Haute école d'ingénierie et d'architecture Fribourg
Hochschule für Technik und Architektur Freiburg

Systèmes d'exploitation

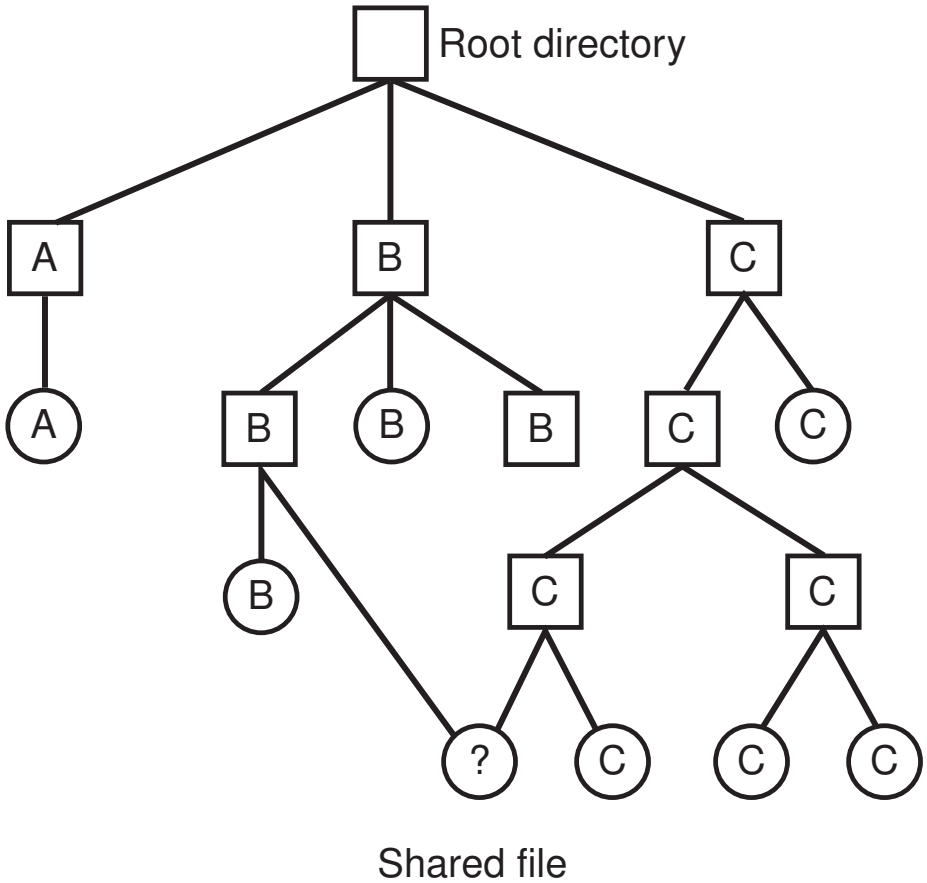
Disques / Systèmes de fichiers

Jacques Supcik | 2018/2019 | T-2adfg

l09-handouts | 2018-12-13

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz

Les fichiers partagés



Système de fichiers contenant un fichier partagé.

Les fichiers partagés

- Avec des fichiers partagés, la structure du répertoire devient un graphe orienté acyclique («directed acyclic graph» ou simplement «DAG»).
- Ca ne fonctionne pas si c'est le «directory» qui contient les attributs du fichiers (tels que la taille ou l'heure de la modification).
- Ca fonctionne très bien avec des i-nodes.

Les fichiers partagés

Avec UNIX, nous avons deux méthodes pour partager des fichiers :

- Les «hard links» (ou liens matériels).
- Les «symbolic links» (ou liens symboliques).

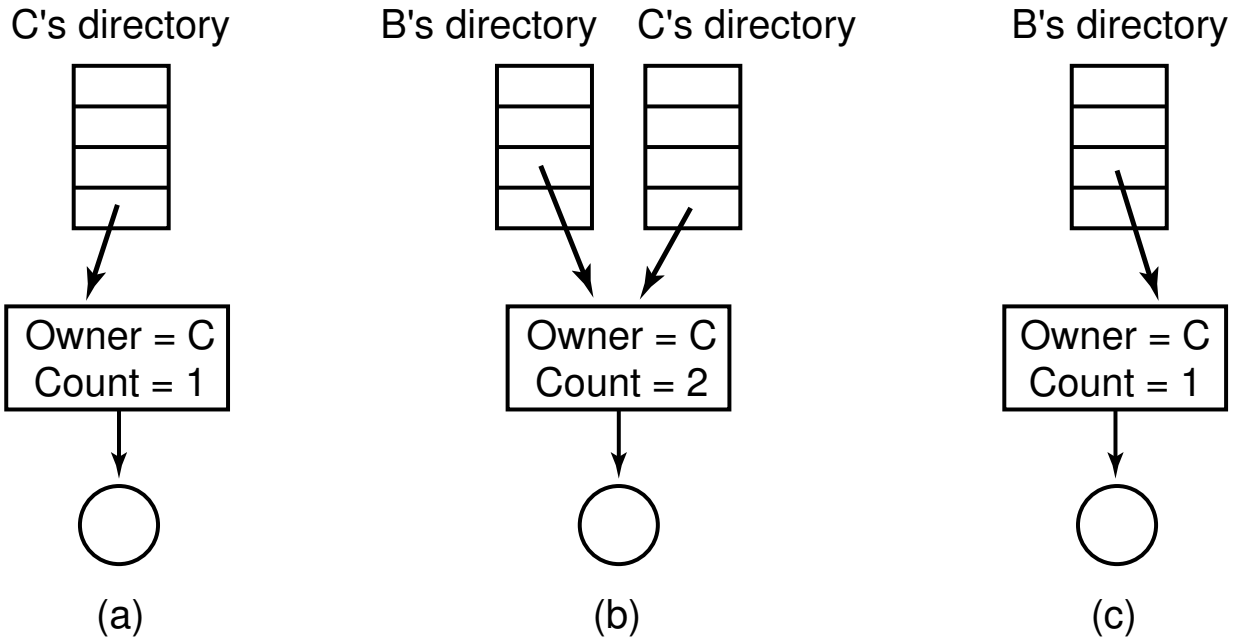
- Commande UNIX «`ln`».
- Le système compte les références vers un fichier donné (visible avec la commande «`ls -l`»).

```
$ ls -l
total 4
-rw-r--r-- 1 user user 12 Dec 12 21:02 file.txt
```

```
$ ln file.txt link.txt
$ ls -l
total 8
-rw-r--r-- 2 user user 12 Dec 12 21:02 file.txt
-rw-r--r-- 2 user user 12 Dec 12 21:02 link.txt
```

Les fichiers partagés

Les liens matériels – Problème avec les quotas



Le propriétaire est stocké dans l'i-node.

(a) Situation avant la création du lien.

(b) Après la création du lien.

(c) Après l'effacement du fichier par le propriétaire.

- Commande UNIX «`ln -s`».
- Plus lent les les liens physiques.
- Risque des liens «flottants» (dangling pointers).
- Nécessite un i-node supplémentaire.
- Permet des liens vers d'autres systèmes de fichiers.

