

# Gestion de clusters de calcul avec Rocks

Anthony Scemama

Laboratoire de Chimie et Physique Quantiques / IRSAMC,  
Toulouse  
scemama@irsamc.ups-tlse.fr

26 Avril 2012

# Outline

## 1 Contexte

## 2 Rocks

- Installation
- Customisation
- Exemples : Clusters du LCPQ

# Qu'est-ce qu'un cluster ?

Ensemble d'ordinateurs (ou *noeuds*) mis en réseau local dans le but de travailler ensemble, tels qu'ils peuvent être vus comme une seule machine.

En général :

- Plusieurs nœuds de calcul
  - Exécutent les calculs
  - OS minimal
  - Programmes, bibliothèques (local ou NFS)
  - Pas d'accès direct par un utilisateur
- Un nœud maître (ou nœud d'administration) :
  - Point d'entrée du réseau (accès interactif)
  - Gestion des comptes utilisateurs
  - Contient localement les `/home` et programmes
  - Serveur NFS pour les `/home` et programmes
  - Serveur NTP, centralise les fichiers de log, etc
  - Configure le cluster
  - Répartit les calculs sur les nœuds de calcul

# Qu'est-ce qu'un cluster ?

Ensemble d'ordinateurs (ou *nœuds*) mis en réseau local dans le but de travailler ensemble, tels qu'ils peuvent être vus comme une seule machine.

En général :

- Plusieurs nœuds de calcul
  - Exécutent les calculs
  - OS minimal
  - Programmes, bibliothèques (local ou NFS)
  - Pas d'accès direct par un utilisateur
- Un nœud maître (ou nœud d'administration) :
  - Point d'entrée du réseau (accès interactif)
  - Gestion des comptes utilisateurs
  - Contient localement les `/home` et programmes
  - Serveur NFS pour les `/home` et programmes
  - Serveur NTP, centralise les fichiers de log, etc
  - Configure le cluster
  - Répartit les calculs sur les nœuds de calcul

# Qu'est-ce qu'un cluster ?

Ensemble d'ordinateurs (ou *noeuds*) mis en réseau local dans le but de travailler ensemble, tels qu'ils peuvent être vus comme une seule machine.

En général :

- Plusieurs nœuds de calcul
  - Exécutent les calculs
  - OS minimal
  - Programmes, bibliothèques (local ou NFS)
  - Pas d'accès direct par un utilisateur
- Un nœud maître (ou nœud d'administration) :
  - Point d'entrée du réseau (accès interactif)
  - Gestion des comptes utilisateurs
  - Contient localement les `/home` et programmes
  - Serveur NFS pour les `/home` et programmes
  - Serveur NTP, centralise les fichiers de log, etc
  - Configure le cluster
  - Répartit les calculs sur les nœuds de calcul

# Qu'est-ce qu'un cluster ?

Ensemble d'ordinateurs (ou *nœuds*) mis en réseau local dans le but de travailler ensemble, tels qu'ils peuvent être vus comme une seule machine.

En général :

- Plusieurs nœuds de calcul
  - Exécutent les calculs
  - OS minimal
  - Programmes, bibliothèques (local ou NFS)
  - Pas d'accès direct par un utilisateur
- Un nœud maître (ou nœud d'administration) :
  - Point d'entrée du réseau (accès interactif)
  - Gestion des comptes utilisateurs
  - Contient localement les `/home` et programmes
  - Serveur NFS pour les `/home` et programmes
  - Serveur NTP, centralise les fichiers de log, etc
  - Configure le cluster
  - Répartit les calculs sur les nœuds de calcul

# Qu'est-ce qu'un cluster ?

Ensemble d'ordinateurs (ou *nœuds*) mis en réseau local dans le but de travailler ensemble, tels qu'ils peuvent être vus comme une seule machine.

En général :

- Plusieurs nœuds de calcul
  - Exécutent les calculs
  - OS minimal
  - Programmes, bibliothèques (local ou NFS)
  - Pas d'accès direct par un utilisateur
- Un nœud maître (ou nœud d'administration) :
  - Point d'entrée du réseau (accès interactif)
  - Gestion des comptes utilisateurs
  - Contient localement les `/home` et programmes
  - Serveur NFS pour les `/home` et programmes
  - Serveur NTP, centralise les fichiers de log, etc
  - Configure le cluster
  - Répartit les calculs sur les nœuds de calcul

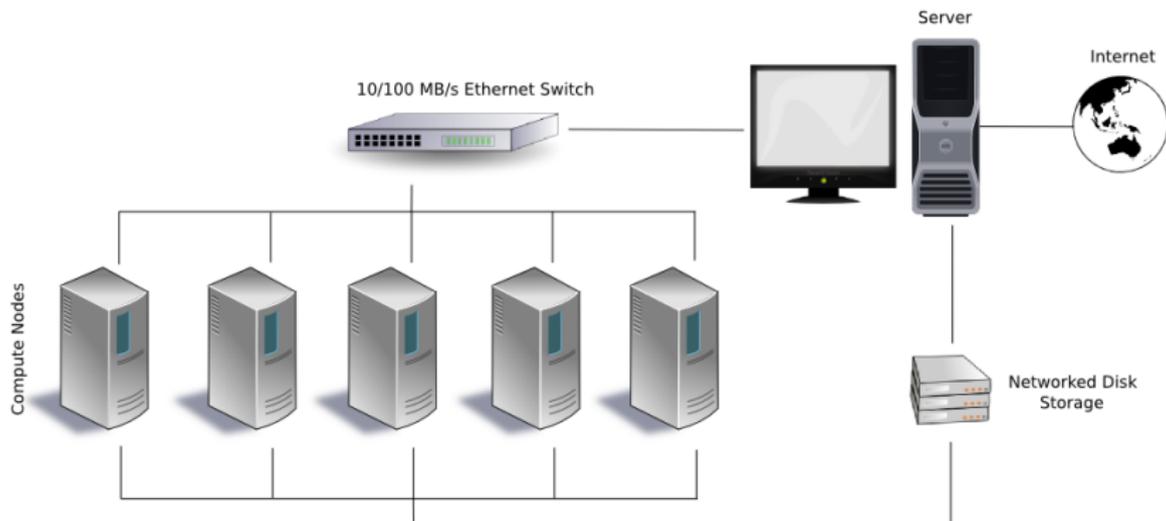


image prise sur  
[http://en.wikipedia.org/wiki/Cluster\\_\(computing\)](http://en.wikipedia.org/wiki/Cluster_(computing))

