Linux Cluster HOWTO

Ram Samudrala (me@ram.org)

v1.1, June 17, 2003

Comment mettre en place un cluster de PC Linux pour le cacul Haute Performance.

---

1. Introduction

Ce document décrit comment mettre en place un cluster de PC sous Linux pour le calcul à haute performance (HPC) dont j'ai eu besoin pour mes recherches.

Utilisez les informations ci-après sous votre entière reponsabilité. Je décline toutes reponsabilités pour tout incident qui pourrait survenir après avoir lu ce HOWTO. La dernière version de ce HOWTO sera toujours disponible à l'adresse http://www.ram.org/computing/linux/linux_cluster.html.

A la différence d'autres documentations qui parlent de la mise en place de cluster de manière générale, ceci est une description spécifique de la manière dont notre laboratoire à installé le cluster, mais aussi les aspects calculs, ainsi que les parties ordinateur de bureau, portable et accès public.

Ceci est principalement fait pour un usage interne, mais j'ai placé ce document sur le web suite à la reception de nombreux mails issuent de questions sur des newsfeed demandant ce type d'information.

Actuellement, j'envisage la mise en place d'un cluster 64 bits, je trouve qu'il y a un manque d'information sur la méthode à suivre pour assembler les composants pour former un noeud qui fonctionne sous Linux et qui inclut, non seulement la description du matériel, mais aussi du logiciel utile pour arriver à un fonctionnement en production dans un enviroennement de recherche.

Le but principal de ce HOWTO est de lister les types de matériels qui fonctionnent bien ou mal avec Linux.

2. Hardware

Cette section couvre nos choix en matière de harware. à part les points notés dans la section des problèmes rencontrés [p 10] , tout ce qui est présenté fonctionne *réellement* bien.

L'installation du matériel est assez simple (les particularitées sont dans les notes), la plupart des informations se trouvent dans les manuels. Pour chaque section, le matériel est listé par ordre d'achat (le plus récent est listé en premier).

2.1 Node hardware

32 machines ont la configurations suivante:

- 2 XEON 2.4GHZ 533FSB CPUs
- Supermicro X5DPR-1G2 motherboard
- 2 512MB PC2100 DDR REG ECC RAM

- 1 40GB SEA 7200 HD
- 1 120GB SEA 7200 HD
- Supermicro Slim 24X CDROM
- CSE-812 400 C/B 1U case

32 machines ont la configuration suivante:

- 2 AMD Palamino MP XP 2000+ 1.67 GHz CPUs
- Asus A7M266-D w/LAN Dual DDR motherboard
- 2 Kingston 512mb PC2100 DDR-266MHz REG ECC RAM
- 1 41 GB Maxtor 7200rpm ATA100 HD
- 1 120 GB Maxtor 5400rpm ATA100 HD
- Asus CD-A520 52x CDROM
- 1.44mb floppy drive
- ATI Expert 2000 Rage 128 32mb
- IN-WIN P4 300ATX Mid Tower case
- Enermax P4-430ATX power supply

32 machines ont la configuration suivante:

- 2 AMD Palamino MP XP 1800+ 1.53 GHz CPUs
- Tyan S2460 Dual Socket-A/MP motherboard
- Kingston 512mb PC2100 DDR-266MHz REG ECC RAM
- 1 20 GB Maxtor UDMA/100 7200rpm HD
- 1 120 GB Maxtor 5400rpm ATA100 HD
- Asus CD-A520 52x CDROM
- 1.44mb floppy drive
- ATI Expert 98 8mb AGP video card
- IN-WIN P4 300ATX Mid Tower case
- Intel PCI PRO-100 10/100Mbps network card
- Enermax P4-430ATX power supply

32 machines ont la configuration suivante:

- 2 Pentium III 1 GHz Intel CPUs
- Supermicro 370 DLE Dual PIII-FCPGA motherboard
- 2 256 MB 168-pin PC133 Registered ECC Micron RAM
- 1 20 GB Maxtor ATA/66 5400 RPM HD
- 1 40 GB Maxtor UDMA/100 7200 RPM HD
- Asus CD-S500 50x CDROM
- 1.4 MB floppy drive
- ATI Expert 98 8 MB PCI video card
- IN-WIN P4 300ATX Mid Tower case