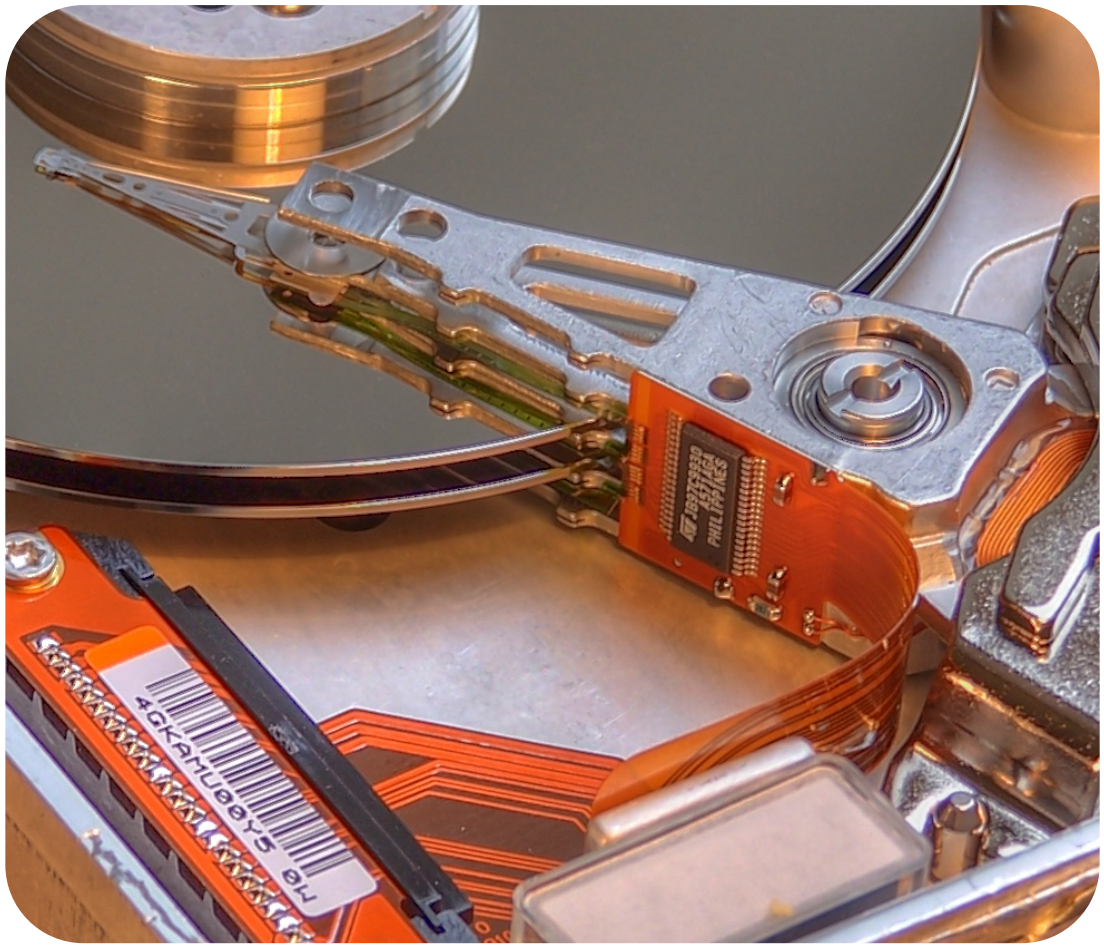
Avec plus de RAM, on peut en utiliser pour implémenter des «caches» et limiter les accès au disque en lecture. Mais que peut-on faire pour optimiser l'écriture ?
Opérations nécessaire pour créer un fichier sous UNIX :

- Écrire l'i-node pour le répertoire.
- Écrire le répertoire.
- Écrire l'i-node pour le fichier.
- Écrire le fichier.

Ca fait beaucoup d'opérations en écriture.

Les systèmes de fichiers LFS

Construction d'un disque



Momentus® XT SSHD

Seagate®

Specifications	750GB
Model Number	ST750LX003
NAND Type/Size	SLC/8GB
Interface	SATA 6Gb/s NCQ
Performance	
Spindle Speed (RPM)	7200
Cache, Multisegmented (MB)	32
SATA Transfer Rates Supported (Gb/s)	6.0/3.0/1.5
Seek Average, Read (ms)	11.0
Seek Average, Write (ms)	13.0
Configuration/Organization	
Heads/Disks	4/2
Bytes per Sector	4096

1 One gigabyte, or GB, equals one billion bytes and one terabyte, or TB, equals one trillion bytes when referring to drive capacity.

Les systèmes de fichiers LFS

En 20 ans, les PC on évolué de la manière suivante :

- La vitesse des CPUs a été multipliée par 100.
- La quantité de RAM a été multipliée par 500.
- La capacité des disques a été multipliée par 1'000.
- Le prix à été divisé par 200.

Mais la vitesse de rotation à peu évolué et le «seek time» est pratiquement resté le même ! Ce sont ces paramètres qui pénalisent la vitesse d'écriture.

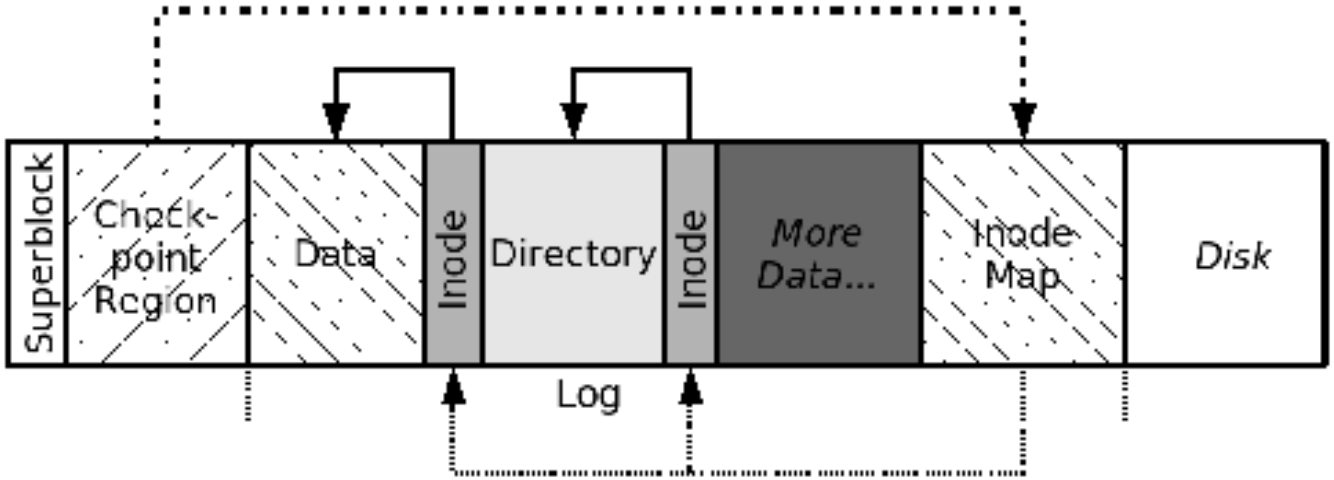
Les systèmes de fichiers LFS

L'écriture d'un bloc sur le disque prend typiquement $50\ \mu s$. Mais cette écriture est précédée du «seek time» ($13\ ms$) et de la «latency» ($4\ ms$)

Un disque tourne à 7200 RPM (tour par minute)
$\implies$ *pour un demi-tour, il faut* $\frac{60}{7200 \cdot 2} \approx 4\ ms$)

