Principes de base de la mise en réseau et du stockage du centre de données

Chapitre 2 - Partie 1 :

Réseaux de stockage : conceptions existantes

Sommaire

- Introduction
- Architectures
- Les disques
- Performances d’un disque
- RAID
- Contrôleurs de stockage
- Logical Unit Numbers
- Gestionnaire de volumes logiques

Introduction

- Aujourd'hui, les centres de données (**Data Center**) sont **essentiels** et font partie intégrante de toute **entreprise**.
- Les **éléments centraux** d'un centre de données sont :
  - Le serveur
  - Le stockage
  - La connectivité (ou réseau)
  - Les applications
  - Et le SGBD

- Ces éléments fonctionnent ensemble pour **traiter et stocker des données**.
- Evolution avec le **VDC** (virtualized data center) :
  - **ressources physiques** sont **regroupées** et **fournies** en tant que **ressources virtuelles**.
  - **créées** à l'aide d'un **logiciel** qui permet un **déploiement** plus **rapide**.
  - **optimiser** l'utilisation de leur infrastructure et **réduire** le **coût** total de **possession**.

- Avec **l'augmentation** de la **criticité** des **actifs informationnels** pour les **entreprises :**
  - le **stockage**, l'un des **éléments essentiels** d'un **centre de données**,
  - **reconnu** comme une **ressource distincte**,
  - le **stockage nécessite** une **attention particulière** pour sa **mise** en **œuvre** et sa **gestion**.

Exemple des architectures hyperconvergées (Hyperconvergence)

Architectures

- Le **réseau de données** disparaît et se transforme en une simple **boîte noire complexe,** composée de :
  - les serveurs,
  - les contrôleurs de stockage,
  - les systèmes de fichiers,
  - les disques
  - et la sauvegarde.
- Les **technologies réseau de stockage utilisent** les **technologies de réseau de données** avec **l'adoption** des réseaux Internet Protocol (**IP**).
- Du point de vue du **réseau de stockage**, les **trois niveaux** sont :
  - les serveurs,
  - le réseau de stockage,
  - et les disques.
- Chaque niveau a sa part de **problèmes d'interopérabilité** et de **compatibilité** :
  - les fournisseurs de **serveurs** : **compatibilité** du **système d'exploitation** avec les **interfaces réseau** et les **contrôleurs** de **disque**,
  - les fournisseurs de **disques** ont **amélioré** la **capacité** et la **vitesse** des disques,
  - et les fournisseurs de **réseaux de stockage** ont dû trouver comment **gérer** les **multiples protocoles** de **stockage** :
    - Fibre Channel (**FC**),
    - Network File System (**NFS**),
    - et Internet Small Computer System Interface (**iSCSI**)