2.2 Server hardware

1 serveur pour utilisation externe (distribution des systèmes) avec la configuration suivante:

- 2 AMD Palamino MP XP 2000+ 1.67 GHz CPUs
- Asus A7M266-D w/LAN Dual DDR
- 4 Kingston 512mb PC2100 DDR-266MHz REG ECC RAM
- Asus CD-A520 52x CDROM
- 1 41 GB Maxtor 7200rpm ATA100 HD
- 6 120 GB Maxtor 5400rpm ATA100 HD
- lecteur de disquette 1.44Mo
- ATI Expert 2000 Rage 128 32mb
- IN-WIN P4 300ATX Mid Tower case
- Enermax P4-430ATX power supply

2.3 Desktop hardware

1 PC desktop avec la configuration suivante:

- 2 AMD XP 2600 MP
- MSI K7D Master-L DUAL MS-6501 motherboard
- 4 1024MB PC2100 DDR REG ECC RAM
- 1 40GB SEA 7200 Maxtor harddisk
- 2 120GB SEA 7200 Maxtor hardidks
- PIONEER DVR-AO5 IDE DVD-RW
- 1.44mb floppy drive
- ATI Expert 2000 Rage 128 32mb video card
- IN-WIN P4 300ATX Mid Tower case
- Intel PCI PRO-100 10/100Mbps network card
- 450W ENERMAX P4-430ATX power supply
- CREATIVE SB 128 5.1 PCI soundcard

2 PC desktop avec la configuration suivante:

- 2 AMD XP 2600 MP
- MSI K7D Master-L DUAL MS-6501 motherboard
- 2 512MB PC2100 DDR REG ECC RAM
- 1 40GB SEA 7200 Maxtor harddisk
- 2 120GB SEA 7200 Maxtor hardidks
- MSI 52X24X52X CR52-A2 CD-RW
- 1.44mb floppy drive
- ATI Expert 2000 Rage 128 32mb video card
- IN-WIN P4 300ATX Mid Tower case
- Intel PCI PRO-100 10/100Mbps network card
- 450W ENERMAX P4-430ATX power supply
- CREATIVE SB 128 5.1 PCI soundcard

1 PC desktop avec la configuration suivante:

- 2 AMD Palamino MP XP 2000+ 1.67 GHz CPUs
- Asus A7M266-D w/LAN Dual DDR
- 2 Kingston 512mb PC2100 DDR-266MHz REG ECC RAM
- Ricoh 32x12x10 CDRW/DVD Combo EIDE
- 1.44mb floppy drive
- 1 41 GB Maxtor 7200rpm ATA100 HD
- 1 120 GB Maxtor 5400rpm ATA100 HD
- ATI Expert 2000 Rage 128 32mb video card
- IN-WIN P4 300ATX Mid Tower case
- Intel PCI PRO-100 10/100Mbps network card
- Enermax P4-430ATX power supply

1 PC desktop avec la configuration suivante:

- 2 Intel Xeon 1.7 GHz 256K 400FS
- Supermicro P4DCE Dual Xeon motherboard
- 4 256mb RAMBUS 184-Pin 800 MHz memory
- 2 120 GB Maxtor ATA/100 5400 RPM HD
- 1 60 GB Maxtor ATA/100 7200 RPM HD
- 52X Asus CD-A520 INT IDE CDROM
- 1.4 MB floppy drive
- Leadtex 64 MB GF2 MX400 AGP
- Creative SB LIVE Value PCI 5.1
- Microsoft Natural Keyboard
- Microsoft Intellimouse Explorer
- Supermicro SC760 full-tower case with 400W PS

2 PC desktop avec la configuration suivante:

- 2 AMD K7 1.2g/266 MP Socket A CPU
- Tyan S2462NG Dual Socket A motherboard
- 4 256mb PC2100 REG ECC DDR-266Mhz
- 3 40 GB Maxtor UDMA/100 7200 RPM HD
- 50X Asus CD-A520 INT IDE CDROM
- 1.4 MB floppy drive
- Chaintech Geforce2 MX200 32mg AGP
- Creative SB LIVE Value PCI
- Microsoft Natural Keyboard
- Microsoft Intellimouse Explorer
- Full-tower case with 300W PS