L'efficacité de l'écriture est donc de $\frac{50\ \mu s}{17\ ms} \approx 0.3\%$

Les systèmes de fichiers LFS

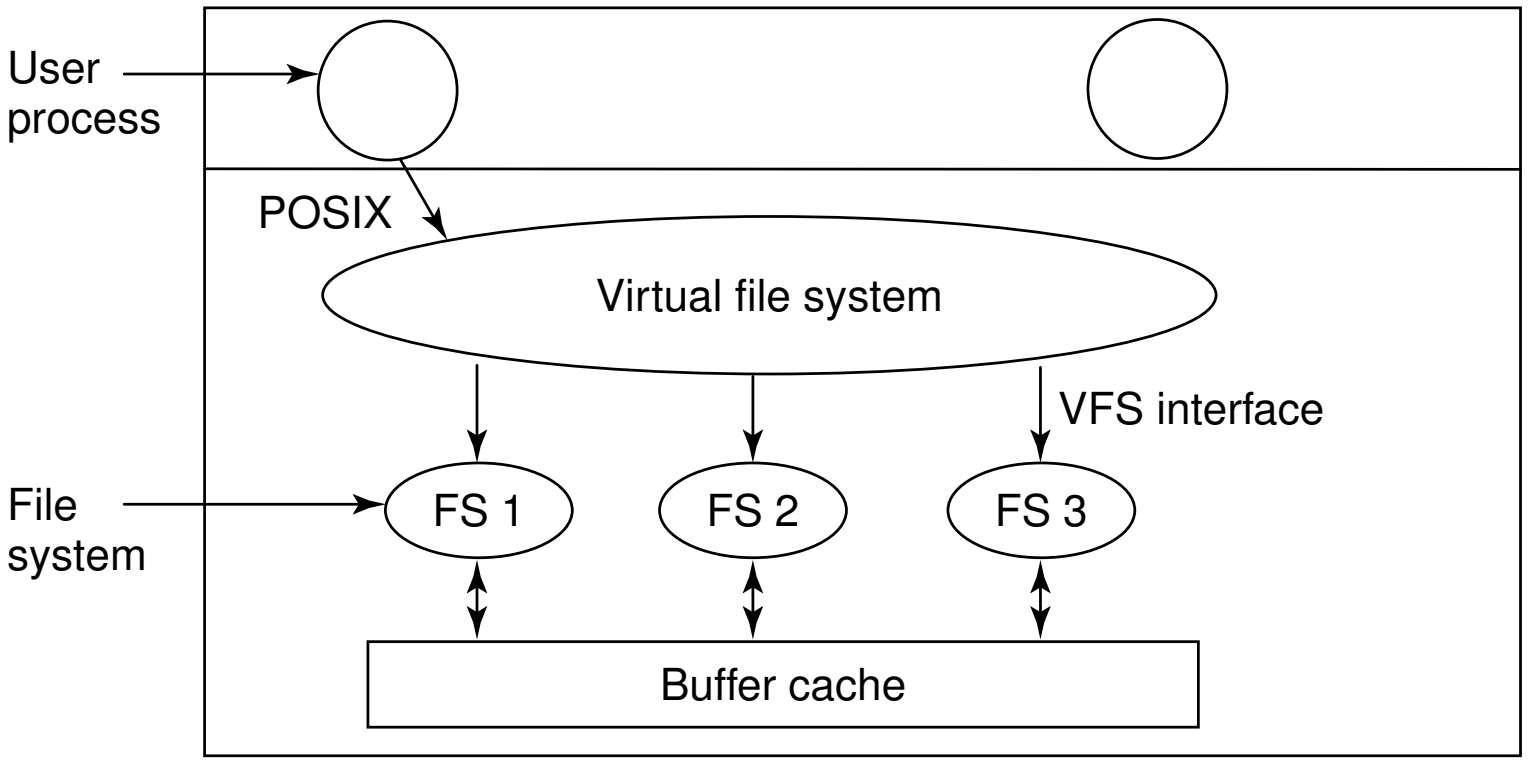


L'idée du LFS est d'améliorer l'efficacité en sérialiser les opérations d'écriture. Les i-node sont écrits entre les blocs de données et ne sont plus à des endroits fixe. L'inconvénient de cette méthode est qu'elle nécessite un «mapping» des i-nodes et qu'elle est plus compliquée à implémenter.

Les systèmes de fichiers journalisés

- Les opérations sont écrites séquentiellement dans un journal.
- La mise-à-jour du système de fichier se fait par la suite.
- Ce système en deux phases augmente la fiabilité du système en cas de panne de courant.
- Par contre, la performance n'est pas améliorée (les données sont écrites 2 fois).

Les systèmes de fichiers virtuels



Position du système de fichiers virtuel.

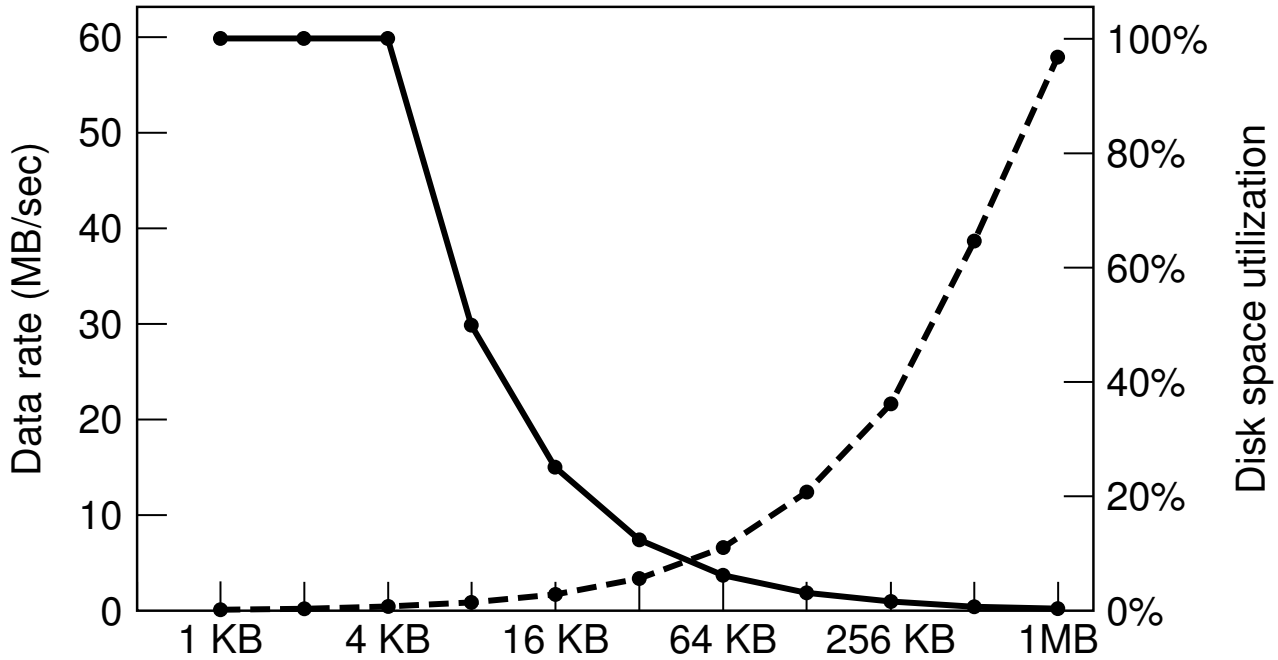
Length	VU 1984	VU 2005	Web
1	1.79	1.38	6.67
2	1.88	1.53	7.67
4	2.01	1.65	8.33
8	2.31	1.80	11.30
16	3.32	2.15	11.46
32	5.13	3.15	12.33
64	8.71	4.98	26.10
128	14.73	8.03	28.49
256	23.09	13.29	32.10
512	34.44	20.62	39.94
1 KB	48.05	30.91	47.82
2 KB	60.87	46.09	59.44
4 KB	75.31	59.13	70.64
8 KB	84.97	69.96	79.69

Length	VU 1984	VU 2005	Web
16 KB	92.53	78.92	86.79
32 KB	97.21	85.87	91.65
64 KB	99.18	90.84	94.80
128 KB	99.84	93.73	96.93
256 KB	99.96	96.12	98.48
512 KB	100.00	97.73	98.99
1 MB	100.00	98.87	99.62
2 MB	100.00	99.44	99.80
4 MB	100.00	99.71	99.87
8 MB	100.00	99.86	99.94
16 MB	100.00	99.94	99.97
32 MB	100.00	99.97	99.99
64 MB	100.00	99.99	99.99
128 MB	100.00	99.99	100.00

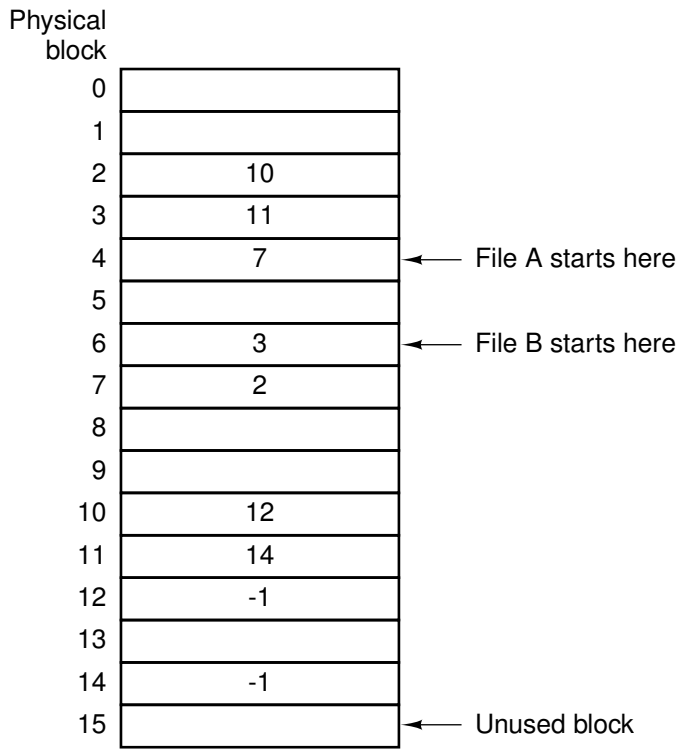
Pourcentage des fichiers qui ont une taille inférieure à une valeur donnée (en octet).

L'étude montre aussi qu'avec une taille de bloc de 4 Ko, 93% des blocs du disque sont utilisés par 10% des gros fichiers. $\Longrightarrow$ perdre un peu d'espace à la fin de chaque fichier n'a pas de grande importance.

Des petits blocs nécessitent plus de blocs par fichiers, donc plus de «seek time» et donc plus lent.

