

Présentation de la technologie RAID, un article de « Comment ça marche »

La technologie RAID (acronyme de *Redundant Array of Inexpensive Disks*, parfois *Redundant Array of Independent Disks*, traduisez *Ensemble redondant de disques indépendants*) permet de constituer **une** unité de stockage à partir de plusieurs disques durs. L'unité ainsi créée (appelée **grappe**) a donc une grande tolérance aux pannes (haute disponibilité), ou bien une plus grande capacité/vitesse d'écriture. La répartition des données sur plusieurs disques durs permet donc d'en augmenter la sécurité et de fiabiliser les services associés.

Cette technologie a été mise au point en 1987 par trois chercheurs (*Patterson, Gibson et Katz*) à l'Université de Californie (Berkeley). Depuis 1992 c'est le RAID Advisory Board qui gère ces spécifications. Elle consiste à constituer un disque de grosse capacité (donc coûteux) à l'aide de plus petits disques peu onéreux (c'est-à-dire dont le **MTBF**, *Mean Time Between Failure*, soit le temps moyen entre deux pannes, est faible).

Les disques assemblés selon la technologie RAID peuvent être utilisés de différentes façons, appelées **Niveaux RAID**. L'Université de Californie en a défini 5, auxquels ont été ajoutés les niveaux 0 et 6. Chacun d'entre-eux décrit la manière de laquelle les données sont réparties sur les disques :

- **Niveau 0**: appelé *striping*
- **Niveau 1**: appelé *mirroring, shadowing* ou *duplexing*
- **Niveau 2**: appelé *striping with parity* (obsolète)
- **Niveau 3**: appelé *disk array with bit-interleaved data*
- **Niveau 4**: appelé *disk array with block-interleaved data*
- **Niveau 5**: appelé *disk array with block-interleaved distributed parity*
- **Niveau 6**: appelé *disk array with block-interleaved distributed parity*

Chacun de ces niveaux constitue un mode d'utilisation de la grappe, en fonction :

- des performances
- du coût
- des accès disques

Niveau 0

Le niveau RAID-0, appelé **striping** (traduisez *entrelacement* ou *agrégat par bande*, parfois injustement appelé *stripping*) consiste à stocker les données en les répartissant sur l'ensemble des disques de la grappe. De cette façon, il n'y a pas de redondance, on ne peut donc pas parler de tolérance aux pannes. En effet en cas de défaillance de l'un des disques, l'intégralité des données réparties sur les disques sera perdue.

Toutefois, étant donné que chaque disque de la grappe a son propre contrôleur, cela constitue une solution offrant une vitesse de transfert élevée.

Le RAID 0 consiste ainsi en la juxtaposition logique (agrégation) de plusieurs disques durs physiques. En mode RAID-0 les données sont écrites par "bandes" (en anglais *stripes*) :

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

On parle de facteur d'entrelacement pour caractériser la taille relative des fragments (*bandes*) stockés sur chaque unité physique. Le débit de transfert moyen dépend de ce facteur (plus petite est chaque bande, meilleur est le débit).

Si un des éléments de la grappe est plus grand que les autres, le système de remplissage par bande se trouvera bloqué lorsque le plus petit des disques sera rempli. La taille finale est ainsi égale au double de la capacité du plus petit des deux disques :

- deux disques de 20 Go donneront un disque logique de 40 Go.
- un disque de 10 Go utilisé conjointement avec un disque de 27 Go permettra d'obtenir un disque logique de 20 Go (17 Go du second disque seront alors inutilisés).



Il est recommandé d'utiliser des disques de même taille pour faire du RAID-0 car dans le cas contraire le disque de plus grande capacité ne sera pas pleinement exploité.

Niveau 1

Le niveau 1 a pour but de dupliquer l'information à stocker sur plusieurs disques, on parle donc de *mirroring*, ou *shadowing* pour désigner ce procédé.

Disque1	Disque2	Disque3
Bande 1	Bande 1	Bande 1
Bande 2	Bande 2	Bande 2
Bande 3	Bande 3	Bande 3

On obtient ainsi une plus grande sécurité des données, car si l'un des disques tombe en panne, les données sont sauvegardées sur l'autre. D'autre part, la lecture peut être beaucoup plus rapide lorsque les deux disques sont en fonctionnement. Enfin, étant donné que chaque disque possède son propre contrôleur, le serveur peut continuer à fonctionner même lorsque l'un des disques tombe en panne, au même titre qu'un camion pourra continuer à rouler si un de ses pneus crève, car il en a plusieurs sur chaque essieu...

En contrepartie la technologie RAID1 est très onéreuse étant donné que seule la moitié de la capacité de stockage n'est effectivement utilisée.