Architectures

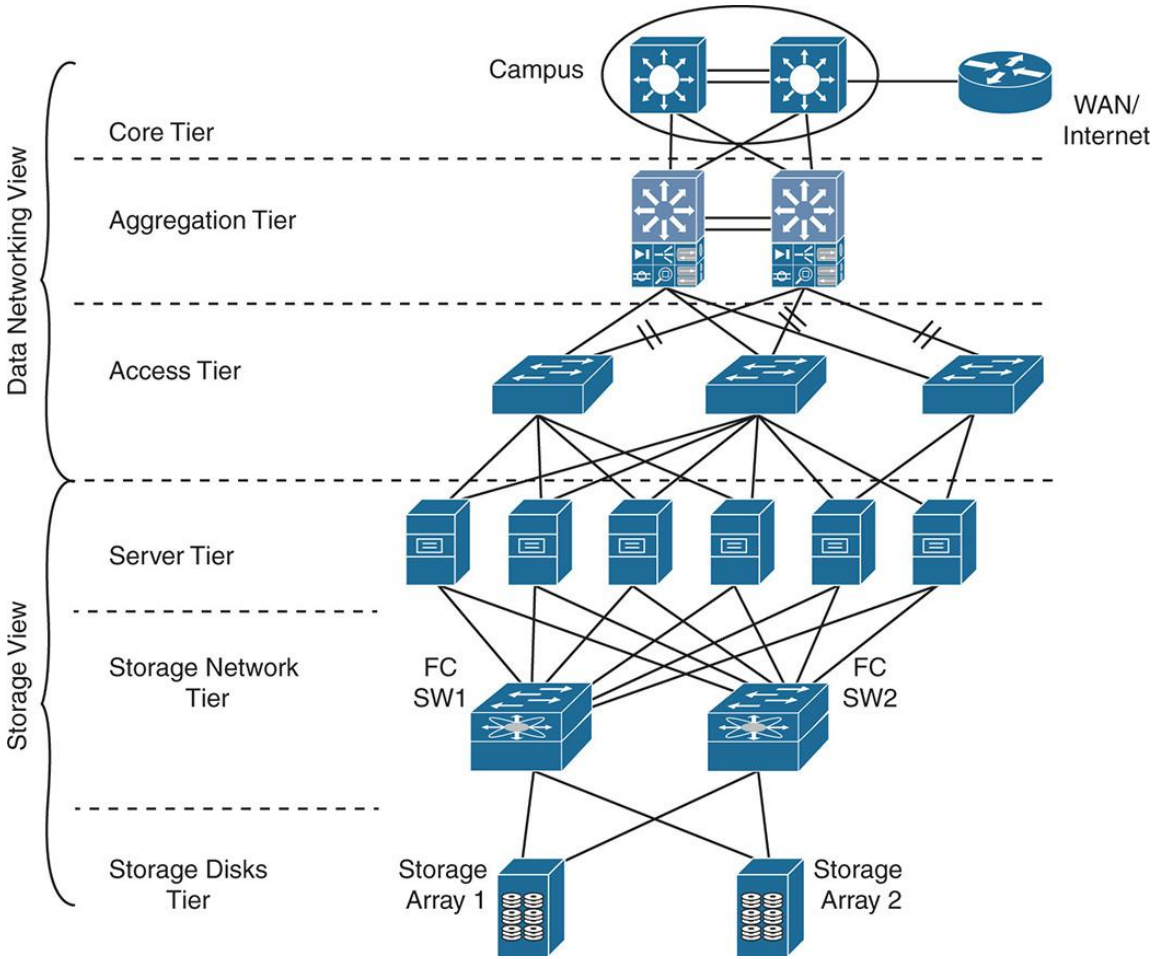

Figure 2-1 Architecture de stockage multiniveau

Architectures : Les disques

- **Disques SATA** :
  - conviennent aux **applications** qui nécessitent **moins** de **performances** de disque et moins de **charge** de travail.
  - **Vitesse** du plateau SATA est de **7200 RPM**.
  - Utilisent la **commande** Advanced Technology Attachment (**ATA**) pour le transfert de données

- **Disques SAS**:
  - conviennent aux **applications** qui nécessitent des **performances** de disque **plus élevées** et une **charge** de travail plus **élevée**.
  - **Vitesse** du plateau SAS est de **15 000 tr/min** maximum.
  - Utilisent le jeu de **commandes SCSI**.
  - Les **disques SAS** et les **disques SATA** peuvent se **connecter** à un **fond de panier SAS** ; **SAS** ne **peut pas se brancher sur** un **fond de panier SATA**.
  - Les disques SAS **SED**, **permettent** le **chiffrement** et le **déchiffrement automatiques** des données par le contrôleur de disque et **n'affecte pas** les **performances**.

- **Disque FC** :
  - conviennent aux **applications de serveur d'entreprise**.
  - **performances** similaires aux disques **SAS**
  - **FC** est une **méthode de connectivité** pour les réseaux de stockage (**SAN**) et **transporte** le jeu de commandes **SCSI** sur le **réseau**.
  - La **connectivité SAN** est **dominée** par **Fibre Channel**, mais les **disques dur**s réels des **baies SAN** sont des **SAS**.

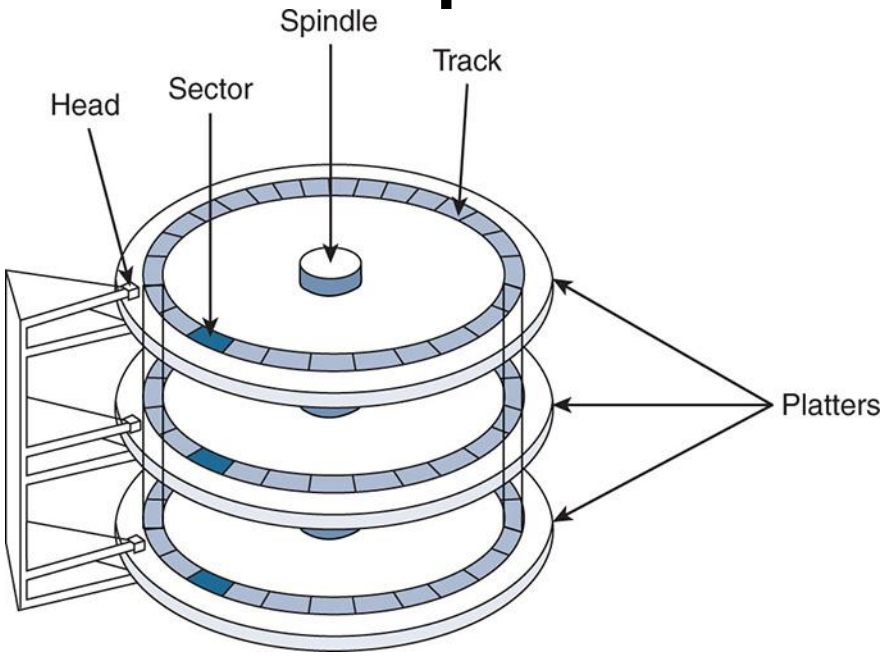

Figure 2-2 Pistes et secteurs du disque dur

- **Disques SSD**:
  - pas de moteurs, de têtes de lecture ou de plateaux tournants, **construits sans pièces mobiles**.
  - Beaucoup **plus chers** que les **disques durs** et offrent des **performances** bien **supérieures**, car **reposent** sur un **processeur intégré** pour stocker, récupérer et effacer les données.
  - **Moins** de **capacités** que les disques durs.
  - se **connectent** aux serveurs via une **interface SATA** ou **SAS**.

