le cnam

Les systèmes de stockage

NSY 104

Introduction

Besoins

- Stocker des informations de manière fiable et pérenne.
- Retrouver des informations de manière efficace.

SGF (système de gestion de fichier)

- Fait partie du SE

Concept de fichier

- Ils sont identifiés grâce
  - à leurs noms
  - à leurs extensions
  - à leurs en-têtes.

Les systèmes de fichiers

Différents types de structure :

- Suite d'octets

  Le système ne gère que des octets (DOS-UNIX). Le SGF ne gère pas la nature du contenu des fichiers.

- Suite d'enregistrements

  Le système ne gère que des enregistrements de taille fixe.

- Structure arborescente

  Arbre d'enregistrements de tailles variables, qui possèdent chacun une clé de position.

Les types de fichiers

- Fichiers spéciaux
  - bloc : modélisent les disques
  - caractère : liés aux E/S et permettent de modéliser les périphériques d'E/S série tels que les terminaux, imprimantes et les réseaux

- Les catalogues (*directories*) : fichiers systèmes qui maintiennent la structure du système de fichier

- Fichier ordinaire : contient les informations des programmes utilisateurs
  - Binaire
  - ASCII

Les fichiers ASCII

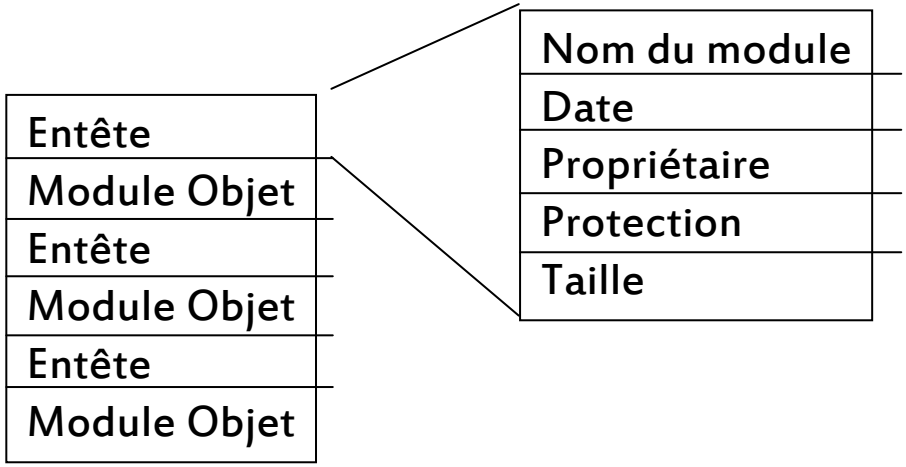
- Editable avec un éditeur de texte standard
- Visualisable et imprimable
- Possibilité de faire communiquer la sortie d'un programme avec l'entrée d'un autre programme. Exemple : le tube (pipe).

Les fichiers binaires

- Leur impression donne une suite de caractères de signes incompréhensible.
- Structure interne forte

Structure interne des fichiers binaires

Un fichier exécutable UNIX



Un fichier archive

Attributs des Fichiers

Protection	Qui peut accéder au fichier et de quelle façon
Mot de passe	Mot de passe requis pour accéder au fichier
Créateur	Personne qui a créé le fichier
Propriétaire	Propriétaire courant du fichier
Indicateur lecture seule	0 pour lecture / écriture, 1 pour lecture seule
Indicateur fichier caché	0 pour fichier normal, 1 pour fichier non affiché dans listage
Indicateur fichier système	0 pour fichier normal, 1 pour fichier système
Indicateur d'archivage	0 le fichier a été archivé, 1 il doit être archivé
Indicateur Ascii / Binaire	0 pour fichier Ascii, 1 pour fichier binaire
Indicateur accès aléatoire	0 pour accès séquentiel, 1 pour accès aléatoire
Indicateur fichier temporaire	0 pour fichier normal, 1 pour fichier a supprimer lorsque le processus utilisateur sera terminé
Indicateur verrouillage	0 pour fichier non verrouillé, 1 pour fichier verrouillé
Longueur d'enregistrement	Nombre d'octets dans un enregistrement
Position de la clé	Position relative de la clé dans chaque enregistrement
Longueur de la clé	Nombre d'octets du champ de la clé
Date de création	Date et heure de la création du fichier
Date du dernier accès	Date et heure du dernier accès au fichier
Date de modification	Date et heure de la dernière modification du fichier
Taille courante	Nombre d'octets dans le fichier
Taille maximale	Taille maximale autorisée pour le fichier

Opérations sur les fichiers

Accès séquentiel : les octets sont lus dans l'ordre depuis le début du fichier.

Accès aléatoire : les octets peuvent être lus indépendamment de leurs emplacement.

Chaque ouverture de fichier provoque la création par le système d'un objet contenant un pointeur sur la position courante dans le fichier. Ce pointeur est initialisé au début du fichier et chaque accès au fichier le fait passer à l'élément suivant.

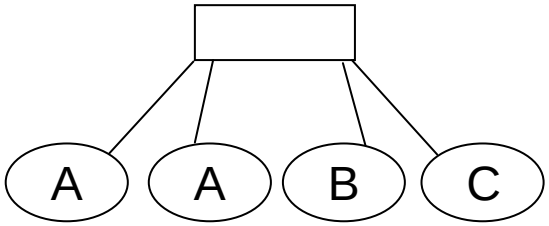
Opérations d'accès :

- Read : Lecture du fichier depuis la position spécifiée ou courante.
- Write : Ecriture du fichier à la position spécifiée ou courante.
- Seek : Modification de la position courante dans le fichier.

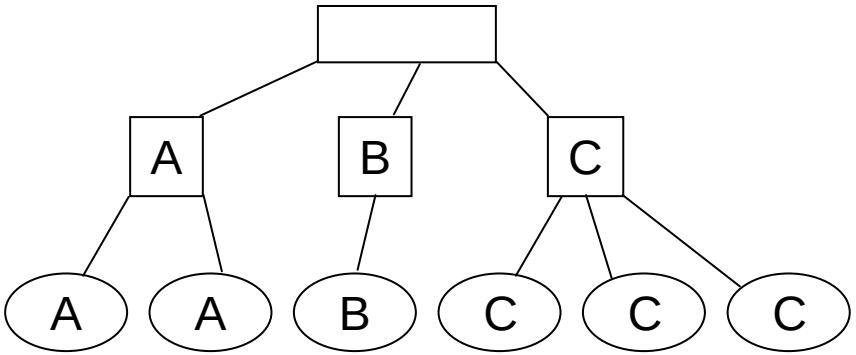
Opérations sur les fichiers

- CREATE : Le fichier est créé sans données. Cet appel renvoie un identifiant système du fichier.
- DELETE : Le fichier est détruit et l'espace disque récupéré.
- OPEN : Le fichier doit être ouvert pour pouvoir être utilisé. Cet appel renvoie un identifiant système du fichier.
- CLOSE : Le fichier est fermé, cela permet une libération des tables et structures utilisées par le système pour gérer les accès aux fichiers.
- READ / WRITE : Lecture du fichier – Ecriture du fichier.
- SEEK : Spécification de la position courante de lecture ou d'écriture dans le fichier, ou lecture de cette position.
- ATTRIBUTES : Lecture ou modification des attributs du fichier.
- RENAME : Changement du nom du fichier spécifié.
- MAP / UNMAP : Mappage du fichier dans un espace d'adressage virtuel d'un processus.