# Utilisation d'un cluster

- 1 L'utilisateur se connecte par SSH au nœud maître
- 2 Il écrit un script shell qui contient la séquence à effectuer sur le nœud de calcul
- 3 Il ajoute à son script une demande de réservation de ressources (temps CPU, RAM, espace disque, jetons de licences . . .)
- 4 Il soumet le script à un "gestionnaire de batch"
- 5 L'utilisateur peut se déconnecter et faire autre chose
- 6 Le gestionnaire de batch place le script dans une queue
- 7 Lorsque les ressources sont disponibles, elles sont allouées au script et le calcul tourne en utilisant les ressources demandées sur un (ou plusieurs) nœud(s) de calcul
- 8 Le calcul se termine et les ressources sont libérées
- 9 L'utilisateur reçoit éventuellement une alerte par email que son calcul est terminé
- 10 Il se connecte au nœud maître par SSH et interprète son résultat

# Utilisation d'un cluster

- 1 L'utilisateur se connecte par SSH au nœud maître
- 2 Il écrit un script shell qui contient la séquence à effectuer sur le nœud de calcul
- 3 Il ajoute à son script une demande de réservation de ressources (temps CPU, RAM, espace disque, jetons de licences ...)
- 4 Il soumet le script à un "gestionnaire de batch"
- 5 L'utilisateur peut se déconnecter et faire autre chose
- 6 Le gestionnaire de batch place le script dans une queue
- 7 Lorsque les ressources sont disponibles, elles sont allouées au script et le calcul tourne en utilisant les ressources demandées sur un (ou plusieurs) nœud(s) de calcul
- 8 Le calcul se termine et les ressources sont libérées
- 9 L'utilisateur reçoit éventuellement une alerte par email que son calcul est terminé
- 10 Il se connecte au nœud maître par SSH et interprète son résultat

# Utilisation d'un cluster

- 1 L'utilisateur se connecte par SSH au nœud maître
- 2 Il écrit un script shell qui contient la séquence à effectuer sur le nœud de calcul
- 3 Il ajoute à son script une demande de réservation de ressources (temps CPU, RAM, espace disque, jetons de licences ...)
- 4 Il soumet le script à un "gestionnaire de batch"
- 5 L'utilisateur peut se déconnecter et faire autre chose
- 6 Le gestionnaire de batch place le script dans une queue
- 7 Lorsque les ressources sont disponibles, elles sont allouées au script et le calcul tourne en utilisant les ressources demandées sur un (ou plusieurs) nœud(s) de calcul
- 8 Le calcul se termine et les ressources sont libérées
- 9 L'utilisateur reçoit éventuellement une alerte par email que son calcul est terminé
- 10 Il se connecte au nœud maître par SSH et interprète son résultat

# Utilisation d'un cluster

- 1 L'utilisateur se connecte par SSH au nœud maître
- 2 Il écrit un script shell qui contient la séquence à effectuer sur le nœud de calcul
- 3 Il ajoute à son script une demande de réservation de ressources (temps CPU, RAM, espace disque, jetons de licences . . .)
- 4 Il soumet le script à un "gestionnaire de batch"
- 5 L'utilisateur peut se déconnecter et faire autre chose
- 6 Le gestionnaire de batch place le script dans une queue
- 7 Lorsque les ressources sont disponibles, elles sont allouées au script et le calcul tourne en utilisant les ressources demandées sur un (ou plusieurs) nœud(s) de calcul
- 8 Le calcul se termine et les ressources sont libérées
- 9 L'utilisateur reçoit éventuellement une alerte par email que son calcul est terminé
- 10 Il se connecte au nœud maître par SSH et interprète son résultat

# Utilisation d'un cluster

- 1 L'utilisateur se connecte par SSH au nœud maître
- 2 Il écrit un script shell qui contient la séquence à effectuer sur le nœud de calcul
- 3 Il ajoute à son script une demande de réservation de ressources (temps CPU, RAM, espace disque, jetons de licences . . .)
- 4 Il soumet le script à un "gestionnaire de batch"
- 5 L'utilisateur peut se déconnecter et faire autre chose
- 6 Le gestionnaire de batch place le script dans une queue
- 7 Lorsque les ressources sont disponibles, elles sont allouées au script et le calcul tourne en utilisant les ressources demandées sur un (ou plusieurs) nœud(s) de calcul
- 8 Le calcul se termine et les ressources sont libérées
- 9 L'utilisateur reçoit éventuellement une alerte par email que son calcul est terminé
- 10 Il se connecte au nœud maître par SSH et interprète son résultat

# Utilisation d'un cluster

- 1 L'utilisateur se connecte par SSH au nœud maître
- 2 Il écrit un script shell qui contient la séquence à effectuer sur le nœud de calcul
- 3 Il ajoute à son script une demande de réservation de ressources (temps CPU, RAM, espace disque, jetons de licences . . .)
- 4 Il soumet le script à un "gestionnaire de batch"
- 5 L'utilisateur peut se déconnecter et faire autre chose
- 6 Le gestionnaire de batch place le script dans une queue
- 7 Lorsque les ressources sont disponibles, elles sont allouées au script et le calcul tourne en utilisant les ressources demandées sur un (ou plusieurs) nœud(s) de calcul
- 8 Le calcul se termine et les ressources sont libérées
- 9 L'utilisateur reçoit éventuellement une alerte par email que son calcul est terminé
- 10 Il se connecte au nœud maître par SSH et interprète son résultat