Architectures : Performances du disque

- Les **performances** du disque pour les **lectures** et les **écritures dépendent** de facteurs :
  - support de disque (HDD ou SSD)
  - file d'attente (disk queue)
  - interconnexion entre le disque et le serveur
- Cette **section traite** des **facteurs** qui **affectent** les **performances** du disque et de la **terminologie** utilisée dans l'industrie.
- **Examinons** ensuite les **vitesses de transfert** en termes de débit, de **latence** et d'opérations d'entrée/sortie par seconde (**IOPS**).

Débit ou vitesse de transfert

- la **vitesse** à laquelle les **données** sont **transférées vers et depuis** le support de **disque** dans une certaine **période de temps**.

Temps d'accès

- **Mesure** globale, normalement en **millisecondes**, du **temps** nécessaire pour **démarrer** l'**opération de transfert de données**.
- **HDD** sont de l'ordre de **10 millisecondes** ou moins.
- **SDD** sont de l'ordre de **100 microsecondes**.

iNTERFACE	vitesse de transfert
SATA 1	1,5 Gbit/s
SATA 2	3 Gbit/s
SATA 3	6 Gbit/s
SAS 1	3 Gbps
SAS 2	6 Gbps
SAS 3	12 Gbps
SAS 4	22,5 Gbps

Architectures : Performances du disque

Latence et IOPS (Input/output operations per second)

- Les **IOPS** ont en général une **relation inverse** avec la **latence** du point de vue de l'application :
  - une **latence** globale de **0,01 ms**, les **IOPS** du disque doivent être comprises entre **1/0,01 = 100 000 IOPS**.
- Les disques **SSD** fournissent beaucoup **plus d'IOPS (**atteignant même un **million d'IOPS)** que les disques durs **SATA** ou **SAS**
- Autres **facteurs affectant les performances** :
  - La vitesses de l'interface.
  - Le type des données : séquentielles ou aléatoires.
  - Le type d'opération : lecture ou écriture.
  - Le nombre de disques
  - Le niveau de tolérance aux pannes dans les baies de disques
  - La façon dont les données sont lues ou écrites sur le disquela façon dont les données sont lues ou écrites sur le disque :
    - Si vous li**sez** ou **écrivez** de **gros fichiers** dans des **blocs de grande taille** de manière **séquentielle**, il n'y a **pas beaucoup** de **mouvement de la tête** de disque, donc les **temps d'accès** sont **plus courts**.
    - Si les **données** sont **stockées** sur le disque **par petits blocs** et de manière **aléatoire**, les **temps d'accès** en lecture et en écriture sont beaucoup **plus longs**. Dans les **modèles aléatoires**, il y a **beaucoup** de **recherches** de **disque** avec des disques durs qui **réduisent** les **performances**.
  - Les **performances** des disques **dépendent également** de **l'interface** du disque, telle que **SATA** et **SAS :**
    - **Exemple** : un disque **SSD gérant 200 000 IOPS** pour des blocs de données **de 4 Ko** (comme dans 4 Ko) **transfère** en fait 200 000 × 4 Ko = 800 000 Ko/s, soit **800 Mo/s** or la **vitesse** de **transfert théorique** d'une interface **SATA 3** à 6 Gbps, soit **750 Mbps**.

    => L'interface devient un goulot d'étranglement : il est important que des IOPS plus rapides soient associées à des vitesses d'interface plus rapides.

    D'autre part, les **IOPS** sont **liées** au **nombre** de **disques** dans une **matrice redondante** de **système** de disques indépendants (ou peu coûteux) (**RAID**), comme indiqué ci-après.

RAID

- RAID (Redundant Array of Independent Disks) :
  - **Idée** : **plusieurs disques** plus petits et **peu coûteux** ont de **meilleures performances** qu'un **disque volumineux** et **coûteux**.
  - Suppose une **répartition** des **données entre** les multiples **disques**.
  - RAID **offrent** une plus grande **tolérance aux pannes**.
  - **Différents niveaux** RAID en **fonction** de **l'équilibre** requis entre **performances**, **fiabilité** et **disponibilité** : RAID 0, 1, 1+0 et 0+1, 5 et 6.

RAID 0

- Combine tous les disques en un seul disque.
- Amélioration des performances.
- Pas de tolérance aux pannes : si un disque tombe en panne, toutes les données sont perdues

RAID 1

- Mise en miroir de disque
- Performance : les vitesses de lecture/écriture sont très bonnes.
- Tolérance aux pannes : si un disque tombe en panne, il existe une copie miroir prête de ce disque.
- Inconvénient du RAID 1 est que la moitié de la capacité est perdue.