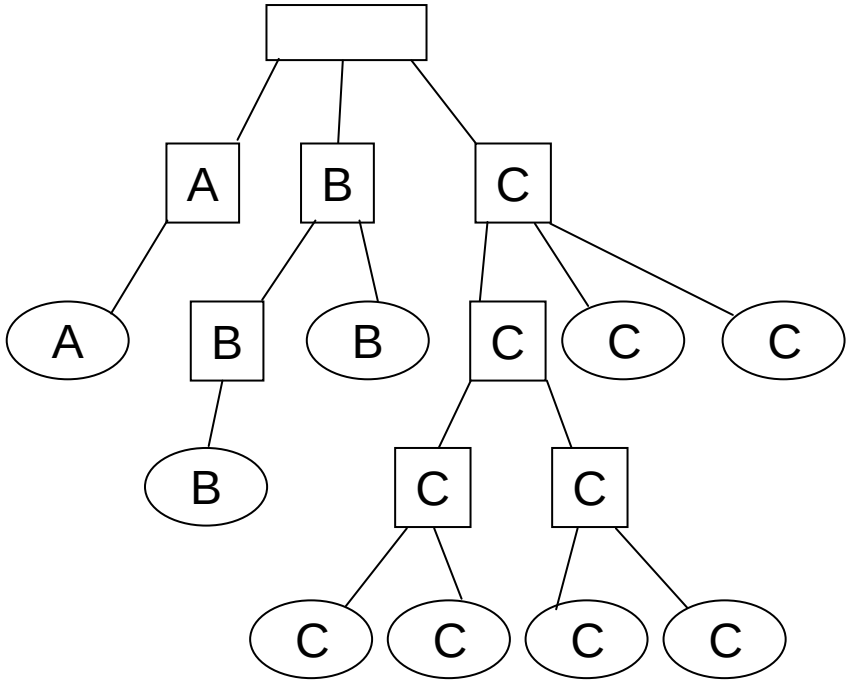
Architectures de Catalogues



Tous les fichiers utilisateurs sont mis dans le même catalogue

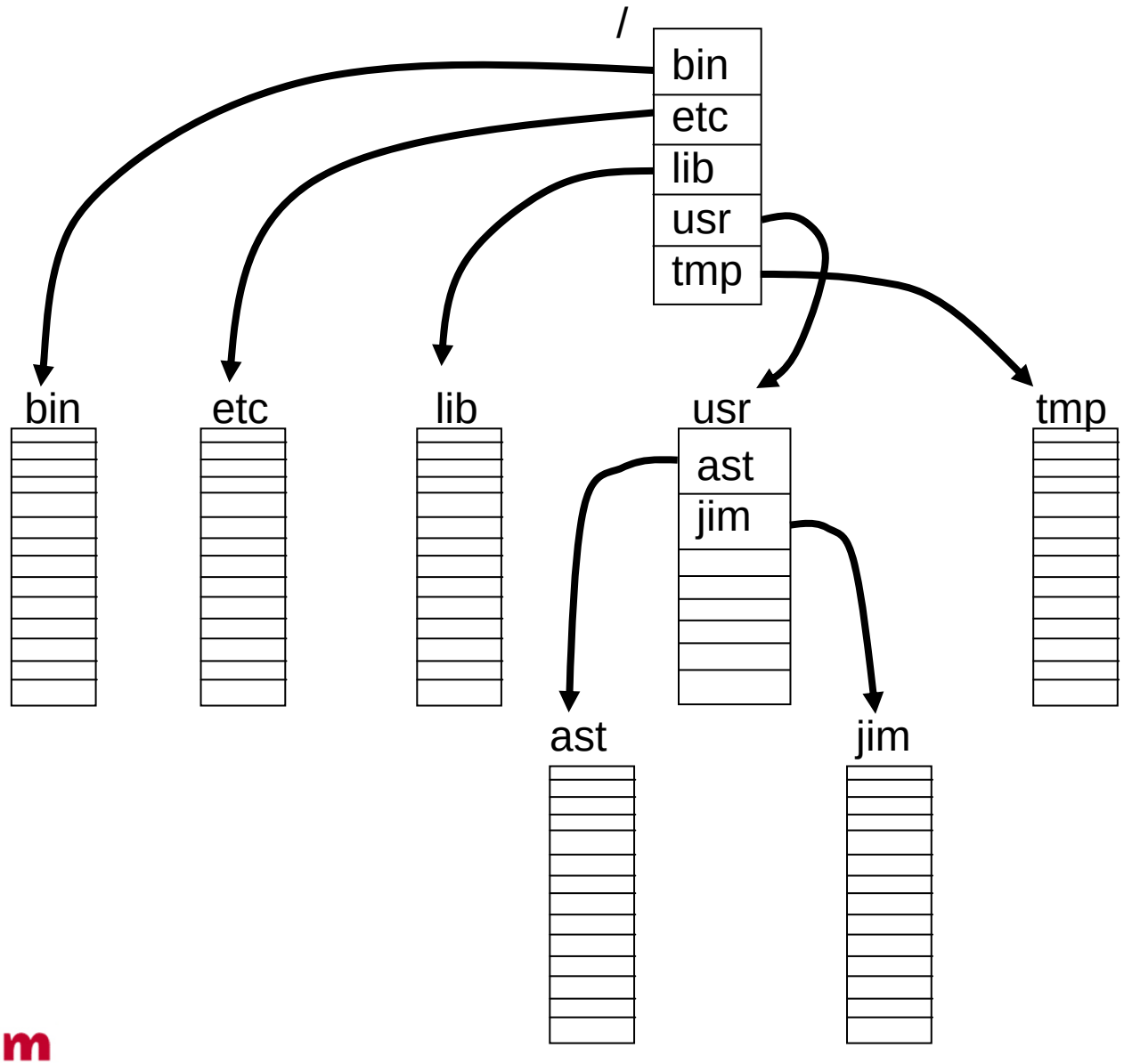


Chaque utilisateur possède son catalogue

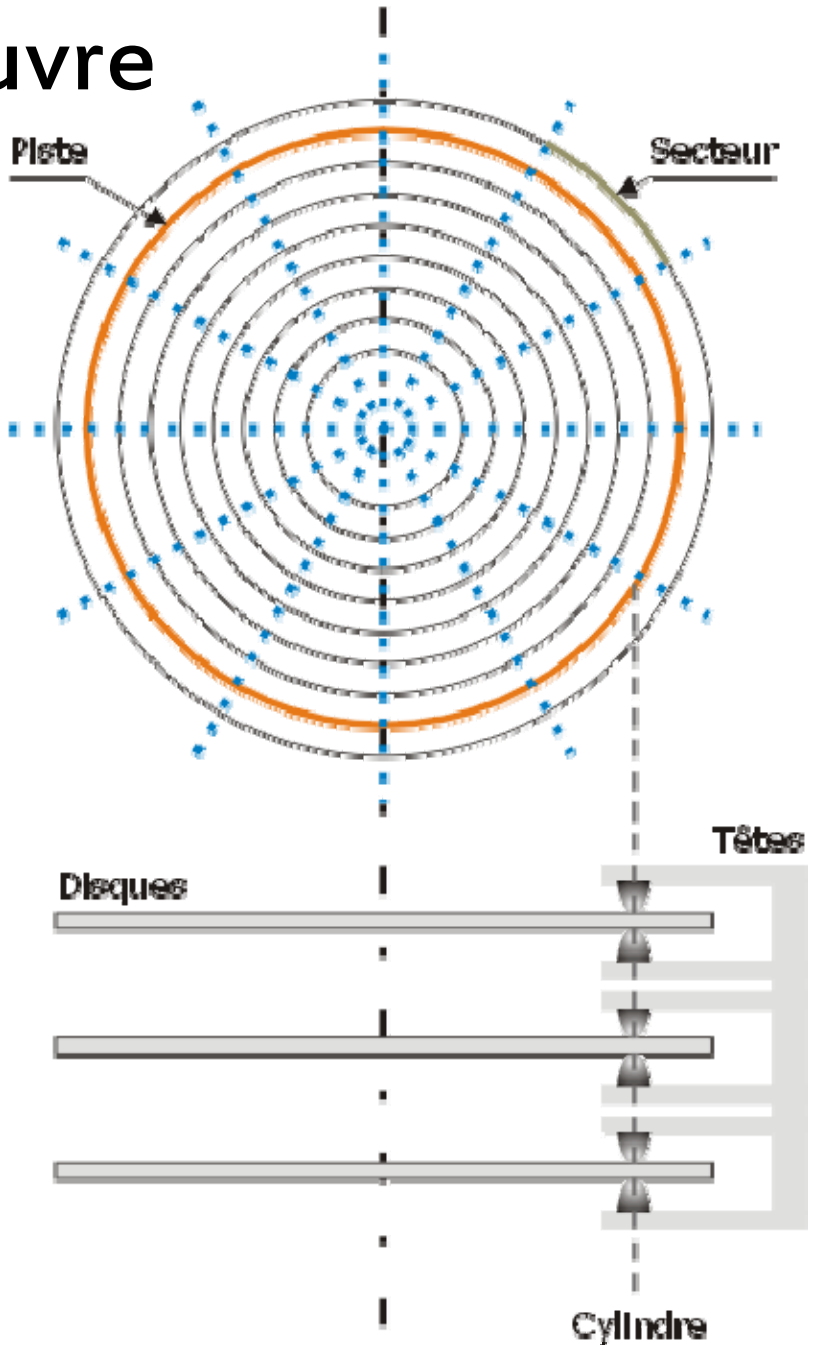


Une arborescence de fichiers pour chaque utilisateur

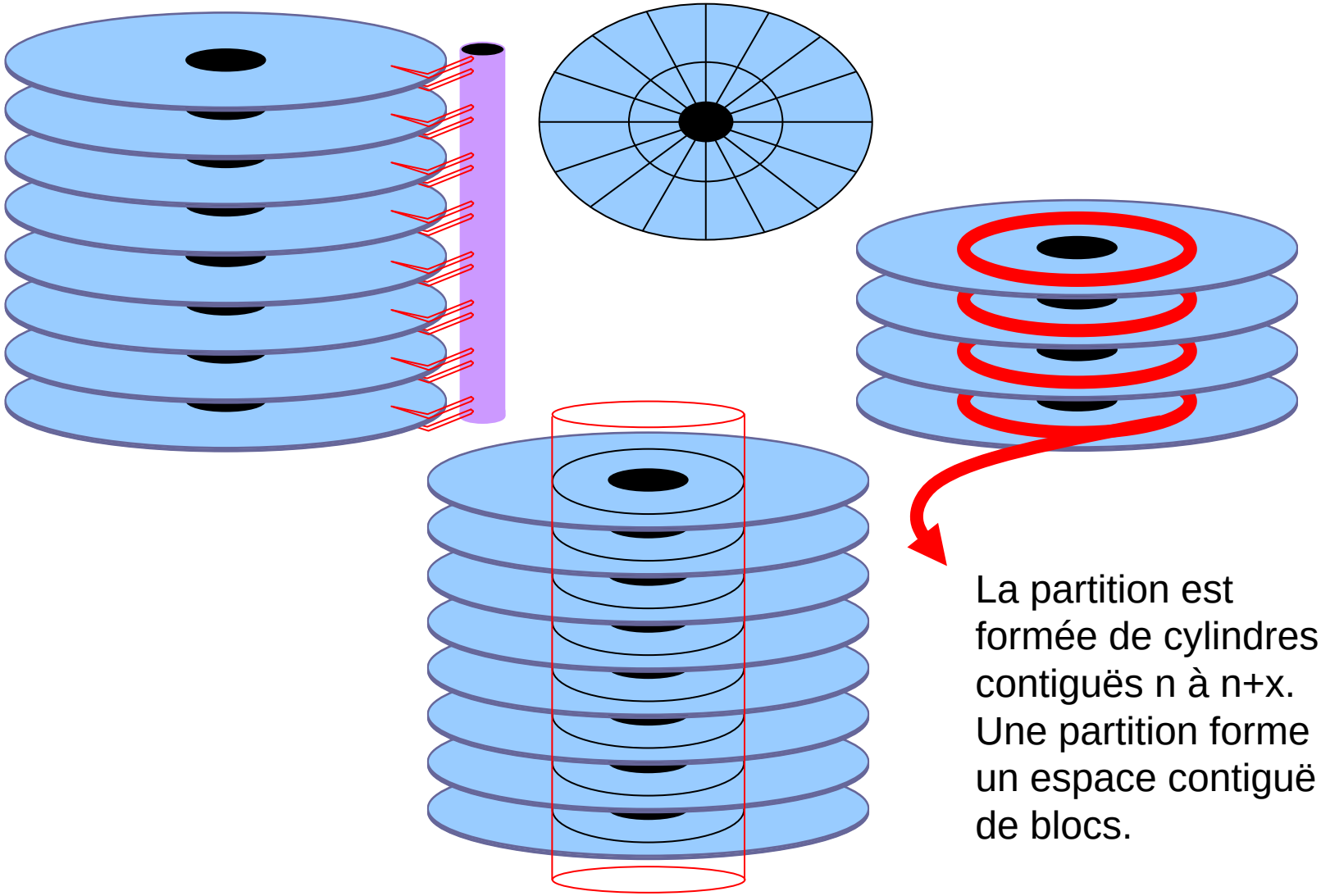
Arborescence de Catalogue - Unix



Mise en œuvre



Mise en œuvre



La partition est formée de cylindres contiguës n à n+x. Une partition forme un espace contiguë de blocs.

Mise en œuvre

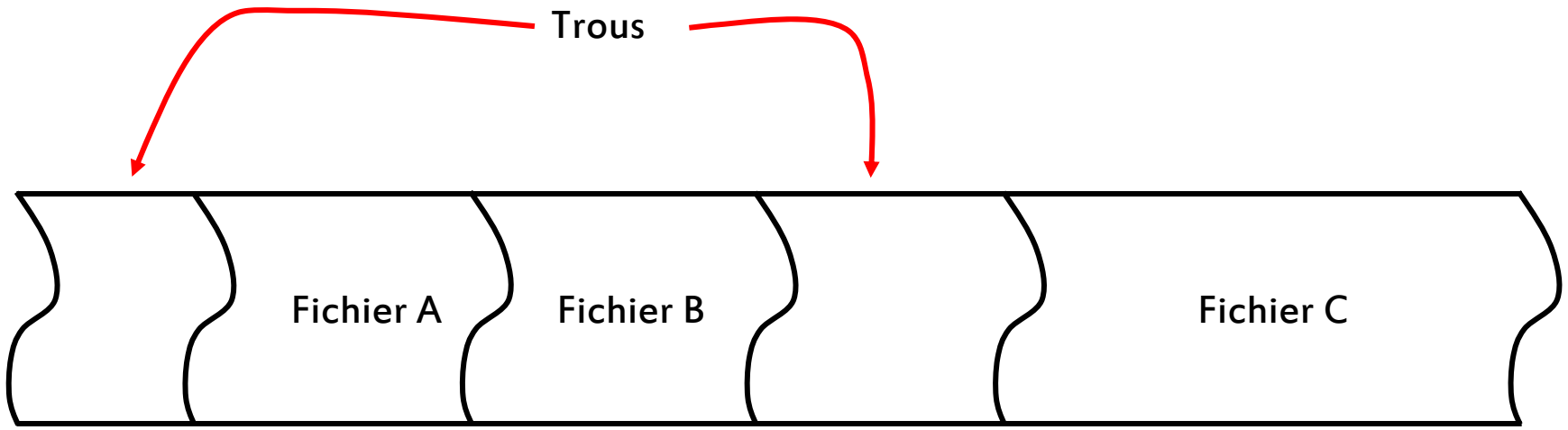
Le concept fondamental est

la mémorisation des adresses des blocs de chaque fichier

- Allocation de blocs contigus.
- Allocation de blocs sous forme de liste chaînée.
- Allocation de blocs sous forme de liste chaînée indexée.
- Allocation de nœuds d'informations.

Mise en œuvre

Allocation de blocs contigus :



Les données sont stockées dans des blocs contigus, il suffit de mémoriser le premier bloc pour lire tous les blocs du fichier. Mais la taille du fichier doit être connu lors de sa création.

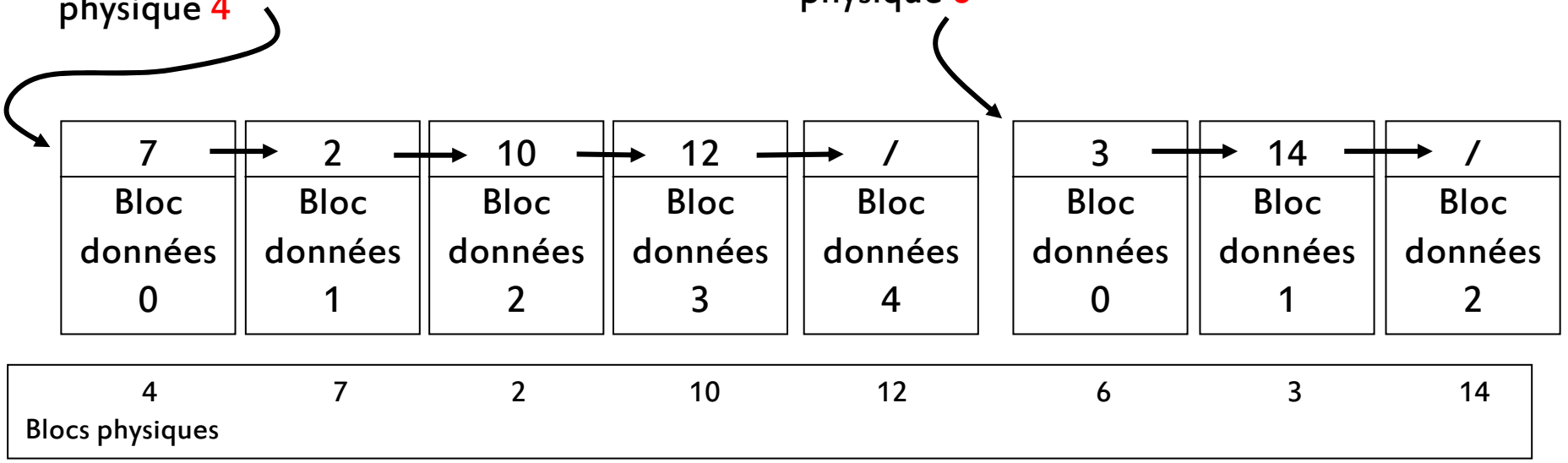
Mise en œuvre

Allocation de blocs sous forme de liste chaînée :

Le premier mot de chaque bloc est un pointeur sur le bloc suivant.

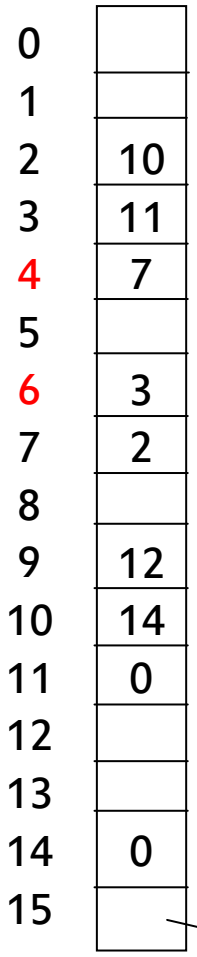
Le fichier répertoire Contenant le fichier A indique que celui-ci commence au bloc physique 4

Le fichier répertoire contenant le fichier B indique que celui-ci commence au bloc physique 6



Mise en œuvre

Allocation de blocs sous forme de liste chaînée indexée :



Le répertoire contenant le fichier A indique que celui-ci commence au bloc 4.