# Utilisation d'un cluster

- 1 L'utilisateur se connecte par SSH au nœud maître
- 2 Il écrit un script shell qui contient la séquence à effectuer sur le nœud de calcul
- 3 Il ajoute à son script une demande de réservation de ressources (temps CPU, RAM, espace disque, jetons de licences . . .)
- 4 Il soumet le script à un "gestionnaire de batch"
- 5 L'utilisateur peut se déconnecter et faire autre chose
- 6 Le gestionnaire de batch place le script dans une queue
- 7 Lorsque les ressources sont disponibles, elles sont allouées au script et le calcul tourne en utilisant les ressources demandées sur un (ou plusieurs) nœud(s) de calcul
- 8 Le calcul se termine et les ressources sont libérées
- 9 L'utilisateur reçoit éventuellement une alerte par email que son calcul est terminé
- 10 Il se connecte au nœud maître par SSH et interprète son résultat

# Utilisation d'un cluster

- 1 L'utilisateur se connecte par SSH au nœud maître
- 2 Il écrit un script shell qui contient la séquence à effectuer sur le nœud de calcul
- 3 Il ajoute à son script une demande de réservation de ressources (temps CPU, RAM, espace disque, jetons de licences . . .)
- 4 Il soumet le script à un "gestionnaire de batch"
- 5 L'utilisateur peut se déconnecter et faire autre chose
- 6 Le gestionnaire de batch place le script dans une queue
- 7 Lorsque les ressources sont disponibles, elles sont allouées au script et le calcul tourne en utilisant les ressources demandées sur un (ou plusieurs) nœud(s) de calcul
- 8 Le calcul se termine et les ressources sont libérées
- 9 L'utilisateur reçoit éventuellement une alerte par email que son calcul est terminé
- 10 Il se connecte au nœud maître par SSH et interprète son résultat

# Utilisation d'un cluster

- 1 L'utilisateur se connecte par SSH au nœud maître
- 2 Il écrit un script shell qui contient la séquence à effectuer sur le nœud de calcul
- 3 Il ajoute à son script une demande de réservation de ressources (temps CPU, RAM, espace disque, jetons de licences . . .)
- 4 Il soumet le script à un "gestionnaire de batch"
- 5 L'utilisateur peut se déconnecter et faire autre chose
- 6 Le gestionnaire de batch place le script dans une queue
- 7 Lorsque les ressources sont disponibles, elles sont allouées au script et le calcul tourne en utilisant les ressources demandées sur un (ou plusieurs) nœud(s) de calcul
- 8 Le calcul se termine et les ressources sont libérées
- 9 L'utilisateur reçoit éventuellement une alerte par email que son calcul est terminé
- 10 Il se connecte au nœud maître par SSH et interprète son résultat

# Utilisation d'un cluster

- 1 L'utilisateur se connecte par SSH au nœud maître
- 2 Il écrit un script shell qui contient la séquence à effectuer sur le nœud de calcul
- 3 Il ajoute à son script une demande de réservation de ressources (temps CPU, RAM, espace disque, jetons de licences . . .)
- 4 Il soumet le script à un "gestionnaire de batch"
- 5 L'utilisateur peut se déconnecter et faire autre chose
- 6 Le gestionnaire de batch place le script dans une queue
- 7 Lorsque les ressources sont disponibles, elles sont allouées au script et le calcul tourne en utilisant les ressources demandées sur un (ou plusieurs) nœud(s) de calcul
- 8 Le calcul se termine et les ressources sont libérées
- 9 L'utilisateur reçoit éventuellement une alerte par email que son calcul est terminé
- 10 Il se connecte au nœud maître par SSH et interprète son résultat

## Pour une bonne utilisation du cluster

Les utilisateurs doivent avoir une idée de comment ça fonctionne, sinon, ils peuvent faire des erreurs.

Par exemple, il vaut mieux :

- Connaître les ressources nécessaires pour un calcul. Moins on en demande, plus le script passe vite en queue.
- Connaître le hardware des nœuds de calcul : par ex, 8 cœurs physiques apparaissent comme 16 cœurs avec HyperThreading.
- Connaître des différents réseaux : éviter de faire du MPI sur le réseau ethernet si il y a de l'Infiniband
- Utiliser les bibliothèques installées par les administrateurs. La plupart du temps, elles sont optimisées et configurées pour le hardware du cluster.

# Pour une bonne utilisation du cluster

Les utilisateurs doivent avoir une idée de comment ça fonctionne, sinon, ils peuvent faire des erreurs.

Par exemple, il vaut mieux :

- Connaître les ressources nécessaires pour un calcul. Moins on en demande, plus le script passe vite en queue.
- Connaître le hardware des nœuds de calcul : par ex, 8 cœurs physiques apparaissent comme 16 cœurs avec HyperThreading.
- Connaître des différents réseaux : éviter de faire du MPI sur le réseau ethernet si il y a de l'Infiniband
- Utiliser les bibliothèques installées par les administrateurs. La plupart du temps, elles sont optimisées et configurées pour le hardware du cluster.

# Pour une bonne utilisation du cluster

Les utilisateurs doivent avoir une idée de comment ça fonctionne, sinon, ils peuvent faire des erreurs.

Par exemple, il vaut mieux :

- Connaître les ressources nécessaires pour un calcul. Moins on en demande, plus le script passe vite en queue.
- Connaître le hardware des nœuds de calcul : par ex, 8 cœurs physiques apparaissent comme 16 cœurs avec HyperThreading.
- Connaître des différents réseaux : éviter de faire du MPI sur le réseau ethernet si il y a de l'Infiniband
- Utiliser les bibliothèques installées par les administrateurs. La plupart du temps, elles sont optimisées et configurées pour le hardware du cluster.

## Pour une bonne utilisation du cluster

Les utilisateurs doivent avoir une idée de comment ça fonctionne, sinon, ils peuvent faire des erreurs.

