Architecture

des systèmes

R.A.I.D.

Julien VEHENT

Master Sécurité des Systèmes Industriels et des Systèmes d'Informations

1ére année - décembre 20005

Introduction

Ce dossier a pour objectif de présenter succintement les différents types d'architectures RAID existants et leurs avantages/inconvénients. Nous commencerons tout d'abord par un bref rappel de ce qu'est le RAID, des différents niveaux de RAID et des différents types de connectiques. Nous verrons ensuite plusieurs contrôleurs qui répondent à des besoins différents.

Redundant Array of Independent /Inexpensive Disks

Le terme de RAID (*Redundant Array of Independent/Inexpensive Disks, c'est-à-dire un groupe de disques redondants et indépendants/bon marché*) désigne une architecture matérielle (et parfois logicielle) permettant d'accélérer, de sécuriser et/ou de fiabiliser les accès aux données stockées sur disques durs. Cette architecture est basée sur la multiplication des disques durs, par opposition à la méthode SLED (*Single Large Expensive Disk*) où toutes les données sont rassemblées sur un seul disque de prix élevé.

La première description de cette architecture date d'une publication de 1987, dans une publication de Patterson, Gibson & Katz (*3 chercheurs de l'Université de Berkeley*). Cet article comparait le RAID au sled et proposait cinq niveaux différents de RAID, chacun d'eux ayant ses avantages et ses inconvénients.

Les différents types d'architectures RAID sont numérotés à partir de 0 et peuvent se combiner entre eux (*on parlera alors de RAID 0+1, 1+0, etc...*).

RAID 0: Dans cette configuration, chaque octet est divisé en autant de morceaux (*généralement de 32Ko*) qu'il y a de disques.

Exemple : avec un RAID 0 composé de quatre disques, si l'on veut écrire l'octet 00111001, le stockage des différents bits composant cet octet se fera de la façon suivante :

- disque 1 : 00
- disque 2 : 11
- disque 3 : 10
- disque 4 : 01

Ainsi, sur un RAID 0 de *n* disques, chaque disque ne doit lire et écrire que 1/*n* des données, ce qui a pour effet de décupler les taux de transfert des données entre le CPU et les disques, et donc d'accélérer les traitements.

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, ce qui fait du RAID 0 une solution moins fiable que l'utilisation d'un seul disque de stockage, puisque la probabilité de défaillance d'un des disques du RAID est largement supérieure à la probabilité de défaillance d'un disque unique.

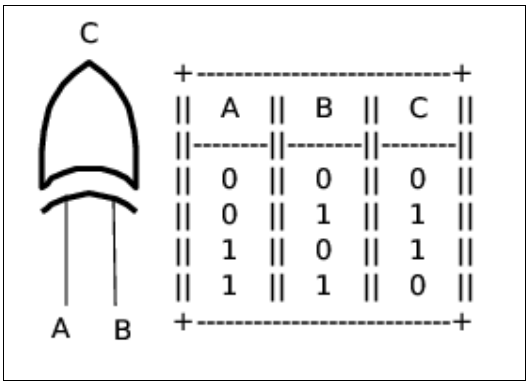
RAID 1: Le RAID 1 consiste en l'utilisation de disques redondants, c'est-à-dire *n* disques (*en général deux*) sur lesquels sont copiées exactement les mêmes données. Si cette solution n'apporte aucun gain de performance, elle permet en revanche de sécuriser les données en cas de défaillance d'un des disques. Il est à noter que dans ce type de RAID, la perte de capacité de stockage liée à l'utilisation des disques, est égale à 50% de la capacité totale des disques utilisés.

RAID 3 : Le RAID 3 nécessite une matrice de trois disques au minimum. *n*-1 disques contiennent les données (*type RAID 0*) et le dernier disque stocke la parité. En effet si un des disques de données tombe en panne, il est possible de reconstruire l'information avec le disque de parité et dans le cas où celui-ci tombe en panne, le système devient alors un RAID 0. Il est important que le disque de parité soit de bonne qualité car il est à tout instant sollicité à l'écriture. Ce dernier point est le limitation du RAID 3.