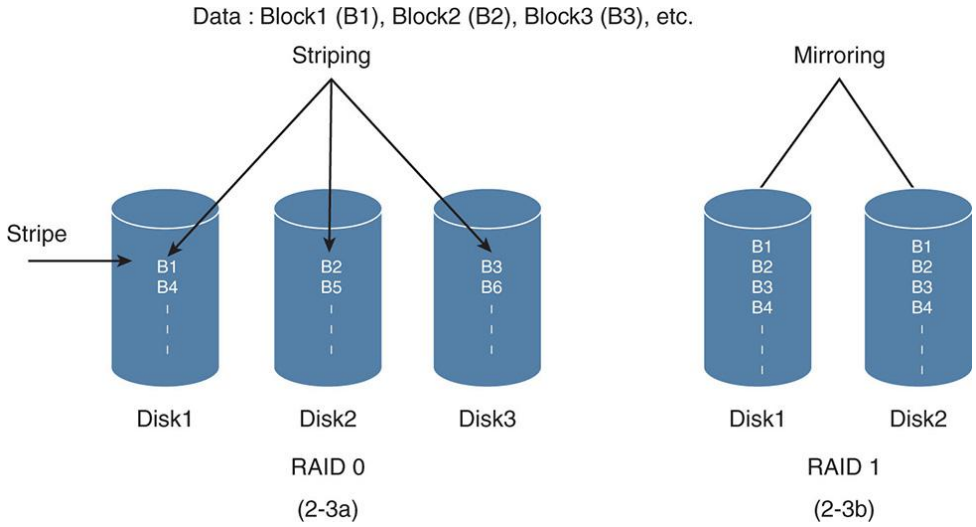

Figure 2-4 Exemples RAID 1+0 (gauche) et RAID 0+1 (droite)

RAID

RAID 1 + 0

- Combinaison d'un RAID 0 et d'un RAID 1
- Nécessite un minimum de quatre disques
- Bonnes performances en lecture/écriture
- Capacité de protection contre les pannes de disque : si un ensemble de disques en miroir échoue, les données sont par être perdues.
- La moitié de la capacité de la baie est perdue.

RAID 0 + 1

- Combinaison d'un RAID 1 et d'un RAID 0
- Nécessite un minimum de quatre disques
- Bonnes performances en lecture/écriture
- Capacité de protection contre les pannes de disque : **offre une meilleure protection que RAID 1+0**.
- La moitié de la capacité de la baie est perdue.

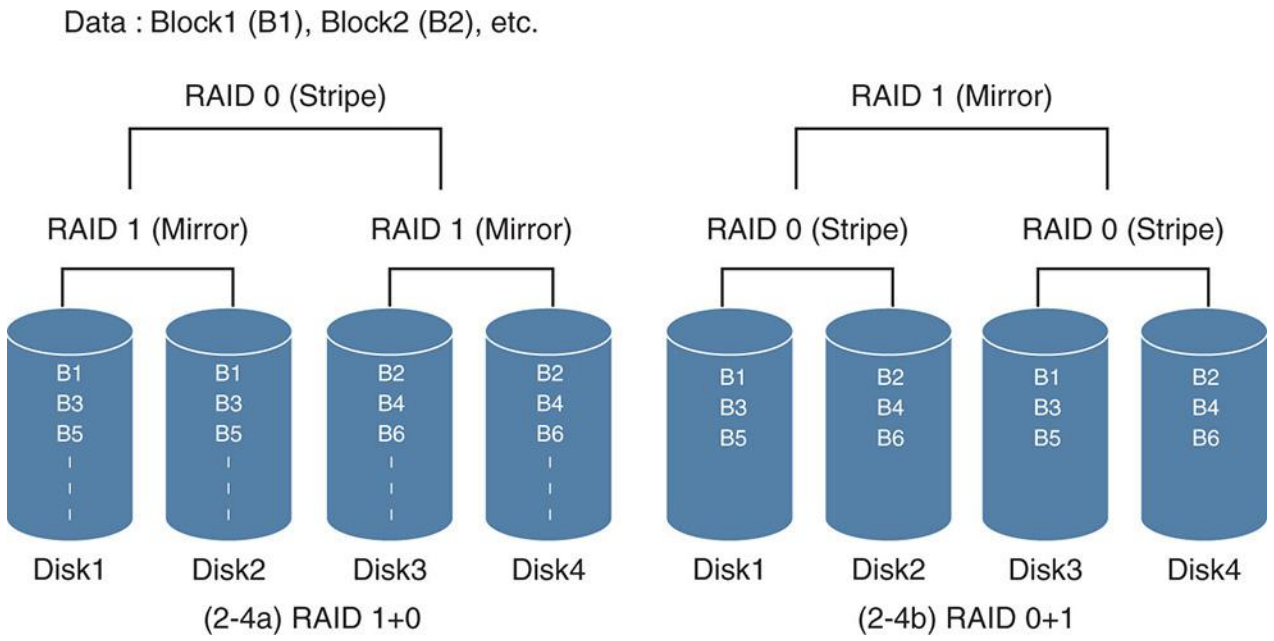

Figure 2-4 Exemples RAID 1+0 (gauche) et RAID 0+1 (droite)