Par exemple, il vaut mieux :

- Connaître les ressources nécessaires pour un calcul. Moins on en demande, plus le script passe vite en queue.
- Connaître le hardware des nœuds de calcul : par ex, 8 cœurs physiques apparaissent comme 16 cœurs avec HyperThreading.
- Connaître des différents réseaux : éviter de faire du MPI sur le réseau ethernet si il y a de l'Infiniband
- Utiliser les bibliothèques installées par les administrateurs. La plupart du temps, elles sont optimisées et configurées pour le hardware du cluster.

# Pour une bonne utilisation du cluster

- Avoir une idée de la fiabilité du `/home` : Y a-t-il un RAID-1/5/6 ? Y a-t-il des backups ? A quelle fréquence ?
- Savoir que les `/home` se remplissent très vite. Toujours récupérer ses données sur sa propre machine et faire ses propres backups

## NE JAMAIS :

- écrire les gros fichiers temporaires dans le `/home` : très mauvaises perfs, effondrement du serveur NFS (le nœud maître), remplissage des disques,...
- Faire des milliers d'ouverture/fermeture de fichiers chaque seconde sur un système de fichiers Lustre : effondrement des serveurs et plantage de tout le cluster

# Pour une bonne utilisation du cluster

- Avoir une idée de la fiabilité du `/home` : Y a-t-il un RAID-1/5/6 ? Y a-t-il des backups ? A quelle fréquence ?
- Savoir que les `/home` se remplissent très vite. Toujours récupérer ses données sur sa propre machine et faire ses propres backups

## NE JAMAIS :

- écrire les gros fichiers temporaires dans le `/home` : très mauvaises perfs, effondrement du serveur NFS (le nœud maître), remplissage des disques,...
- Faire des milliers d'ouverture/fermeture de fichiers chaque seconde sur un système de fichiers Lustre : effondrement des serveurs et plantage de tout le cluster

# Pour une bonne utilisation du cluster

- Avoir une idée de la fiabilité du `/home` : Y a-t-il un RAID-1/5/6 ? Y a-t-il des backups ? A quelle fréquence ?
- Savoir que les `/home` se remplissent très vite. Toujours récupérer ses données sur sa propre machine et faire ses propres backups

## NE JAMAIS :

- écrire les gros fichiers temporaires dans le `/home` : très mauvaises perfs, effondrement du serveur NFS (le nœud maître), remplissage des disques, . . .
- Faire des milliers d'ouverture/fermeture de fichiers chaque seconde sur un système de fichiers Lustre : effondrement des serveurs et plantage de tout le cluster

## Pour une bonne utilisation du cluster

- Avoir une idée de la fiabilité du `/home` : Y a-t-il un RAID-1/5/6 ? Y a-t-il des backups ? A quelle fréquence ?
- Savoir que les `/home` se remplissent très vite. Toujours récupérer ses données sur sa propre machine et faire ses propres backups

### NE JAMAIS :

- écrire les gros fichiers temporaires dans le `/home` : très mauvaises perfs, effondrement du serveur NFS (le nœud maître), remplissage des disques, . . .
- Faire des milliers d'ouverture/fermeture de fichiers chaque seconde sur un système de fichiers Lustre : effondrement des serveurs et plantage de tout le cluster

# Outline

## 1 Contexte

## 2 Rocks

- Installation
- Customisation
- Exemples : Clusters du LCPQ

# Outline

## 1 Contexte

## 2 Rocks

- Installation
- Customisation
- Exemples : Clusters du LCPQ



Rocks (<http://www.rocksclusters.org>) est une distribution open-source Linux basée sur CentOS (copie de RedHat), avec des ajouts spécifiques pour les clusters de calcul (MPI, Bibliothèques de calcul BLAS, Lapack, ...). (Plus de 1300 clusters dans le monde utilisent Rocks) Des “super-paquets” appelés *Rolls* peuvent être sélectionnés pour customiser l’installation. Par ex :

- SGE Roll : Gestionnaire de batch Sun Grid Engine (maintenant Oracle)
- Ganglia Roll : Monitoring de l’utilisation du cluster
- Lustre Roll : Utilisation de système de fichier Lustre
- Xen Roll : Virtualisation des nœuds



Rocks (<http://www.rocksclusters.org>) est une distribution open-source Linux basée sur CentOS (copie de RedHat), avec des ajouts spécifiques pour les clusters de calcul (MPI, Bibliothèques de calcul BLAS, Lapack, ...). (Plus de 1300 clusters dans le monde utilisent Rocks) Des “super-paquets” appelés *Rolls* peuvent être sélectionnés pour customiser l’installation. Par ex :

- SGE Roll : Gestionnaire de batch Sun Grid Engine (maintenant Oracle)
- Ganglia Roll : Monitoring de l’utilisation du cluster
- Lustre Roll : Utilisation de système de fichier Lustre
- Xen Roll : Virtualisation des nœuds

# Installation du nœud maître



- Booter sur le DVD

# Installation du nœud maître

**Welcome to Rocks**

**Selected Rolls**

No rolls have been selected.

If you have CD/DVD-based rolls (that is, ISO images that have been burned onto CDs or a DVD), then click the *CD/DVD-based Roll* button. The media tray will eject. Then, place your first roll disk in the tray and click *Continue*. Repeat this process for each roll disk.

