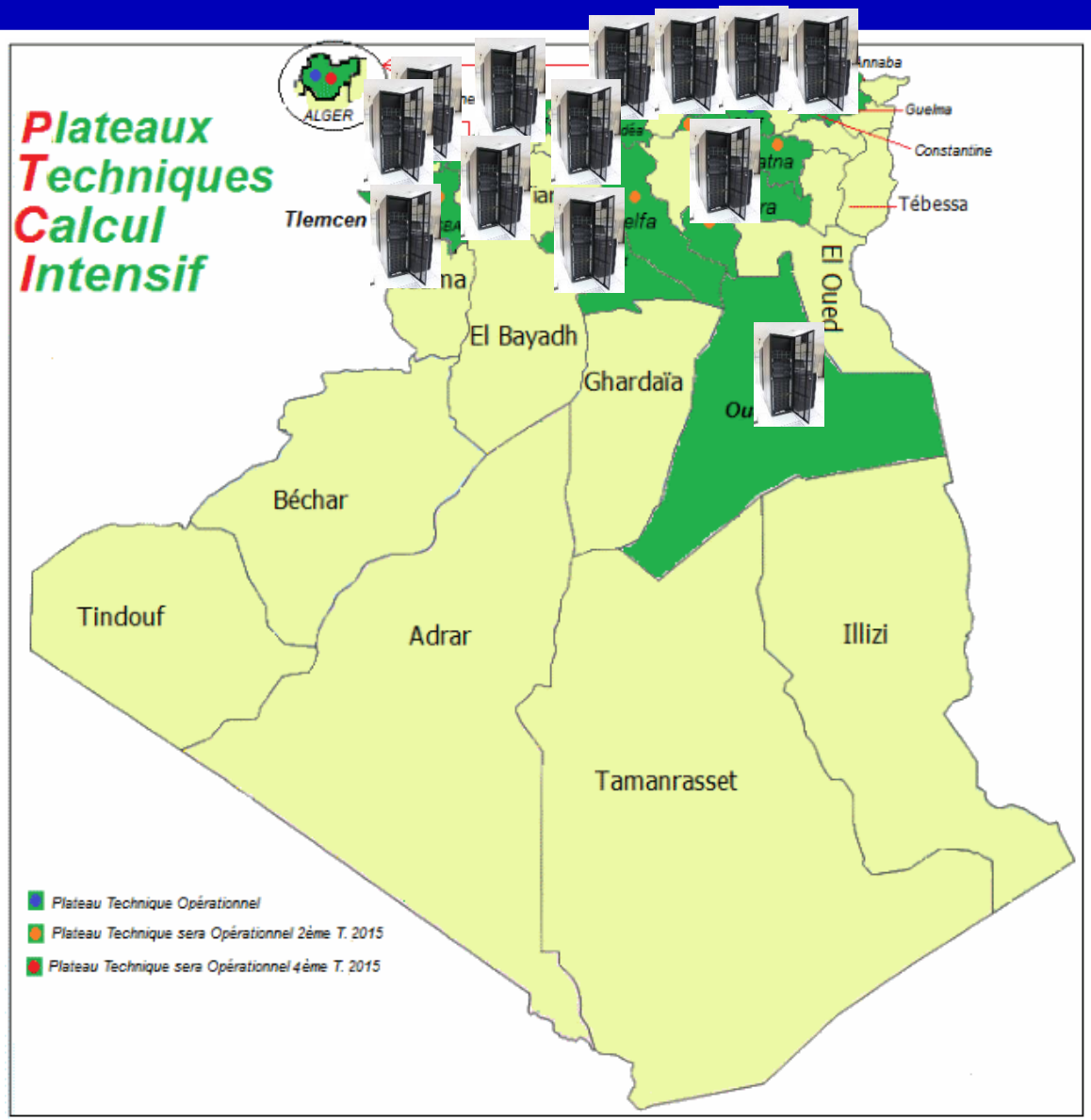
- IBNBADIS
  - Architecture
  - Obtention de compte
  - Charte d'utilisation
  - Première connexion
- Utilisation de Slurm

Les clusters

- Cluster de calcul
  - Ensemble de stations de travail (nœuds)
  - Interconnectées avec un réseau local pour former une seule machine logique.
  - Contrairement aux super-calculateurs, les nœuds d'un même cluster sont moins couplés et utilisent chacun sa propre configuration.
  - Leur faible coût de mise en œuvre par rapport au coût exorbitant des supercalculateurs.

