RAID

Historique

En 1978, un employé d'IBM, Norman Ken Ouchi, déposa un brevet[3] concernant un « Système de récupération de données stockées dans une unité de stockage défectueuse », et dont la description était ce que deviendrait plus tard le RAID 5. Ce brevet fait également mention du miroitage (mirroring) de disque (qui sera appelé plus tard RAID 1), ainsi que de la protection avec une parité dédiée (qui sera appelé plus tard RAID 3 et 4).
La technologie RAID a été élaborée par un groupe de chercheurs de l'Université de Berkeley (Californie) en 1987. Ces derniers étudièrent la possibilité de faire reconnaître deux disques durs ou plus comme une seule entité par le système. Ils obtinrent pour résultat un système de stockage aux performances bien meilleures que celles des systèmes à disque dur unique, mais doté d'une très mauvaise fiabilité. Les chercheurs s'orientèrent alors vers des architectures redondantes, afin d'améliorer la tolérance aux pannes du système de stockage.
En 1988, les différents RAID, de type 1 à 5, étaient formellement définis par David Patterson, Garth Gibson et Randy Katz dans la publication intitulée « *A Case for Redundant Arrays of Inexpensive Disks (RAID)*[4] ». Cet article introduisait le terme « RAID », dont l'industrie du disque s'est immédiatement emparée, dont elle proposait cinq niveaux différents, en les comparant au « SLED », chacun d'eux ayant ses avantages et ses inconvénients.

Parité et redondance

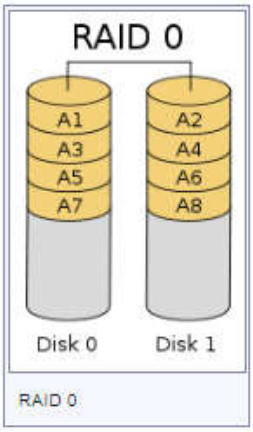
Le miroitage s'avère être une solution onéreuse, puisqu'il est nécessaire d'acquérir les périphériques de stockage en plusieurs exemplaires. Aussi, partant du principe que plusieurs unités de stockage ont une faible probabilité de tomber en panne simultanément, d'autres systèmes ont été imaginés, dont ceux permettant de régénérer les données manquantes à partir des données restant accessibles et d'une ou plusieurs données supplémentaires, dites de redondance.
Le système de redondance le plus simple et le plus largement utilisé est le calcul de parité. Ce système repose sur l'opération logique XOR (OU exclusif) et consiste à déterminer si sur n bits de données considérés, le nombre de bits à l'état 1 est pair ou impair. Si le nombre de 1 est pair, alors le bit de parité vaut 0. Si le nombre de 1 est impair, alors le bit de parité vaut 1. Lorsque l'un des $n+1$ bits de données ainsi formés devient indisponible, il est alors possible de régénérer le bit manquant en appliquant à nouveau la même méthode sur les n éléments restants. Cette technique est utilisée dans les systèmes RAID 5.
Il existe des systèmes de redondance plus complexes et capables de générer plusieurs éléments de redondance afin de supporter l'absence de plusieurs éléments. Le RAID 6 utilise par exemple une technique de calcul de parité fondée sur des polynômes.

Le système RAID est :

- soit un système de redondance qui donne au stockage des données une certaine tolérance aux pannes matérielles (ex : RAID1).
- soit un système de répartition qui améliore ses performances (ex : RAID0).
- soit les deux à la fois mais avec une moins bonne efficacité (ex : RAID5).

Le système RAID est donc capable de gérer d'une manière ou d'une autre la répartition et la cohérence de ces données. Ce système de contrôle peut être purement logiciel ou utiliser un matériel dédié.

RAID 0 : volume agrégé par bandes



RAID 0

Le RAID 0, également connu sous le nom d'« entrelacement de disques » ou de « volume agrégé par bandes » (*striping* en anglais) est une configuration RAID permettant d'augmenter significativement les performances de la grappe en faisant travailler n disques durs en parallèle (avec $n \geq 2$).

- Capacité :
  La capacité totale est égale à celle du plus petit élément de la grappe multiplié par le nombre d'éléments présent dans la grappe, car le système d'agrégation par bandes se retrouvera bloqué une fois que le plus petit disque sera rempli (voir schéma). L'espace excédentaire des autres éléments de la grappe restera inutilisé. Il est donc conseillé d'utiliser des disques de même capacité.
- Fiabilité :
  Le défaut de cette solution est que la perte d'un seul disque entraîne la perte de toutes ses données.
- Coût :
  Dans un RAID 0, qui n'apporte aucune redondance, tout l'espace disque disponible est utilisé (tant que tous les disques ont la même capacité).