RAID 4 : Le RAID 4 est sensiblement semblable au RAID 3 sauf sur le point où il travaille par blocs et non par octets ce qui ne nécessite plus de synchronisation entre les disques.

RAID 5 : Le RAID 5 associe le *striping* et un système à *parité répartie*, il permet une bonne disponibilité (même en cas de défaillance d'un des périphériques de stockage.

Exemple pratique : soit trois disques durs de taille identique A, B et C. Le système va enregistrer sur les disques A et B les données (*Strip*) comme en mode RAID 0 et, sur le disque C, le résultat de l'opération *ou-exclusif* entre A et B (*A xor B*).

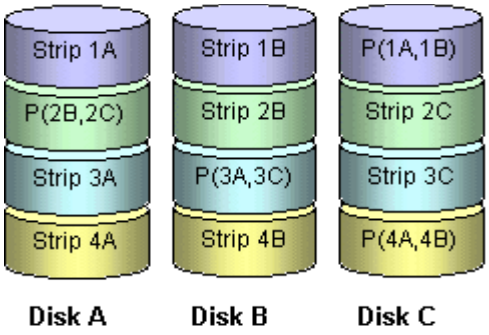


Fonction XOR

Ainsi, en cas de défaillance du disque A, les données qui y étaient accessibles le sont toujours avec les disques B et C, par l'opération B xor C (*B xor C = A*).

Il en va de même pour le disque B. Et si le disque C tombe en panne, les informations sont toutes sur A et B.

Important : les disques doivent être de même taille. On ne stocke que sur les deux tiers de la place totale des disques (*le dernier tiers étant la parité*). .



Répartition des données sur un RAID 5

Les hybrides : Sur ces 5 types de RAID, on peut jouer avec les configurations pour construire des architectures de disques rapides et sûres. Les plus communéments utilisées sont:

- RAID 0+1 : Deux groupes de *n* disques en RAID 0, ces deux groupes sont eux en RAID 1.

- RAID 10 (1+0) : *n* groupes de deux disques en RAID 1, tous ceux-ci en RAID 0. Il faut que deux disques d'un même groupe rendent l'âme pour que le tout soit perdu, ce qui réduit la probabilité.

- RAID 50 (5+0) : n groupes de trois disques en RAID 5, tous ceux-ci en RAID 0. Il faut que deux disques d'un même groupe rendent l'âme pour que le tout soit perdu, ce qui réduit encore la probabilité. Un des meilleurs compromis lorsque l'on cherche la rapidité !

Implémentation

Il existe trois grandes familles de RAID: logiciels, matériels internes et matériels externes.

Les RAID logiciels font appels au système d'exploitation pour découper et disperser les données sur les disques. C'est donc le CPU qui travaille, ce qui pose des problèmes de performances sur les systèmes exigeants.

Les RAID matériels internes sont implémentés par des contrôleurs, typiquement des cartes filles connectées en PCI. Ainsi les contrôleurs travaillent indépendamment de la charge CPU et de façon constante. Les données relatives à la configuration RAID se trouvent sur tous les disques durs du système RAID et sont ainsi protégées même si l'un des disques ou le contrôleur venait à tomber en panne. Lors de l'échange du contrôleur ou d'un disque les données de configuration RAID sont reconstruites grâce aux informations stockées sur les disques.

RAID Externe – 4*300Go et Controleur raid 0,1,5

Les RAID matériels externes (*souvent appelés Network Attached Storage [NAS]*) représentent la solution RAID dite de haut de gamme. Ici, les contrôleurs et les disques durs sont logés dans un boîtier externe autonome par rapport à l'ordinateur. La connexion est assurée par un câble SCSI ou Fibre Channel.