En Algérie ?

- 25 Clusters
- Cluster déjà opérationnels depuis 2013
- Clusters acquis et en attente de mise en service
- Clusters en cours d'acquisition
- Caractéristiques



IBNBADIS

IBNBADIS

Baie de stockage
36 TB

32 X R4246-E3

2 X E5-2650, 2.0 GHz,
8 cores
32 x 2 x 8 = 512 cœurs

➔500 fois plus rapide qu'un simple PC à un cœur

- **Puissance théorique :**
  - **8 Tflops (8000 Milliards d'opérations flotantes par seconde)**
- **Linpack :**
  - **7.8 Tflops**

IBNBADIS (Hardware)

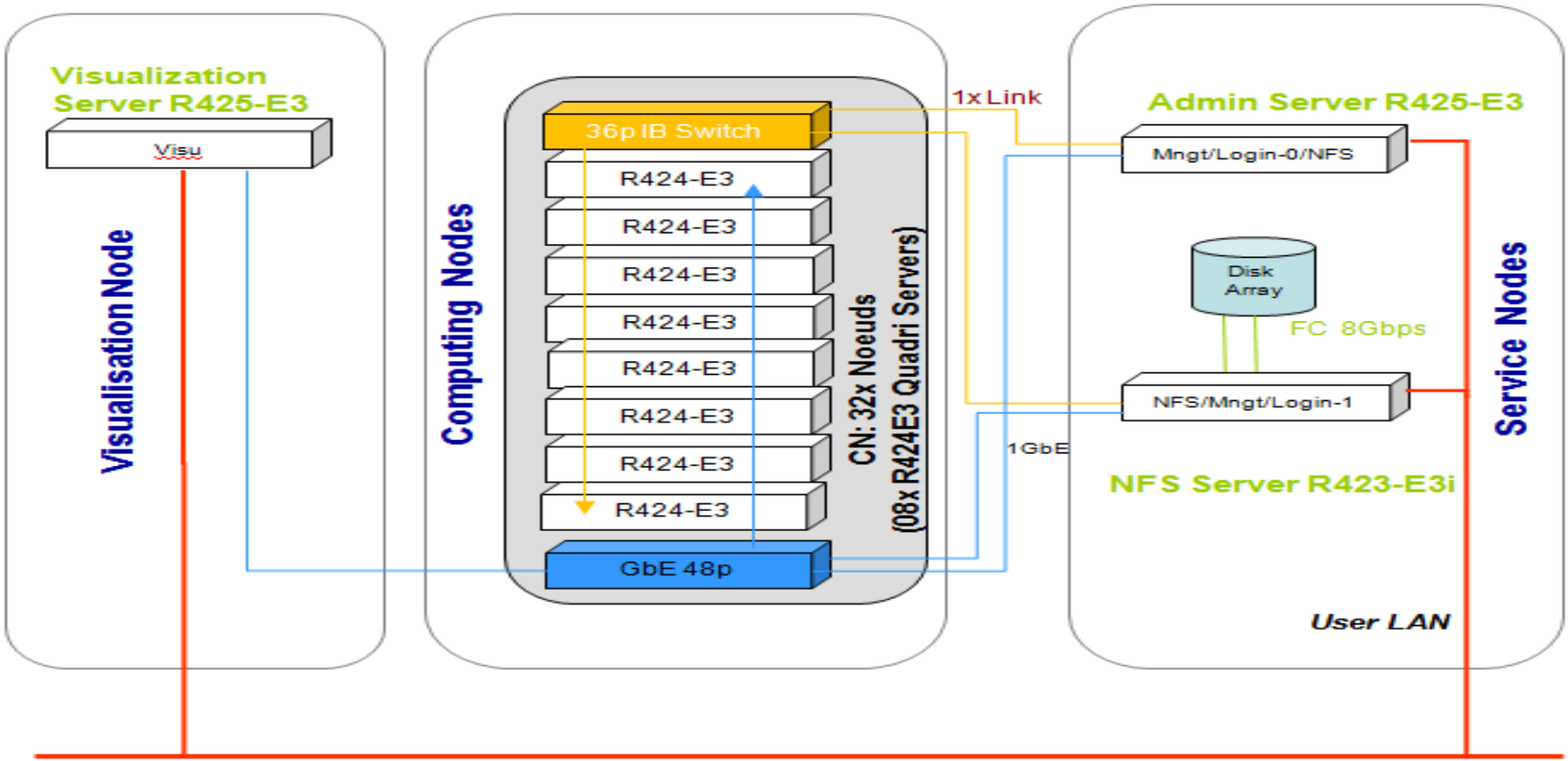
Baie de stockage
36 TB

32 X R4246-E3

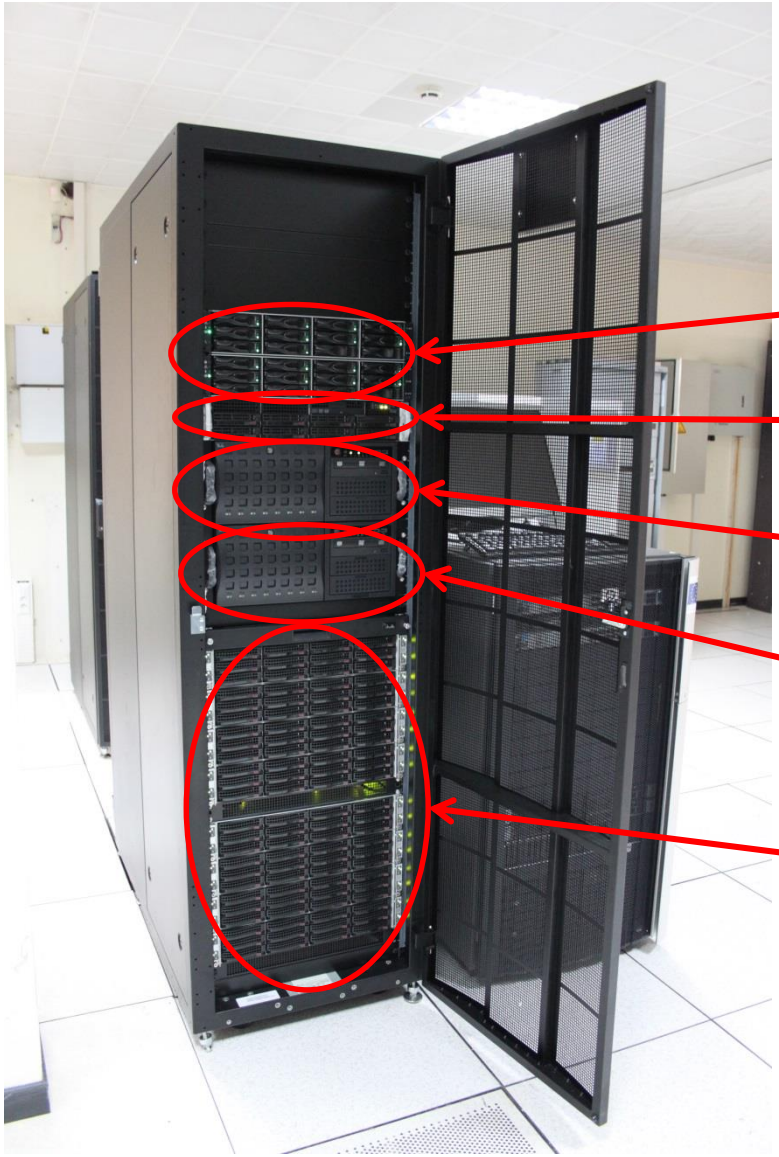
2 X E5-2650, 2.0 GHz,
8 cores
32 x 2 x 8 = 512 cœurs
➔ 500 fois plus rapide
qu'un simple PC à un cœur

- **Puissance théorique :**
  - **8 Tflops (8000 Milliards d'opérations flotantes par seconde)**
- **Linpack :**
  - **7.8 Tflops**

IBNBADIS (Architecture)



IBNBADIS (Architecture)