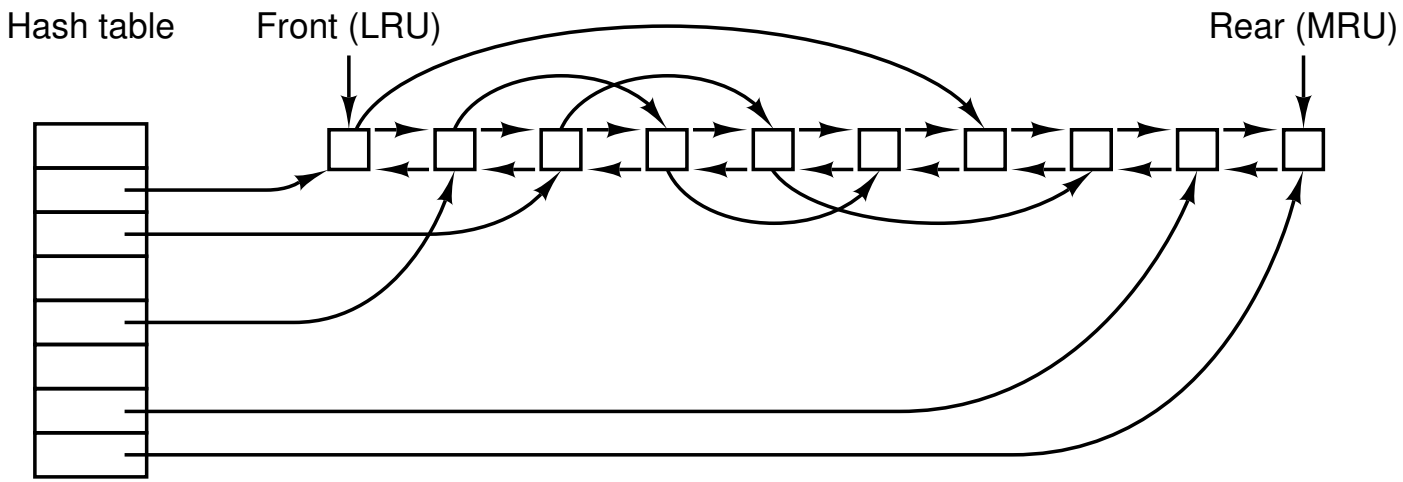


La courbe en pointillé (échelle de gauche) donne la vitesse de transfert. La courbe en trait continu (échelle de droite) indique le taux de remplissage. Tous les fichiers sont de 4 Ko.



(a) Conservation des blocs libres dans une liste chaînée.

(b) Une table de bits.

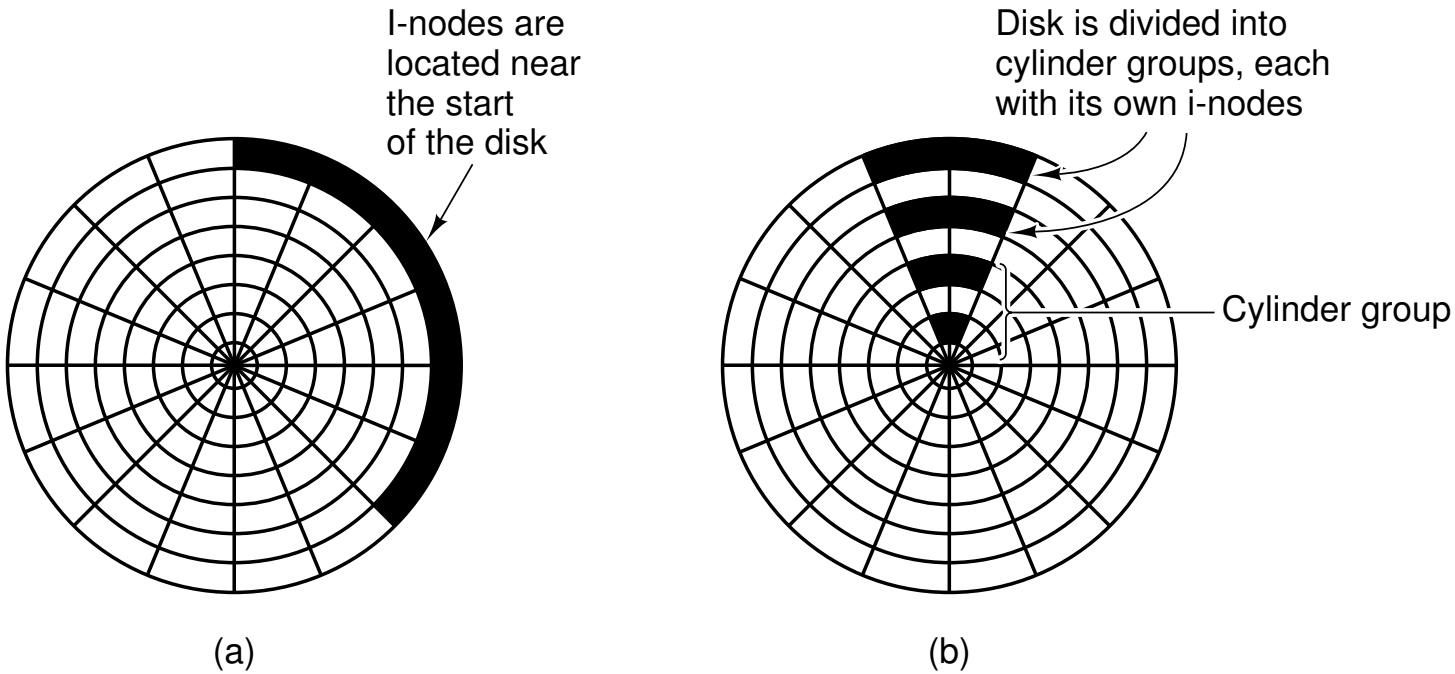


Structure de donnée d'une mémoire cache.

- Certains blocs, tels que les blocs d'i-node, sont rarement référencés deux fois de suite dans un intervalle de temps court.
- L'accès au cache est relativement peu fréquent (comparé au MMU de la mémoire virtuelle) et il est possible d'implémenter un vrai LRU.
- Considérons une modification de la méthode LRU pour prendre en compte les deux facteurs suivants :
  - Est-il probable que le bloc soit utilisé à nouveau rapidement ?
  - Le bloc est-il important pour la cohérence du système de fichiers ?
- Attention à l'intégrité des données («write through» / «write back»).

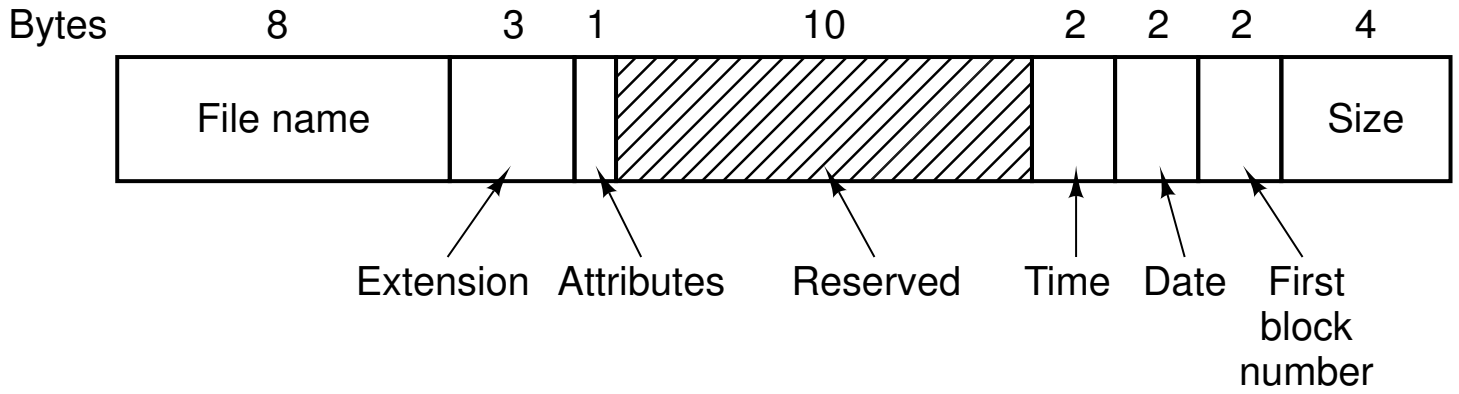
L'efficience du système de fichiers

Réduction du mouvement du bras d'un disque



- (a) Les i-nodes sont placés au début du disque.
- (b) Le disque est divisé en groupes de cylindres qui possèdent chacun leur propre blocs et i-nodes.