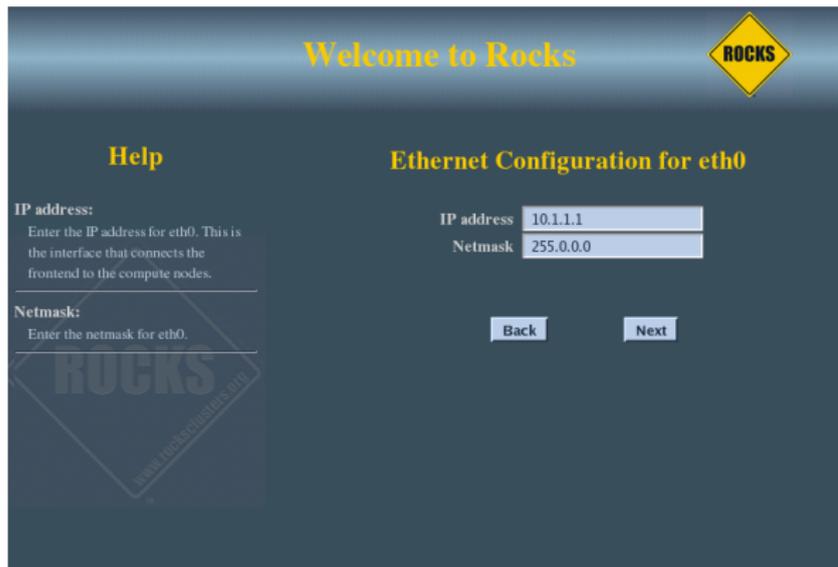
If you are performing a network-based installation (also known as a *central installation*), then input the name of your roll server into the *Hostname of Roll Server* field and then click the *Download* button. This will query the roll server and all the rolls that the roll server has available will be displayed. Click the *selected* checkbox for each roll you will to install from the roll server.

When you have completed your roll selections, click the *Next* button to proceed to cluster input screens (e.g., IP address selection, root password setup, etc.).

Selected	Roll Name	Version	Arch
<input type="checkbox"/>	CentOS	4.3	i386
<input type="checkbox"/>	area51	4.2	i386
<input type="checkbox"/>	base	4.2	i386
<input type="checkbox"/>	bio	4.2	i386
<input type="checkbox"/>	condor	4.2	i386
<input type="checkbox"/>	ganglia	4.2	i386
<input type="checkbox"/>	grid	4.2	i386
<input type="checkbox"/>	hpc	4.2	i386
<input type="checkbox"/>	java	4.2	i386
<input type="checkbox"/>	kernel	4.2	i386
<input type="checkbox"/>	sge	4.2	i386
<input type="checkbox"/>	updates	4.3	i386
<input type="checkbox"/>	viz	4.2	i386
<input type="checkbox"/>	vizagra.rockclusters.org-restore	2006.08.08	i386
<input type="checkbox"/>	web-server	4.2	i386

- Sélectionner les Rolls
- Entrer le *Fully-Qualified Host Name*

# Installation du nœud maître



The screenshot shows the 'Welcome to Rocks' screen with a yellow diamond logo containing the word 'ROCKS'. Below the logo, the title 'Ethernet Configuration for eth0' is displayed. On the left, there is a 'Help' section with two entries: 'IP address:' and 'Netmask:'. The 'IP address:' entry explains that the user should enter the IP address for the eth0 interface. The 'Netmask:' entry explains that the user should enter the netmask for the eth0 interface. In the center, there are two input fields: 'IP address' with the value '10.1.1.1' and 'Netmask' with the value '255.0.0.0'. Below these fields are two buttons: 'Back' and 'Next'. A large, semi-transparent 'ROCKS' logo is overlaid on the bottom left of the screen.

- Configuration du LAN cluster (eth0)
- Configuration du WAN (eth1)
- Configuration du DNS, passerelle etc
- Mot de passe root
- NTP

# Installation du nœud maître

www.rocksclusters.org



## Disk Setup

Choose where you would like Rocks to be installed.

If you do not know how to partition your system or if you need help with using the manual partitioning tools, refer to the product documentation.

If you used automatic partitioning, you can either accept the current partition settings (click **Next**), or modify the setup using the manual partitioning tool.

If you are manually partitioning your system, you can see your current hard drive(s) and partitions displayed below. Use the partitioning tool to add, edit,

Drive /dev/hda (76317 MB) (Model: WDC WD800BB-22 JHC0)

hda1	hda2	hda5
8001 MB	4000	63318 MB

New Edit Delete Reset RAID LVM

Device	Mount Point/ RAID/Volume	Type	Format	Size (MB)	Start	End
▼ Hard Drives						
▼ /dev/hda						
/dev/hda1	/	ext3	✓	8001	1	1020
/dev/hda2	/var	ext3	✓	4001	1021	1530
/dev/hda3		swap		996	1531	1657
▼ /dev/hda4						
/dev/hda5	/export	ext3		63319	1658	9729

Hide RAID device/LVM Volume Group members

Hide Help Release Notes Back Next

- Partitionnement auto/manuel
- /var Utilisé pour MySQL (gestion de Rocks)
- /export exports NFS. Contient les /home

# Installation du nœud maître

www.rocksclusters.org

**Installing Packages**

We have gathered all the information needed to install Rocks on the system. It may take a while to install everything, depending on how many packages need to be installed.



Remaining time: 6 minutes

Installing perl-5.8.5-36.RHEL4.i386 (40 MB)  
The Perl programming language.

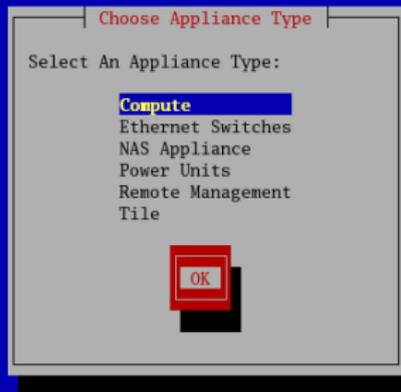
Status: Installing...

Hide Help Release Notes Back Next

- Les paquets s'installent...
- Le maître est installé !