Baie de stockage 36 TB

Nœud NFS
ibnbadis1

Nœud de visualisation
ibnbadis10

Nœud de management
ibnbadis0

Nœuds de calcul (32)
ibnbadis11-ibnbadis42

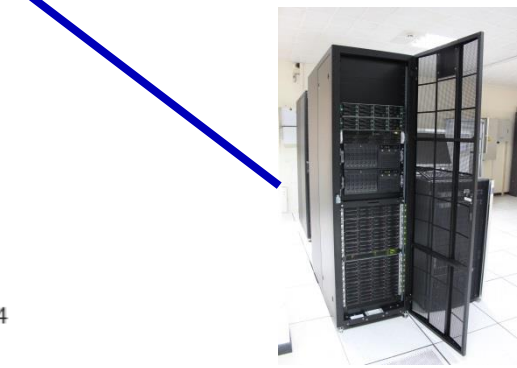
IBNBADIS vs TOP500

www.top500.org

500000 cores
17.5 PF/S

Tianhe2
31 petaflops
3 Millions cores

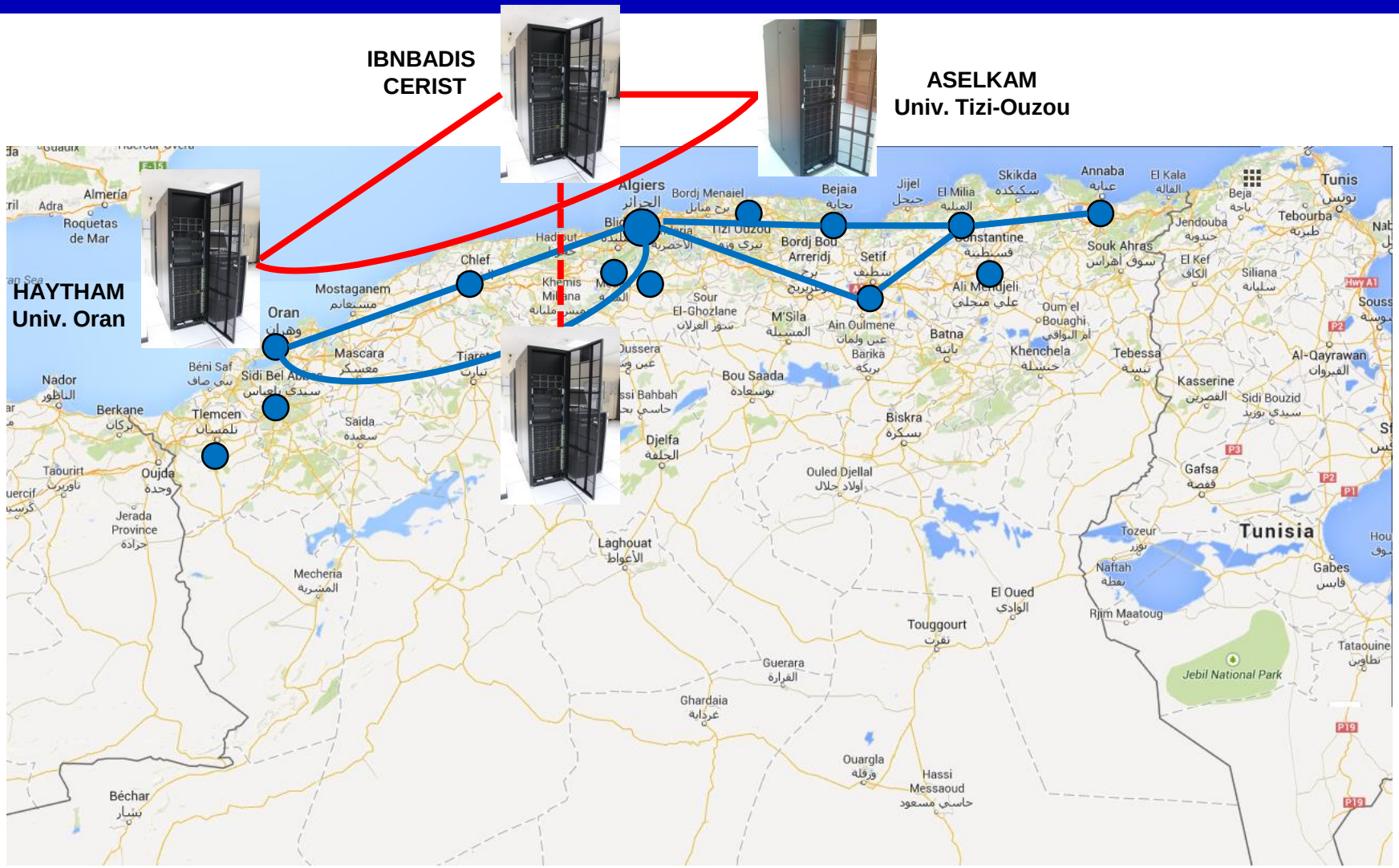
16.32 PFLOP/S
1,572,864 cores



7.8 TFLOP/S
512 cores

Et si on forme une grille !?

Et si on formait une grille !?



Ensemble vs TOP500 !

www.top500.org

500000 cores
17.5 PF/S

16.32 PFLOP/S
1,572,864 cores

Tianhe2
3 Millions cores

IBNBADIS
171.6 TFLOP/S
11264 cores
TOP400

IBNBADIS
7.8 TFLOP/S
512 cores

IBNBADIS (Obtenir un compte)

- Obtenir un compte ?
  - Site web de IBNBADIS : www.ibnbadis.cerist.dz
  - Portail National du Calcul Intensif : www.rx-racim.cerist.dz
  - Par envoi de mail : ahcene.bendjoudi@gmail.com , abendjoudi@cerist.dz
  - Dans les trois cas **joindre un résumé du travail que vous comptez effectuer sur le cluster**
- Procédure
  - Réception de la requête par mail
  - Vérification des conditions de création de compte (vérification de l'email, organisme, étudiants ?, tuteur … etc)
  - Création du compte
  - Envoie d'un mail avec les détails du compte
  - … Suivi de la charte d'utilisation à signer et renvoyer par mail

Email envoyé pour chaque nouveau compte

Bonjour,

Voici les détails de votre compte utilisateur sur le cluster IBNBADIS :

Login : **user**

passwd : **user** (merci de le changer à la première connexion en utilisant la commande passwd)

Vous pouvez vous connecter à partir de l'Algérie en ssh comme suit : **ssh user@193.194.81.82**.

Vous atterrissez sur le nœud de management ibnbadis0. Les nœuds de calcul sont **ibnbadis11 - ibnbadis42**.

Veuillez trouver ci-joint la charte d'utilisation et un manuel d'utilisation du cluster. Merci de lire attentivement la charte et de la signer.

Prières de ne pas lancer des calculs sur le nœud **ibnbadis0**. Les calculs devront être lancés sur les nœuds ibnbadis11 à ibnbadis42. Référencez-vous au manuel d'utilisation.

Bonne utilisation

Charte d'utilisation