2 PC desktop avec la configuration suivante:

- 2 Pentium III 1 GHz Intel CPUs
- Supermicro 370 DLE Dual PIII-FCPGA motherboard
- 4 256 MB 168-pin PC133 Registered ECC Micron RAM
- 3 40 GB Maxtor UDMA/100 7200 RPM HD

- Asus CD-S500 50x CDROM
- 1.4 MB floppy drive
- Jaton Nvidia TNT2 32mb PCI
- Creative SB LIVE Value PCI
- Microsoft Natural Keyboard
- Microsoft Intellimouse Explorer
- Full-tower case with 300W PS

2 PC desktop avec la configuration suivante:

- 2 Pentium III 1 GHz Intel CPUs
- Supermicro 370 DLE Dual PIII-FCPGA motherboard
- 4 256 MB 168-pin PC133 Registered ECC Micron RAM
- 3 40 GB Maxtor UDMA/100 7200 RPM HD
- Mitsumi 8x/4x/32x CDRW
- 1.4 MB floppy drive
- Jaton Nvidia TNT2 32mb PCI
- Creative SB LIVE Value PCI
- Microsoft Natural Keyboard
- Microsoft Intellimouse Explorer
- Full-tower case with 300W PS

1 PC desktop avec la configuration suivante:

- 2 Pentium III 1 GHz Intel CPUs
- Supermicro 370 DE6 Dual PIII-FCPGA motherboard
- 4 256 MB 168-pin PC133 Registered ECC Micron RAM
- 3 40 GB Maxtor UDMA/100 7200 RPM HD
- Ricoh 32x12x10 CDRW/DVD Combo EIDE
- Asus CD-A520 52x CDROM
- 1.4 MB floppy drive
- Asus V7700 64mb GeForce2-GTS AGP video card
- Creative SB Live Platinum 5.1 sound card
- Microsoft Natural Keyboard
- Microsoft Intellimouse Explorer
- Full-tower case with 300W PS

3 PC desktop avec la configuration suivante:

- 2 Pentium III 1 GHz Intel CPUs
- Supermicro 370 DE6 Dual PIII-FCPGA motherboard
- 4 256 MB 168-pin PC133 Registered ECC Micron RAM
- 3 40 GB Maxtor UDMA/100 7200 RPM hard disk
- Ricoh 32x12x10 CDRW/DVD Combo EIDE
- 1.4 MB floppy drive
- Asus V7700 64mb GeForce2-GTS AGP video card
- Creative SB Live Platinum 5.1 sound card
- Microsoft Natural Keyboard
- Microsoft Intellimouse Explorer

- Full-tower case with 300W PS

2.4 Firewall/gateway hardware

Un firewall avec la configuration suivante:

- AMD Palamino XP 1700+ 1.47GHz CPU
- MSI KT3 Ultra2 KT333 MS-6380E motherboard
- 512 MB PC2100 DDR-266MHz DIMM RAM
- 40GB Seagate 7200rpm ATA/100 hard disk
- Asus 52X CD-A520 INT IDE cdrom
- 1.44 MB floppy drive
- ATI Expert 2000 Rage 128 32mb video card
- 3 Intel Pro/1000T Gigabit Server ethernet cards
- 4U Black Rackmount Steel case

Une passerelle avec la configuration suivante. LA passerelle est un système mirroir du firewall pour le cas ou le firewall sera dégradé.

- AMD Palamino XP 1800+ 1.57GHz CPU
- MSI KT3 Ultra2 KT333 MS-6380E motherboard
- 512 MB PC2100 DDR-266MHz DIMM RAM
- 40GB Seagate 7200rpm ATA/100 hard disk
- Asus 52X CD-A520 INT IDE cdrom
- 1.44 MB floppy drive
- ATI Expert 2000 Rage 128 32mb video card
- 3 Intel Pro/1000T Gigabit Server ethernet cards
- 4U Black Rackmount Steel case

2.5 Divers matériels/accessoires

Sauvegarde:

- 2 lecteurs Sony 20/40 GB DSS4 SE LVD DAT

Moniteurs:

- 2 moniteurs 17" Viewsonic VE700 LCD
- 1 moniteurs 20.1" Viewsonic VP201M LCD
- 1 moniteurs 22" Viewsonic P220F 0.25-0.27m
- 4 moniteurs 21" Sony CPD-G500 .24mm
- 2 moniteurs 18" Viewsonic VP181 LCD
- 1 moniteurs 17" Viewsonic VE170 LCD
- 2 moniteurs Sun monitors

Imprimantes:

- HP colour laserject 4600dn

2.6 Relier toute la configuration ensemble

Nous avons utilisé un switch KVM avec un petit écran pour se connecter et "examiner" toutes les machines:

- Moniteur 15" .28dp XLN CTL
- 3 Belkin Omniview 16-Port Pro Switches
- Belkin Omniview 2-Port Switch
- APC AR203 netshelter rack unit

Pour parfaire tout cela et pour en faire une jolie solution, nous autions besoin d'un petit PDA que nous pourrions connecter à l'arrière des PC (utilisable avec un stylet, comme les Palm).

Je n'envisage pas d'utiliser d'avantage de connecteurs dans le switch KVM.

Le reseau est important:

- 2 Netgear FS750NA 48 port/1 git network switch
- 1 Netgear FSM750S 48 port/2 git network switch
- 1 Netgear FS517TS 16 port/1 git network switch
- 1 Netgear FS524 24 port network switch
- 1 Cisco Catalyst 3448 XL Enterprise Edition 48 port network switch
- 1 Netgear ME102NA Wireless Access Point
- 1 Netgear MA401NA Wireless PCMCIA network card

2.7 Couts

Notre vendeur est Hard Drives Northwest (http://www.hdnw.com). Pour chaque noeud dans notre cluster (contenant 2 CPU chacun), nous avons payé entre 1500 et 2000 $, en incluant les taxes. Généralement, notre but est de garder le cout de chaque processeur en dessous des 1000 $ (en incluant l'emplacement).

3. Logiciel

3.1 Système d'exploitation, Linux, bien sur !

Les version de Kernels et des distributions que nous avons utilisés :

- Kernel 2.2.16-22, distribution KRUD 7.0
- Kernel 2.4.9-7, distribution KRUD 7.2
- Kernel 2.4.18-10, distribution KRUD 7.3
- Kernel 2.4.20-13.9, distribution KRUD 9.0

Ces distributions fonctionne bien pour nous, les mise a jour nous sont transmises sur CD et il n'y a aucune connexion avec le reseau externe. Elles ont semblés plus "propre" que les distributions standard RedHat, et la configuration est extrèmement stable.

3.2 Logiciel reseau

Nous utilisons Shorewall 1.3.14a ((http://www.shorewall.net) pour le firewall.

3.3 Environnement parallèle

Nous utilisons nos propres logiciels pour la parallélisation des applications mais nous avons expérimenté PVM et MPI. A mon avis l'overhead généré par ces environnement est trop important. Je recommande d'écrire son propre code pour les taches que vous voulez remplir (c'est ma vue personnelle). (NDLR je recommande à l'inverse l'utilisation de MPI, qui est très portable sur toute sortes de plateforme, et qui permet de se détacher de l'architecture et de l'écriture du logiciel pour se consacrer à son propre problème).

3.4 Coûts

Linux et la plupart des logiciels qui tourne sous Linux sont librement copiable.

4. Démarrage, configuration, et maintenance

4.1 Configuration disques

Cette section décrit la stratégie de partitionnement disques.

```
ferme/cluster machines:

hda1 - swap   (2 * RAM)
hda2 - /      (le reste de l'espace disque disponible)
hdb1 - /maxa  (totalité disque)

PC desktops (sans windows):

hda1 - swap   (2 * RAM)
hda2 - /      (4 GB)
hda3 - /spare (le reste de l'espace disque disponible)
hdb1 - /maxa  (totalité disque)
hdd1 - /maxb  (totalité disque)

desktops (sans windows):

hda1 - /win   (totalité disque)
hdb1 - swap   (2 * RAM)
hdb2 - /      (4 GB)
hdb3 - /spare (le reste de l'espace disque disponible)
hdd1 - /maxa  (totalité disque)

laptops (un seul disque):

hda1 - /win   (la moitié de la taille du disque)
hda2 - swap   (2 * RAM)
hda3 - /      (le reste de l'espace disque disponible)
```

4.2 Configuration de l’environnement

Installer un minimum de packages dans la ferme de PC. Les utilisateurs sont autorisés à configurer les PC desktops comme ils le désirent.

