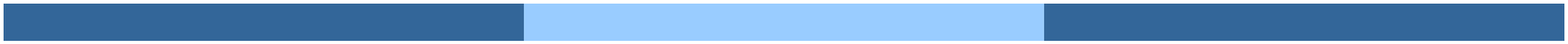
Mémoire en masse



DIRO

Menu

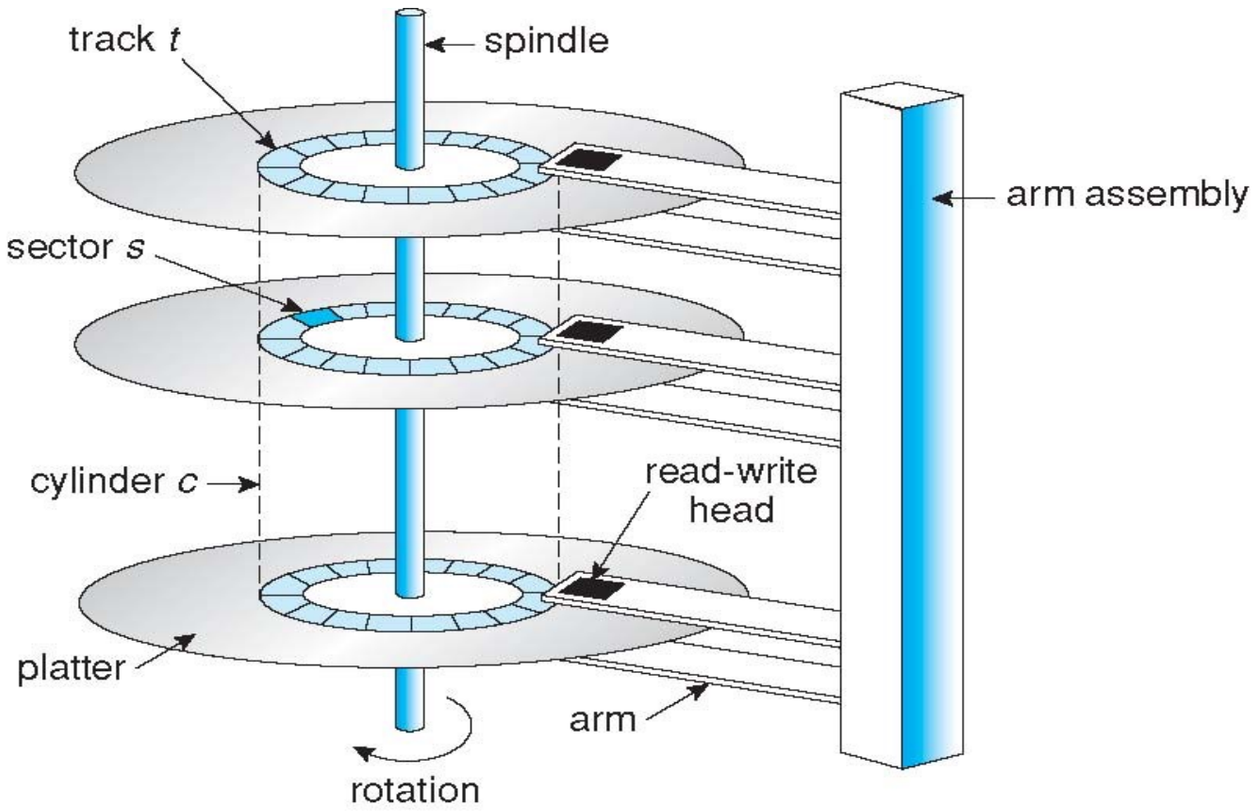
- Structure du disque et performance
- Ordonnancement de disques
- Gestion de disques
- RAID
- Autres périphériques de stockage

Menu

- **Structure du disque et performance**
- Ordonnancement de disques
- Gestion de disques
- RAID
- Autres périphériques de stockage