Charte de sécurité et de bon usage de la plate-forme de calcul haute performance IBNBADIS du CERIST

1. Charte de sécurité

Le CERIST met à la disposition de la communauté scientifique algérienne la plate-forme de Calcul Haute Performance (IBNBADIS) qui leur fourni la puissance de calcul nécessaire pour les expérimentations large échelle de leurs travaux de recherche. Cette plate-forme est destinée à des utilisateurs de divers domaines et appartenant à diverses institutions algériennes.

Tout utilisateur de IBNBADIS est responsable de l'utilisation qu'il fait de la puissance de calcul qu'elle fournie et de son espace de stockage et s'engage à ne pas effectuer des opérations qui pourraient avoir des conséquences néfastes sur son fonctionnement ou sur autrui.

Tout utilisateur devra strictement :

- Ne pas accéder au compte d'un autre utilisateur et à ses informations sans l'autorisation de celui-ci
- Ne pas modifier ou détruire des informations appartenant à d'autres utilisateurs et ceci sans leur autorisation
- Ne pas développer des outils mettant sciemment en cause l'intégrité de la plate-forme
- Ne pas utiliser ses ressources pour des fins commerciales
- Ne pas développer des outils mettant en danger la vie d'autrui ou en relation avec une quelconque pratique interdite par la loi en vigeur
- Ne pas nuire à l'image du CERIST par une mauvaise utilisation de la ressource

Charte d'utilisation (Suite)

Règles de bon usage :

- User raisonnablement de toutes les ressources partagées (puissance de calcul, espace disque, logiciels, bande passante sur le réseau)
- Ne jamais communiquer son mot de passe à un tiers
- Signaler aux administrateurs tout dysfonctionnement de la plate-forme
- Signaler aux administrateurs toute suspicion de violation d'accès au compte
- Signaler aux administrateurs toute installation frauduleuse ou de logiciels, de logiciels malsains ou commerciaux

Toute suspicion d'utilisation malsaine de la plate-forme peut engager son auteur à une interdiction d'accès et à la suppression de son compte et autres mesures légales selon le cas.