Le système de fichiers de MS-DOS



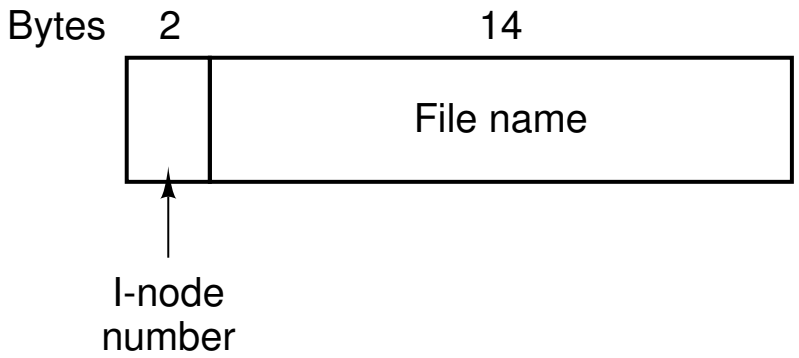
Entrée d'un répertoire MS-DOS

Le système de fichiers de MS-DOS

Block size	FAT-12	FAT-16	FAT-32
0.5 KB	2 MB		
1 KB	4 MB		
2 KB	8 MB	128 MB	
4 KB	16 MB	256 MB	1 TB
8 KB		512 MB	2 TB
16 KB		1024 MB	2 TB
32 KB		2048 MB	2 TB

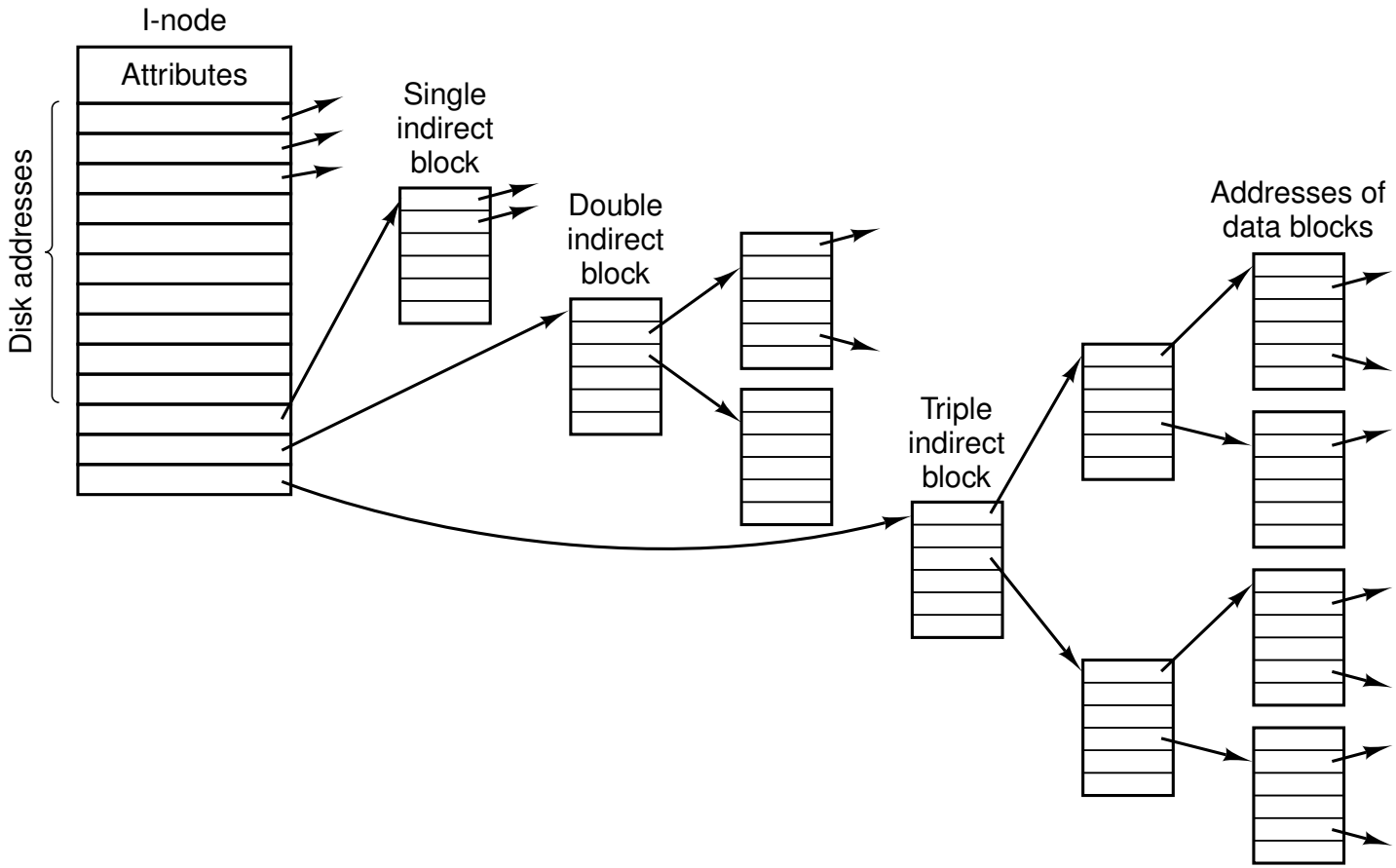
Tailles de partition maximales pour différentes tailles de blocs. Les emplacements vides représentent des combinaisons interdites.

Le système de fichiers d'UNIX V7



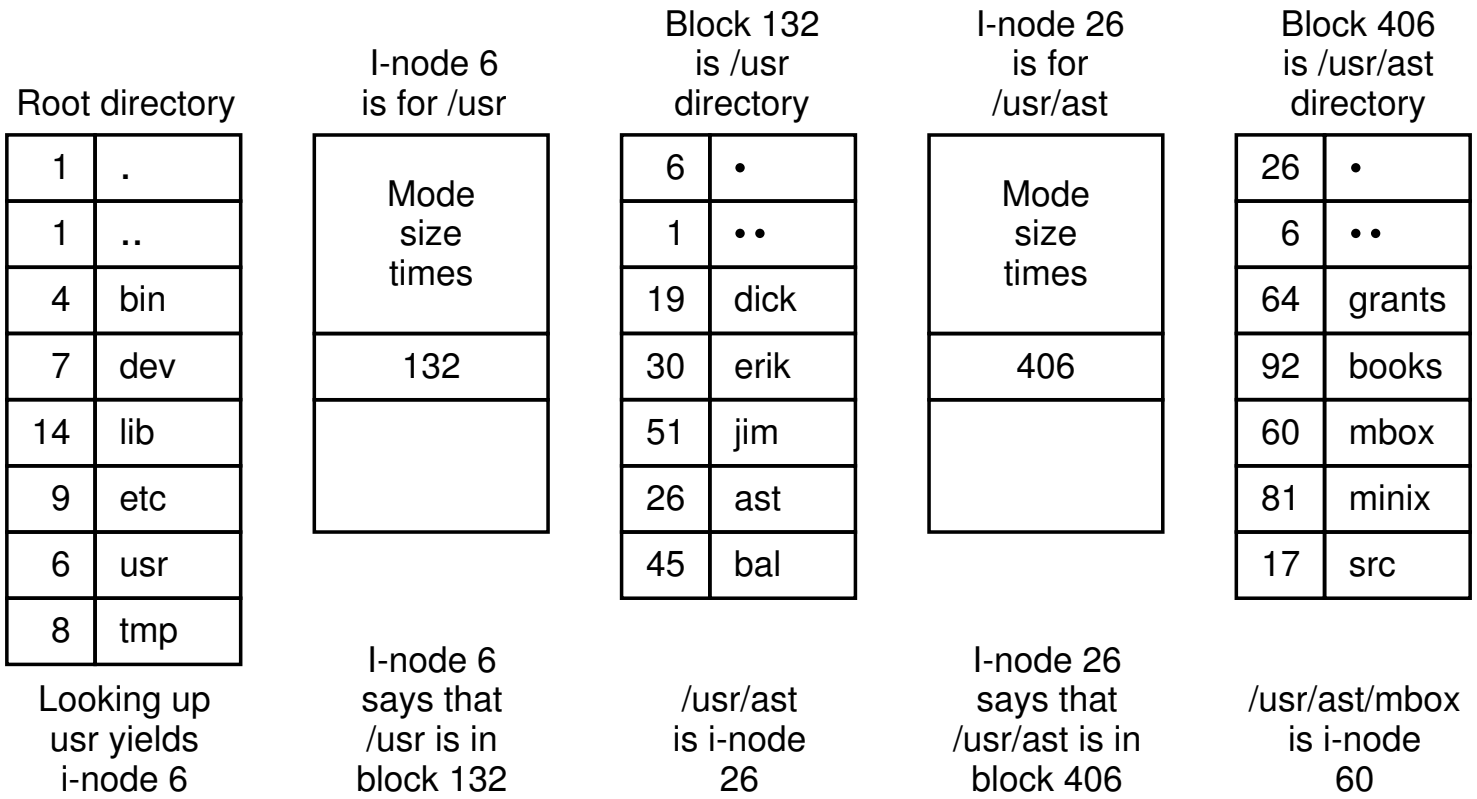
Entrée d'un répertoire d'UNIX V7

Le système de fichiers d'UNIX V7



Un i-node d'UNIX

Le système de fichiers d'UNIX V7



Déroulement de la recherche de `/usr/ast/mbox`

Les Sauvegardes

Il y a deux sortes d'individus sur terre :

- Ceux qui font des sauvegardes de leur données.
- Ceux qui vont, un jour, perdre des données.

