

Disques et systèmes de fichiers

Administration Système et Réseaux, Sécurité

Disques et systèmes de fichiers

Philippe Harrand

¹Département Informatique
Pôle Sciences et Technologie

²Direction Territoriale Sud Ouest
France Télécom

16 septembre 2007

Philippe Harrand (Université de La Rochelle)

16 septembre 2007 1 / 36

Disques

RAID

Systèmes de Fichiers

Locaux
Distants

Philippe Harrand (Université de La Rochelle)

16 septembre 2007 2 / 36

Disques

- ▶ IDE (Integrated Drive Electronic)
 - ▶ ATA (Advanced Technology Attachment)
 - ▶ DMA => UDMA jusqu'à 133 Mo/s
 - ▶ ATAPI ATA adapté aux cd/dvd
- ▶ SCSI (Small Computer System Interface)
 - ▶ Jusqu'à 32 périphériques
 - ▶ 640 Mo/s
- ▶ Serial ATA
 - ▶ 600 Mo/s à terme
 - ▶ Hot plug

Philippe Harrand (Université de La Rochelle)

16 septembre 2007 3 / 36

Partitionnement

- ▶ 4 partitions «primaires»
- ▶ Dont une peut-être « étendue » et contenir
 - ▶ 63 partitions logiques en IDE
 - ▶ 15 partitions logiques en SCSI

Philippe Harrand (Université de La Rochelle)

16 septembre 2007 4 / 36

Organisation disque

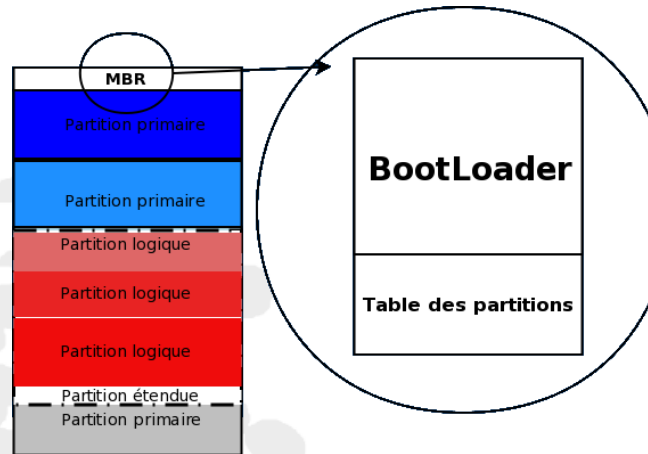


Table des partitions

Adresse	Champ	taille
1BE	1ère entrée dans la table de partition	16 octets
1CE	2ème entrée dans la table de partition	16 octets
1DE	3ème entrée dans la table de partition	16 octets
1EE	4ème entrée dans la table de partition	16 octets
1FE	AA55 (code d'identification)	2 octets

Table des partitions (suite)

Adresse	Contenu	Taille
0	État de la partition : 00 : partition non active 80 : partition active	1 octet
01	Tête où commence la partition	1 octet
02	Secteur et cylindre où commence la partition	2 octets
04	Type de partition	1 octet
05	Tête où finit la partition	1 octet
06	Secteur et cylindre où finit la partition	2 octets
08	Distance en secteurs entre MBR et secteur de boot de la partition	4 octets
0C	Nombre de secteurs de la partition	4 octets

Nommage des disques et partitions

	LINUX	*BSD
IDE0 maître	/dev/hda	/dev/wd0
IDE0 esclave	/dev/hdb	/dev/wd1
IDE1 maître	/dev/hdc	/dev/wd2
IDE1 esclave	/dev/hdd	/dev/wd3
Premier SCSI	/dev/sda	/dev/sd0
Second SCSI	/dev/sdb	/dev/sd1

Windows

- ▶ Windows considère les partitions, pas les disques
- ▶ Chaque partition porte une lettre
- ▶ Les lettres peuvent être attribuées manuellement

Linux

- ▶ Les partitions primaires sont numérotées de 1 à 4
- ▶ Les partitions logiques sont numérotées de 5 à n