Le répertoire contenant le fichier B indique que celui-ci commence au bloc 6.

Le fichier A contient donc les blocs physique suivants : 4, 7, 2, 10, et 14.

Le fichier B contient donc les blocs physique suivants : 6, 3 et 11.

La liste doit toujours être parcourue pour trouver un déplacement donné dans le fichier, mais elle réside entièrement en mémoire et peut être parcourue sans accéder au disque.

Bloc non utilisé, libre pour être affecté à un fichier.

Mise en œuvre

Allocation de nœuds d'informations :

- Le nœud est une petite table (*i-node*) qui contient les attributs et les adresses des blocs du fichier.

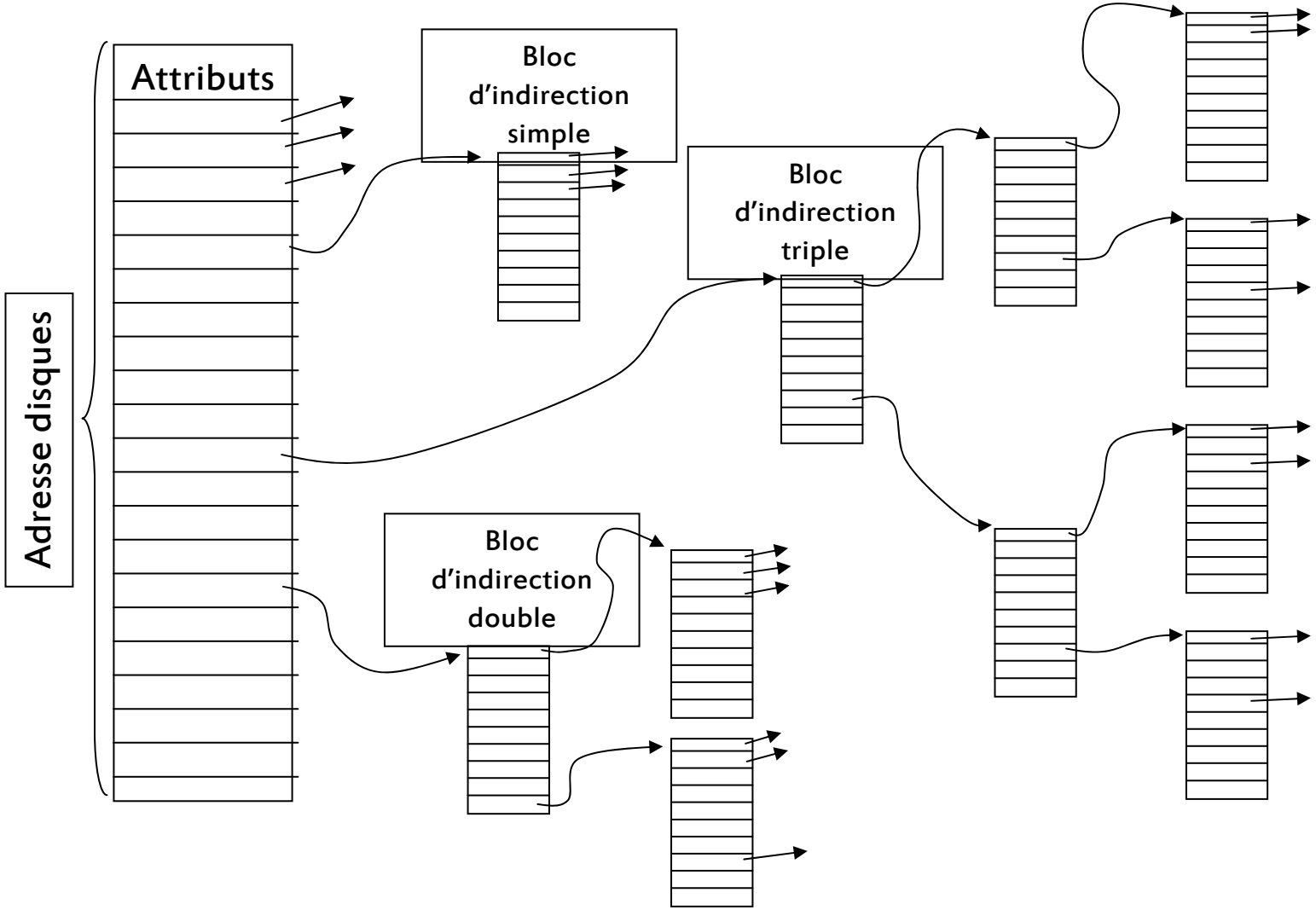
- Les premières adresses sont contenues dans le nœud d'information de façon à pouvoir gérer efficacement les petits fichiers.

- Si le fichier est trop gros, le nœud contient alors des adresses de blocs appelées blocs d'indirection. Ces derniers contiennent les adresses des autres blocs.

- Il peut y avoir jusqu'à trois niveaux d'indirection. Chaque bloc d'indirection pointe sur quelques centaines de blocs de données.

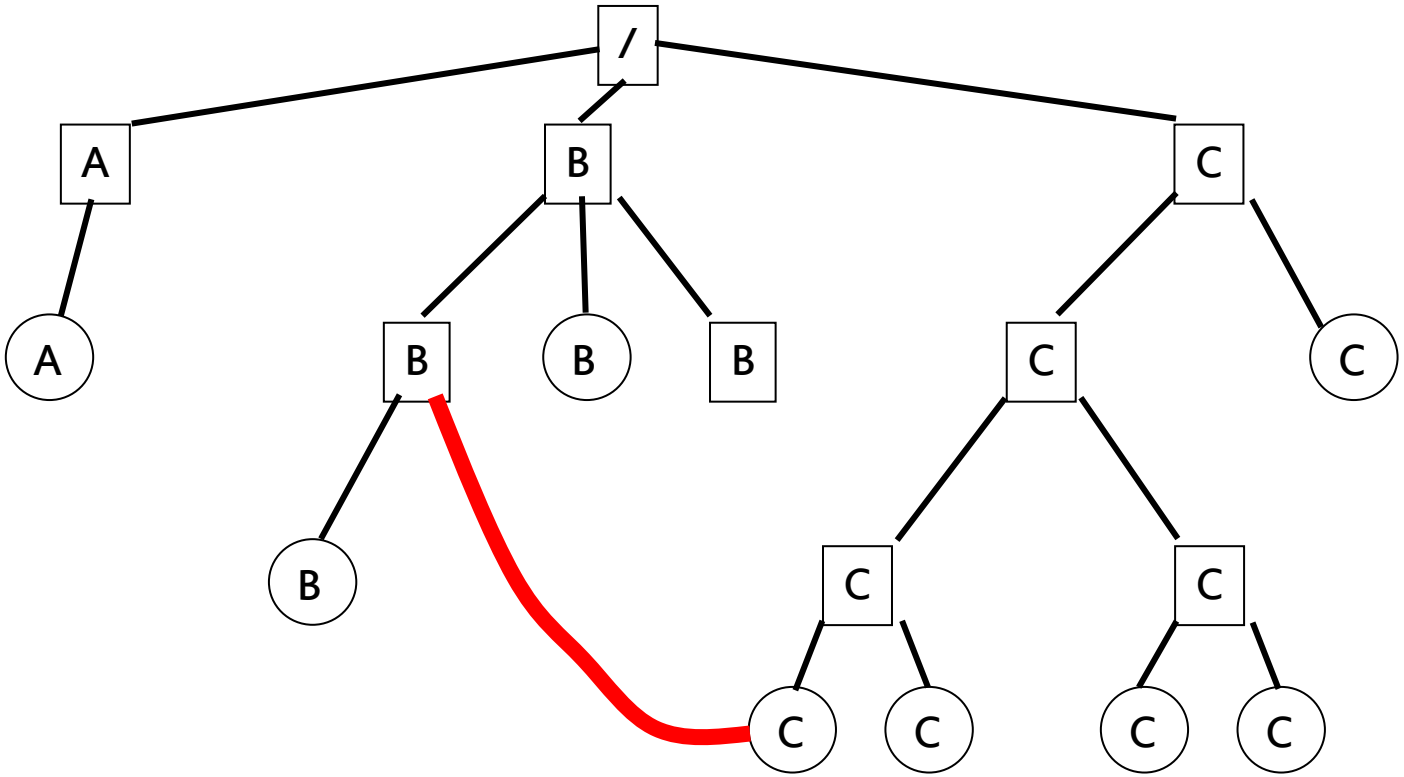
Mise en œuvre

Allocation de nœuds d'informations :



Les Fichiers Partagés

Il peut être souhaitable de voir apparaître le même fichier dans plusieurs catalogues simultanément. Cette fonctionnalité peut être obtenue par l'utilisation du mécanisme de lien.



FAT

Le système de fichier FAT est composé de trois grandes sections :

- Le secteur de boot (BPB: Bios Parameter Block)
  C'est le premier secteur de la partition
  !! Différent du MBR : 1er secteur du DD !!
- Les tables d'allocation (FATs)
  C'est une carte du disque.
- Le répertoire racine (Root directory)
  C'est une liste des fichiers présents à la racine du disque.

- Un cluster est un groupe de secteurs – C'est la taille minimale allouable
  Il sert d'unité d'allocation aux fichiers.
- Chaque cluster stocke donc les données d'un fichier.

Pour un fichier de 9 000 octets, sur un disque utilisant des clusters de 8 192 octets (16 secteurs de 512 octets)

- *2 clusters sont utilisés*
- *Le dernier n'utilise que 808 octets (9 000 - 8 192).*

FAT - table d'allocation

Valeurs numériques permettant de décrire l'allocation des clusters d'une partition (structure de tableau)

0×0000	Cluster libre.
0×0001	Cluster réservé.
0×0002-0xFFEF	Cluster utilisé. *Indice du prochain cluster du fichier.*
0xFFF0-0xFFF6	Valeurs réservées.
0xFFF7	Cluster défectueux.
0xFFF8	Cluster utilisé. Dernier cluster d'un fichier.

FAT – root directory

Format d'une entrée du root directory

Offset	Taille	Description
0x00	8	Nom du fichier (peut être '.' ou '..')
0x08	3	Extension (rempli par des espaces)
0x0B	1	Attributs du fichier
0x0C	1	Réservé, utilisé par NT
0x0D	1	Heure de création
0x0E	2	Heure de création
0x10	2	Date de création
0x12	2	Date du dernier accès
0x14	2	Index EA ou 2 octets poids fort du numéro du 1er *cluster* (FAT32)
0x16	2	Heure de dernière modification
0x18	2	Date de dernière modification
0x1A	2	Numéro du premier *cluster* du fichier (FAT12 et FAT16) ou 2 octets de poids faible de ce numéro (FAT32).
0x1C	4	Taille du fichier

Quand le fichier est un répertoire (voir bits attributs), le contenu est une structure identique au format précédent, et commence par les deux entrées '.' et '..'.

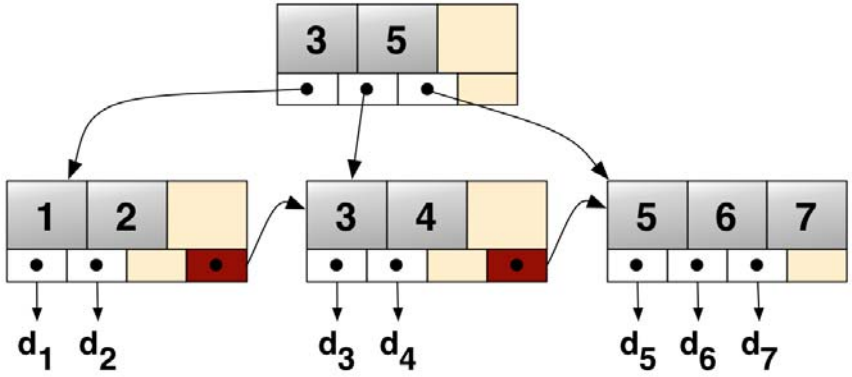
exFAT ?

NTFS (1993)

Le système de fichier NTFS est un système journalisé basé sur une structure :

- MFT (Master File Table)
- La MFT représente donc une structure de stockage des métadonnées des fichiers de la partition.