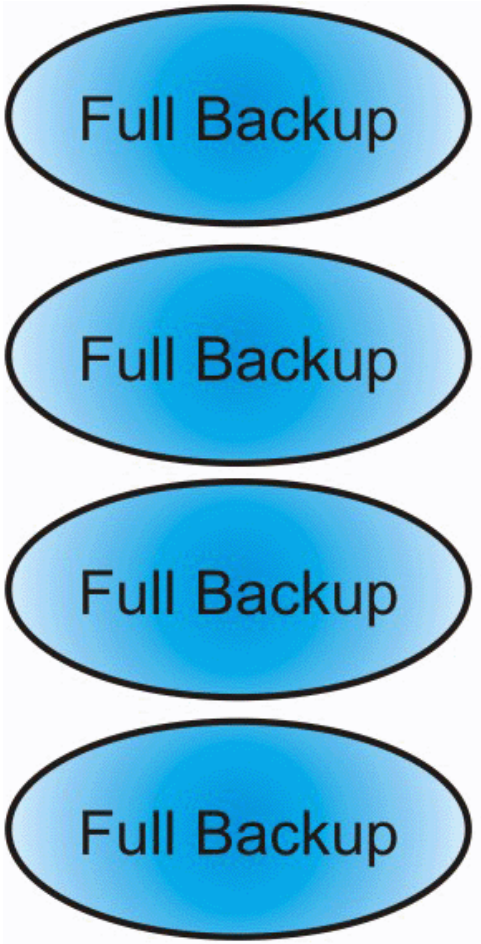
Les Sauvegardes

Les sauvegardes servent généralement à régler les deux problèmes potentiels suivants :

- Réparer un «désastre».
- Réparer une «bêtise».

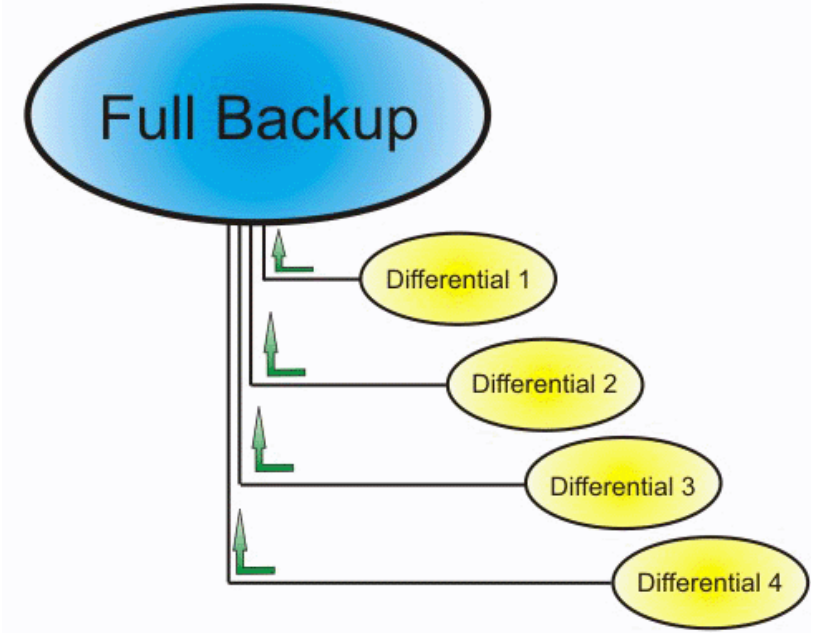
Les Sauvegardes

Le backup «complet» (Full Backup)



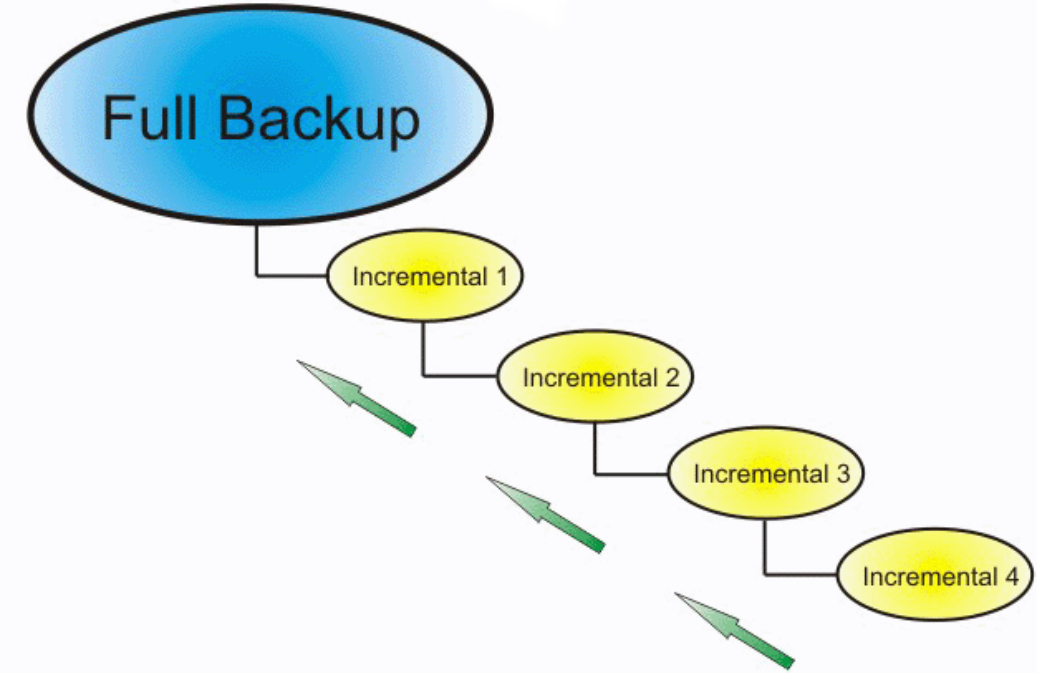
Les Sauvegardes

Le backup «différentiel» (Differential Backup)



Les Sauvegardes

Le backup «incrementiel» (Incremental Backup)

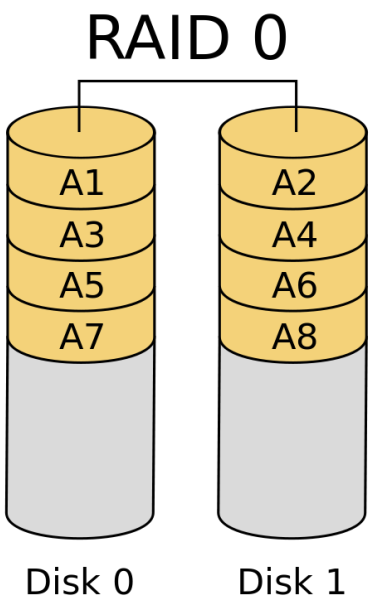


Les RAIDs

- RAID = «Redundant Array of Inexpensive Disks».
- L'industrie n'aime pas vraiment le concept de «bon marché» et RAID est aussi utilisé pour «Redundant Array of Independent Disks».
- Le contraire est le SLED (Single Large Expensive Disk).

Les RAIDs

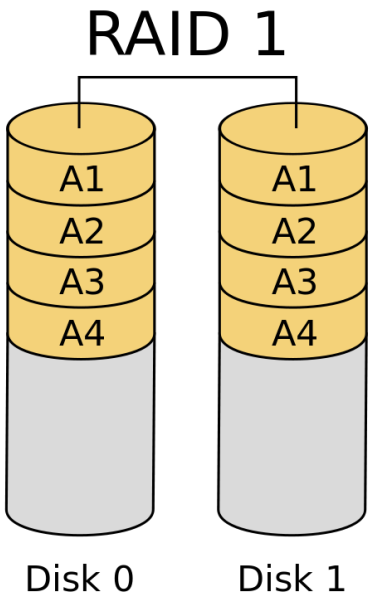
Raid Level 0 (striping)

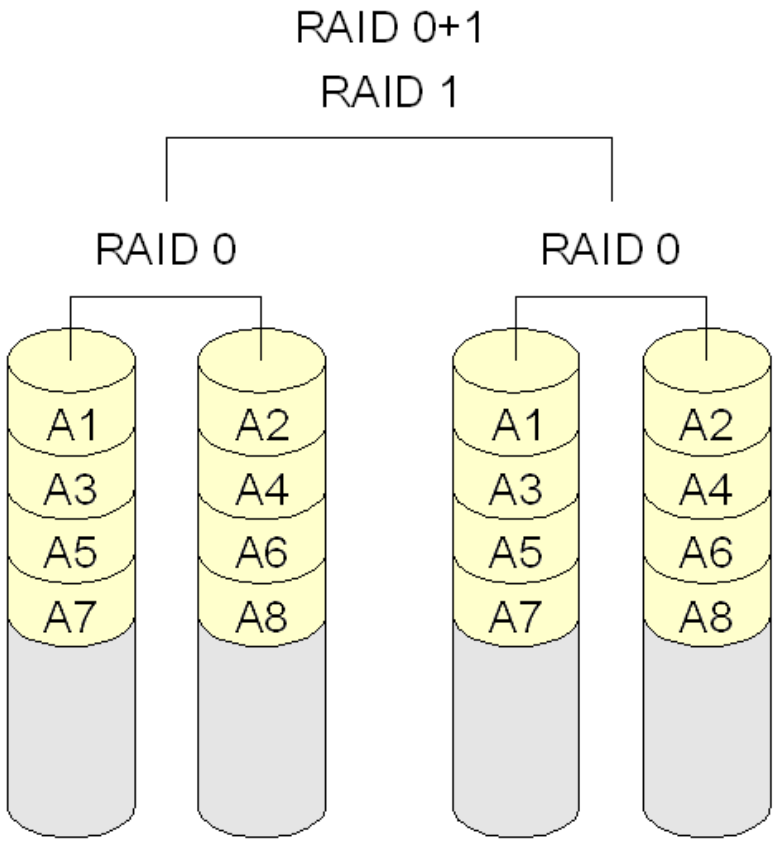


- «Stripe» = bloc = séquence de secteurs.
- Plus rapide pour les gros fichiers.
- Moins robuste car *n* fois plus de risque de panne.