L'utilisation d'un contrôleur RAID redondant augmente le niveau de sécurité. Lorsque le contrôleur primaire tombe en panne, l'ensemble RAID commute automatiquement sur le second contrôleur, sans perte de temps et de données. Le contrôleur défectueux peut alors être remplacé en cours de fonctionnement.

Les RAID matériels externes travaillent, comme le contrôleur RAID PCI, indépendamment de la charge CPU. Les données de l'adaptateur SCSI sont directement transmises au contrôleur RAID via le bus SCSI. La totalité des données est alors chargée dans le cache afin de libérer au maximum le bus. Le contrôleur du RAID matériel répartit les données sur les différents disques et calcule la parité en fonction du niveau RAID choisi. Un cache peut contribuer à augmenter considérablement la performance en écriture. La lecture des données se fait de façon identique dans le sens inverse. Comparés aux contrôleurs PCI RAID, les matériels RAID peuvent être installés indépendament d'un quelconque serveur.

Connectiques

Le type de connectiques des périphériques d'un système RAID à un impact immédiat sur les performances de celui-ci. Présentons succintement trois d'entre elles.

La connectique IDE, acronyme de Integrated Drive Electronics, est la plus répandue au sein des ordinateurs personnels. Son utilisation dans les systèmes RAID est rare, hormis dans le cas d'un RAID logiciel que l'on peut retrouver dans des stations personnelles de calculs (*imagerie numérique,*

animations de synthèse, etc...).

La plus implémentée des connectiques RAID est sans contexte SCSI (*Small computer System Interface*). Ce standard définit un bus permettant de relier un ordinateur à des périphériques ou bien même à un autre ordinateur.

Ce qui différencie ce bus des autres est qu'il déporte l'intelligence vers le périphérique lui-même. De ce fait les commandes envoyées au périphérique peuvent être complexes, c'est sous le contrôle du périphérique qu'elles seront décomposées en sous-tâches plus simples, ce qui est extrêmement avantageux si l'on travaille avec des systèmes d'exploitations multi-tâche.

Cette interface est donc plus rapide, plus universelle et plus complexe que l'interface IDE dont le principal inconvénient est d'accaparer un pourcentage non négligeable du processeur,ce qui constitue un handicap quand de nombreux flux de données sont simultanément ouverts.

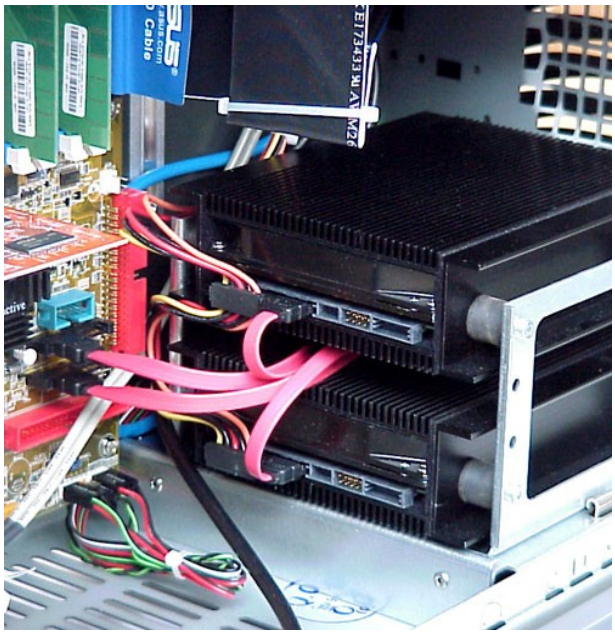
Norme	Interface	Vitesse Bus (MO/s)	Fréq. de bus (Mhz)	Taille Bus (bits)
SCSI-1	SCSI	5	5	8
SCSI-2	Fast SCSI	10	10	8
	Fast Wide SCSI	20	10	16
SCSI-3	Ultra SCSI	20	20	8
	Ultra Wide SCSI	40	20	16
	Ultra2 SCSI	40	40	8
	Ultra2 Wide SCSI	80	40	16
	Ultra3 SCSI	80	80	16
	Ultra-160 SCSI	160	80	16
	Ultra-320 SCSI	320	160	16
	Ultra-640 SCSI	640	160	16