4.3 Installation et maintenance des systèmes d’exploitation

Clonage et mintenance des packages

FAI

FAI (http://www.informatik.uni-koeln.de/fai/) est un système automatisé pour installer le système Debian GNU/Linux sur un cluster. Vous pouvez prendre un ou plusieurs PC vierges, les allumer et après quelques minutes Linux est installé, configuré et en état de fonctionner sur la totalité du cluster, sans qu’aucune interaction ne soit nécessaire.

SystemImager

SystemImager (http://systemimager.org) est un logiciel qui automatise l’installation, la distribution et le déploiement de Linux.

Stratégie personnelle de clonage

Je crois dans un système complètement distribué. Ceci veux dire que chaque machine contient une copie du système d’exploitation. Installer un système d’exploitation sur chaque machine manuellement est pénible. Pour optimiser ce processus, j’ai d’abord installé et paramétré le système sur une machine. J’ai ensuite créé un fichier tar (que j’ai zippé (gzip)) du système tout entier. J’ai placé ce fichier sur un CDROM qui m’a ensuite servi plour le clonage de chaque machine dans mon cluster.

Les commandes que j’ai utilisé pour crééer le fichier tar sont les suivantes :

```
tar -czvlps --same-owner --atime-preserve -f /maxa/slash.tgz /
```

J’ai utilisé un script apellé `go` qui reçoit comme paramètre le nom de la machine et l’adresse IP, puis détarre le fichier `slash.tgz` sur le CD-ROM, enfin remplace le nom de la machine et l’adresse IP aux endroits appropriés. Une version du script `go` et du fichier d’entrée peuvent être trouvés à l’adresset: http://www.ram.org/computing/linux/linux/cluster/. Ce script devra être édité pour correspondre au design de votre cluster.

Pour faire fonctionner tout cela, j’ai aussi utilisé le Tom’s Root Boot package (http://www.toms.net/rb/) pour booter la machine et cloner le système. Le script `go` peut être placé sur un CDROM, ou sur une disquette contenant le Tom’s Root Boot package (vous devrez effacer quelques programmes car la disquette est relativement limité en place libre).

Plus commodément, vous pouvez graver un CDROM bootable contenant le Tom’s Root Boot package, incluant le script `go`, et le fichier tgz contenant le système à cloner. Vous pouvez aussi éditer le fichier init du boot de manière à ce qu’il execute le script `go` (vous devrez quand même positionner l’adresse IP si vous n’utilisez pas DHCP).

Vous pouvez crééer de manière alternative votre propre disque (comme un disque de secours) qui contiennent le kernel et les outils que vous voulez. Il y a de nombreux documents qui décrivent comment faire cela, incluant le Linux Bootdisk HOWTO (http://www.linuxdoc.org/HOWTO/Bootdisk-HOWTO/), qui contient lui aussi des liens vers des images de disques bootable.

Ainsi, vous pouvez développer un système ou tout ce que vous avez à faire est d'insérer un CDROM, allumer la machine, prendre un café (ou une canette de coca) (NDLR: buvez de l'eau, c'est meilleur pour la santé ;-)) et retourner vous assoir pour constater un clonage complet. Vous pouvez répeter cette procédure pour autant de machines que vous le désirez. Cette procédure à extrèmement bien focntionné pour moi, et si de plus, vous trouvez quelqu'un (pour insérer et retirer les CDROM !) c'est idéal.

Rob Fantini (rob@fantinibakery.com) a contribué aux modifications du script cité si-dessus pour cloner la Mandrake 8.2 qui est accessible à l'adresse http://www.ram.org/computing/linux/cluster/fantini_contribution.tgz.

J'avais travaillé sur une procédure ou tout ce que vous aviez à faire était d'insérer un CD, démarrer la machine, et tout était cloné. Je mettrai cela à disposition dans un futur proche.

DHCP vs. adresse IP codées en dur

Si vous avez DHCP déjà en focntionnement, alors vous n'aurez pas à changer l'adresse IP et cette partie pourra être retirée du script `go`.

