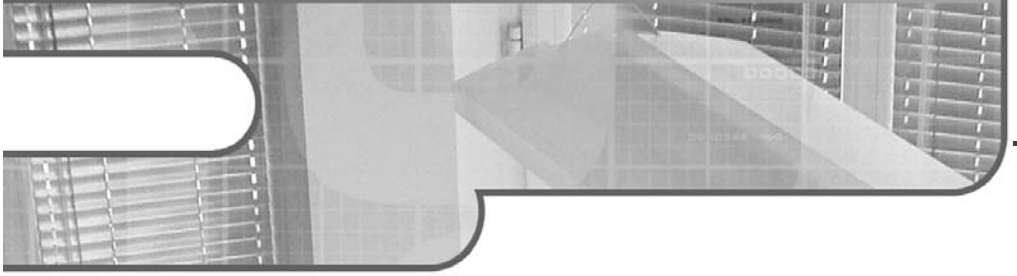

Chapitre 3
Gestion des périphériques de stockage

1. Gestion des périphériques de stockage

Ce chapitre traite de la gestion des disques RAID, de la configuration bas-niveau des disques, des disques réseau iSCSI, ainsi que du gestionnaire de volumes logiques LVM.

1.1 Configuration des disques RAID

L'objectif de cette section est de vous apprendre à :

– configurer et implémenter du RAID logiciel. Cela inclut les niveaux de RAID 0, 1 et 5.

1.1.1 Compétences principales

– Fichiers de configuration et utilitaires de gestion du RAID logiciel.

1.1.2 Éléments mis en œuvre

– `mdadm.conf`
– `mdadm`
– `/proc/mdstat`
– Partition de type `0xFD`

1.2 Optimiser l'accès aux périphériques de stockage

L'objectif de cette section est de vous apprendre à :

– configurer le noyau pour gérer différents types de disques ;
– connaître les outils logiciels pour lister et modifier le paramétrage des périphériques iSCSI.

1.2.1 Compétences principales

– Outils et commandes pour configurer le DMA pour des périphériques IDE, y compris ATAPI et SATA.
– Outils et commandes pour configurer des disques SSD (*Solid State Drive*), y compris AHCI et NVMe.
– Outils et commandes pour configurer ou analyser les ressources système (par exemple les interruptions).
– Connaissances de base de la commande `sdparm` et de son utilisation.
– Outils et commandes de gestion des périphériques iSCSI.
– Connaissances de base du SAN, y compris les protocoles spécifiques (AoE, FCoE).

1.2.2 Éléments mis en œuvre

– `hdparm, sdparm`
– `nvme`
– `tune2fs`
– `fstrim`
– `sysctl`
– `/dev/hd*, /dev/sd*, /dev/nvme*`
– `iscsiadm, scsi_id, iscsid` et `iscsid.conf`
– `WWID, WWN, n° LUN`

1.3 Logical Volume Manager

L'objectif de cette section est de vous apprendre à :

– créer et supprimer des volumes logiques, des groupes de volumes et des volumes physiques. Cet objectif inclut les instantanés (*snapshots*) et le redimensionnement des volumes logiques.

1.3.1 Compétences principales

- Outils de la suite LVM.
- Redimensionner, renommer, créer, supprimer des volumes logiques, des groupes de volumes et des volumes physiques.
- Créer et maintenir des instantanés (*snapshots*).
- Activer des groupes de volumes.

1.3.2 Éléments mis en œuvre

- `/sbin/pv*`
- `/sbin/lv*`
- `/sbin/vg*`
- `mount`
- `/dev/mapper/`
- `lvm.conf`

2. Configuration des disques RAID

La technologie RAID (*Redundant Array of Independent Disks*) permet de combiner différents périphériques pour qu'ils soient vus comme un seul espace de stockage par les applications. On peut ainsi améliorer les temps d'accès et/ou la fiabilité des périphériques de stockage. Les différentes techniques mises en œuvre sont définies par leur niveau de RAID. Les niveaux les plus courants sont RAID 0, 1 et 5.

Le RAID peut être géré de façon matérielle, par des contrôleurs de disques spécialisés, ou logicielle, au niveau du système d'exploitation.

Linux implémente un pilote de gestion logicielle du RAID, le pilote `md` (*Multiple Device driver*), qui gère les niveaux les plus courants de RAID, 0, 1 et 5.