- Sensible à la casse
- Tolère les noms de fichier longs



- Améliorations de FAT
  - Arbre équilibré pour localiser les fichiers (B+ tree)
    - Plus performant
  - Métadonnées (propriétaire, dernier accès en lecture, droits de groupe) plus riches orientées multiutilisateur

Ext3

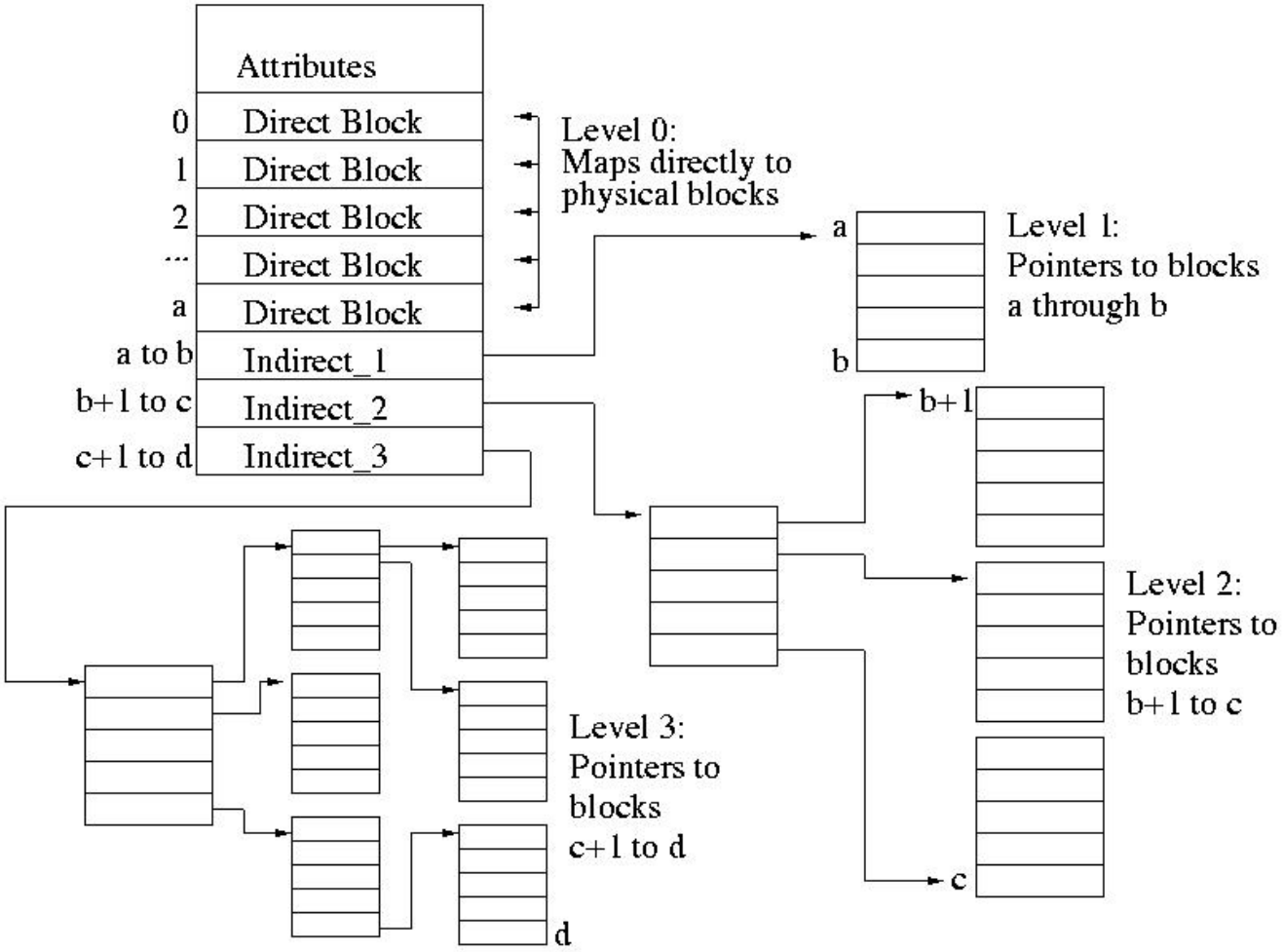
- Le système de fichier ext3 est lui aussi journalisé (évolution majeure de ext2)

- ext3 alloue les blocs libres juste à côté des autres blocs utilisés par le fichier
  - minimise l'espace physique entre les blocs.
  - reste néanmoins fragmenté (mais moins que les systèmes MS)

- 3 niveaux de journalisation:
  - Métadonnées et données
  - Métadonnées seulement
    - Avec écriture préalable effective
    - Sans écriture préalable assurée

iNode

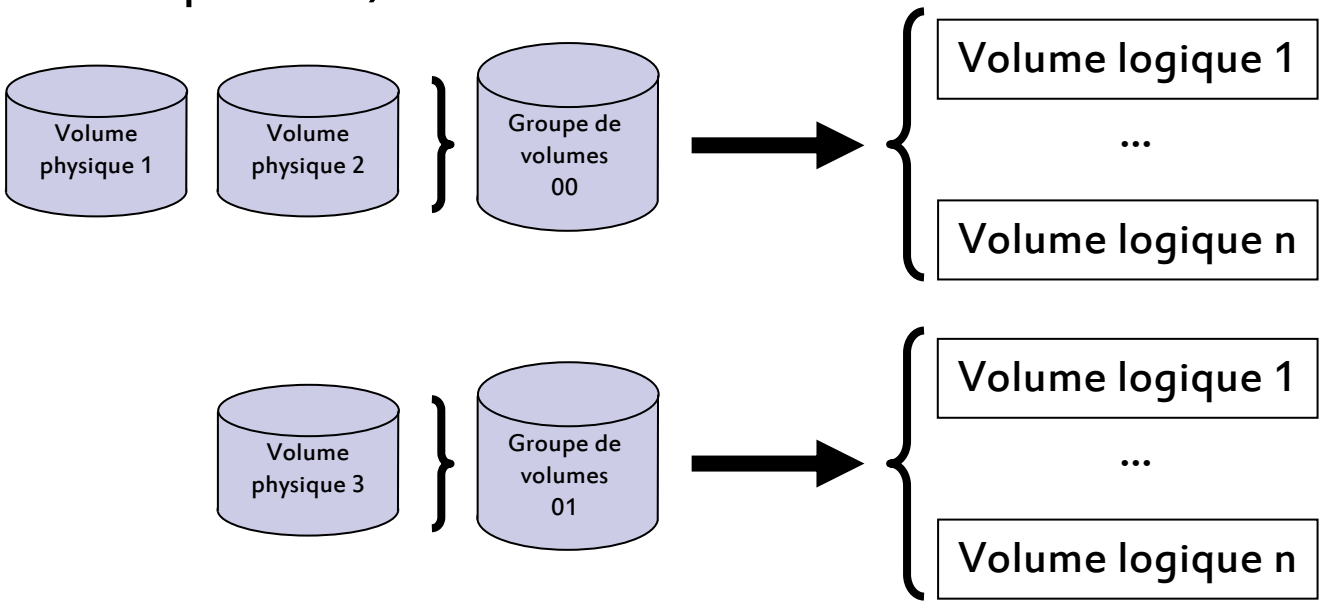
Métadonnées d'un fichierd'un fichier. Contient environ 64 champs, dont 13 pointent vers les blocs de données du fichier, soit de manière directe (les 10 premiers), soit de manière indirecte (les 3 suivants).



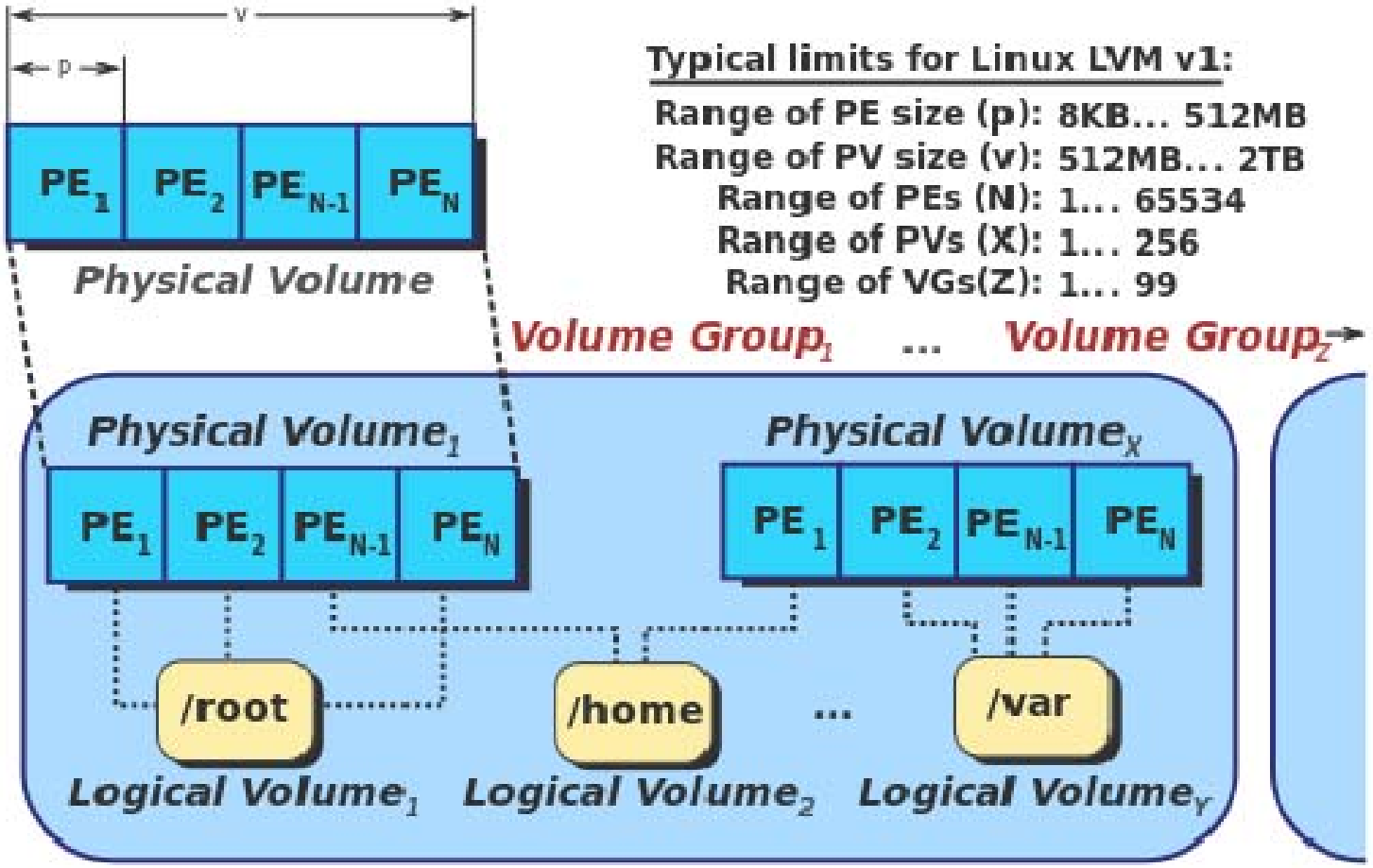
Les volumes logiques

Avec les volumes logiques (LVM : *Logical Volume Manager*) :

- Les disques (physiques) s'appellent des volumes physiques.
- Il est possible de rassembler des volumes physique pour former un groupe de volumes.
- Dans ce groupe de volumes, il est possible de déclarer un ou plusieurs volumes logiques (représentant un ensemble de blocs contiguës de données, c'est l'équivalent d'une partition).

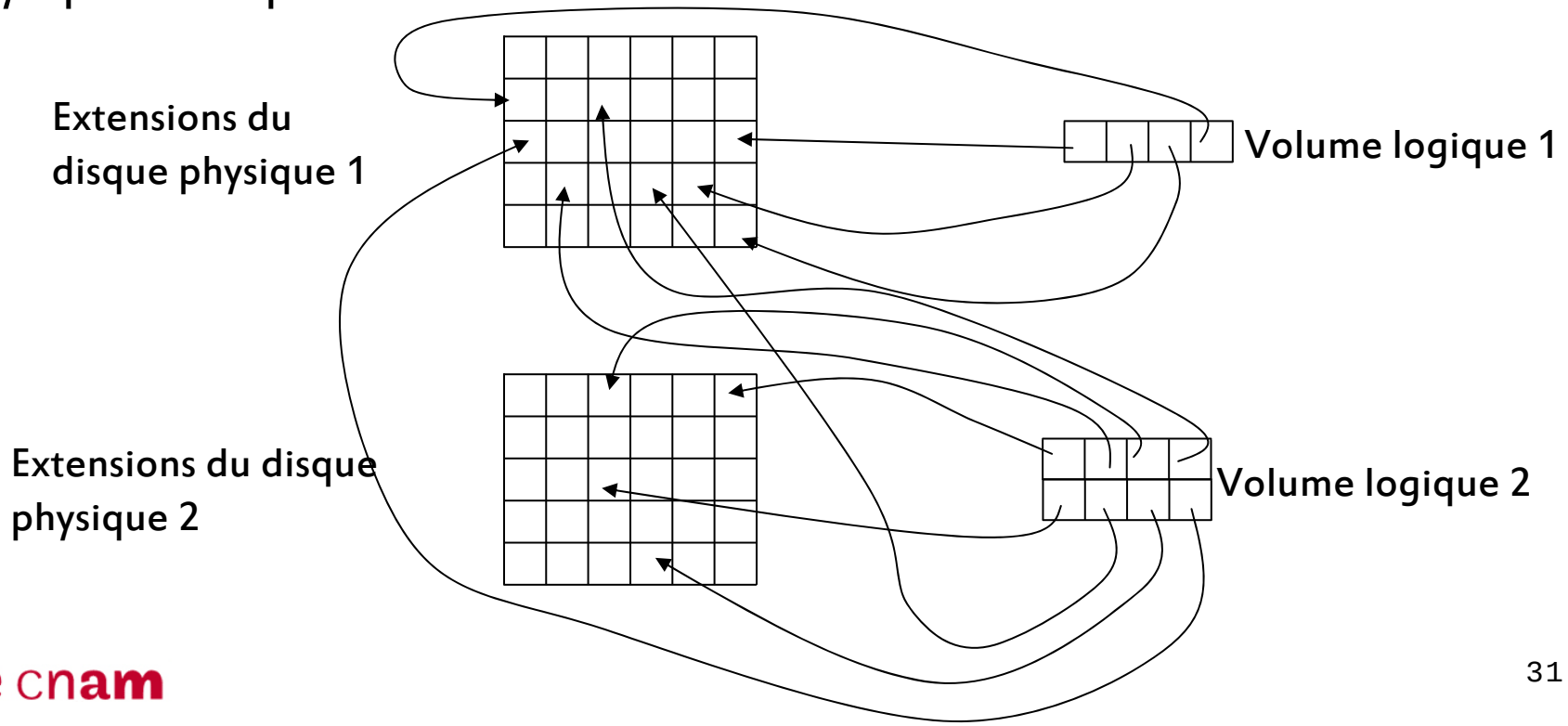


Les volumes logiques



Les Volumes Logiques

- Les disques sont découpés en petites unités (4 Mo) appelées extensions physiques (*physical extends*) PE.
- Les volumes logiques sont une structure regroupant les extensions logiques (*logical extends*) LE qui sont parcourues séquentiellement.
- Chaque extension logique est un pointeur vers une extension physique. L'écriture ou la lecture d'un bloc se fait en suivant ce pointeur pour retrouver la partie physique du disque où la donnée a été écrite.