Les bus SCSI

Le Serial ATA est un bus principalement conçu pour le transfert de données entre le CPU et un disque dur (*a la différence de SCSI et IDE qui sont plus généraux*). C'est l'évolution du standard Advanced Technology Attachment (*ATA ou IDE*). Le S-ATA a de multiples avantages par rapport à l'IDE, les trois principaux étant sa vitesse, la gestion des câbles et le Hot-Plug.

Les premiers modèles de S-ATA, apparus en 2003 permettent un débit théorique de 150 Mo/s mais il a été conçu pour aller bien plus vite. Le S-ATA 2 double sa vitesse à 300 Mo/s, puis les 600 Mo/s sont prévus pour 2007, rattrapant ainsi les 640mo/s du Ultra-640 SCSI,

Physiquement les câbles utilisés sont le plus grand changement du S-ATA. Les données sont transmises par un fil flexible de sept conducteurs avec des connecteurs de 8 mm à chaque extrémité. Le Sata utilise l'encodage 8b/10b pour effectuer des transferts (*technique d'encodage permettant une haute vitesse de transmission*),



connectique S-ATA

Ces connectiques répondent à des besoins différents. L'IDE pour la bureautique, le S-ATA pour des RAID peu coûteux (*pour le moment*) et le SCSI pour les systèmes qui ont de gros besoins en bande passante. Mais à l'avenir, il y a de fortes chances pour que le SATA remplissent ces trois cas de figures à lui seul...

Contrôleurs

La multiplicité des normes oblige a restreindre le champ d'étude à quelques produits significatifs. Nous verrons donc le contrôleurs Adaptec SCSI RAID 2200S, le contrôleur LSI Logic MegaRAID S-ATA 300-8X et enfin le contrôleurs RAID logiciel du noyau Linux.

Adaptec SCSI RAID 2200S

Adaptec est un des leader mondiaux des contrôleurs RAID. Ce modèle utilise une interface SCSI-3 Ultra-320 et gére les niveaux de RAID 0,1,5,10,50 et JBOD (*addition de disques disparates sans tolérance de pannes*). La Strip-size (*taille des blocs de données*) est configurable de 16Ko à 32Ko. Il dispose de 2 connecteur internes et 2 externes et d'une mémoire cache de 64Mo. Enfin, avec un coût d'achat d'environ 850€ HT, on peut dire de ce contrôleur qu'il est tout à fait représentatif de ceux utilisés aux seins des serveurs de fichiers.

Contrôleur Adaptec SCSI RAID 2200S

Fonctionnalités avancées:

- Batterie de secours pour maintenir le cache en cas de coupure de courant.
- Strip-Size variable: possibilité de modifier la Strip-Size aprés la création du RAID et sans destruction de la grappe.

Performances:

- RAID0: lecture=110,1 Mo/s|| écriture=62,6 Mo/s
- RAID5: lecture=83,6 Mo/s || écriture=56,8 Mo/s
- Disques Durs de tests

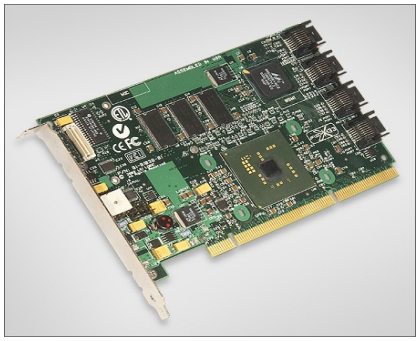
Maxtor Atlas U320 - 36 Go - 15000 tr/min

Administration:

- Interface web sécurisée par SSL
- Interface ligne de commande

LSI Logic MegaRAID SATA 300-8X

LSI Logic est également un important constructeur de contrôleurs RAID. Ce modèle-ci représente la derniére génération de ce que permet d'atteindre la technologie S-ATA.