Remarque

D'autres solutions peuvent être mises en place pour gérer du RAID logiciel sur Linux : RAID LVM ou RAID directement pris en charge par le gestionnaire de système de fichiers ZFS ou Btrfs.

2.1 Les principaux niveaux de RAID

2.1.1 Le RAID 0

Le RAID 0 (agrégat par bandes, *striping*) combine plusieurs disques en un seul ensemble. Les blocs de données sont répartis sur des bandes de taille identique réparties uniformément sur les différents disques. Les opérations d'entrées-sorties peuvent donc être très rapides, car effectuées simultanément par les différents contrôleurs disques.

En revanche, la fiabilité de l'ensemble est fortement diminuée, puisqu'il suffit de perdre un disque pour perdre l'ensemble des données. Il n'y a aucune redondance des données stockées, et la cohérence des volumes logiques est détruite en cas de défaillance d'un disque.

L'espace de stockage utile d'un ensemble RAID 0 est égal à la capacité utile du plus petit des disques, multipliée par le nombre de disques qui le composent, puisqu'il n'y a pas de redondance des données et que les bandes de données sont réparties uniformément sur les disques (chaque disque doit avoir le même nombre de bandes).

Avantages :

– Rapidité de lecture et d'écriture des ensembles de blocs.
– Utilisation optimale de l'espace disque, si les disques sont de même taille.

Inconvénients :

– Pas de redondance des données, donc pas de tolérance de panne.
– La perte d'un disque compromet l'ensemble des données stockées, la fiabilité de l'ensemble est égale à la fiabilité du moins fiable des disques utilisés.

2.1.2 Le RAID 1

Le RAID 1 (disques miroirs, *mirroring*) combine plusieurs disques en un seul ensemble. Chaque bloc de données utile est écrit sur chacun des disques. Cette redondance assure une excellente fiabilité à l'ensemble, d'autant plus grande qu'il y a davantage de disques. Tant qu'il reste un disque opérationnel, les données sont intactes, et tant que le contrôleur de ce disque fonctionne, elles restent accessibles.

Les opérations de lecture peuvent être plus rapides, car elles peuvent être effectuées simultanément par les différents contrôleurs disques.

L'espace de stockage utile d'un ensemble RAID 1 est égal à la capacité utile du plus petit des disques.

Avantages :

– Excellente tolérance de panne, proportionnelle au nombre de disques combinés (et au nombre de contrôleurs de disques pour l'accessibilité).
– Bonnes performances en lecture.

Inconvénients :

– L'espace disque nécessaire est au moins deux fois la taille de l'espace disque utile.
– Les performances en écriture peuvent être impactées, même si en général les écritures se font simultanément sur les différents disques.

2.1.3 Le RAID 5

Le RAID 5 (agrégat par bandes avec parité) combine au moins trois disques en un seul ensemble. Les blocs de données sont répartis sur des bandes de taille identique réparties uniformément sur les différents disques sauf un. Pour chaque ensemble de bandes, une bande de parité est calculée et écrite sur le disque restant. L'emplacement de la bande de parité est réparti à tour de rôle sur les disques.

En cas de perte d'une bande de données, la bande de parité permet de la reconstituer, assurant ainsi la tolérance de panne. Mais ce mécanisme n'est efficace que pour un seul disque inaccessible, la défaillance de deux disques ou plus entraîne la perte des données de l'ensemble des disques. Tant qu'un disque n'est pas opérationnel, il n'y a plus de tolérance de panne pour les nouvelles écritures. C'est pourquoi les ensembles de disques en RAID 5 intègrent généralement un disque de secours (*spare disk*), qui n'est utilisé que pour remplacer un disque défaillant.

Une fois le disque défaillant réparé ou remplacé, il faut reconstruire l'ensemble RAID 5 en reconstituant les données et les bandes de parité pour les écrire sur le disque remplaçant.

Les opérations de lecture peuvent être très rapides, car effectuées simultanément par les différents contrôleurs disques. Les écritures peuvent être ralenties, à cause du calcul et de l'écriture de la bande de parité.

L'espace de stockage utile d'un ensemble RAID 5 est égal à la capacité du plus petit des disques, multipliée par le nombre de disques qui le composent, moins 1 à cause des bandes de parité et moins 2 s'il y a un disque de secours (*spare*).

Avantages :

– Tolérance de panne, limitée à un disque. Pendant la défaillance du disque, il n'y a plus de tolérance de panne, sauf avec un disque de secours.