# Installation d'un nouveau nœud de calcul

```
Insert Ethernet Addresses -- version 4.2  
Opened kickstart access to 10.0.0.0/255.0.0.0 network
```



- Sur le nœud maître, sous root, lancer `insert-ethers`
- Configurer le nœud de calcul sur boot PXE

# Installation d'un nouveau nœud de calcul

```
Insert Ethernet Addresses -- version 4.2  
Opened kickstart access to 10.0.0.0/255.0.0.0 network
```



```
Press <F10> to quit, press <F11> to force quit
```

- Attente de la requete DHCP

# Installation d'un nouveau nœud de calcul

```
Insert Ethernet Addresses -- version 4.2
Opened kickstart access to 10.0.0.0/255.0.0.0 network
```



```
Discovered a new appliance with MAC (00:13:72:ba:c8:df)
```

```
Press <F10> to quit, press <F11> to force quit
```

- Détection de l'adresse MAC du nouveau nœud
- Association MAC-IP-nom dans la base de données (compute-0-0.local)

# Installation d'un nouveau nœud de calcul

```
Insert Ethernet Addresses -- version 4.2
Opened kickstart access to 10.0.0.0/255.0.0.0 network

+-----+-----+-----+-----+
|                Inserter Appliances                |
+-----+-----+-----+-----+
| 00:13:72:ba:c8:df   compute-0-0   ( )   #         |
+-----+-----+-----+-----+

Press <F10> to quit, press <F11> to force quit
```

- Attente de réception de fichier kickstart par le nœud

# Installation d'un nouveau nœud de calcul

Nœud de calcul installé :

- Comptes utilisateurs synchronisés (ssh sans mot de passe)
- `/home` et `/share/apps` montés sur le nœud
- Ressources accessibles par le gestionnaire de batch
- Le nœud est monitoré dans Ganglia

# Outline

## 1 Contexte

## 2 **Rocks**

- Installation
- **Customisation**
- Exemples : Clusters du LCPQ

Toute la config est stockée dans une base de données MySQL.  
(réseaux, firewall, partitions, MACs, boot, routage, etc)  
L'accès est simplifié via la commande "rocks".

```
[scemama@lqqsv3 ~]$ rocks list host interface
HOST      SUBNET  IFACE  MAC                IP                NETMASK          MODULE NAME      VLAN  OPTIONS  CHANNEL
lqqsv3:   private eth0  00:1b:21:76:8f:78  10.1.1.1          255.255.0.0     ----- lqqsv3          ----
lqqsv3:   ipmi    eth0:1  -----            10.2.1.1          255.255.0.0     ----- ipmi-frontend   ----
lqqsv3:   public  eth1  84:28:2b:74:c4:c9  130.120.229.23  255.255.252.0   ----- lqqsv3          ----
lqqsv3:   ----- eth2  84:28:2b:74:c4:c9  -----          -----          -----          -----
compute-0-5: ----- eth0  00:1b:21:72:ad:dd  -----          -----          -----          -----
compute-0-5: private eth1  f0:4d:a2:3d:ab:af  10.1.255.250    255.255.0.0     ----- compute-0-5     ----
compute-0-5: ----- eth2  f0:4d:a2:3d:ab:b1  -----          -----          -----          -----
compute-0-5: ipmi    ipmi    -----            10.2.255.250     255.255.0.0     1      compute-0-5     ---- 1
compute-0-1: ----- eth0  00:1b:21:76:bf:74  -----          -----          -----          -----
compute-0-1: private eth1  f0:4d:a2:3d:ff:40  10.1.255.251    255.255.0.0     ----- compute-0-1     ----
compute-0-1: ----- eth2  f0:4d:a2:3d:ff:42  -----          -----          -----          -----
compute-0-1: ipmi    ipmi    -----            10.2.255.251     255.255.0.0     1      compute-0-1     ---- 1
compute-0-0: ----- eth0  00:1b:21:76:bf:51  -----          -----          -----          -----
compute-0-0: private eth1  14:fe:b5:c7:b9:67  10.1.255.252    255.255.0.0     ----- compute-0-0     ----
compute-0-0: ----- eth2  14:fe:b5:c7:b9:69  -----          -----          -----          -----
compute-0-0: ----- eth3  14:fe:b5:c7:b9:6b  -----          -----          -----          -----
compute-0-0: ----- eth4  14:fe:b5:c7:b9:6d  -----          -----          -----          -----
compute-0-0: ipmi    ipmi    -----            10.2.255.252     255.255.0.0     1      compute-0-0     ---- 1
nas-0-1:   private ----- 84:2b:2b:6b:cc:8a  10.1.255.253    255.255.0.0     ----- nas-0-1         ----
nas-0-0:   private ----- 84:2b:2b:5b:b8:61  10.1.255.254    255.255.0.0     ----- nas-0-0         ----
compute-0-8: ----- eth0  00:1b:21:76:bf:a2  -----          -----          -----          -----
compute-0-8: private eth1  00:25:64:fc:56:04  10.1.255.249    255.255.0.0     ----- compute-0-8     ----
compute-0-8: ----- eth2  00:25:64:fc:56:06  -----          -----          -----          -----
compute-0-8: ipmi    ipmi    -----            10.2.255.249     255.255.0.0     1      compute-0-8     ---- 1
compute-0-4: ----- eth0  00:1b:21:76:bf:db  -----          -----          -----          -----
compute-0-4: private eth1  f0:4d:a2:3d:ff:13  10.1.255.248    255.255.0.0     ----- compute-0-4     ----
compute-0-4: ----- eth2  f0:4d:a2:3d:ff:15  -----          -----          -----          -----
compute-0-4: ipmi    ipmi    -----            10.2.255.248     255.255.0.0     1      compute-0-4     ---- 1
compute-0-7: ----- eth0  00:1b:21:d6:96:7e  -----          -----          -----          -----
compute-0-7: ----- eth1  00:1b:21:d6:96:7f  -----          -----          -----          -----
compute-0-7: private eth2  d0:67:e5:f9:18:33  10.1.255.246    255.255.0.0     ----- compute-0-7     ----
compute-0-7: ----- eth3  d0:67:e5:f9:18:35  -----          -----          -----          -----
compute-0-7: ipmi    ipmi    -----            10.2.255.246     255.255.0.0     1      compute-0-7     ---- 1
[scemama@lqqsv3 ~]$
```