Structure d'un disque

- Les **disques magnétiques** fournissent l'essentiel du stockage secondaire des ordinateurs modernes
  - **Transfer rate** (R_s) est le taux auquel les données circulent entre le lecteur et l'ordinateur
  - **seek time** (T_a): temps de déplacement de la tête.
  - **rotational latency** (T_l): temps de déplacement du bloc par rotation
  - Têtes volent sur les plateaux
    - **Head crash** résultat de la tête en contact avec la surface du disque (pas bon!)
- Piste = plateau ∩ cylindre



Performance des disques

- **IOPS** (I/O Operations per second) La mesure est généralement utilisée pour décrire la capacité d'un périphérique de stockage à traiter des demandes d'entrée-sortie aléatoires.
  - Une "opération" est équivalente au transfert d'un bloc entier
- **Throughput (bandwidth)** est une mesure du nombre de bytes pouvant être transférés par unité de temps
- **Response time** (T_r) est le temps nécessaire à un périphérique de stockage pour traiter une demande d'opération d'entrée-sortie
- **Block size** - quantité de données transférées pendant une opération d'E / S (= N secteurs - typiquement secteur = 512B, bloc = 4KB bloc size - equal to page size)

$$\text{IOPS} = \text{throughput} / \text{block\_size}$$
$$T_{io} = T_a + T_l + T_t = 1/\text{IOPS}$$
$$T_t = \text{block\_size} / R_s$$
$$\text{throughput} = R_s * T_t / T_{io}$$
$$\text{cas optimal: } T_r = T_{io}$$
$$\text{en réalité: } T_r = (Q+1) * T_{io} \text{ moyen}$$

Disques magnétique

- Capacité: de 30GB à 3TB, en blocs de 4kB

- Performance
  - Transfer rate - théorique - 6 GB/s
    - “Effective Transfer Rate” - réel - 1 Gb/s
  - Seek time: de 3ms à 12ms - 9ms commun pour les lecteurs de bureau
  - Latence basée sur la vitesse du spindle
    - 1 / (RPM / 60)
    - Latence moyenne = ½ temps de latence

Spindle [rpm]	Average latency [ms]
4200	7.14
5400	5.56
7200	4.17
10000	3
15000	2

Exemple de performance d'un disque

- **Latence d'accès** = **temps d'accès moyen** = seek time moyen + latence moyen
  - Disque plus vite 3ms + 2ms = 5ms
  - Disque plus lent 9ms + 5.56ms = 14.56ms
- Temps moyen d'E/S (T_{io}) = temps d'accès moyen + (montant à transferrer / taut de transfer)

- Ex: transferrer un bloc de 4KB à 7200 RPM avec 5ms seek time moyen, 1GB/sec taut de transfert:

5ms + 4.17ms + 4KB / 1Gb/sec =

9.17ms + 4 / 131,072 sec =
9.17ms + 0.12ms =
9.29ms

Structure de disque

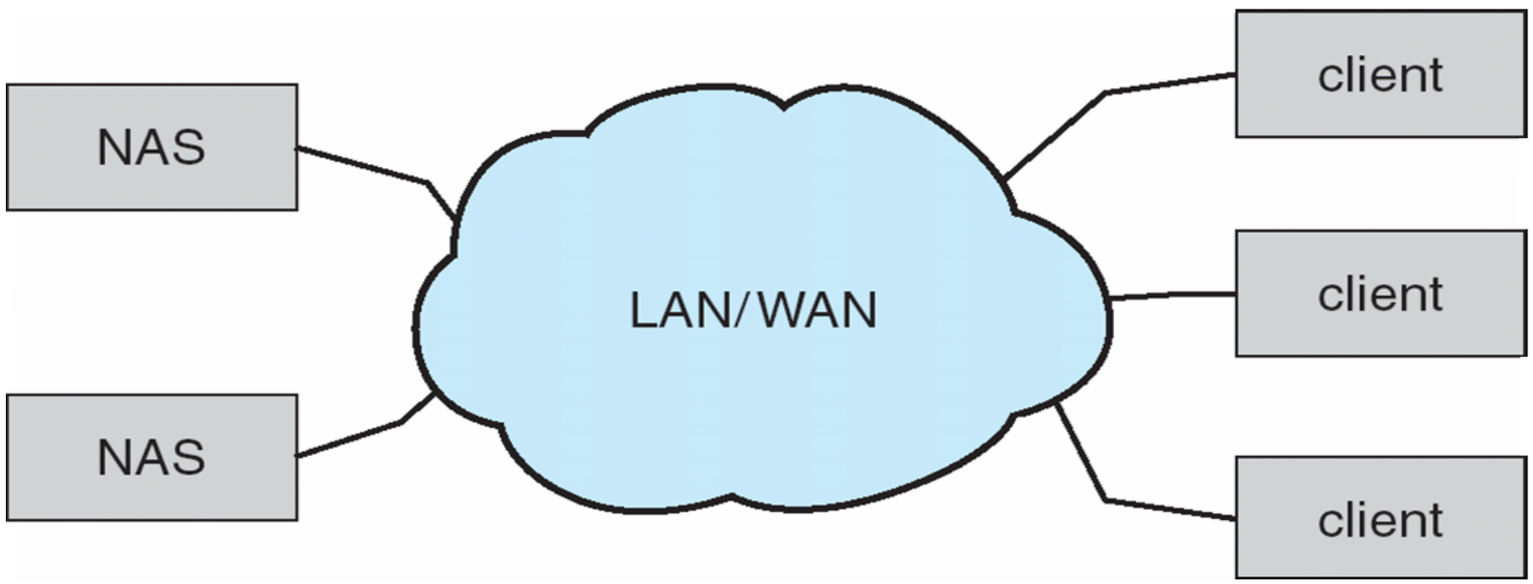
- Le SE ne voit qu'un tableau de N blocs logiques de taille fixe
- Reçoit des requêtes read et write sur ces blocs
- Accès séquentiels censés obtenir meilleure performance
- Bloc endommagés cachés par remapping
- Vitesse linéaire constante vs. Vitesse angulaire constante