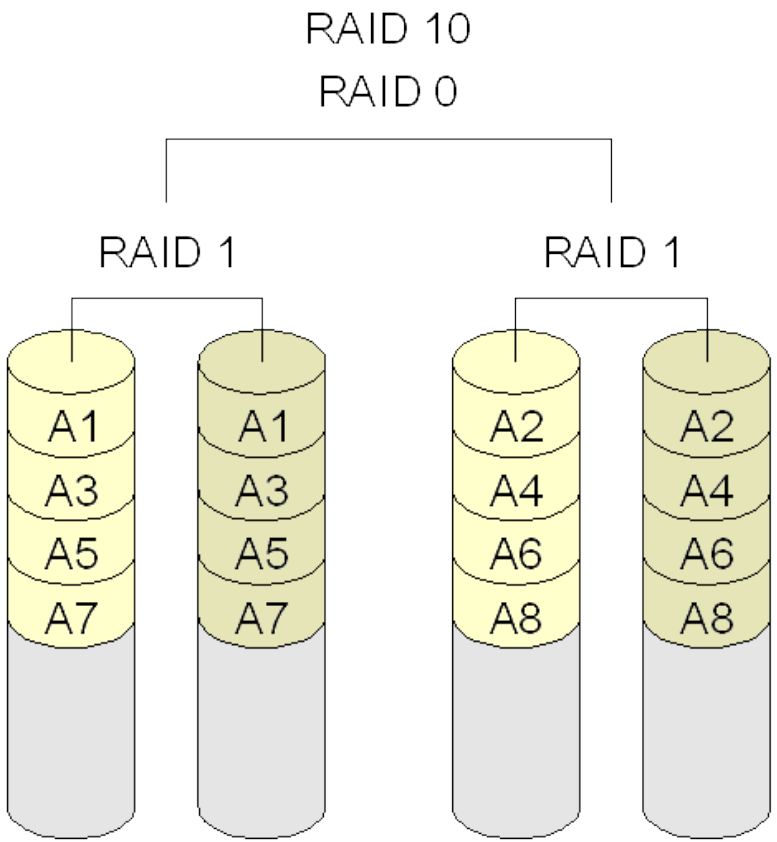
Les RAIDs

Raid Level 1 (mirroring)



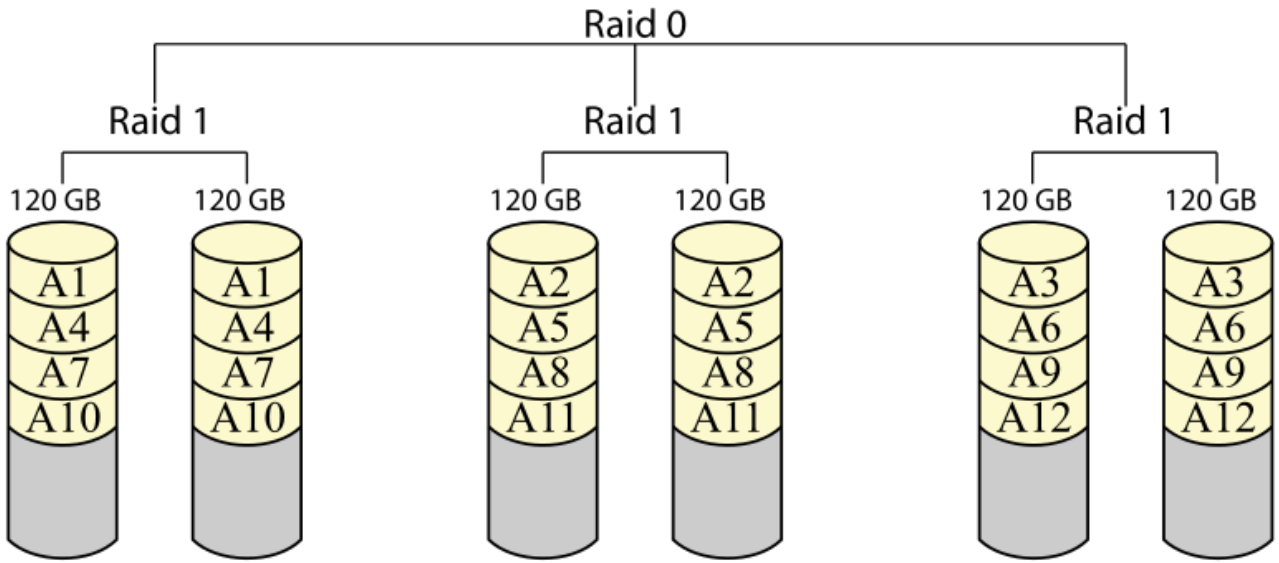


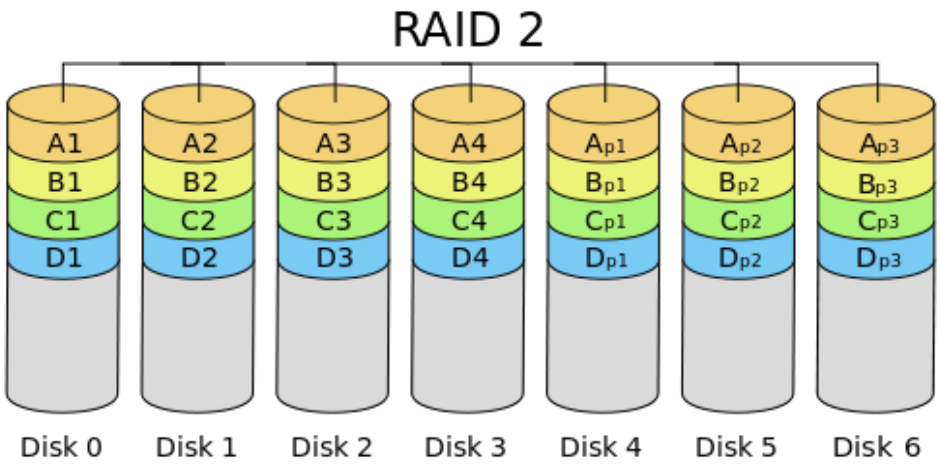
- Minimum 4 disques.