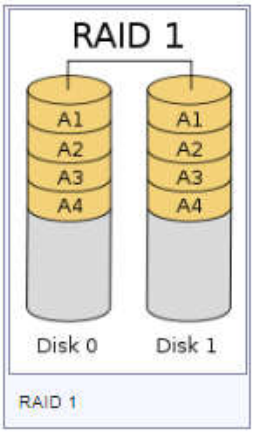
Dans cette configuration, les données sont réparties par bandes (*stripes* en anglais) d'une taille fixe. Cette taille est appelée granularité (voir plus loin la section granularité).

> Exemple : avec un RAID 0 ayant une bande de 64 Kio et composé de deux disques (disque *Disk 0* et disque *Disk 1*), si l'on veut écrire un fichier **A** de 500 Kio, le fichier sera découpé en 8 bandes (car $7 < \frac{500}{64} \leq 8$), appelons-les 1, 2, 3, 4, 5, 6, 7 et 8, qui seront réparties sur l'ensemble des disques de la façon suivante :
> Disk 0 : 1, 3, 5, 7
> Disk 1 : 2, 4, 6, 8
>
> Ainsi l'écriture du fichier pourra être effectuée simultanément sur chacun des disques en un temps équivalent à l'écriture de 256 kio.

Ainsi, sur un RAID 0 de n disques (avec $n \geq 2$), chaque disque ne doit lire et écrire que $\frac{1}{n}$ des données, ce qui a pour effet de diminuer les temps d'accès (lecture et écriture) aux données; les disques se partageant le travail, les traitements se trouvent accélérés. Ce type de RAID est parfait pour des applications requérant un traitement rapide d'une grande quantité de données. Mais cette architecture n'assure en rien la sécurité des données ; en effet, si l'un des disques tombe en panne, la totalité des données du RAID est perdue.

RAID 1 : Disques en miroir

Le RAID 1 consiste en l'utilisation de n disques redondants (avec $n \geq 2$), chaque disque de la grappe contenant à tout moment exactement les mêmes données, d'où l'utilisation du mot « miroir » (*mirroring* en anglais).



RAID 1

- Capacité :
  La capacité totale est égale à celle du plus petit élément de la grappe. L'espace excédentaire des autres éléments de la grappe restera inutilisé. Il est donc conseillé d'utiliser des éléments identiques.
- Fiabilité :
  Cette solution offre un excellent niveau de protection des données. Elle accepte une défaillance de $n - 1$ éléments.
- Coût :
  Les coûts de stockage sont élevés et directement proportionnels au nombre de miroirs utilisés alors que la capacité totale reste inchangée. Plus le nombre de miroirs est élevé, et plus la sécurité augmente, mais plus son coût devient prohibitif.

Les accès en lecture du système d'exploitation se font sur le disque le plus facilement accessible à ce moment-là.[réf. nécessaire] Les écritures sur la grappe se font de manière simultanée sur tous les disques, de façon à ce que n'importe quel disque soit interchangeable à tout moment.

Lors de la défaillance de l'un des disques, le contrôleur RAID désactive, de manière transparente pour l'accès aux données, le disque incriminé. Une fois le disque défectueux remplacé, le contrôleur RAID reconstitue, soit automatiquement, soit sur intervention manuelle, le miroir. Une fois la synchronisation effectuée, le RAID retrouve son niveau initial de redondance.

Enfichage à chaud (hotplug/hotswap)

On parle abusivement de disques pouvant être enfichés à chaud (*hotplug/hotswap* en anglais), alors qu'en réalité, c'est la baie de disques du système ainsi que le contrôleur qui doivent être conçus de manière à permettre le retrait ou l'insertion de disques durs alors que le système est sous tension.
Cette fonctionnalité n'est pas disponible avec toutes les technologies :

- Bien qu'il n'y ait généralement pas de dommages physiques, les disques IDE ne gèrent pas cette fonctionnalité.
- Cette fonctionnalité est gérée par des disques SATA (sous réserve que le contrôleur le gère également).
- Cette fonctionnalité est gérée par des disques SCSI (sous réserve que le contrôleur le gère également) bien que le bus puisse être perturbé au moment de l'échange.