Attachement de disque

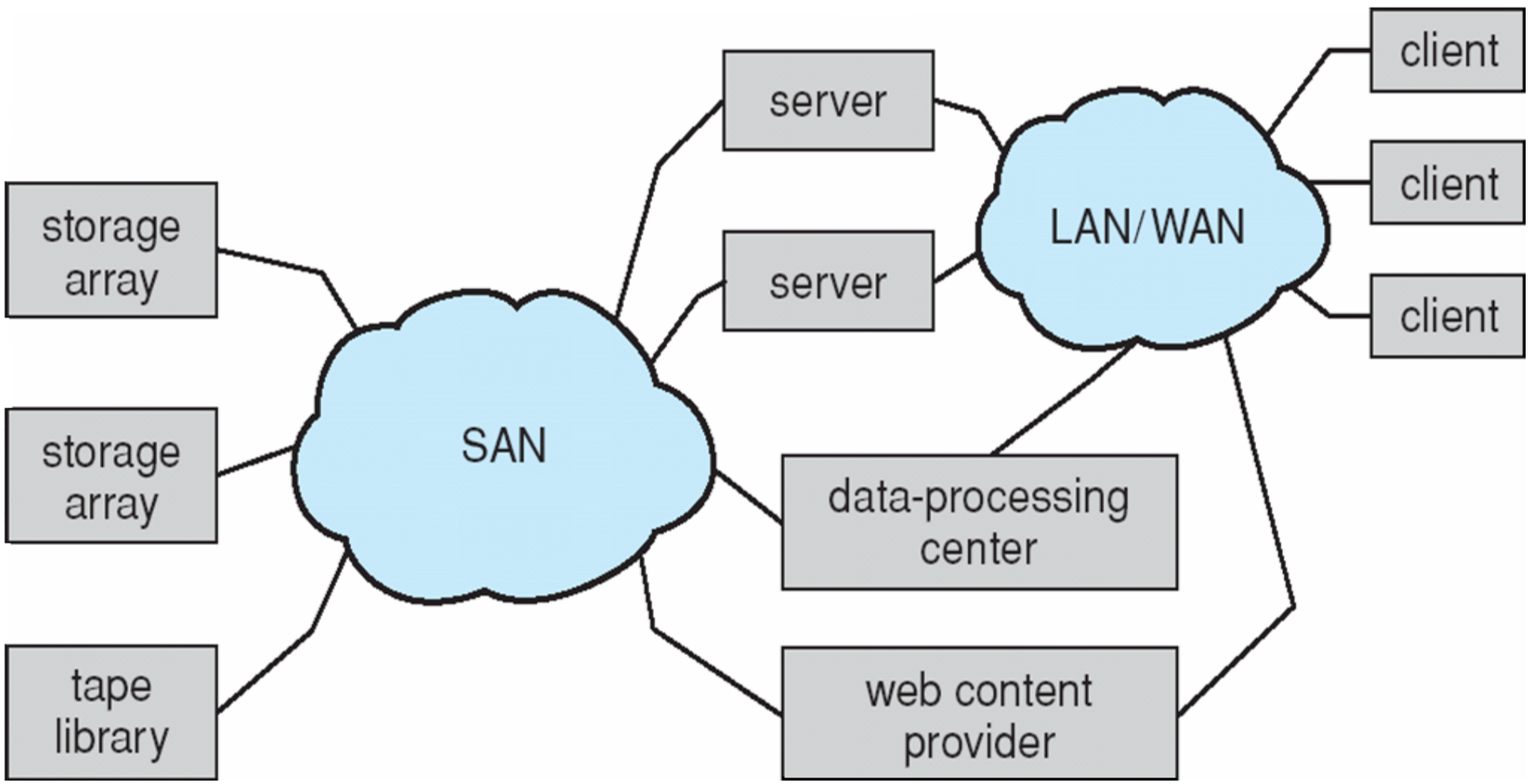
- Host-attached storage
  - Communication par bus d'E / S
    - e.g., ATA, FC
  - Disque dur, CD-ROM, RAID arrays
- Storage array
  - contrôleur connecté à plusieurs/beaucoup de disques
  - Vu de l’extérieur comme un autre ensemble de disques
- Network-attached storage
  - Disque vus soit comme des disque ou des système de fichiers
  - RPCs sur un Réseau standard
- Storage area network (SAN)
  - Réseau spécialisé pour connecter des disques et des machine
  - Disques dédiés. Facilité d’ajouter ou enlever disques

Network-Attached Storage

- Le network-attached storage (NAS) est un stockage disponible sur un réseau plutôt que sur une connexion locale (comme un bus)
- NFS = network file system
- RPC entre l'hôte et le stockage sur TCP ou UDP sur réseau IP