Le CERIST ne pourra être tenu pour responsable pour utilisation malsaine du fait d'un utilisateur ne s'étant pas conformé à l'engagement qu'il a signé.

Charte d'utilisation (Suite 2)

2. Charte de bon usage

IBNBADIS est une ressource partagée par plusieurs utilisateurs appartenant à diverses institutions avec des besoins variés en puissance de calcul. Afin d'assurer un usage efficace et équilibré, les règles ci-dessous doivent être respectées :

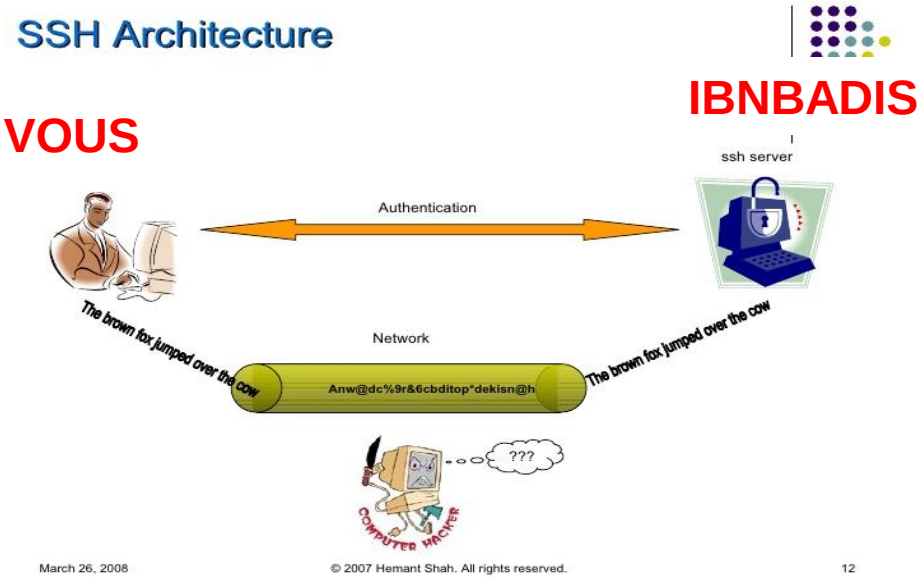
- Ne pas faire de réservations de nœuds à long terme (une semaine)
- Ne pas utiliser plus de la moitié de la totalité des nœuds pendants les heures de travail (08h à 18h)
- Les expérimentations dépassant la moitié des nœuds doivent être programmées entre 18h et 08h.
- Mentionner (IBNBADIS/CERIST) dans toutes les publications présentant des résultats obtenus en utilisant cette plate-forme.

Je soussigné(e), ..., en ma qualité de (Enseignant ☐ /Etudiant ☐ /Chercheur ☐ /Ingénieur ☐/Autre ☐ . . . . . .) à (Université ☐/ Centre de recherche ☐/ Autre ☐) . . . . . utilisateur de la plate-forme de calcul haute performance IBNBADIS du CERIST, déclare avoir pris connaissance de la présente charte de sécurité et de bon usage de la plate-forme et m'engage à la respecter.

Lu et approuvé + signature, à : . . . . . le : . . . . . .
. . . . . .