Les Volumes Logiques

La gestion de ce mécanisme nécessite des espaces sur les disques destinés à stocker des informations telles que :

- Physical Volume Reserved Area.
  - Cette structure contient les informations LVM spécifiques aux volumes physiques.
- Volume Group Reserved Area.
  - Cette structure contient les informations LVM spécifiques aux groupes de volumes entier. Elle est redondante sur chaque volume physique du groupe de volume.
- User Data Area.
  - Cette zone contient les systèmes de fichiers destinés aux utilisateurs.
- Boot Disk.
  - Cette zone contient des structures supplémentaires destinées au démarrage de la machine.

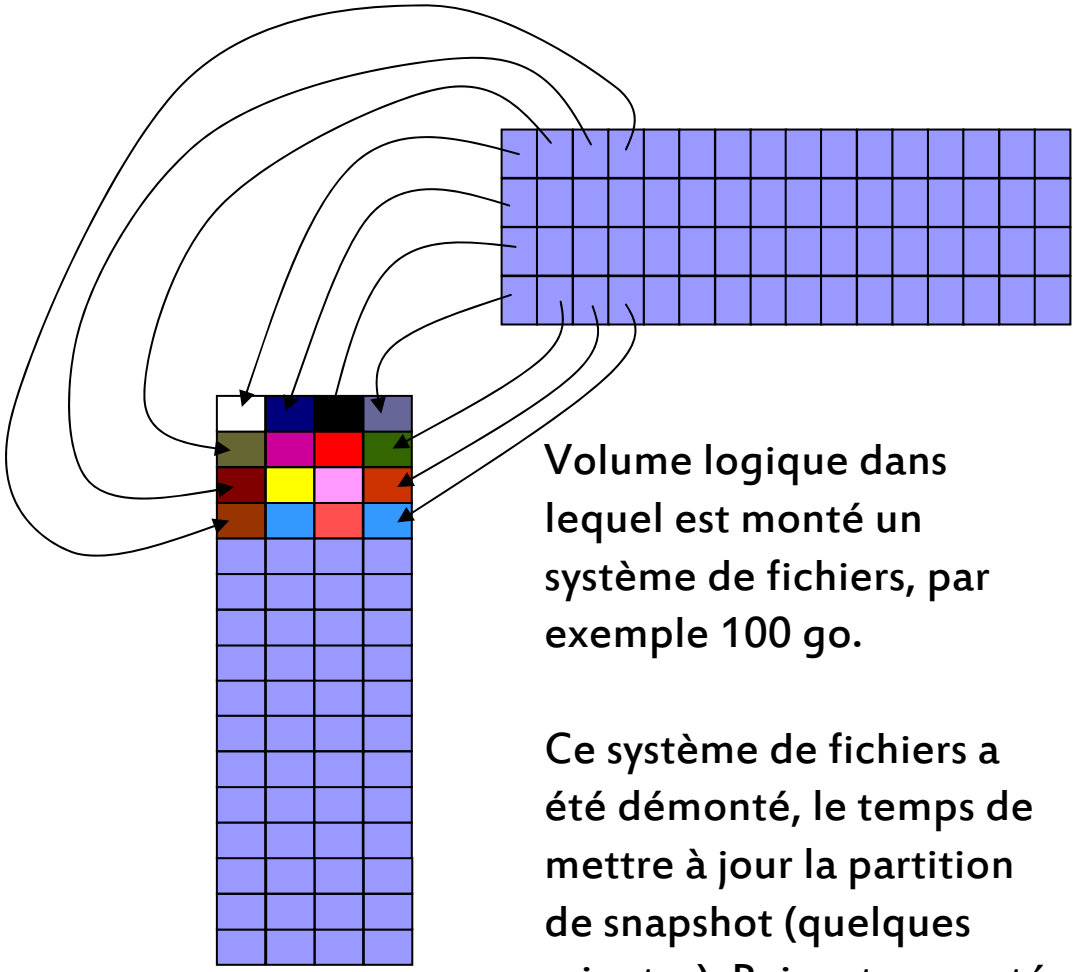
Les Volumes Logiques

Le LVM permet aussi l'utilisation de *snapshot*.

Ce mécanisme permet de sauvegarder un système de fichier en même temps que son utilisation.

Cela est très utile pour les sauvegardes des bases de données qui peuvent prendre beaucoup de temps.

Les Volumes Logiques



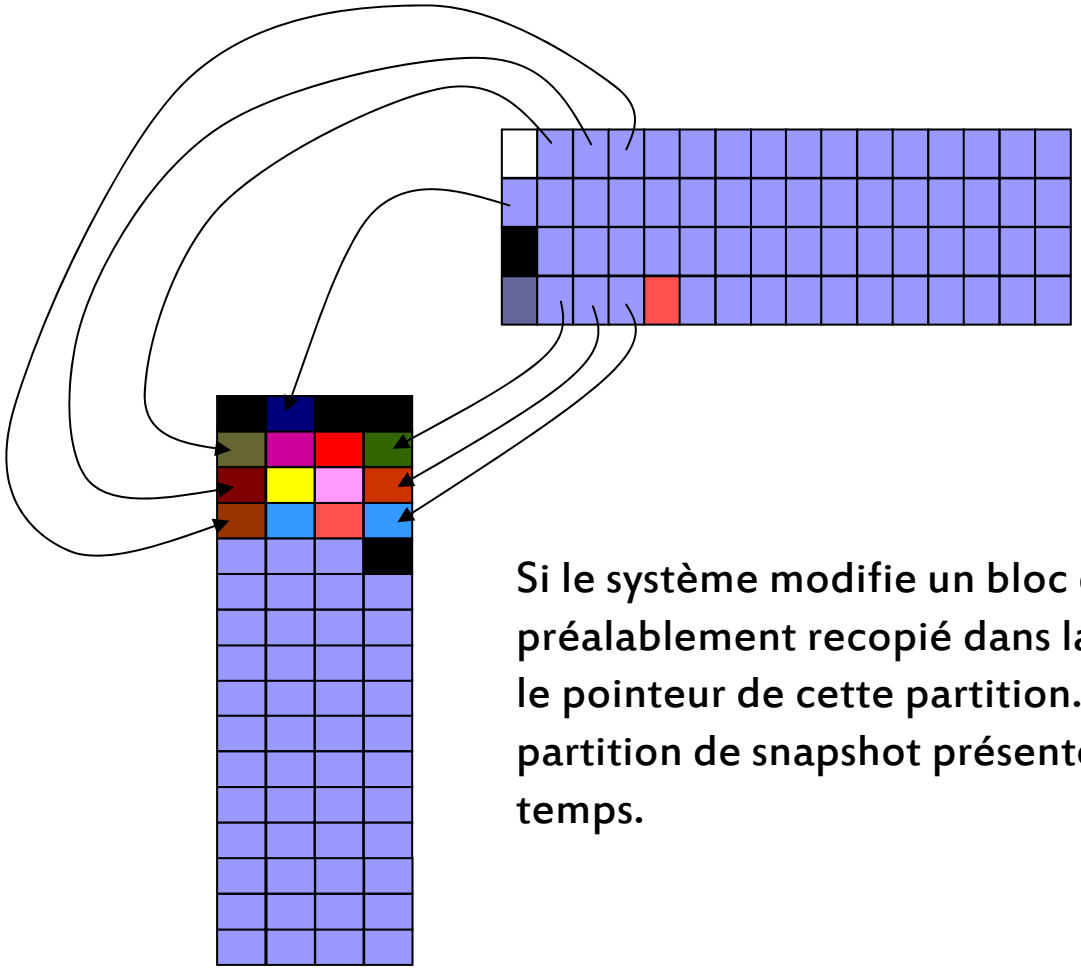
Partition de snapshot, chaque bloc contient un pointeur vers les extensions physiques des disques constituant les volumes physiques.

Volume logique dans lequel est monté un système de fichiers, par exemple 100 go.

Ce système de fichiers a été démonté, le temps de mettre à jour la partition de snapshot (quelques minutes). Puis est remonté et utilisé par le système informatique qui y lit et y écrit des données.

On tente de sauvegarder la partition de snapshot. Pour retrouver les données, on suit les pointeurs vers les disques.

Les Volumes Logiques



Si un bloc est trouvé à la place du pointeur, alors ce bloc est pris pour le bloc à sauvegarder. Un bloc ne peut être recopié qu'une seule fois dans la partition de snapshot.

Si le système modifie un bloc de données, celui-ci est préalablement recopié dans la partition de snapshot et remplace le pointeur de cette partition. De cette façon, le parcours de la partition de snapshot présente un système de fichier figé dans le temps.

Sécurité

- Les systèmes de fichiers permettent de stocker toutes les données du système, programmes et flux d'entrées – sorties.

- La plupart des objets du système tel que les drivers, objets de synchronisation, de communication inter – tâches sont représentés par des fichiers de type spéciaux.

LE SYSTEME DE FICHIERS EST LE POINT CLE D'UN SYSTEME D'EXPLOITATION

Il faut donc veiller à sa cohérence, sa sécurité et sa protection :

- Stratégie de sauvegarde.
- Parer les attaques.

Le RAID

- Acronyme de Redundant Array of Independent Disks (matrice redondante de disques indépendants)

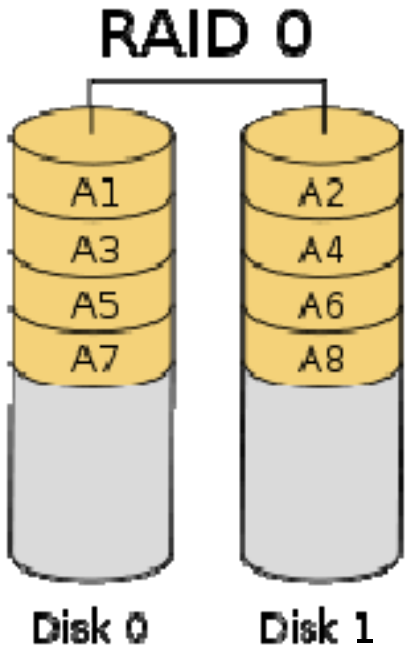
- Plusieurs configurations de RAID

- Le RAID peut être logiciel, pseudo matériel, matériel avec, pour chacun, des avantages et des inconvénients.

Le RAID

RAID 0

- Le RAID 0 est une configuration permettant d'augmenter les performances d'une grappe de disque en faisant travailler n disques en parallèle avec n supérieur ou égale à 2 (stripping).



- Ce type de RAID est parfait pour des applications requérant un traitement rapide d'une grande quantité de données.
- Cette architecture n'assure en rien la sécurité des données. En effet, si l'un des disques tombe en panne, la totalité des données de la grappe est perdue.

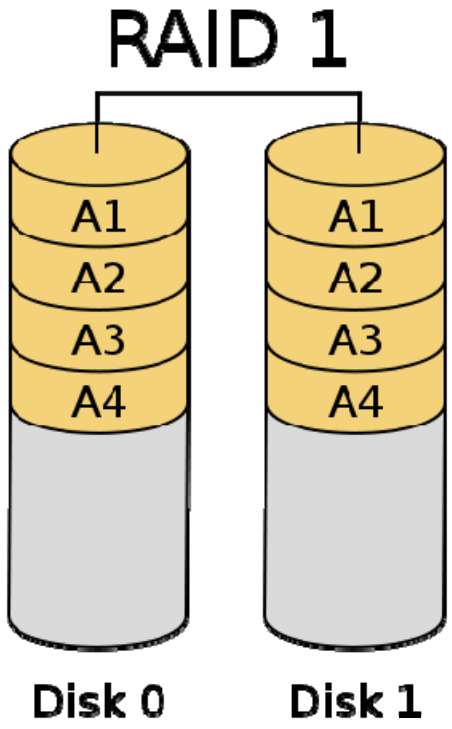
Performances ++
Sécurité --

Le RAID

RAID 1

o Le raid 1 consiste en l'utilisation de n disques redondants (n >= 2). Chaque disque de la grappe contenant à tout moment exactement les mêmes données, d'où l'utilisation du terme *mirroring*.