Storage Area Network



Menu

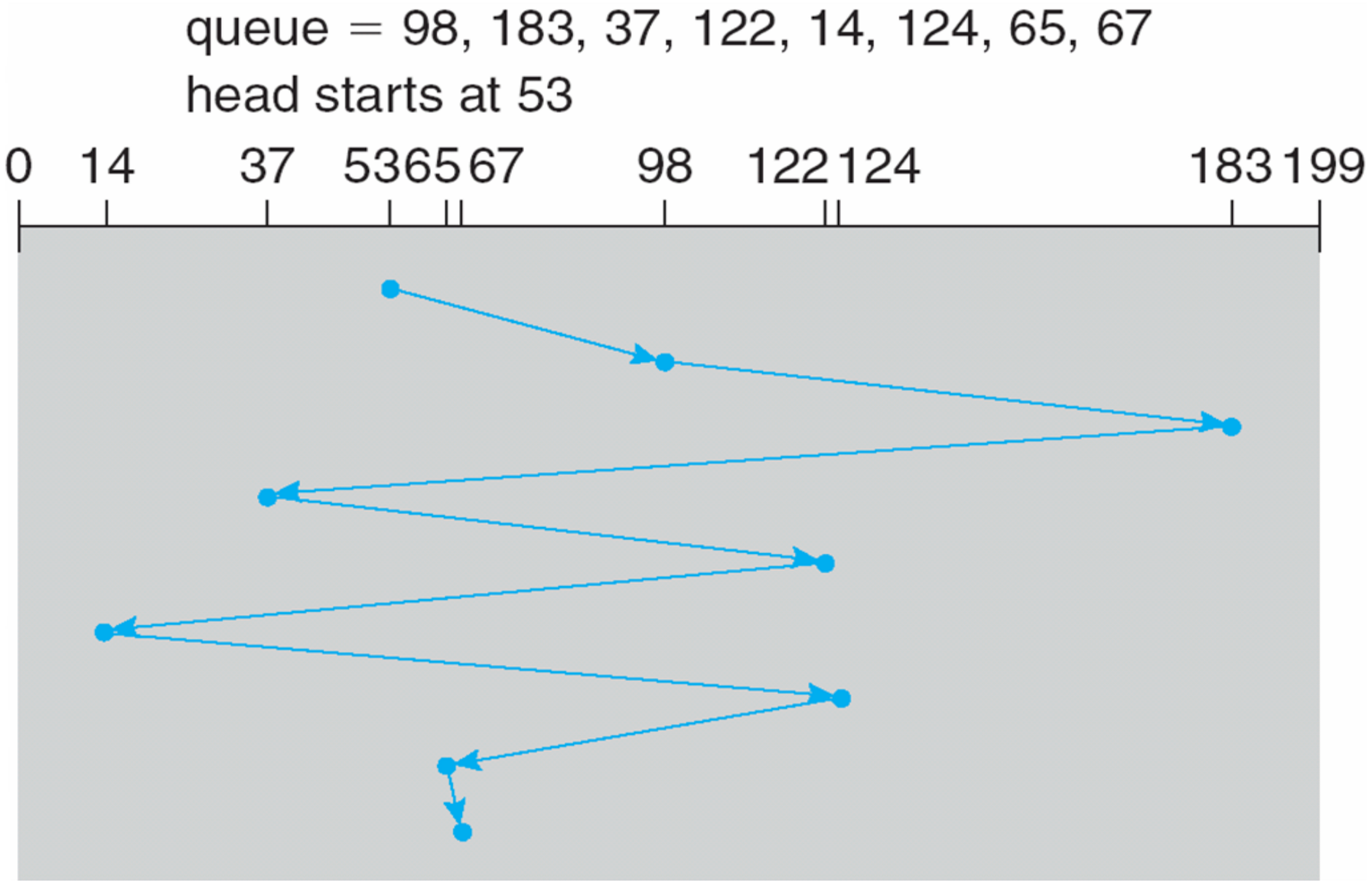
- Structure du disque et performance
- **Ordonnancement de disques**
- Gestion de disques
- RAID
- Autres périphériques de stockage

Ordonnancement de disques

- Le S/E est responsable de l'utilisation efficace du matériel - pour les unités de disque, cela signifie un temps d'accès rapide et bandwidth disque
- Veut minimisé le “seek time”
- Le “seek time” est lié à la distance entre les blocs sur les disques (les déplacement de têtes)

- Un queue de requêtes par disque (dans le SE et/ou le disque)
- Si la queue est vide: pas de différence
- Sinon, choix d’algorithmes d’ordonnancement

First-come first-served (FCFS)