RAID

RAID 5

- Fonctionne avec une répartition au niveau des blocs et une parité distribuée
- Nécessite au moins trois disques
- Les blocs de données sont répartis sur tous les disques de la même couche, à l'exception du dernier disque qui stocke les données de parité
- Les données de parité permettent aux données d'être reconstruites au niveau de la couche
- Les données de parité se déplacent entre les couches
- Bonnes performances en lecture
- Mauvaises performances en écriture
- Bonne tolérance aux pannes
- La capacité d'un disque complet de la matrice est perdue.

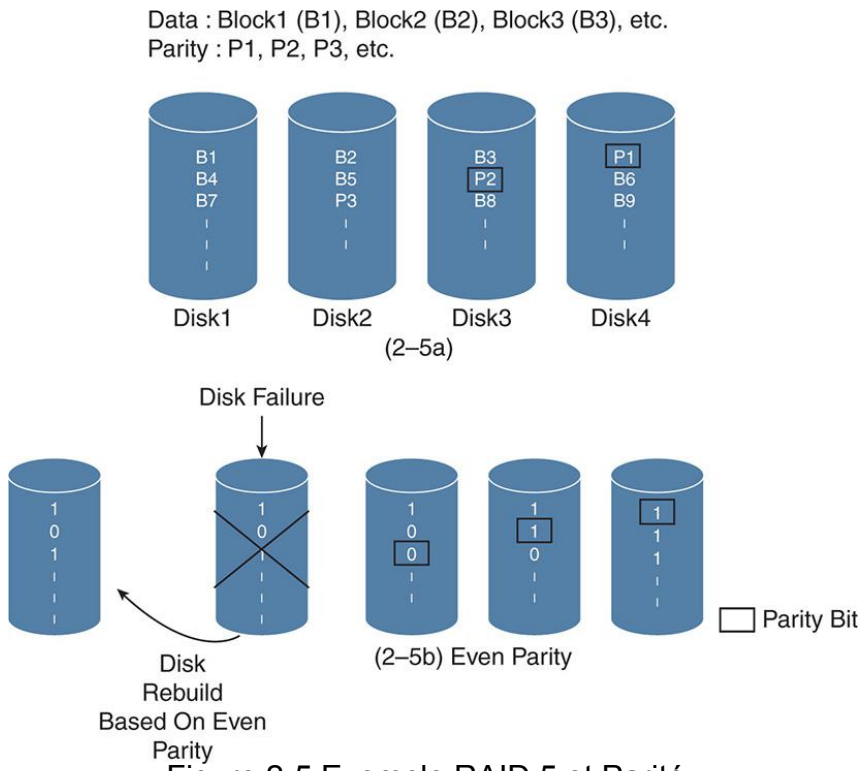

Figure 2-5 Exemple RAID 5 et Parité

RAID

RAID 6

- Fonctionne exactement comme RAID 5, mêmes avantages et inconvénients
- Supporte deux pannes de disque au lieu d'une panne de disque comme dans RAID 5
- Les informations de parité sont écrites sur deux disques de la matrice à chaque couche
- On perd la capacité totale de deux disques de la matrice

- Discutons de la manière dont les IOPS sont affectées lorsqu'il s'agit de plusieurs disques, comme dans les matrices RAID.
- Les systèmes RAID encourent une “pénalité RAID” chaque fois que la parité est lue ou écrite :
  - Pénalité est de 4 pour le RAID 5
  - Pénalité est de 6 pour le RAID 6
- La formule de calcul des IOPS dans un système RAID est la suivante :

```
Raw IOPS = Disk Speed IOPS * Number of Disks
Functional IOPS = (RAW IOPS * Write % / RAID
Penalty) + (RAW IOPS * Read %)
```

- Soit un **disque SSD** avec un **Disk Speed IOPS de 100 000**
- Et **cinq disques** de ce type dans une matrice **RAID 5**
- Supposons que **40 %** des opérations d'E/S sont une **lecture** et **60 %** sont une **écriture**
- On a :

Raw IOPS = 100,000 * 5 = 500,000 IOPS
Functional IOPS = (500,000 * 0.6/4) + (500,000 * 0.4) = 275,000 IOPS

- Notez que les **IOPS du système RAID** sont **presque la moitié** des **IOPS RAW attendues**.
- si **l'application nécessite 500 000 IOPS** de la baie RAID, vous avez **besoin** d'au **moins cinq disques à 200 000 IOPS**