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



Sécurité ++
Performances =
Coût --

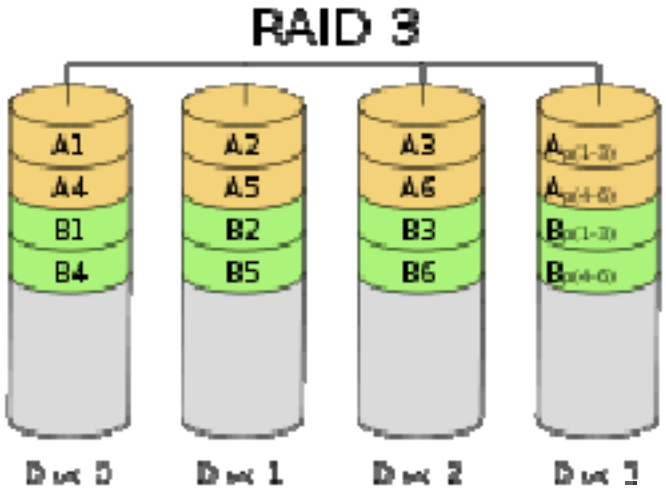
Le RAID

RAID 3 & 4

o Le RAID 3, ou 4, fonctionnent avec, au minimum, 3 disques. Le dernier disque de la grappe contient un code de parité sur les données des autres disques. Les deux niveaux de RAID sont semblables, sauf que le niveau 3 travaille par octets alors que le niveau 4 travaille par blocs.

Suite à la défaillance d'un disque, le contrôleur RAID peut reconstruire ce disque qu'il s'agisse d'un disque de données ou du disque de parité.

Architecture sécuritaire mais limite des capacités de flux de données (et parité) car la parité est uniquement sur le dernier disque. Le système est limité par la bande passante de ce disque sollicité à chaque instant.

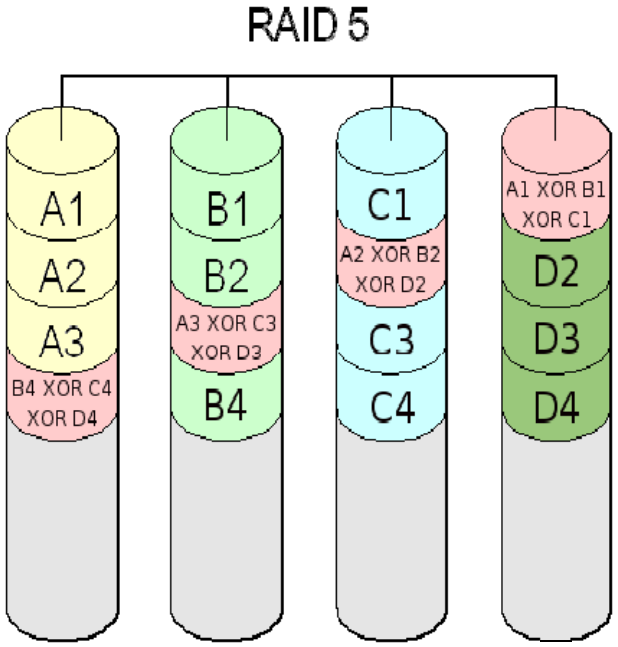


Le RAID

RAID 5

o Le RAID utilise une parité répartie. La parité, qui est incluse avec chaque écriture se retrouve répartie circulairement sur les différents disques de la grappe. Ce système nécessite l'utilisation de 3 disques, au minimum. Pour un système de X disques, la capacité est X-1.

Suite à la défaillance d'un disque, le contrôleur RAID peut reconstruire ce disque. C'est le niveau de RAID le plus utilisé.

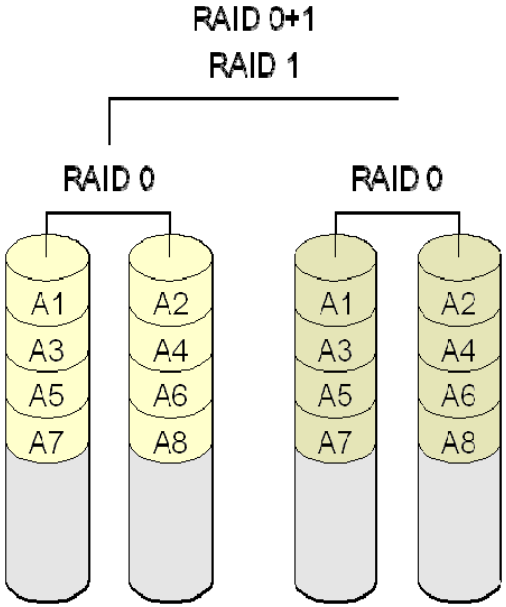


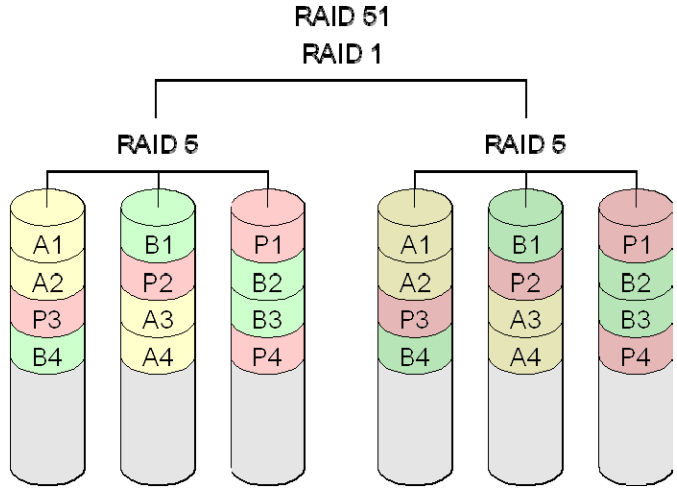
Architecture sécuritaire et très performante mais

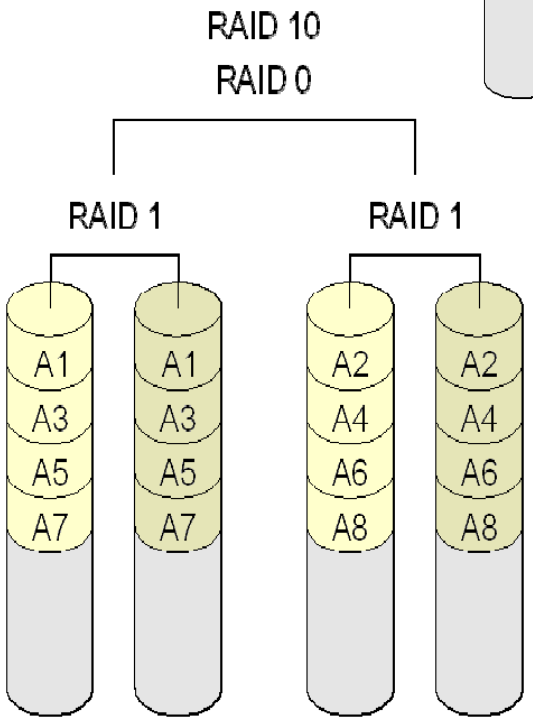
- ne supporte la perte que d'un seul disque.
- pénalité en écriture du fait du calcul de la parité
- pas aussi sécuritaire qu'on le suppose.

Le RAID

Combinaisons possibles des niveaux de RAID :







Le RAID

Le RAID ne dispense pas d'effectuer des sauvegardes régulières. En effet, des défaillances à plusieurs disques sont fréquentes car :

- Généralement les disques s'usent simultanément, ils vont donc tomber en panne a peu près en même temps.

- Les disques ne sont pas lus en totalité. Conséquence ils peuvent avoir des secteurs défectueux non détectés. Lors le défaillance d'un disque, le contrôleur RAID va parcourir l'ensemble et découvrir les secteurs défectueux. Le disque ayant des secteurs défectueux sera alors considéré par le contrôleur RAID comme HS, ce qui revient a dire qu'il y aura plus d'un disque en panne et donc impossibilité de reconstruire la grappe.

- Le RAID ne protège pas contre : les surtensions, les incendies, les inondations, les fichiers supprimés par inadvertance.

Les Baies de Disques

Dans les systèmes professionnels, l'utilisation de baies de disques est extrêmement courante.

oCes baies possèdent certaines caractéristiques :

o Plusieurs chemins d'accès afin de pouvoir partager des disques entre machines

o Alimentations redondantes

o Technologie RAID matériel embarquée dans les baies

o Des interfaces variées : SCSI => Fibre Channel

o Placement des data-center (hors des couloirs aériens)

Les Baies de Disques

Les Baies de Disques

Les baies de grande capacité peuvent être partagées selon 2 techniques :

- Par réseau: les NAS (Network Attached Storage)

- Par des machines se connectant sur le réseau de disques : les SAN (Storage Area Network).

Le NAS

Le **serveur** NAS a pour vocation d'être accessible depuis des serveurs à travers le réseau pour y stocker des données.

Le composant informatique principal de ce type de serveur est le disque dur. L'interface SCSI, IDE, SAS, SATA ou Fibre Channel utilisé est choisie en fonction du rapport coût/performance recherché.

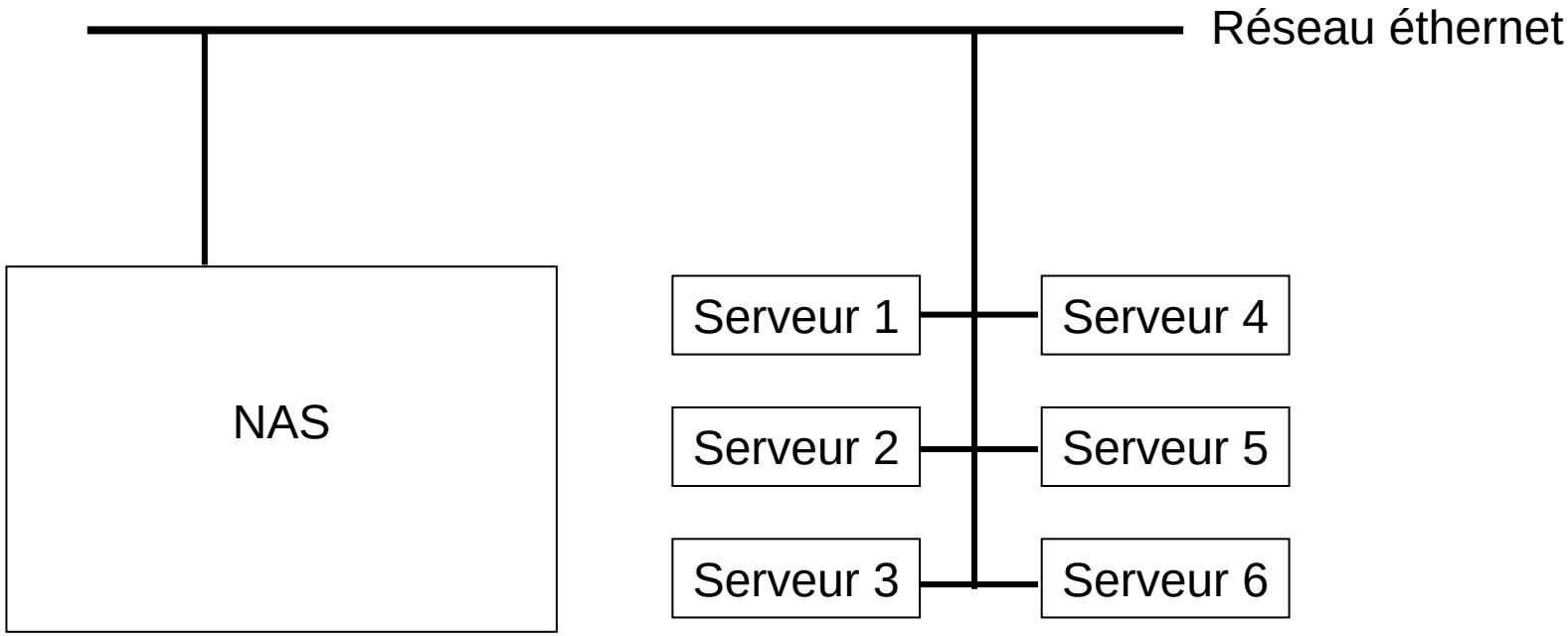
La technologie RAID est employée pour sécuriser les données stockées contre la défaillance d'un (ou plusieurs) disques.

Le NAS intègre un serveur de fichier de type :

- NFS
- CIFS (disques Windows)
- AFP (AppleShare File Protocol).

Des outils permettent de se connecter au firmware du NAS afin de l'administrer.