Première connexion au cluster

- Les connexions à IBNBADIS se font via ssh
  - ssh user@193.194.81.82



  - Saisir le mot de pass

```
ahcene@193.194.81.82's password:
Last login: Tue Apr 19 17:59:15 2016 from 193.194.91.130
[ahcene@ibnbadis0 ~]$ ls
archives                files                                   intel         machines
Bull_Diag_Tool_v1.3.1   haytham-shadows                         jdk1.8.0_05   mp2p.c
cmake-2.8.12.2          ibnbadis0_BullxR_2015-03-03_1127.tgz    jjj           nodes
Diag                    ibnbadis10_BullxR_2015-03-03_1137.tgz   lacle         NVIDIA-Linux-x86_64-346.59.run
dossiertest             _Inline                                 logs          pths
[ahcene@ibnbadis0 ~]$
```

Rappel Linux (Système de fichier)

Linux

Répertoire	Contenu
/	Répertoire racine : Toutes les données accessibles par le système
/home	Les répertoires personnels des utilisateurs
/bin	Binaires exécutables des commandes de bases (cd, ls, mkdir, ?)
/lib	Librairies partagées et modules du noyeau
/usr	Ressources accessibles par les utilisateurs
/etc	Fichiers de configuration (profile, passwd,fstab...)
/tmp	Données temporaires
/dev	Fichiers spéciaux correspondants aux périphériques
/mnt	Points de montage des périphériques
/var	Fichiers de log ou fichiers changeant fréquemment
/root	Répertoire personnel de l'administrateur

Windows

Documents
Téléchargements
Utilisateurs
Windows
Etc.

Rappel Linux (Quelques commandes de base)

- **pwd** : permet d'afficher le chemin absolu du répertoire courant
- **cd** : pour changer de répertoire

  cd sans option ni argument permet de se déplacer vers le répertoire personnel de l'utilisateur courant (home directory)
- **ls** : permet de lister le contenu d'un répertoire

  - **ls** sans option ni argument affiche le contenu du répertoire courant
  - **ls -a** permet d'afficher de plus les fichiers cachés dont les noms commencent par un point « **.** »
  - **ls -l** permet un affichage long (type de fichier, droit, propriété, date de modification, taille du fichier, etc.)
  - **ls -i** permet d'afficher le numéro d'inode auquel est rattaché le fichier

Rappel Linux (Quelques commandes de base)

- **mkdir** : création d'un nouveau répertoire.
  - -p créer les répertoires parents en amont
  - -m choisit le mode d'accès pour le nouveau répertoire

Exemple :

mkdir -m 444 perso créer un répertoire en lecture seule

mkdir -p travail/cours/questions

- **rm** : supprime un fichier ou un répertoire
  - rm test.old supprime le fichier test.old
  - rm -d rep1 supprime le répertoire rep1 même s'il n'est pas vide
  - rm -f test forcer la suppression

Rappel Linux (Quelques commandes de base)

- **cp** : pour copier le contenu d'un fichier syntaxe :

  cp [option] source destination

- **mv** : pour déplacer ou renommer un fichier syntaxe :

  mv [option] source cible

- **ln** : pour effectuer un lien sur un fichier

  ln fichier-source fichier-destination (lien matériel)

  ln -s fichier-source fichier-destination (lien symbolique)

Rappel Linux (Quelques commandes de base)

- **man :** affiche le manuel d'une commande
- **whatis** : Affiche une ligne de description des page de manuel
- **history :** affiche les commandes précédemment lancées par l'utilisateur courant;utilise le fichier **.bash_history**

Rappel Linux (Quelques commandes de base)

- Lancement : **$vim**
- Ouvrir un fichier: **$vim *filename***
  - Si *filename* n'existe pas, il sera créé
- Insérer du texte
  - Basculer en mode insertion (touche « i »)
  - Taper au clavier
  - Notez le message -- INSERT -- en bas de l'écran
- Quitter le mode insertion
  - Touche « Echap »

SLURM

- SLURM : Simple Linux Utility for Resource Management
  - Gestionnaire des ressources : Processeurs, cœurs, mémoire, stockage, files d'attente, jobs des utilisateurs,
  - Le serveur Slurm se trouve sur le nœud de managementibnbadis0
  - Chaque nœud de calcul contient un deamon slurm qui répond aux requêtes du serveur
  - Plus d'info sur http://slurm.schedmd.com

sinfo display characteristics of partitions
squeue display jobs and their state
scancel cancel a job or set of jobs.
scontrol display and changes characteristics of jobs, nodes, partitions.
sstat show status of running jobs.
sacct display accounting information on jobs.
sprio show factors that comprise a jobs scheduling priority
smap graphically show information on jobs, nodes, partitions