Cela permet :

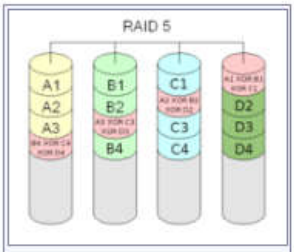
- d'ajouter des disques de manière dynamique, de sorte qu'il soit possible de faire évoluer le système de stockage de données.
- de remplacer un matériel défectueux sans qu'il soit nécessaire d'interrompre le fonctionnement du système informatique.

L'utilisation de systèmes de connexion à chaud permet donc d'éviter l'indisponibilité durant une opération de maintenance.

Disques de rechange (spare/hotspare)

Les disques de rechange (*spare/hotspare* en anglais) permettent de limiter la vulnérabilité d'une solution.
Un disque complémentaire est affecté à une unité RAID mais n'est pas utilisé au quotidien. Il est appelé disque de rechange. Lorsqu'un disque de la grappe vient à défaillir, le disque de rechange prend immédiatement et automatiquement son relais. Ce disque est alors reconstruit à partir des données présentes sur les autres disques, ce qui peut durer plusieurs heures en fonction de la quantité de données. Une fois le disque reconstruit, le système revient à un niveau optimal de sécurité et de performances.
Une fois le disque de rechange mis en service, il faut procéder à l'échange physique du disque en panne par un nouveau disque qui pourra jouer le rôle de nouveau disque de rechange.

RAID 5 : volume agrégé par bandes à parité répartie



RAID 5

Le RAID 5 combine la méthode du volume agrégé par bandes (*striping*) à une parité répartie. Il s'agit là d'un ensemble à redondance $N+1$. La parité, qui est incluse avec chaque écriture se retrouve répartie circulairement sur les différents disques. Chaque bande est donc constituée de N blocs de données et d'un bloc de parité. Ainsi, en cas de défaillance de l'un des disques de la grappe, pour chaque bande il manquera soit un bloc de données soit le bloc de parité. Si c'est le bloc de parité, ce n'est pas grave, car aucune donnée ne manque. Si c'est un bloc de données, on peut calculer son contenu à partir des $N-1$ autres blocs de données et du bloc de parité. L'intégrité des données de chaque bande est préservée. Donc non seulement la grappe est toujours en état de fonctionner, mais il est de plus possible de reconstruire le disque une fois échangé à partir des données et des informations de parité contenues sur les autres disques.
On voit donc que le RAID 5 ne supporte la perte que d'un seul disque à la fois. Ce qui devient un problème depuis que les disques qui composent une grappe sont de plus en plus gros (1 To et plus). Le temps de reconstruction de la parité en cas de disque défaillant est allongé. Il est généralement de 2 h pour des disques de 300 Go contre une dizaine d'heures pour 1 To. Pour limiter le risque il est courant de dédier un disque dit de *spare*. En régime normal il est inutilisé. En cas de panne d'un disque il prendra automatiquement la place du disque défaillant. Cela nécessite une phase communément appelée "recalcul de parité". Elle consiste pour chaque bande à recréer sur le nouveau disque le bloc manquant (données ou parité).
Bien sûr pendant tout le temps du recalcul de la parité le disque est disponible normalement pour l'ordinateur qui se trouve juste un peu ralenti.

> Exemple pratique : Considérons quatre disques durs A, B, C et D, de tailles identiques. Le système va enregistrer les premiers blocs en les répartissant sur les disques A, B et C comme en mode RAID 0 (*striping*) et, sur le disque D, le résultat de l'opération OU exclusif entre les autres disques (ici A xor B xor C). Ensuite le système va enregistrer les blocs suivants en les répartissant sur les disques D, A et B, puis la parité (soit D xor A xor B) sur le disque C, et ainsi de suite en faisant permuter circulairement les disques, à chaque bloc. La parité se trouve alors répartie sur tous les disques.
>
> En cas de défaillance d'un disque, les données qui s'y trouvaient pourront être reconstituées par l'opération xor. En effet, l'opération XOR ($\oplus$) a la propriété suivante : si on considère N blocs de taille identique $A_1, A_2 \cdots A_N$ et si $A_1 \oplus A_2 \oplus \cdots A_N = X$ alors $X \oplus A_2 \oplus \cdots A_N = A_1$, et de façon générale, $A_1 \oplus \cdots A_{k-1} \oplus X \oplus A_{k+1} \oplus \cdots A_N = A_k$.
> C'est-à-dire que n'importe quel bloc de données A_k perdu à cause d'un disque défaillant sur un RAID 5 de $N+1$ disques peut-être récupéré grâce au bloc X de données de contrôle.
>
> On voit donc que si on veut écrire dans un bloc, il faut lire le bloc à modifier. Lire le bloc de parité de la bande. Écrire le bloc de données et le bloc de parité. L'opération xor permet heureusement de calculer la nouvelle parité sans avoir besoin de lire les N blocs de données de la bande. Augmenter le nombre de disque d'une grappe RAID 5 n'allonge donc pas le temps de lecture ou d'écriture. Cependant si plusieurs processus veulent écrire simultanément dans un ou plusieurs blocs de données d'une même bande la mise à jour du bloc de parité devient un point de blocage. Les processus concurrents sont suspendus à la libération du bloc de parité et de fait cela limite le débit d'écriture. Plus le nombre de disque d'une grappe RAID 5 augmente plus le temps de reconstruction d'un disque défaillant augmente. Puisque pour reconstituer le bloc manquant d'une bande il faut lire tous les autres blocs de la bande et donc tous les autres disques.