Le NAS


Réseau éthernet
NAS
Serveur 1
Serveur 2
Serveur 3
Serveur 4
Serveur 5
Serveur 6

Le SAN

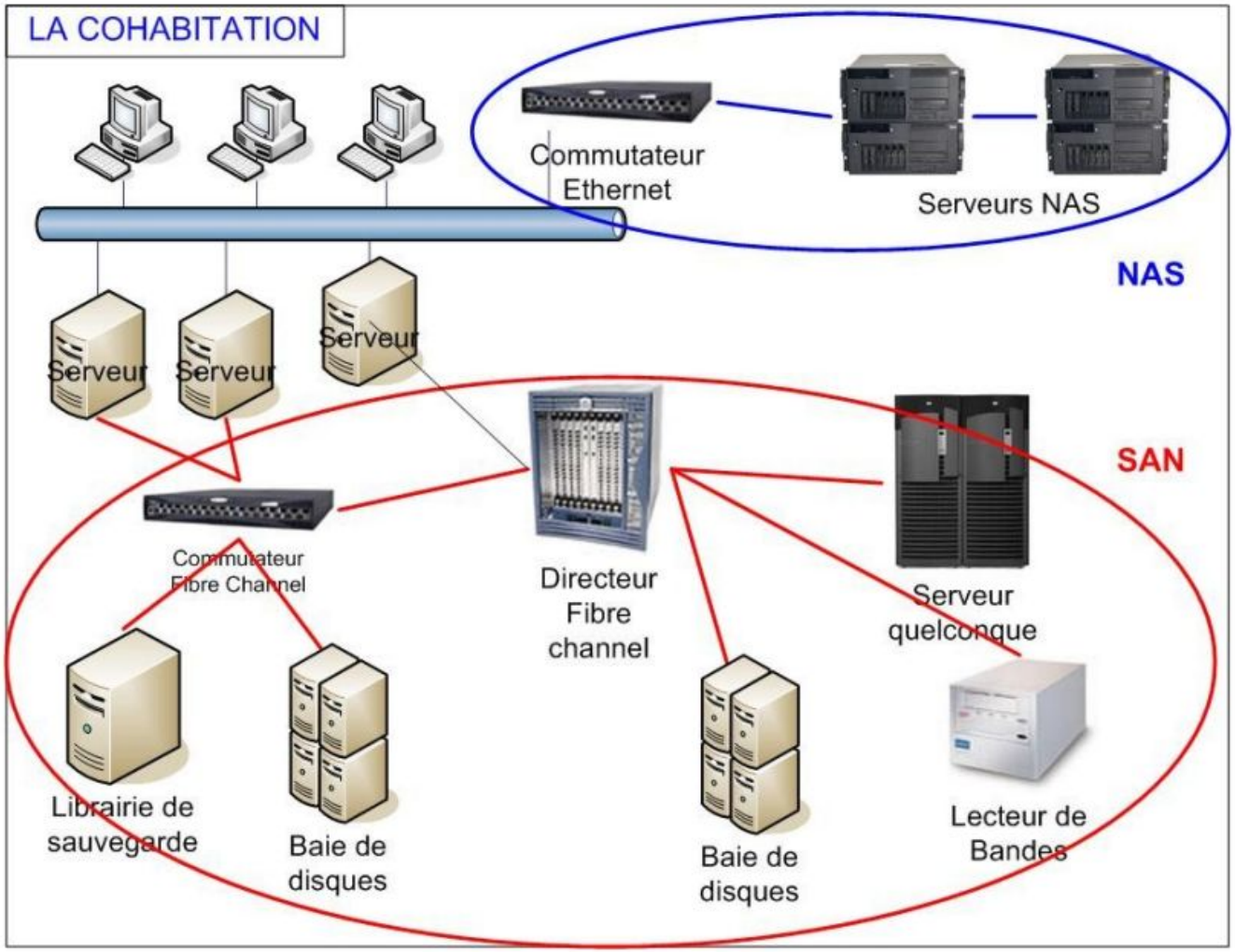
Un SAN se distingue des autres systèmes de stockage par un accès bas niveau aux disques. C'est une mutualisation des ressources de stockage.

Dans le cas du SAN, les baies de stockage n'apparaissent pas comme des volumes partagés sur le réseau. Elles sont directement accessibles en mode bloc par les systèmes de fichiers des serveurs. Chaque serveur voit l'espace disque d'une baie SAN à laquelle il a accès comme celui de ses propres disques durs.

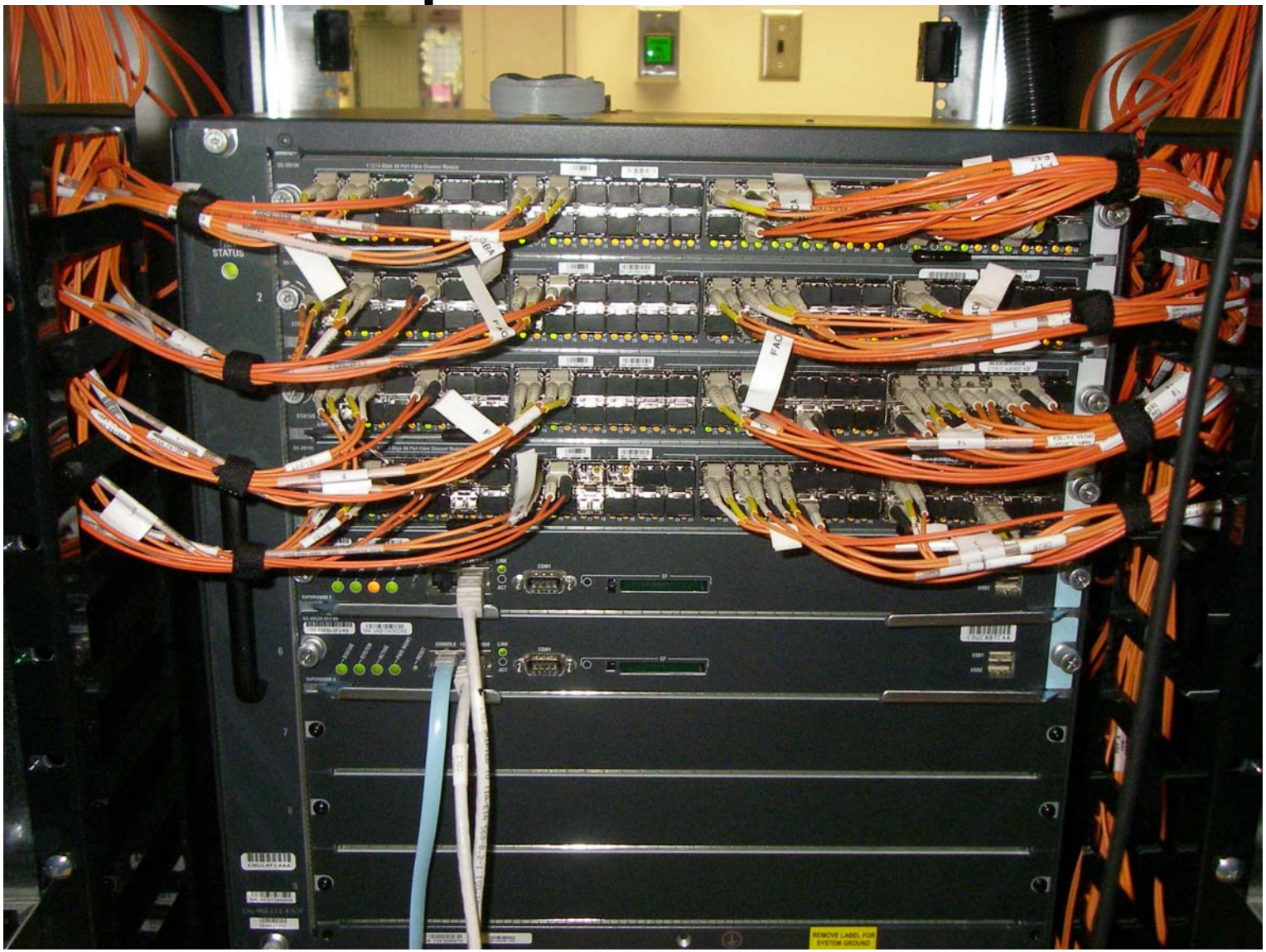
L'administrateur doit donc définir très précisément les LUNs (unités logiques), le masking et le zoning pour qu'un serveur unix n'accède pas aux mêmes ressources qu'un serveur Windows car les deux utilisent des systèmes de fichiers différents.

LUN
Masking
Zoning

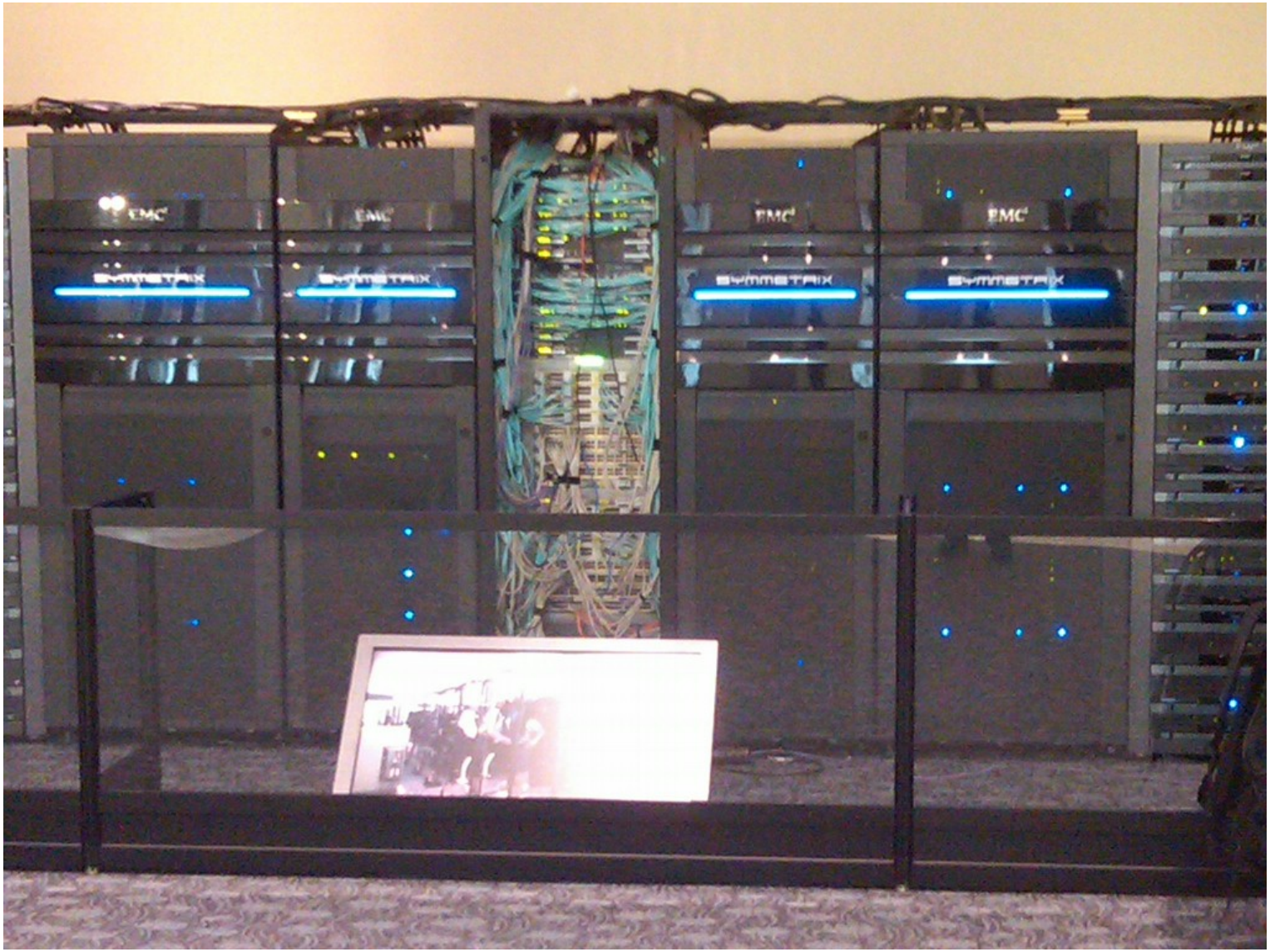
Le SAN


LA COHABITATION
Commutateur Ethernet
Serveurs NAS
NAS
Serveur
Serveur
Serveur
SAN
Commutateur Fibre Channel
Directeur Fibre channel
Serveur quelconque
Librairie de sauvegarde
Baie de disques
Baie de disques
Lecteur de Bandes

Les baies de disques : Le Connectrix EMC



Symmétrix EMC



Symmétrix EMC

Symmétrix EMC

La sauvegarde

Pour sauvegarde la volumétrie des baies, il est nécessaire d'utiliser des robots de sauvegarde et d'automatiser au maximum la gestion des sauvegardes.

Les robots de sauvegarde sont des baies remplies de cartouches magnétiques sur lesquelles sont écrites les données des baies. Une bonne technique consiste a extraire certaines cartouches contenant les données essentielles afin de les stocker sur un site différent et ceci à intervalle de temps réguliers.