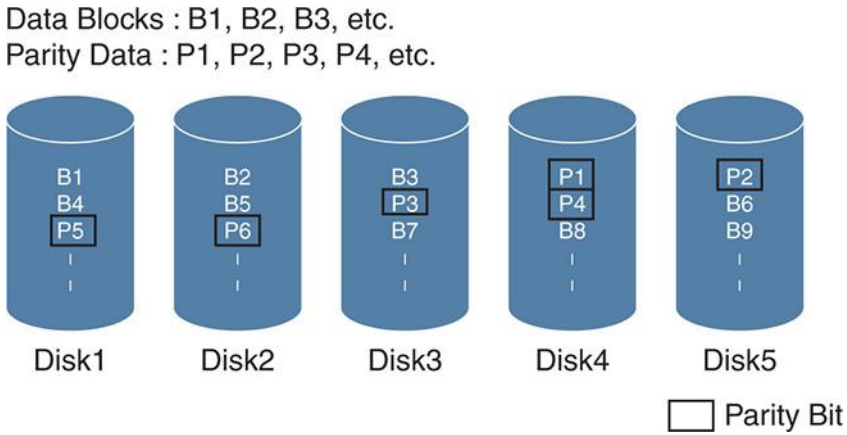

Figure 2-6 Exemple RAID 6, Double Parité

TP

TP : RAID 5

Contrôleurs de stockage

- Les **contrôleurs de stockage** sont **essentiels** pour **exécuter** des **fonctions** de **stockage** :
  - **gestion** du **volume** de **disque**
  - **présentation** des **disques** en tant que différents niveaux de **matrices RAID**
  - prennent un ensemble de **disques physiques** et les **configurent** en tant que **partitions physiques appelées** disques logiques ou **disques virtuels**
  - Le contrôleur prend ensuite les **disques virtuels** et **crée** des numéros d'unité logique (**LUN**) qui sont **présentés** au système d'exploitation (**OS**) en tant que **volumes**
- Distinguons la terminologie :
  - **HBA** (Host Bus Adapters) :
    - est un **adaptateur matériel** qui se **connecte** au **bus PCI** (Peripheral Component Interconnect) du **serveur** ou **PCI Express** (PCIe)
    - Le **HBA abrite** les **ports** de **stockage** tels que **SATA** ou **SAS** ou **FC** pour la **connectivité** des **disques**.

  - Le **contrôleur RAID hôte logiciel** :
    - fait **partie** de la carte mère du serveur (de **l'OS** ?)
    - **exécute** les **fonctions RAID**
  - **Contrôleur RAID** hôte basé sur le **matériel** :
    - plus **performant** que le **RAID logiciel**
    - **adaptateur matériel** qui se **connecte** à la **carte mère** du serveur
    - **décharge** les fonctions de stockage du **processeur** hôte principal
    - Lorsque des **disques** sont **connectés** au **contrôleur**, celui-ci les **détecte** et **présente** au **système d'exploitation** les multiples disques en tant **qu'unités logiques**
  - **Processeurs** de **stockage** :
    - appelés **contrôleurs RAID externes**
    - fournis avec les **baies de disques**
    - L'architecture **SAN repose** principalement sur la **connexion d'un** ou **plusieurs processeurs** de **stockage** à la structure SAN de manière **redondante**