Ce système nécessite impérativement un minimum de trois disques durs. Ceux-ci doivent généralement être de même taille, mais un grand nombre de cartes RAID modernes autorisent des disques de tailles différentes.
La capacité de stockage utile réelle, pour un système de X disques de capacité c identiques est de $(X-1) \times c$. En cas d'utilisation de disques de capacités différentes, le système utilisera dans la formule précédente la capacité minimale.
Ainsi par exemple, trois disques de 100 Go en RAID 5 offrent 200 Go utiles ; dix disques, 900 Go utiles.
Ce système allie sécurité (grâce à la parité) et bonne disponibilité (grâce à la répartition de la parité), même en cas de défaillance d'un des périphériques de stockage.
Il existe une variante : le « RAID 5 varientable » où chaque disque a son propre contrôle. Toutes les autres fonctionnalités sont identiques.
On a souvent tendance à croire qu'un système RAID 5 est totalement fiable. Il est en effet généralement admis que la probabilité de défaillance simultanée de plusieurs disques est extrêmement faible — on parle évidemment d'une défaillance entraînant la perte de données définitive sur plusieurs disques et non d'une simple indisponibilité de plusieurs disques. Cela est vrai pour une défaillance générale d'une unité de disque. Cependant, cela est faux si l'on considère comme "défaillance" un seul secteur devenu illisible.
En effet, dans la pratique, il est très rare que toutes les données d'un volume soient lues régulièrement. Et quand bien même ce serait le cas, la cohérence de la parité n'est que très rarement vérifiée pour des raisons de performances. Il est donc probable que des défauts tels que des secteurs de parité illisibles ne soient pas détectés pendant une très longue période. Lorsque l'un des disques devient réellement défectueux, la reconstruction nécessite de parcourir l'intégralité des disques restants. On peut alors découvrir des défauts qui étaient restés invisibles jusque-là.
Tout ceci pourrait ne pas être bien grave et occasionner la perte d'une quantité de données minime (un secteur de disque), cependant, l'extrême majorité des contrôleurs RAID est incapable de gérer les défaillances partielles : ils considèrent généralement qu'un disque contenant un secteur illisible est totalement défaillant. À ce moment-là, 2 disques sont considérés défaillants simultanément et le volume RAID 5 devient inutilisable. Il devient extrêmement difficile de récupérer les données, et extrêmement coûteux.
Un système RAID 5 doit donc être vérifié et sauvegardé très périodiquement pour s'assurer que l'on ne risque pas de tomber sur ce genre de cas. D'autre part, en cas de défaillance, il est nécessaire de disposer de matériel très coûteux pour espérer récupérer les données, ce qui rend le RAID 5 très peu recommandable aux particuliers et aux petites entreprises.

- Avantages :

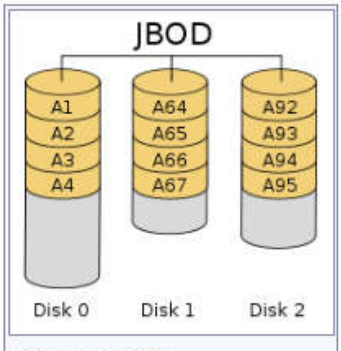
  performances en lecture aussi élevées qu'en RAID 0 et sécurité accrue
  surcoût minimal (capacité totale de $n-1$ disques sur un total de n disques)

- Inconvénients :

  pénalité en écriture du fait du calcul de la parité
  minimum de 3 disques

NRAID (ou JBOD - Just a Bunch Of Disks) : concaténation de disques

NRAID : Near/Non Redundant Array of Inexpensive/Independent Disk
La concaténation de disques consiste à additionner les capacités de plusieurs disques durs en un volume logique d'une taille équivalente à la somme des tailles des disques durs. Cette méthode utilise une méthode d'écriture séquentielle : les données ne sont écrites sur le disque dur suivant que lorsqu'il ne reste plus de place sur le précédent.
Le NRAID n'est pas à proprement parler un RAID, et il ne permet d'ailleurs aucune redondance de données. Sa tolérance aux pannes est celle de disques utilisés isolément : on ne perd que ce qui est sur le disque défectueux. On le rencontre souvent sous le nom de JBOD (*Just a Bunch Of Disks*).
Le NRAID est aussi représenté comme "Volume Simple" sous Windows 2000, XP, 2003, Vista, 2008 et 7[5].