# Installation de programmes/bibliothèques

- Chaque utilisateur peut compiler un programme dans son /home
- Installation dans un répertoire NFS sous /share/apps
- Pour installer des paquets RPM sur tous les nœuds, 2 possibilités :
  - 1 Des fichiers XML permettent de configurer les fichiers kickstart des nœuds. Pour installer des paquets RPM sur tous les nœuds, il suffit d'ajouter une ligne dans ce fichier et de ré-installer les nœuds. Par exemple :

```
<package> gcc44-gfortran </package>
```
  - 2 La commande `tentakel` permet d'exécuter la même commande sur tous les nœuds. Par exemple, sous root `tentakel yum install -y gcc44-gfortran`

# Installation de programmes/bibliothèques

- Chaque utilisateur peut compiler un programme dans son /home
- Installation dans un répertoire NFS sous /share/apps
- Pour installer des paquets RPM sur tous les nœuds, 2 possibilités :
  - 1 Des fichiers XML permettent de configurer les fichiers kickstart des nœuds. Pour installer des paquets RPM sur tous les nœuds, il suffit d'ajouter une ligne dans ce fichier et de ré-installer les nœuds. Par exemple :

```
<package> gcc44-gfortran </package>
```
  - 2 La commande `tentakel` permet d'exécuter la même commande sur tous les nœuds. Par exemple, sous root `tentakel yum install -y gcc44-gfortran`

# Installation de programmes/bibliothèques

- Chaque utilisateur peut compiler un programme dans son `/home`
- Installation dans un répertoire NFS sous `/share/apps`
- Pour installer des paquets RPM sur tous les nœuds, 2 possibilités :
  - 1 Des fichiers XML permettent de configurer les fichiers kickstart des nœuds. Pour installer des paquets RPM sur tous les nœuds, il suffit d'ajouter une ligne dans ce fichier et de ré-installer les nœuds. Par exemple :

```
<package> gcc44-gfortran </package>
```
  - 2 La commande `tentakel` permet d'exécuter la même commande sur tous les nœuds. Par exemple, sous `root`

```
tentakel yum install -y gcc44-gfortran
```

# Installation de programmes/bibliothèques

- Chaque utilisateur peut compiler un programme dans son `/home`
- Installation dans un répertoire NFS sous `/share/apps`
- Pour installer des paquets RPM sur tous les nœuds, 2 possibilités :
  - 1 Des fichiers XML permettent de configurer les fichiers kickstart des nœuds. Pour installer des paquets RPM sur tous les nœuds, il suffit d'ajouter une ligne dans ce fichier et de ré-installer les nœuds. Par exemple :

```
<package> gcc44-gfortran </package>
```
  - 2 La commande `tentakel` permet d'exécuter la même commande sur tous les nœuds. Par exemple, sous root 

```
tentakel yum install -y gcc44-gfortran
```

# Outline

## 1 Contexte

## 2 Rocks

- Installation
- Customisation
- Exemples : Clusters du LCPQ

# Cluster Standard

## Machines de générations différentes

- AMD(Opteron) 2 sockets, 2.4 GHz
- 4 Gb RAM
- 2 disques SCSI
- Xeon(Core) 2 sockets, quad core, 2.66–3.2 GHz
- 16–48 Gb de RAM
- 2 ou 4 disques SATA pour le scratch
- Réseau Gb

## Config SGE :

- Queues de machines à même fréquence pour le MPI
- Queues pour calculs openMP
- Un calcul peut avoir l'exclusivité sur un disque physique

# Cluster Standard

## Machines de générations différentes

- AMD(Opteron) 2 sockets, 2.4 GHz
- 4 Gb RAM
- 2 disques SCSI
- Xeon(Core) 2 sockets, quad core, 2.66–3.2 GHz
- 16–48 Gb de RAM
- 2 ou 4 disques SATA pour le scratch
- Réseau Gb

## Config SGE :

- Queues de machines à même fréquence pour le MPI
- Queues pour calculs openMP
- Un calcul peut avoir l'exclusivité sur un disque physique

# Cluster Standard

HOSTNAME	ARCH	NCPU	LOAD	MEMTOT	MEMUSE	SWAPTO	SWAPUS
global	-	-	-	-	-	-	-
compute-0-19	lx26-amd64	2	1.00	3.9G	758.7M	2.0G	22.8M
compute-0-20	lx26-amd64	2	1.01	3.9G	751.7M	2.0G	640.0K
compute-0-21	lx26-amd64	2	0.98	3.9G	738.6M	2.0G	20.2M
compute-0-22	lx26-amd64	2	0.01	3.9G	104.8M	2.0G	29.0M
compute-0-23	lx26-amd64	2	0.00	3.9G	117.8M	2.0G	88.0K
compute-0-24	lx26-amd64	2	0.00	3.9G	123.6M	2.0G	25.8M
compute-0-25	lx26-amd64	2	1.01	3.9G	816.1M	2.0G	88.0K
compute-0-31	lx26-amd64	2	1.03	3.9G	786.2M	2.0G	88.0K
compute-1-0	lx26-amd64	8	1.02	15.7G	1013.2M	4.0G	12.5M
compute-1-1	lx26-amd64	8	1.03	15.7G	462.3M	4.0G	13.0M
compute-1-2	lx26-amd64	8	1.00	15.7G	515.7M	4.0G	1.3M
compute-1-3	lx26-amd64	8	1.00	15.7G	557.2M	4.0G	164.0K
compute-2-0	lx26-amd64	8	1.01	15.7G	600.3M	2.0G	27.1M
compute-2-1	lx26-amd64	8	1.00	11.7G	394.4M	2.0G	25.6M
compute-2-3	lx26-amd64	8	1.00	15.7G	683.0M	2.0G	428.0K
compute-2-4	lx26-amd64	8	0.00	15.7G	148.1M	2.0G	26.0M
compute-2-5	lx26-amd64	8	1.02	15.7G	888.3M	2.0G	136.0K
compute-2-6	lx26-amd64	8	1.00	15.7G	1.2G	2.0G	676.0K
compute-2-7	lx26-amd64	8	1.01	15.7G	744.5M	2.0G	14.3M
compute-3-0	lx26-amd64	8	1.00	31.4G	1.5G	2.0G	26.4M
compute-3-1	lx26-amd64	8	0.97	31.4G	1.2G	2.0G	14.3M
compute-3-2	lx26-amd64	8	1.00	31.4G	4.2G	2.0G	140.0K
compute-3-3	lx26-amd64	8	1.00	31.4G	1.0G	1.0G	13.8M
compute-3-4	lx26-amd64	8	1.02	31.4G	9.2G	2.0G	14.7M
compute-3-5	lx26-amd64	8	1.00	31.4G	5.0G	2.0G	14.5M
compute-3-6	lx26-amd64	8	1.00	31.4G	786.2M	1.0G	13.3M
compute-3-7	lx26-amd64	8	4.04	31.4G	491.9M	2.0G	0.0
compute-3-8	lx26-amd64	8	1.00	47.2G	910.2M	2.0G	11.9M
compute-3-9	lx26-amd64	8	0.00	47.2G	196.8M	2.0G	4.0K

```
[scemama@lpqsv9 ~]$
```