Contrôleurs de stockage

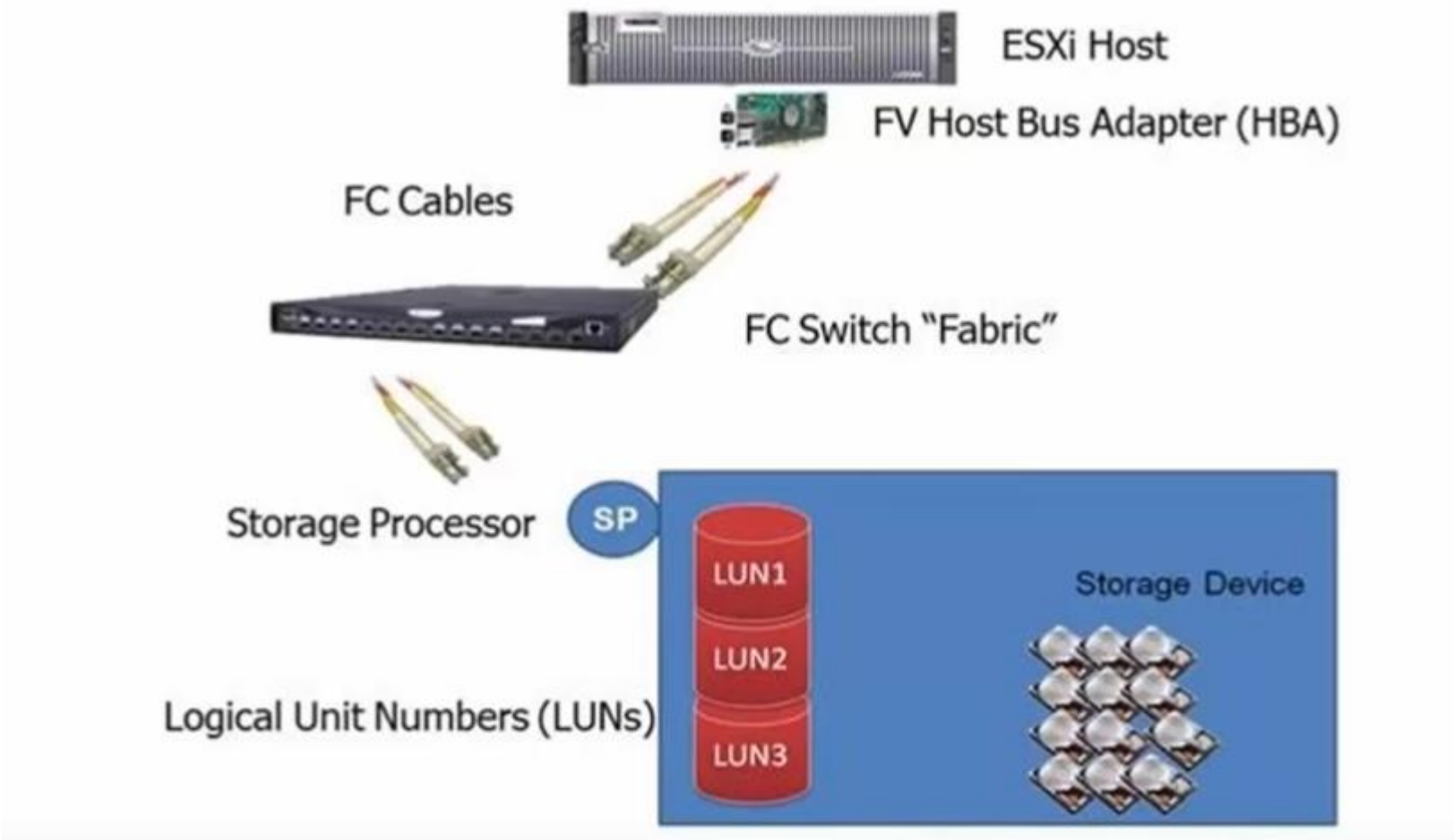

Schéma de principe : Host + HBA + Fabric + HBA + baie de stockage

Contrôleurs de stockage

Zoom sur la baie de stockage et les processeurs de stockage

Logical Unit Numbers

- Les **LUN** sont une **représentation logique** ou un point de référence logique vers un **disque physique** ou des **partitions** d'un disque dans des **matrices RAID**.
- **permettent** un **contrôle** plus **facile** des **ressources** de **stockage** dans un **SAN** en masquant l'aspect physique du disque
- Un **LUN** est **identifié** par un **nombre hexadécimal** de **16 bits**.
- Les **LUN fonctionnent différemment** avec **FC** et **iSCSI** :
  - Dans le cas de **FC** :
    - **l'initiateur** est normalement un **serveur avec** un **HBA FC initie** la connexion à un ou plusieurs **ports** sur le **système** de **stockage**
    - Les **cibles** sont les **ports** du **système** de **stockage** sur lesquels vous **accédez** aux **volumes**
    - Ces **volumes** ne sont autres que les **LUN**.
  - Dans le cas **d'iSCSI** :
    - **l'initiateur** est le **serveur** avec l'**adaptateur hôte iSCSI** ou le **logiciel iSCSI**
    - La **cible** iSCSI est telle que l'**adresse IP d'un système de stockage** connecté au réseau
    - La **cible** iSCSI doit **gérer** la **connexion** entre **l'initiateur** et le **LUN**