3 disques en JBOD

Les niveaux de RAID combinés

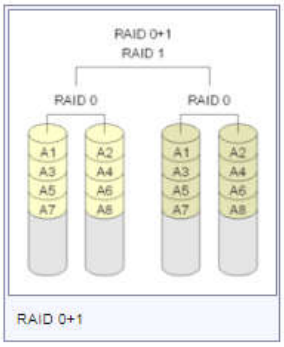
Fondamentalement, un niveau de RAID combiné est l'utilisation d'un concept de RAID classique sur des éléments constitutifs qui sont eux-mêmes le résultat d'un concept RAID classique. Le concept utilisé peut être le même ou différent.
La syntaxe est encore un peu floue mais on peut généralement considérer que le premier chiffre indique le niveau de raid des "grappes" et que le second indique le niveau de raid global. Dans l'absolu rien n'empêche d'imaginer des RAID combinés à 3 étages ou plus mais cela reste pour l'instant plus du domaine de la théorie et de l'expérimentation.
Le nombre important (et croissant) de permutations possibles fait qu'il existe une multitude de raid combinés et nous n'en ferons pas l'inventaire. Nous pouvons cependant présenter les avantages et les faiblesses des plus courants.
Pour les calculs suivants, on utilise les variables suivantes :

- G : nombre de grappes ;
- N : nombre de disques par grappe;
- C : capacité d'un disque (tous les disques sont supposés identiques) ;
- V : vitesse d'un disque.

Les seuils de mise en défaut indiqués ci-dessous indiquent le nombre minimal de disques en panne pouvant entraîner une mise en défaut de l'ensemble du RAID (ie. en dessous de ce nombre de disques en panne le RAID ne peut pas être en défaut). En pratique il est possible qu'un RAID ayant plus que ce nombre de disques en panne fonctionne toujours mais il est recommandé de changer les disques défectueux le plus rapidement possible.

le RAID 01 (ou RAID 0+1)

Il permet d'obtenir du *mirroring* rapide puisqu'il est basé sur des grappes en striping. Chaque grappe contenant au minimum 2 éléments, et un minimum de 2 grappes étant nécessaire, il faut au minimum 4 unités de stockage pour créer un volume RAID0+1.
La fiabilité est moyenne car un disque défectueux entraîne le défaut de toute la grappe qui le contient. Par ailleurs, cela allonge beaucoup le temps de reconstruction et dégrade les performances pendant la reconstruction. L'intérêt principal est que dans le cas d'un miroir à 3 grappes ou plus, le retrait volontaire d'une grappe entière permet d'avoir une sauvegarde "instantanée" sans perdre la redondance.

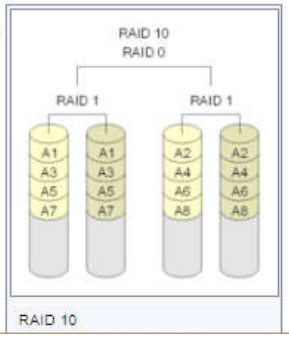


RAID 0+1

Capacité totale : $C_t = G \times C$
Vitesse maximale : $V_m = G \times V$
Seuil de mise en défaut : G disques

RAID 10 (ou RAID 1+0)

Il permet d'obtenir un volume agrégé par bande fiable (puisqu'il est basé sur des grappes répliquées). Chaque grappe contenant au minimum 2 éléments et un minimum de 2 grappes étant nécessaire, il faut au minimum 4 unités de stockage pour créer un volume RAID10.
Sa fiabilité est assez grande puisqu'il faut que tous les éléments d'une grappe soient défectueux pour entraîner un défaut global. La reconstruction est assez performante puisqu'elle ne mobilise que les disques d'une seule grappe et non la totalité.



RAID 10

Capacité totale : $C_t = G \times C$
Vitesse maximale : $V_m = G \times V$
Seuil de mise en défaut : N disques

RAID 05