Commandes de base Slurm

- **sinfo :** Donne l'état des ressources de calcul « nœuds » (**idle :** nœud libre, **alloc :** nœud occupé, **down :** nœud en panne, **drain :** nœud gelé)

```
[ahcene@ibnbadis0 ~]$ sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
visu        up   infinite      1   idle ibnbadis10
r424*       up   infinite      2  down* ibnbadis[12,25]
r424*       up   infinite      1  drain ibnbadis31
r424*       up   infinite     19  alloc ibnbadis[11,13-24,26-30,41]
r424*       up   infinite     10   idle ibnbadis[32-40,42]
[ahcene@ibnbadis0 ~]$ sinfo -n ibnbadis12
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
visu        up   infinite      0    n/a
r424*       up   infinite      1  down* ibnbadis12
```

- **squeue :** Elle permet de voir les jobs encours d'exécution, en attente et

```
[ahcene@ibnbadis0 ~]$ squeue
JOBID PARTITION     NAME     USER  ST       TIME  NODES NODELIST(REASON)
83479     r424 1200K_hi   hzenia   R 2-10:11:50      1 ibnbadis11
83484     r424      hss krezoual   R 2-04:18:19     16 ibnbadis[13-24,26-29]
83485     r424 script.s   gchahi   R 1-11:08:31      1 ibnbadis30
```

PS : Vous pouvez toujours utiliser la commande de linux « man » pour avoir plus d'info sur les commandes slurm

Commandes de base de Slurm

- **scancel :** Permet de stoper un job

  Syntaxe : **scancel** [OPTIONS...] [job_id]

- **srun :** permet de lancer un job (faire une requête d'utilisation de ressources pour un programme de utilisateur)

  Syntaxe : **srun** [OPTIONS...] executable [args...]

```
[ahcene@ibnbadis0 ~]$ srun -N3 -l hostname    ← -N : Nœuds
0: ibnbadis26
1: ibnbadis28
2: ibnbadis30

[ahcene@ibnbadis0 ~]$ srun -n3 -l hostname    ← -n : cœurs
0: ibnbadis26
1: ibnbadis26
2: ibnbadis26
```

- **Essayez les options : --begin, --cores-per-socket, --sockets-per-node=2**

PS : Vous pouvez toujours utiliser la commande de linux « man » pour avoir plus d'info sur les commandes slurm

Commandes de base de Slurm

- Exemple d'une boucle qui calcule la somme des N premiers nombres en « c »

```
#include <stdio.h>
#include <stdlib.h>
int main(int argc, char** argv)
{
        double i,somme;
        i=0;
        somme=0;
        for (i=0;i<10000000000;i++){
                somme+=i;
        }
        printf("La somme des %f premiers nombres = %f\n",i,somme);
}
```

- srun -N3 -l --begin=21:08 ./boucle.exe
- srun -N3 --sockets-per-node=2 ./boucle.exe
- srun -N3 –cores-per-node=8 ./boucle.exe

Commandes de base de Slurm

- **salloc** : Permet de réserver des nœuds et travailler dessus en ligne avec **srun**
- Syntaxe : **salloc** [options] [<command> [command args]]
- Les mêmes options que srun
- Exemple :

```
[ahcene@ibnbadis0 JCIA2016]$ salloc -N3 --cores-per-socket=2
salloc: Granted job allocation 83609
[ahcene@ibnbadis0 JCIA2016]$ srun ./boucle.exe
La somme des 10000000000.000000 premiers nombres = 49999999990067863552.000000
La somme des 10000000000.000000 premiers nombres = 49999999990067863552.000000
La somme des 10000000000.000000 premiers nombres = 49999999990067863552.000000
La somme des 10000000000.000000 premiers nombres = 49999999990067863552.000000
La somme des 10000000000.000000 premiers nombres = 49999999990067863552.000000
La somme des 10000000000.000000 premiers nombres = 49999999990067863552.000000
```

- sbatch : permet de lancer un script en offline. Equivalent à srun mais en offline. Le contenu du script doit impérativement contenir les options de srun.

Syntaxe : **sbatch** [options] script [args...]

- Exemple :

Commandes de base de Slurm

- Création de comptes
- Première connexion ssh
- Modifier le mot de pass
- Copie de fichier
- Se connecter à d'autres comptes