Niveau 2

Le niveau RAID-2 est désormais obsolète, car il propose un contrôle d'erreur par code de Hamming (codes **ECC** - *Error Correction Code*), or ce dernier est désormais directement intégré dans les contrôleurs de disques durs.

Cette technologie consiste à stocker les données selon le même principe qu'avec le RAID-0 mais en écrivant sur une unité distincte les bits de contrôle *ECC* (généralement 3 disques ECC sont utilisés pour 4 disques de données).

La technologie RAID 2 offre de piètres performances mais un niveau de sécurité élevé.

Niveau 3

Le niveau 3 propose de stocker les données sous forme d'octets sur chaque disque et de dédier un des disques au stockage d'un bit de parité.

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

De cette manière, si l'un des disques venait à défaillir, il serait possible de reconstituer l'information à partir des autres disques. Après "reconstitution" le contenu du disque défaillant est de nouveau intègre. Par contre, si deux disques venaient à tomber en panne simultanément, il serait alors impossible de remédier à la perte de données.

Niveau 4

Le niveau 4 est très proche du niveau 3. La différence se trouve au niveau de la parité, qui est faite sur un secteur (appelé *bloc*) et non au niveau du bit, et qui est stockée sur un disque dédié. C'est-à-dire plus précisément que la valeur du facteur d'entrelacement est différente par rapport au RAID 3.

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

Ainsi, pour lire un nombre de blocs réduits, le système n'a pas à accéder à de multiples lecteurs physiques, mais uniquement à ceux sur lesquels les données sont effectivement stockées. En contrepartie le disque hébergeant les données de contrôle doit avoir un temps d'accès égal à la somme des temps d'accès des autres disques pour ne pas limiter les performances de l'ensemble.

Niveau 5

Le niveau 5 est similaire au niveau 4, c'est-à-dire que la parité est calculée au niveau d'un secteur, mais répartie sur l'ensemble des disques de la grappe.

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

De cette façon, RAID 5 améliore grandement l'accès aux données (aussi bien en lecture qu'en écriture) car l'accès aux bits de parité est réparti sur les différents disques de la grappe.

Le mode RAID-5 permet d'obtenir des performances très proches de celles obtenues en RAID-0, tout en assurant une tolérance aux pannes élevée, c'est la raison pour laquelle c'est un des modes RAID les plus intéressants en termes de performance et de fiabilité.



L'espace disque utile sur une grappe de n disques étant égal à $n-1$ disques, il est intéressant d'avoir un grand nombre de disques pour "rentabiliser" le RAID-5.

Niveau 6

Le niveau 6 a été ajouté aux niveaux définis par Berkeley. Il définit l'utilisation de 2 fonctions de parité, et donc leur stockage sur deux disques dédiés. Ce niveau permet ainsi d'assurer la redondance en cas d'avarie simultanée de deux disques. Cela signifie qu'il faut au moins 4 disques pour mettre en œuvre un système RAID-6.

Comparaison

Les solutions RAID généralement retenues sont le RAID de niveau 1 et le RAID de niveau 5.

Le choix d'une solution RAID est lié à trois critères :

- **la sécurité** : RAID 1 et 5 offrent tous les deux un niveau de sécurité élevé, toutefois la méthode de reconstruction des disques varie entre les deux solutions. En cas de panne du système, RAID 5 reconstruit le disque manquant à partir des informations stockées sur les autres disques, tandis que RAID 1 opère une copie disque à disque.
- **Les performances** : RAID 1 offre de meilleures performances que RAID 5 en lecture, mais souffre lors d'importantes opérations d'écriture.
- **Le coût** : le coût est directement lié à la capacité de stockage devant être mise en œuvre pour avoir une certaine capacité effective. La solution RAID 5 offre un volume utile représentant 80 à 90% du volume alloué (le reste servant évidemment au contrôle d'erreur). La solution RAID 1 n'offre par contre qu'un volume disponible représentant 50 % du volume total (étant donné que les informations sont dupliquées).

Mise en place d'une solution RAID

Il existe plusieurs façons différentes de mettre en place une solution RAID sur un serveur :

- **de façon logicielle** : il s'agit généralement d'un driver au niveau du système d'exploitation de l'ordinateur capable de créer un seul volume logique avec plusieurs disques (SCSI ou IDE).
- **de façon matérielle**
 - **avec des matériels DASD** (*Direct Access Stockage Device*) : il s'agit d'unités de stockage externes pourvues d'une alimentation propre. De plus ces matériels sont dotés de connecteurs permettant l'échange de disques à chaud (on dit généralement que ce type de disque est *hot swappable*). Ce matériel gère lui-même ses disques, si bien qu'il est reconnu comme un disque SCSI standard.
 - **avec des contrôleurs de disques RAID** : il s'agit de cartes s'enfichant dans des slots PCI ou ISA et permettant de contrôler plusieurs disques durs.