Inconvénients :

– Une partie de l'espace disque n'est pas utilisable pour les données.
– Les performances en écriture peuvent être impactées, à cause du calcul de la parité.

2.2 Configuration du RAID

Le pilote md est un module du noyau qui prend en charge le RAID logiciel sur un ensemble de périphériques de stockage, disques durs complets et/ou partitions de disques durs.

La commande `mdadm` permet de configurer des volumes RAID et de les gérer. Elle fait partie du paquet `mdadm`.

2.2.1 Création d'un volume RAID

Un volume RAID est composé de plusieurs espaces de stockage, qui peuvent être des disques durs entiers ou des partitions de disque dur.

La création d'un volume RAID se fait par l'option `-C` de la commande `mdadm`. Il faut spécifier le nom du nouveau volume ou son numéro, le niveau de RAID à mettre en œuvre et la liste des espaces de stockage à lui allouer.

Remarque

Le fichier de configuration de la commande, généralement `/etc/mdadm/mdadm.conf`, est devenu facultatif dans les versions récentes et n'est pas créé à l'installation du paquet.

Syntaxe

```
mdadm -C FicSpecVol -l|--level=Niveau -n|--raid-devices=NbDevRaid
[ -x|--spare-devices=NbSecours ] FicSpec1 ... FicSpecN
```

Principaux paramètres

`-C FicSpecVol`	Fichier spécial du volume RAID créé.
`-l|--level=Niveau`	Niveau de RAID.
`-n|--raid-devices=NbDevRaid`	Nombre d'espaces de stockage actifs.
`-x|--spare-devices=NbSecours`	Nombre d'espaces de stockage de secours.
`FicSpec1 ... FicSpecN`	Espaces de stockage.

Description

L'option `-C` crée un nouveau volume RAID. Le fichier spécial qui lui sera associé, `FicSpecVol`, est généralement de la forme `/dev/mdX`, `X` étant un chiffre, mais ce n'est pas obligatoire.

L'option `-n` spécifie le nombre d'espaces de stockage à utiliser pour le volume. Il doit être égal ou supérieur au nombre d'éléments de la liste `FicSpec1 ...`, `FicSpecN` moins le nombre d'espaces de stockage de secours, indiqué par l'option `-x`.

Une fois créé, le volume RAID peut être immédiatement utilisé. Il est vu comme un périphérique en mode bloc, on peut donc y créer un système de fichiers ou en faire un volume physique LVM.

Exemples

On utilise deux partitions de disques durs, `/dev/sda4` et `/dev/sdd1`, pour créer un volume `RAID` de niveau `1` (miroir). Comme les partitions contiennent des systèmes de fichiers et sont de tailles différentes, la commande affiche un avertissement et demande une confirmation :

```
mdadm -C /dev/md0 -l 1 -n 2 /dev/sda4 /dev/sdd1
mdadm: /dev/sda4 appears to contain an ext2fs file system
       size=5237760K  mtime=Thu Jan  1 01:00:00 1970
mdadm: Note: this array has metadata at the start and
   may not be suitable as a boot device.  If you plan to
   store '/boot' on this device please ensure that
   your boot-loader understands md/v1.x metadata, or use
   --metadata=0.90
mdadm: largest drive (/dev/sdd1) exceeds size (5232640K) by more than 1%
Continue creating array? y
mdadm: Defaulting to version 1.2 metadata
mdadm: array /dev/md0 started.
```

Le volume RAID est créé :

```
ls -l /dev/md0
brw-rw----. 1 root disk 9, 0 10 mars  18:08 /dev/md0
Le fichier /dev/md0 est un fichier spécial bloc.
```

La commande `blkid` donne des informations sur les deux espaces de stockage composant le volume RAID :

```
blkid /dev/sda4 /dev/sdd1
/dev/sda4: UUID="760b8ab1-9041-5cd1-7e1e-1308fa7e75fd" UUID_SUB="679cb515-
2f5c-6078-6829-4b9f0f928907" LABEL="beta:0" TYPE="linux_raid_member"
PARTUUID="12deb3a0-04"
/dev/sdd1: UUID="760b8ab1-9041-5cd1-7e1e-1308fa7e75fd" UUID_SUB="e77f5f03-
448a-1b48-feb6-9bedc62e40b5" LABEL="beta:0" TYPE="linux_raid_member"
PARTUUID="c3072e18-01"
```

Les deux partitions sont de type RAID Linux et ont reçu un label `beta:0`, `beta` étant le nom de la machine.