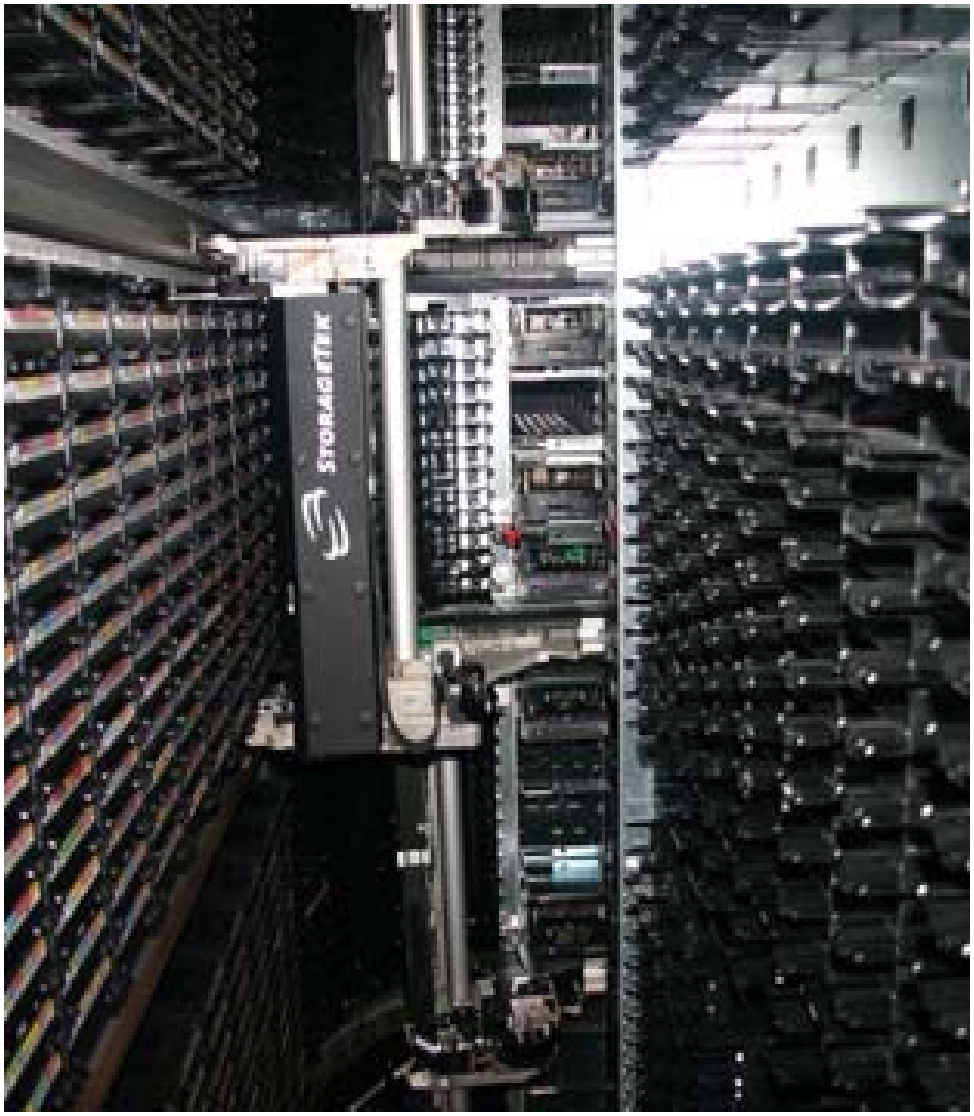
La gestion des sauvegardes s'effectue par un serveur qui pilote le robot de sauvegarde.

Cela permet la gestion des cartouches (par reconnaissance des codes barre collés dessus) en répartissant les cartouches par pool (lot), chaque pool pouvant dédié à certaines applications et ayant des périodes de rétentions différentes...etc.

La sauvegarde

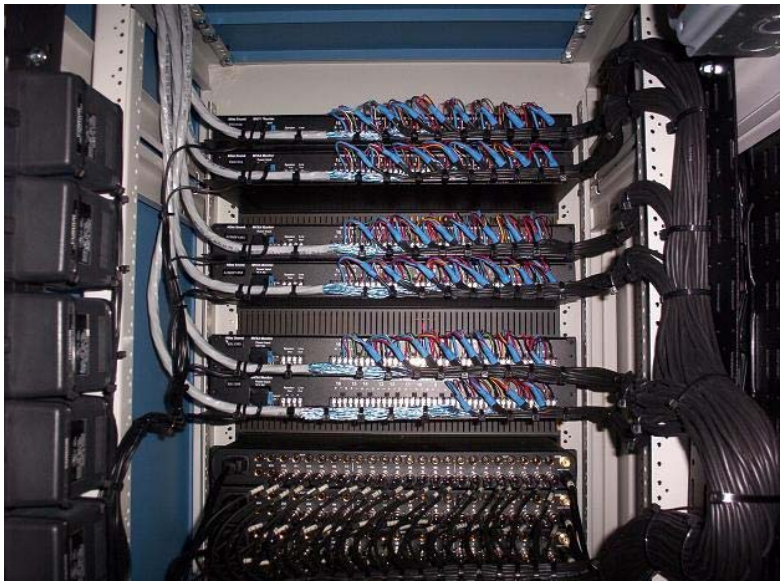
La sauvegarde

La sauvegarde



Une partie de l'intérieur du robot SL8500 (Sun-StorageTek), avec les rayonnages des cartouches et 2 des 4 robots indépendants traitant les cartouches.

- Capacité : de 3 000 à 10 000 cartouches
- Il peut recevoir 64 drives (unités Lecture/écriture).
- Permet des transfert de débit des 120 Mb/s
- Chaque cartouche peut stocker 500 GB non compressé.



Système d'un opérateur télécom

Fonctionnement du système de facturation :

Les données binaires téléphoniques (numéro appelant, numéro appelé, durée de la communication, ...) sont stockés sur des centraux téléphoniques.

A intervalles de temps réguliers, une application va chercher ces données. Elle les transforme dans un format texte et les stocke dans une base de données.

D'autres applications vont utiliser les données de la base afin :

- de créer les factures pour les clients de la société.
- de transférer ces données aux partenaires de la société.
- de faire des études sur l'utilisation du réseau de télécom de la société afin de mieux le gérer.

Vue la volumétrie des données a traiter, les traitements applicatifs peuvent prendre plusieurs heures, voir plusieurs jours.

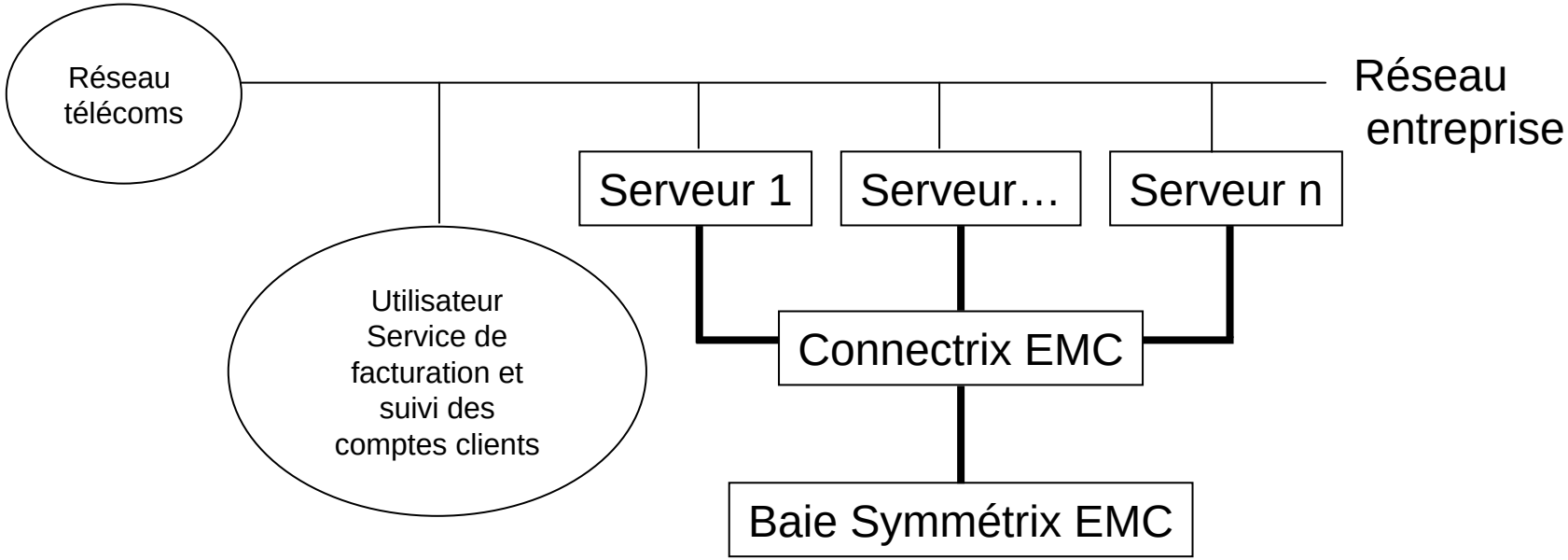
Système d'un opérateur télécom

Architecture matériel du système de facturation :

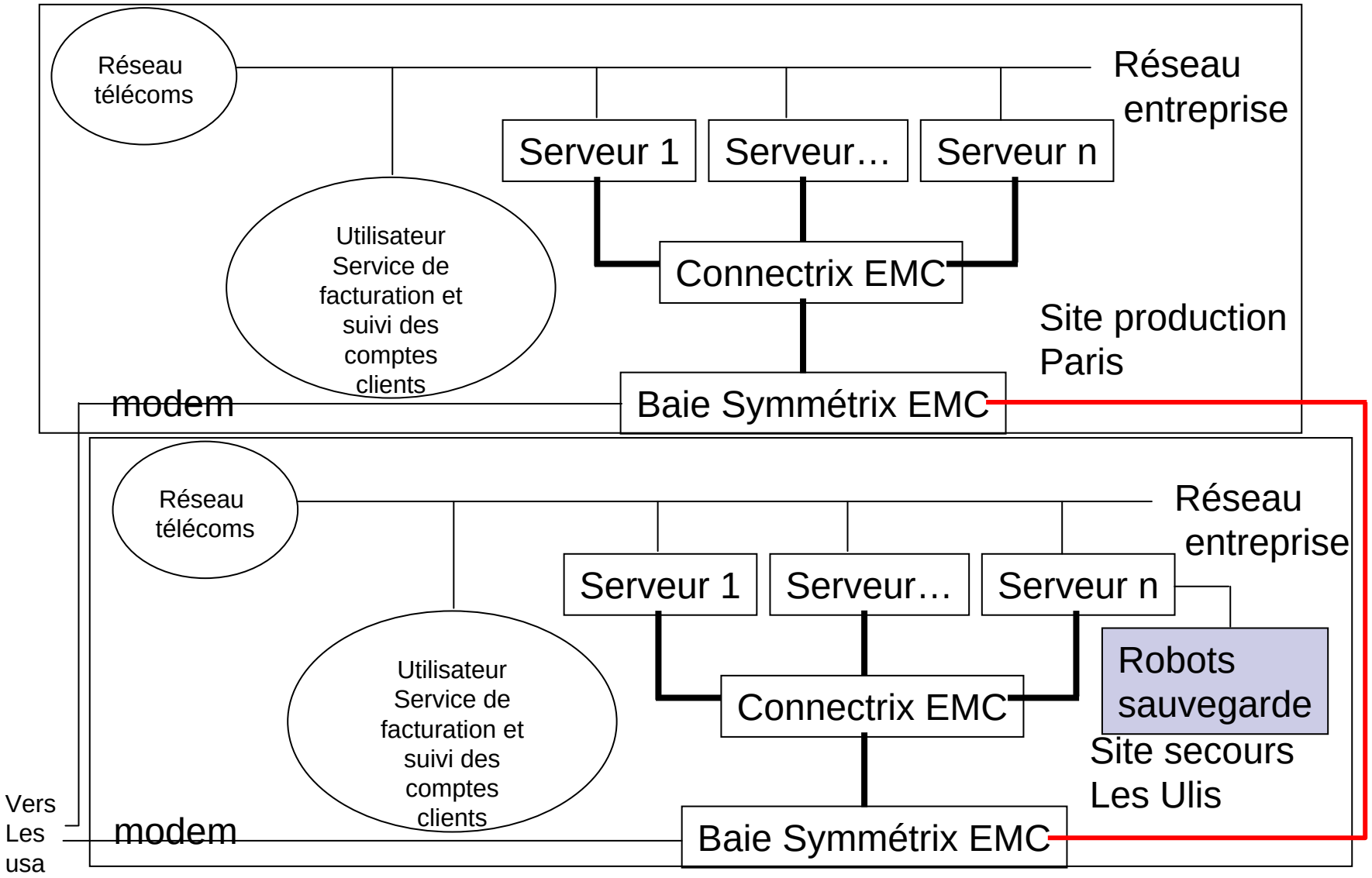
2 centres : Paris & les Ulis.

Chaque centre possède les mêmes serveurs et sur chaque centre on installe une baie symmétix EMC relié par une fibre optique (SRDF).

Les applications sont installées dans chacun des centres. Le site principal est Paris (centre de production), le site des Ulis est le site de secours et de tests de nouvelles applications.



Système d'un opérateur télécom



Système d'un opérateur télécom

Le lien SRDF entre les deux permet une synchronisation des disques d'une baie vers l'autre, ou dans le sens inverse.

Dans l'utilisation faite, la synchronisation s'effectue depuis le centre de production (Paris) vers le centre de secours (Les Ulis).

La baie de secours, outre les disques de réplication, contient des disques spécifiques R2 pouvant être synchronisés a souhait avec les disques de réplication.

Ce qui permet d'effectuer des tests sans mettre en danger la réplication, ou d'effectuer des sauvegardes lourdes des bases.

Chaque baie possède un modem destiné a avertir le centre de surveillance EMC en cas de détection d'une anomalie. Si un tel cas se présente, le centre de surveillance peut se connecter sur la baie via le modem et vérifier s'il faut l'intervention d'un technicien afin de changer le disque, carte ou l'alimentation défectueuse.

Outre la réplique des serveurs de production, le centre des Ulis possède plusieurs serveurs de tests, ainsi qu'un serveur de sauvegarde pilotant un robot.