*BSD

- ▶ Les partitions utilisables sont contenues dans une partition primaire
- ▶ Les *slices* sont numérotées de s1 à sn

Redondancy Array of Inexpensive Disks

- ▶ Buts
 - ▶ Accélérer l'accès aux données
 - ▶ Fiabiliser le stockage de données
 - ▶ Les deux en même temps
- ▶ 2 types d'implémentation
 - ▶ Contrôleur matériel
 - ▶ Module logiciel

Niveaux RAID

- 0 : striping
- 1 : mirroring
- 2 : striping with parity (obsolète)
- 3 : disk array with bit-interleaved data
- 4 : disk array with block-interleaved data
- 5 : disk array with block-interleaved distributed parity
- 6 : disk array with block-interleaved distributed parity

RAID 0

- ▶ Les données sont réparties sur l'ensemble des disques
- ▶ Si les contrôleurs sont différents on accède simultanément à des blocs bien plus gros
- ▶ Utiliser des disques identiques

Disque 1	Disque 2	Disque 3
Bande 1	Bande 2	Bande 3
Bande 4	Bande 5	Bande 6
Bande 7	Bande 8	Bande 9

RAID 1

- ▶ Les mêmes données sont inscrites sur tous les disques
- ▶ Pour un groupe de n disques, la perte de n-1 disques n'affecte pas le fonctionnement
- ▶ On peut constater une accélération ...

Disque 1	Disque 2	Disque 3
Bande 1	Bande 1	Bande 1
Bande 2	Bande 2	Bande 2
Bande 3	Bande 3	Bande 3

RAID 2

- ▶ Des codes de contrôles de Hamming sont stockés sur des disques séparés
- ▶ Piètre performance
- ▶ Obsolète

RAID 3

- ▶ Les données sont stockées par octet et un bit de parité est conservé sur un disque séparé
- ▶ Un disque défaillant peut être reconstitué

Disque 1	Disque 2	Disque 3	Disque 4
Octet 1	Octet 2	Octet 3	Parité 1+2+3
Octet 4	Octet 5	Octet 6	Parité 4+5+6
Octet 7	Octet 8	Octet 9	Parité 7+8+9

RAID 4

- ▶ Très proche de RAID 3
- ▶ La somme de contrôle est calculée sur un secteur au lieu d'un octet

Disque 1	Disque 2	Disque 3	Disque 4
Bloc 1	Bloc 2	Bloc 3	Parité 1+2+3
Bloc 4	Bloc 5	Bloc 6	Parité 4+5+6
Bloc 7	Bloc 8	Bloc 9	Parité 7+8+9

RAID 5

- ▶ Similaire au RAID 4
- ▶ Les sommes de contrôle sont réparties sur tous les disques
- ▶ Très bonnes performances

Disque 1	Disque 2	Disque 3	Disque 4
Bloc 1	Bloc 2	Bloc 3	Parité 1+2+3
Bloc 4	Parité 4+5+6	Bloc 5	Bloc 6
Parité 7+8+9	Bloc 7	Bloc 8	Bloc 9

RAID 6

- ▶ Similaire au RAID 4
- ▶ 2 sommes de contrôle sur 2 disques dédiés
- ▶ Perte de 2 disques sans perte de données

Disque 1	Disque 2	Disque 3	Disque 4	Disque 4bis
Bloc 1	Bloc 2	Bloc 3	Parité 1+2+3	Parité 1+2+3
Bloc 4	Bloc 5	Bloc 6	Parité 4+5+6	Parité 4+5+6
Bloc 7	Bloc 8	Bloc 9	Parité 7+8+9	Parité 7+8+9

Conclusion

Niveau	Avantages	Inconvénients
0	Vitesse	Fiabilité nulle
1	Fiabilité très grande si plus de 3 disques	Très cher
5	Fiabilité, performances	Ecriture plus lente
6	Fiabilité très grande, performances en lecture	Cher, écriture très lente