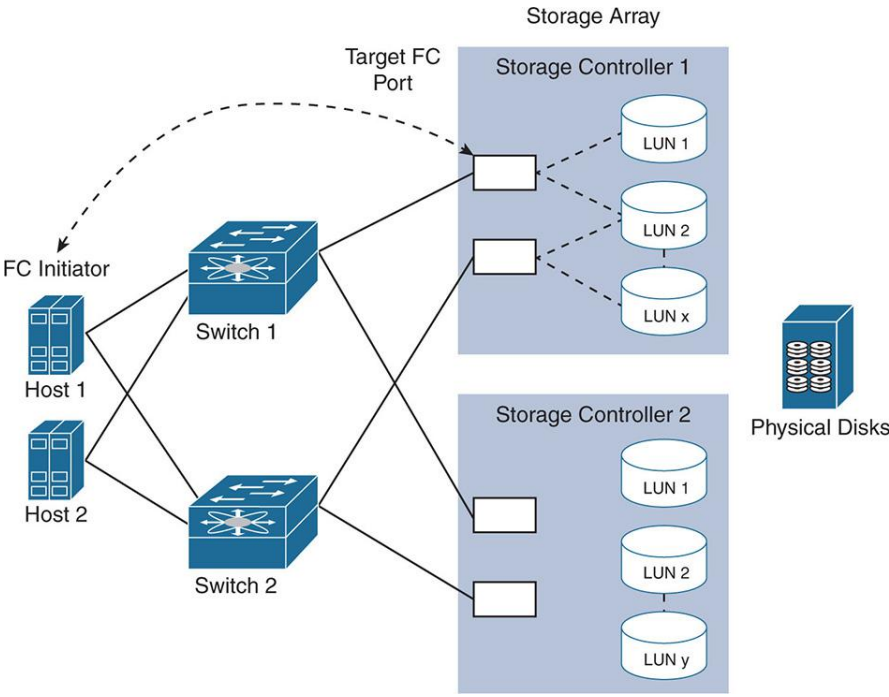

Figure 2-7 Logical Unit Number

Gestionnaire de volumes logiques

- **RAID regroupent plusieurs disques** et **améliorent** les **performances**, la **redondance** et la **tolérance aux pannes** des baies de disques.
- RAID **regroupe** les disques au **niveau physique**
- Les **LUN**, d'autre part, prennent la **matrice RAID résultante** comme point de départ et **offrent** une **vue logique des disques**
- Du **point de vue** du **stockage**, les **LUN** sont **encore** un autre **ensemble** de **disques**
- Du **point de vue du système d'exploitation hôte**, les **LUN ressemblent toujours** à un groupe de **disques** avec un **stockage** au **niveau** des **blocs**
- **Les systèmes d'exploitation fonctionnent** avec des **volumes**, d'où la **nécessité** de **créer** un **autre niveau d'abstraction** à l'aide d'un gestionnaire de volumes logiques (**LVM**).
- Le **LVM présente** les **LUN** au **système d'exploitation** en tant que **volumes**.
- Le **volume** est ensuite **formaté** avec le **système** de **fichiers** approprié.
- Les **baies de disques** sont **partitionnées** à l'aide de différents **niveaux RAID** -> des **LUN** sont **créés** pour **offrir** une **vue logique** des disques -> **LVM présente** les **LUN** comme des **volumes** au **système d'exploitation** hôte -> Les **volumes** sont ensuite **formatés** avec les **systèmes** de **fichiers** appropriés.

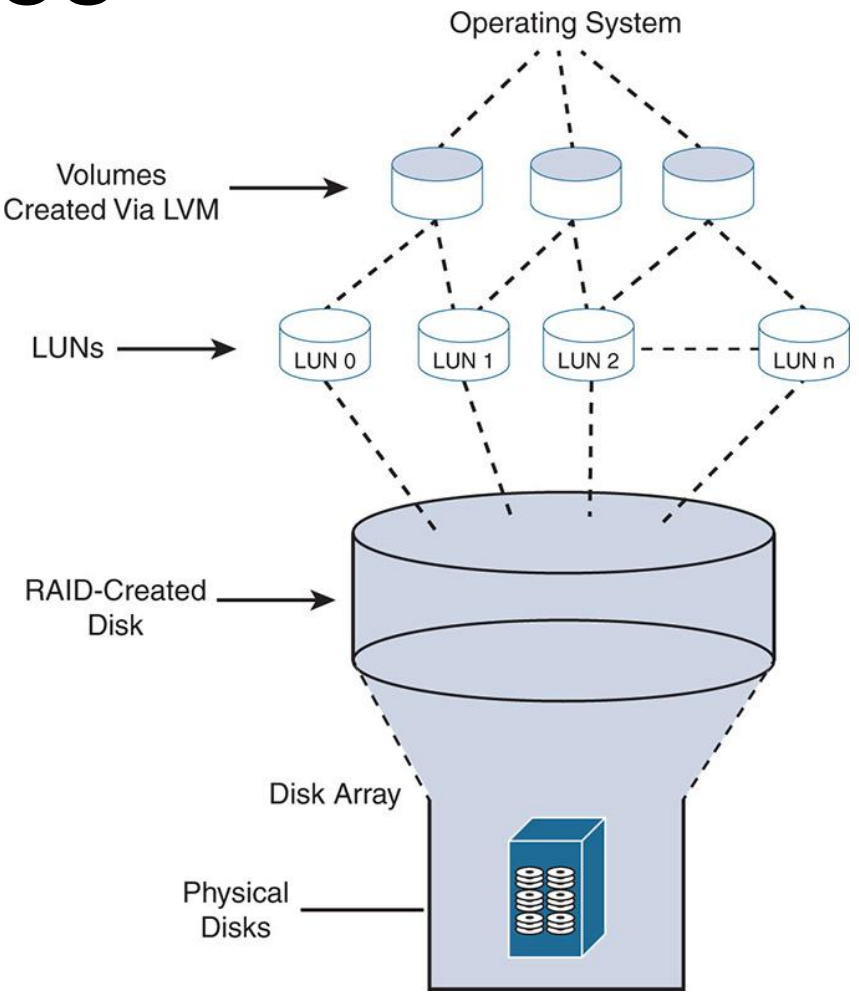

Figure 2-8 Différence entre les volumes et les LUN

TP

TP : LVM