DHCP a l'avantage de ne plus avoir à se préocuper des adresses IP dans la mesure ou le serveur DHCP est correctement configuré.

Il a le désavantage lié à la centralisation (and comme je le disais, j'essaye de répartir les choses le plus possible). En outre, lier l'adresse ethernet de la carte à l'adresse IP peut devenir un inconvénient si vous voulez remplacer des machines, ou changer les noms de machines de manière régulière.

4.4 Particularité du matériel

Le matériel a fonctionné correctement pour nous. Les cas particuliers sont listés ci-dessous :

Les machines bi-processeurs AMD 1.2 GHz chauffent beaucoup. Si on en place deux dans une pièce, la température de celle-ci s'accroit considérablement. En outre, leur utilisation dans un cadre desktop, peux s'avérer correct, mais la témpérarture, et la consommation electrique doivent être pris en considération. La configuration AMD Palmino décrite précédemmentn semble très bien fonctionner pour nous, mais je recommande d'avoir deux ventilateurs au cas ou--ceci resoudra tout problème d'instabilité.

4.5 Particularité du logiciel

Certains commandes tar ne crééent par un fichier tar correct (et notanment en ce qui concerne les liens symboliques) La solution est d'utiliser la commande tar qui se trouve dans la distribution RedHat 7.0 (NDLR: La commande tar GNU fonctionne très bien)

5. Les opérations sur le cluster

Cette section est encore en dévelopement dans la mesure ou l'utilisation de mon cluster évolue, jusqu'ici nous essayons d'écrire nos propres ensemble de routine de Message Passing pour établir la communication entre les processus des différentes machines.

Beaucoup d'applications, en particulier dans les secteurs informatiques de traitement du génome, sont massivement et facilement parallelisable. Cela signifi que la répartition parfaite peut être réalisée en distribuant des tâches de manière homogène entre les machines (par exemple, en analysant un génome entier en utilisant une technique qui travaille sur un seul gène, ou un seule proteine, chaque processeur peut travailler à un gène, ou à une seule proteine à la fois indépendenment de tous les autres processeurs).

Jusqu'ici nous n'avons pas trouvé la nécessité d'employer un système de gestion de file d'attente, mais évidemment ce dépend fortement du type d'applications que vous souhaitez faire tourner. (NDLR: ceal dépend aussi de votre environnement de travail, à savoir si votre cluster est partagé entre plusieurs utilisateurs en concurence ...).

5.1 Benchmarks bruts

Pour le plus important programme que nous faisons tourner (notre *ab initio* programme de simulation de pliage de proteine), en utilisant la machine avec un Pentium 3 à 1GHz comme référence, en moyenne :

```
Athlon 1.2 GHz est environ 16% plus rapide
Xeon   1.7 GHz est environ 27% plus rapide
Athlon 1.5 GHz est environ 38% plus rapide
Athlon 1.7 GHz est environ  46% plus rapide
Xeon   2.4 GHz est environ 62% plus rapide
```

Oui, l'Athlon 1.5 GHz est plus rapide que le Xeon 1.7 GHz car le Xeon execute seulement six instructions par horloge (IPC) alors que l'Athlon en execute neuf IPC (vous faites le calcul!).

5.2 Stabilité

Ces machines sont incroyablement stables, aussi bien en terme de matériel que logiciel une fois débuguées (habituellement les nouveaux batchs sur ls machines ont des problèmes), elles ont fonctionné avec une grosse charge . Un exemple est donné ci-après. Le reboot est généralement arrivé quand un composant electronique a grillé.

```
 2:29pm  up 495 days,  1:04,  2 users,  load average: 4.85, 7.15, 7.72
```

6. Remerciements

Les personnes suivantes ont été d'une grande aide pour réaliser ce HOWTO:

- Michael Levitt (Michael Levitt)

7. Bibliographie

Les documents suivants peuvent vous aider---ce sont des liens vers des sources qui utilisent des clusters pour effectuer du calcul haute performance:

- Page web de RAMBIN
- Page web de RAMP
- Page sur la recherche de Ram Samudrala (qui décrit quel type de recherche est effectué sur ces clusters)