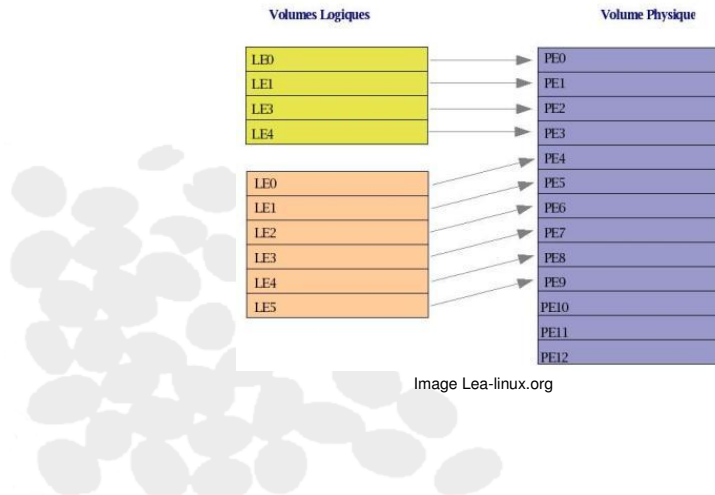
Logical Volume Manager

- ▶ Couche logique entre le RAID et le système
- ▶ Apporte la flexibilité
- ▶ Permet d'ajouter des disques «à chaud»
- ▶ Permet de modifier le partitionnement sans arrêter le système
- ▶ Libère l'administrateur des choix de partitionnement

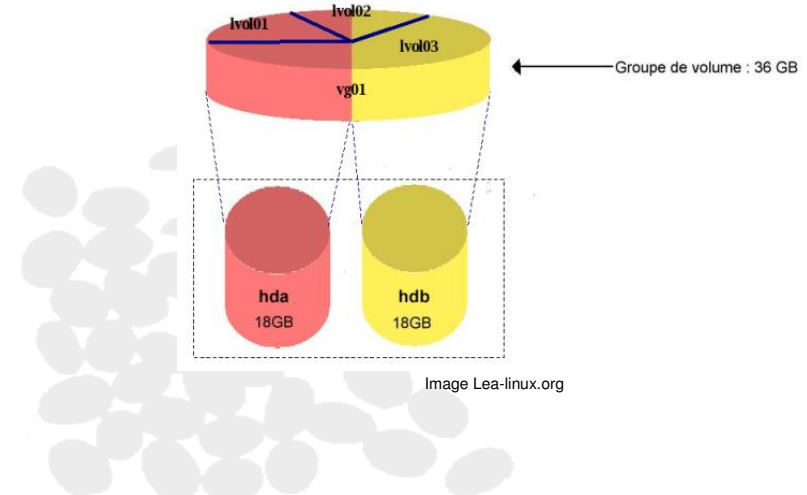
LVM

- ▶ Les volumes physiques (partitions ou disque entier) sont divisés en « Physical Extends »
- ▶ Les PE sont assemblés en «Volume Groups »
- ▶ Les volumes logiques (vus comme des partitions par le système) sont des agrégats de «Logical Extends » (pointeurs vers les PE)
- ▶ Il est possible d'ajouter et/ou de supprimer des LE dans un LV sans arrêter le système

LVM



LVM



Intérêt

- ▶ Gros système
 - ▶ Séparer des groupes d'utilisateurs
 - ▶ Moduler la taille des partitions en fonction des besoins imprévus
- ▶ Petits systèmes
 - ▶ Obtenir une partition importante avec plusieurs petits disques
 - ▶ Augmenter l'espace en faisant les poubelles

LVM

- ▶ Les chargeurs de boot ne connaissent pas encore tous LVM
- ▶ Les PE peuvent être attribués de manière linéaire ou répartie
- ▶ Les utilitaires de partitionnement gèrent le LVM
- ▶ Pour installer une DEBIAN avec LVM booter sur le noyau 2.6 de l'installateur !