LSI Logic MegaRAID SATA 300-8X

La carte est équipé de 8ports SATA2 300 internes qui permettent un débit de 3Gb/s par port, soit 384Mo/s théoriques. Les niveaux supportés sont les mêmes que ceux de l'Adaptec, à l'exception du mode JBOD (*à l'utilité plus que discutable*). Une des principales caractéristiques de ce contrôleurs est qu'il s'appuie sur la nouvelle norme de bus PCI: le PCI Express. Il représente également un investisement plus faible. Du fait de la démocratisation des interfaces SATA, une carte comme celle-ci se négocie aux environ des 450€ TTC.

Fonctionnalités avancées:

- Batterie de secours pour maintenir le cache en

cas de coupure de courant.

- Annonce acoustique des erreurs critiques.
- Migration RAID 0 vers RAID 10 et RAID 5 vers RAID 50

Performances:

- RAID0: lecture=121,1 Mo/s|| écriture=58,9 Mo/s
- RAID5: lecture=77,5 Mo/s || écriture=49,4 Mo/s
- Disques Durs de tests

Western D. Raptor - 74 Go - 10000 tr/min

Administration:

- Ligne de commande uniquement

Raid logiciel du noyau Linux.

Le noyau Linux intégre le système RAID logiciel le plus performant du marché. Il gére les niveaux 0, 1, 4, 5, 6 et 10.

```
[*] Multiple devices driver support (RAID and LVM)
<*>   RAID support
< >     Linear (append) mode
<*>     RAID-0 (striping) mode
<*>     RAID-1 (mirroring) mode
< >     RAID-10 (mirrored striping) mode (EXPERIMENTAL)
< >     RAID-4/RAID-5 mode
< >     RAID-6 mode
< >     Multipath I/O support
< >     Faulty test module for MD
<*>   Device mapper support
<*>     Crypt target support
< >     Snapshot target (EXPERIMENTAL)
< >     Mirror target (EXPERIMENTAL)
< >     Zero target (EXPERIMENTAL)
< >     Multipath target (EXPERIMENTAL)
```

Configuration du noyau 2.6.14-rc3: section Device Drivers

L'utilisation du RAID logiciel apporte certain avantage. Hormis la faiblesse du coût de la solution, on notera que tous les types de connections de disques durs reconnues par le noyau peuvent être utilisées dans la grappe, même si l'homogénéité est fortement recommandée.

Le RAID Logiciel se destine toutefois à des utilisations non critiques. Il est, en effet, impossible d'assurer le même niveau de sécurité et de disponibilité qu'avec un (*ou deux*) contrôleurs RAID matériels. C'est pourquoi ce système n'est jamais utilisé sur des serveurs de fichiers. On le retrouve, par contre, très souvent sur des serveurs à faible coût, comme au sein des PME.

Performances:

- RAID0: lecture=42,3 Mo/s || écriture=30,8 Mo/s

*avec 2*Seagate 120Go SATA1 7200tr/min*

- RAID1: lecture=51,8 Mo/s || écriture=26,4 Mo/s

*avec 2*Western Digital 40Go IDE 7200tr/min*

Administration:

- mdadm: création, maintenance et monitoring des RAID logiciel Linux en commandes shell.

Conclusion

Les systèmes RAID font, depuis une quinzaine d'années, partie intégrante des serveurs professionnels. Ce que l'on peut retenir de cette étude est que la multiplicitée des solutions permet d'utiliser le RAID dans un système à faible coût comme dans une structure de plusieurs dizaines de tera-octets.

La compréhension (*voir la maîtrise*) de cette technologie est indispensable au déploiement de n'importe quelle solution de stockage. Du serveur mail privé à la baie de stockage SAN , tous ces systèmes utilisent, et utiliseront, du RAID pour de nombreuses années encore...

Storage Area Network
Sun Microsystem