# Cluster Parallèle

## 6 Machines

- AMD(Opteron 6100) 4 sockets, 12-core, 2.2 GHz
- 128 Gb de RAM
- Réseau 10Gb
- 2 disques SAS en RAID0 pour le scratch
- scratch sur baie de disques par NFS

## Config SGE :

- Réserveation par multiples de 24 coeurs
- Réserveation de machines entières
- Jusqu'à 192 coeurs/utilisateur

# Cluster Parallèle

## 6 Machines

- AMD(Opteron 6100) 4 sockets, 12-core, 2.2 GHz
- 128 Gb de RAM
- Réseau 10Gb
- 2 disques SAS en RAID0 pour le scratch
- scratch sur baie de disques par NFS

## Config SGE :

- Réservation par multiples de 24 coeurs
- Réservation de machines entières
- Jusqu'à 192 coeurs/utilisateur

# Cluster Parallèle

```
HOSTNAME                ARCH                NCPU  LOAD  MEMTOT  MEMUSE  SWAPT0  SWAPUS
-----
global                  -                  -     -     -        -        -        -
compute-0-0             lx26-amd64         48  0.12  126.0G  417.0M  996.2M  0.0
compute-0-1             lx26-amd64         48  0.10  126.0G  499.6M  996.2M  0.0
compute-0-4             lx26-amd64         48 24.11  126.0G   15.0G  996.2M  0.0
compute-0-5             lx26-amd64         48  0.05  126.0G  377.5M  996.2M  0.0
compute-0-7             lx26-amd64         48  0.06   62.9G  226.6M  996.2M  0.0
compute-0-8             lx26-amd64         48 48.03  126.0G  501.2M  996.2M  0.0
[scemama@lpqsv3 ~]$
```

# Cluster I/O-RAM

## Machines très différentes

- Intel (Core, Nehalem) 8–32 cœurs/nœud
- 48–256 Gb de RAM
- Réseaux Gb et Infiniband
- disques SATA 1Tb, SSD, SAS en RAID0, Lustre (12Tb,800Mb/s)

## Config SGE :

- Un calcul par partition de disque
- Le reste sur Lustre
- Priorités différentes des utilisateurs sur certaines machines (ANR)

# Cluster I/O-RAM

## Machines très différentes

- Intel (Core, Nehalem) 8–32 cœurs/nœud
- 48–256 Gb de RAM
- Réseaux Gb et Infiniband
- disques SATA 1Tb, SSD, SAS en RAID0, Lustre (12Tb,800Mb/s)

## Config SGE :

- Un calcul par partition de disque
- Le reste sur Lustre
- Priorités différentes des utilisateurs sur certaines machines (ANR)

# Cluster I/O-RAM

```
-----
```

HOSTNAME	ARCH	NCPU	LOAD	MEMTOT	MEMUSE	SWAPTOT	SWAPUS
global	-	-	-	-	-	-	-
compute-0-0	lx26-amd64	8	9.10	47.2G	6.5G	4.0G	13.9M
compute-0-1	lx26-amd64	8	3.02	47.2G	2.8G	4.0G	14.7M
compute-0-10	lx26-amd64	24	2.23	188.9G	36.7G	4.0G	27.0M
compute-0-12	lx26-amd64	64	14.78	252.0G	227.1G	4.0G	4.0G
compute-0-2	lx26-amd64	8	8.91	47.2G	4.4G	4.0G	29.6M
compute-0-3	lx26-amd64	8	2.00	47.2G	10.0G	4.0G	30.1M
compute-0-6	lx26-amd64	8	1.25	47.2G	5.6G	4.0G	19.2M
compute-0-7	lx26-amd64	8	4.01	47.2G	23.6G	4.0G	27.4M
compute-0-8	lx26-amd64	8	3.14	47.2G	25.1G	4.0G	188.0K
compute-1-0	lx26-amd64	8	1.01	126.0G	22.3G	3.0G	176.0K
compute-2-7	lx26-amd64	16	5.22	126.0G	82.0G	4.0G	22.0M
compute-3-0	lx26-amd64	16	4.48	47.1G	9.2G	4.0G	374.6M
compute-3-3	lx26-amd64	16	3.42	47.1G	9.7G	4.0G	27.7M
compute-3-4	lx26-amd64	12	3.28	47.2G	28.8G	4.0G	110.9M
compute-3-5	lx26-amd64	16	4.11	94.4G	45.5G	4.0G	6.3M

```
[scemama@lpqsv6 ~]$
```

# Conclusion

Installer un cluster dans un labo peut faire peur, mais il existe des outils open-source qui facilitent les choses !

- **Rocks** : <http://www.rocksclusters.org>
- **xCAT** : <http://xcat.sourceforge.net>
- **OSCAR** : <http://oscar.openclustergroup.org>
- **Perceus** : <http://www.perceus.org>