Rôle d'un système de fichiers

- ▶ Gérer les informations stockées
- ▶ Conserver l'intégrité des données
- ▶ Gérer des méta-données
 - ▶ Dates diverses
 - ▶ Propriétaires
 - ▶ Droits d'accès
 - ▶

Systèmes de fichiers locaux

- ▶ Situés physiquement dans la machine
- ▶ Sur des disques durs
- ▶ Sur des disques souples
- ▶ Sur des supports amovibles
- ▶ En mémoire volatile

Journalisation

Le journal d'un système de fichiers

- ▶ Enregistre les transactions
- ▶ Garantit l'atomicité de celles-ci
- ▶ Permet un redémarrage plus rapide suite à un crash
- ▶ Nécessite une espace disque important (inutilisable sur disquette ou flash-RAM)
- ▶ Ralentit les opérations d'écriture

Quelques systèmes de fichiers

Nom du système de fichiers	Taille maximale d'un fichier	Taille maximale d'une partition	Journalisé ?	Gestion des droits d'accès ?
FAT (File Allocation Table)	2 GiB	2 GiB	Non	Non
FAT32	4 GiB	8 TiB	Non	Non
NTFS (New Technology File System)	Limitée par la taille de la partition	2 TiB	Oui	Oui
ext2fs (Extended File System)	2 TiB	4 TiB	Non	Oui
ext3fs	2 TiB	4 TiB	Oui	Oui
ReiserFS	8 TiB	16 TiB	Oui	Oui
XFS (eXtended File System)	18x10 ⁶ TiB	??	Oui	Oui

TiB = 10¹² octets

Systèmes de fichiers spéciaux

- ▶ ISO9660
 - ▶ joliet ⇒ noms longs en unicode et arborescence > 8 niveaux
 - ▶ Rock Ridge ⇒ noms longs en ASCII et droits Unix
 - ▶ El Torito ⇒ CD bootable
- ▶ UDF
- ▶ ramfs, tmpfs
- ▶ ...

Les Systèmes de fichiers

- ▶ Se montent avec mount
- ▶ Se créent avec mkfs
- ▶ Se réparent avec fsck
- ▶ /etc/fstab indique les SF montables
- ▶ /etc/mtab indique les SF montés
- ▶ En cas de pb, dd est votre ami
- ▶ Dumps2fs permet de visualiser un FS ext2/3

Ext2/3 fs

- ▶ Le superblock
Méta données sur le FS
- ▶ Les inodes
Méta données des fichiers
- ▶ Le directory
Nom ⇔ inode

Les quotas

- ▶ Fixent le nombre maximal de fichiers (inodes) et/ou d'octets pour un utilisateur ou un groupe
- ▶ S'appliquent à un système de fichiers
- ▶ Définissent une limite douce et une limite dure
- ▶ Ne concernent que les systèmes de fichiers montés localement

Particularité SCSI

- ▶ Système de fichier local qui peut être distant !
- ▶ SAN
 - ▶ Fiber Channel
 - ▶ iSCSI

Principaux systèmes de fichiers distants

- ▶ NFS
 - ▶ Standard UNIX
- ▶ SMBFS
 - ▶ Server Message Bloc ou Lan Manager ou « réseau Microsoft »
- ▶ SSHFS
 - ▶ Fonctionne en espace utilisateur
 - ▶ Naturellement chiffré
- ▶ AppleShare
 - ▶ Propriétaire apple

Network File System

- ▶ Utilise les RPC
- ▶ Conservation des droits (basés sur l'UID !)
- ▶ Le serveur exporte un ou plusieurs répertoires vers des machines désignées par leur adresse IP
 - ▶ Tcprawrapper sécurise un peu les connexions
 - ▶ La version 4 de NFS permet le chiffrement

Server Message Block

- ▶ Protocole natif Windows (partage de fichiers)
- ▶ Implémenté sous UNIX par SAMBA
- ▶ Chiffrement des mots de passe facultatif !
- ▶ Le trafic est en broadcast (le plus souvent)
- ▶ Les données sont en clair

Secure SHell File System

- ▶ Basé sur SSH
- ▶ Fonctionne en espace utilisateur grace au module *fuse*
- ▶ Pas besoin d'être root pour le monter
- ▶ Naturellement chiffré