RAID 10
RAID 0
RAID 1
RAID 1
A1
A3
A5
A7
A1
A3
A5
A7
A2
A4
A6
A8
A2
A4
A6
A8

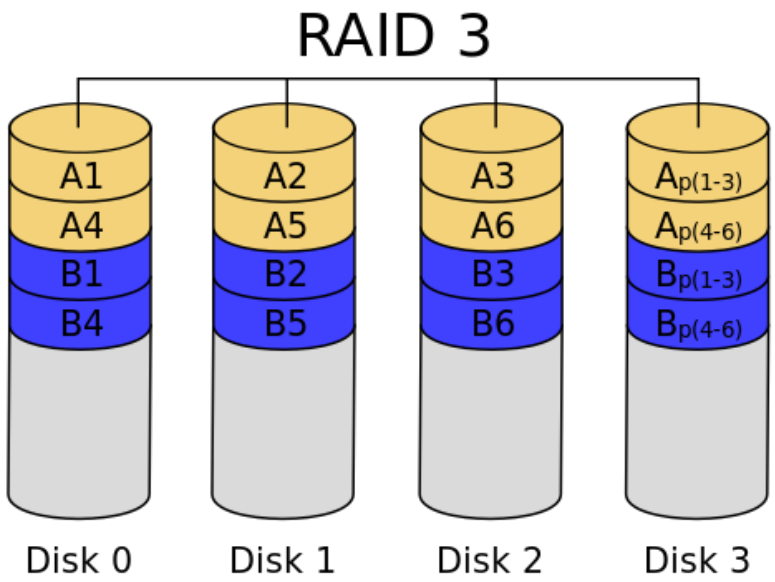
Les RAIDs

Raid Level 10

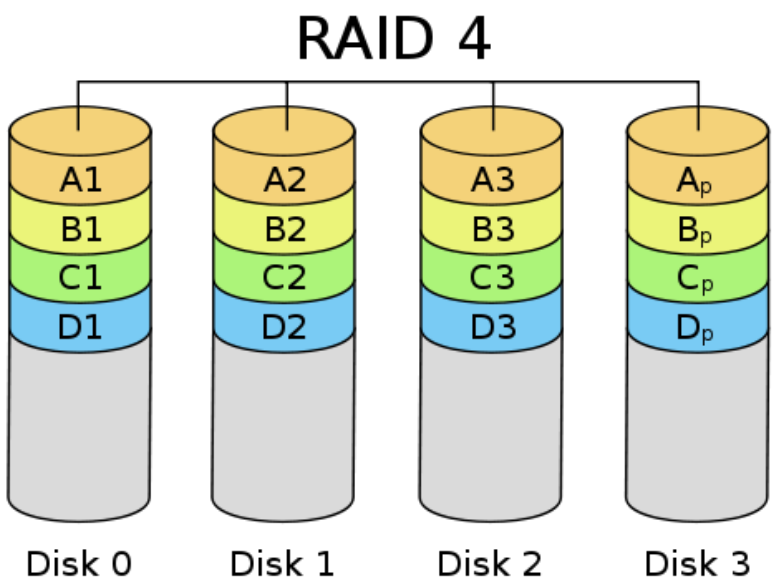




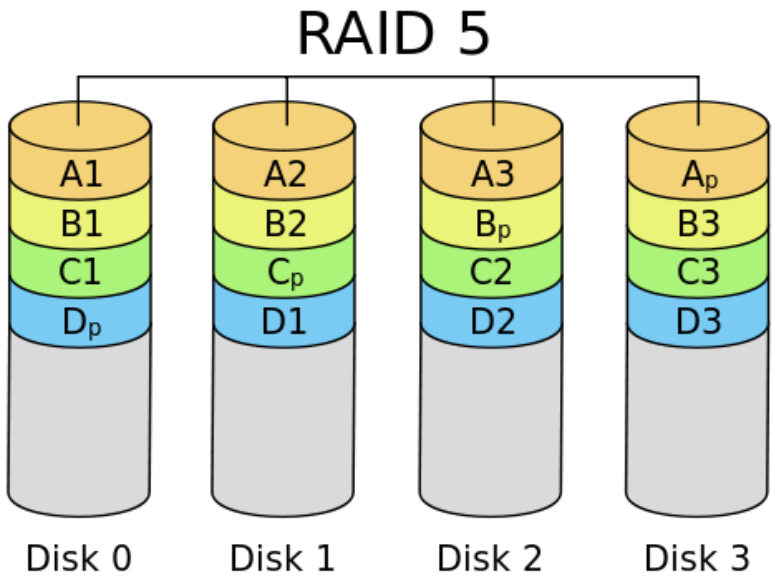
- Bit level striping.
- Correction des erreurs avec les «Hamming Code».
- Nécessite 7 disques (4 data disk + 3 error code disks).
- Plus de vitesse et plus de robustesse.
- Mais pas utilisé en pratique.



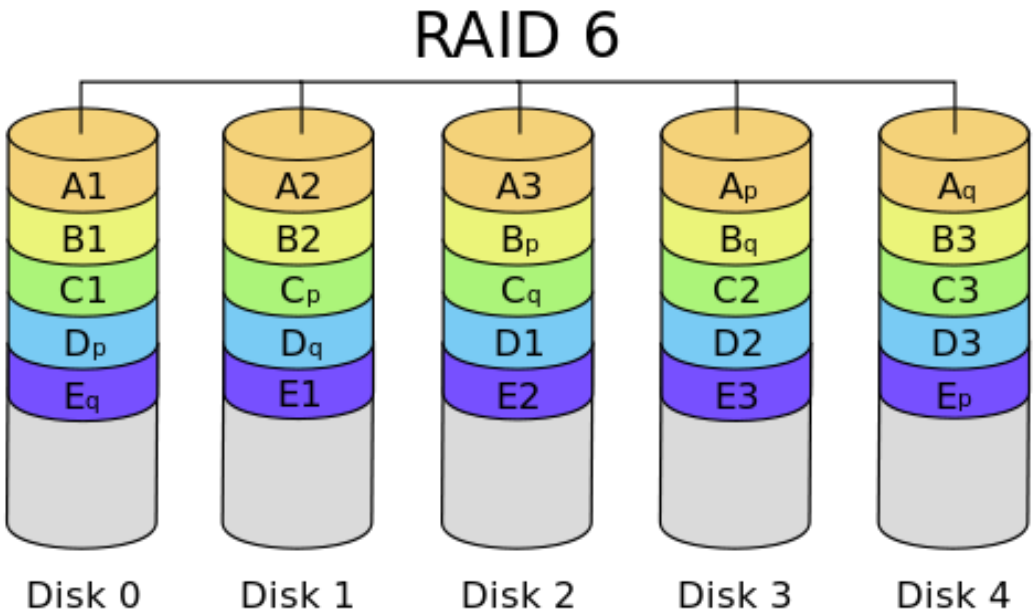
- Byte level striping.
- Nécessite des disques synchronisés.
- Peu utilisé en pratique.



- Bloc level striping.
- Disque dédié pour stocker la parité.



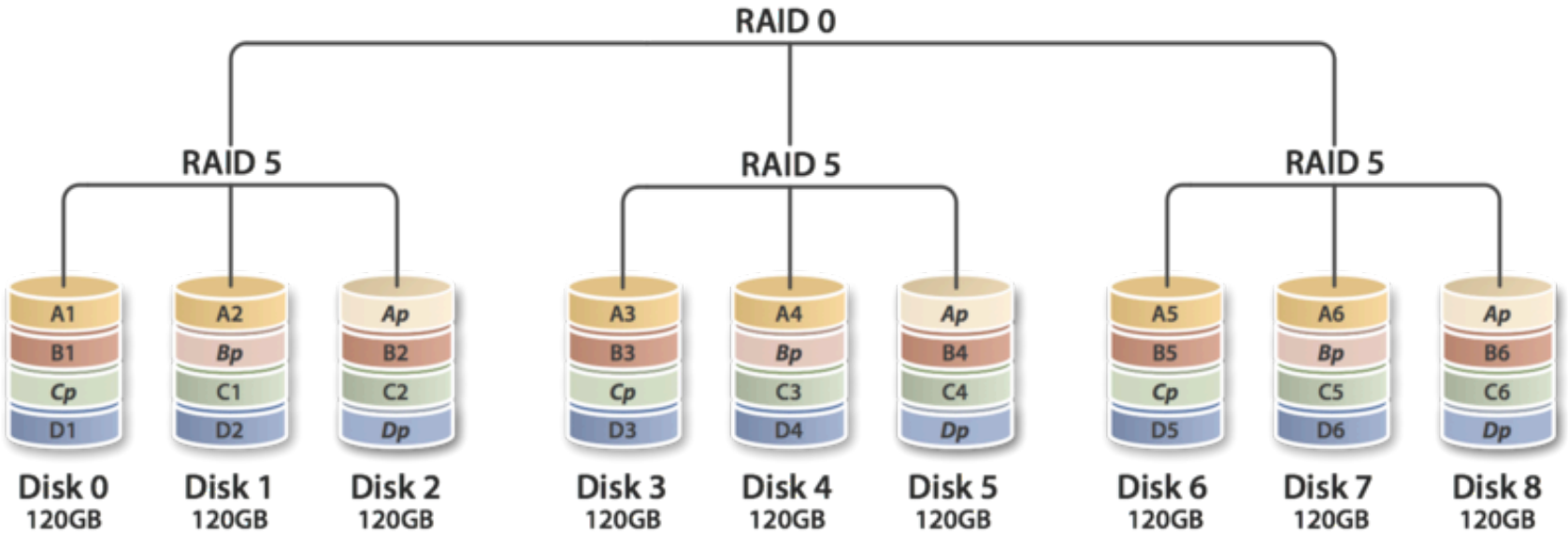
- Parité distribuée.
- Parité calculée pour chaque écriture.
- Le contrôleur est souvent couplé à un cache (avec une batterie) pour améliorer la performance.



- 2 disques de parité distribuée.
- Autorise la perte de 2 disques.

Les RAIDs

Raid Level 50



- Améliore la performance.