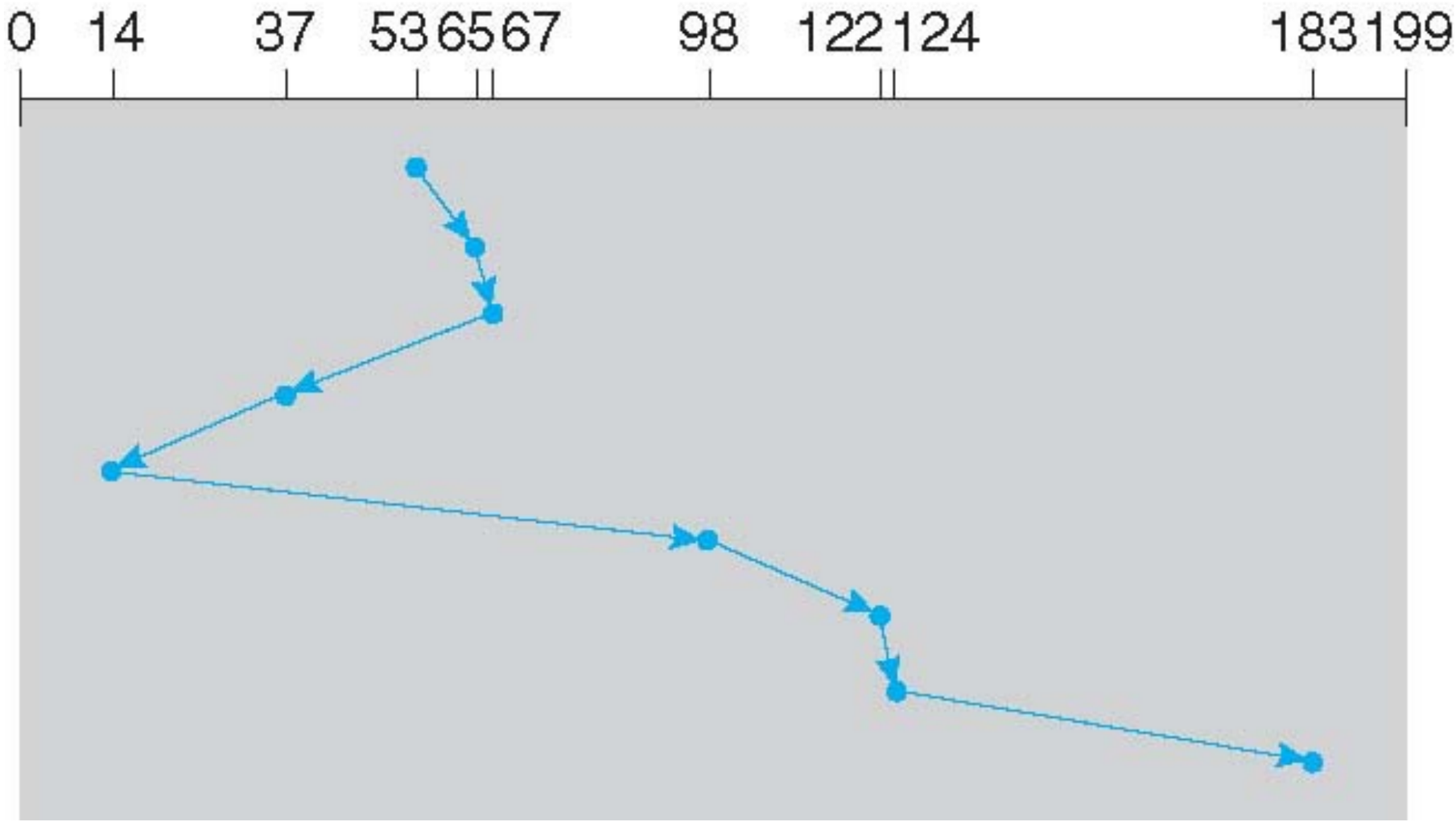


Totale = 640

Shortest Seek Time First (SSTF)

queue = 98, 183, 37, 122, 14, 124, 65, 67

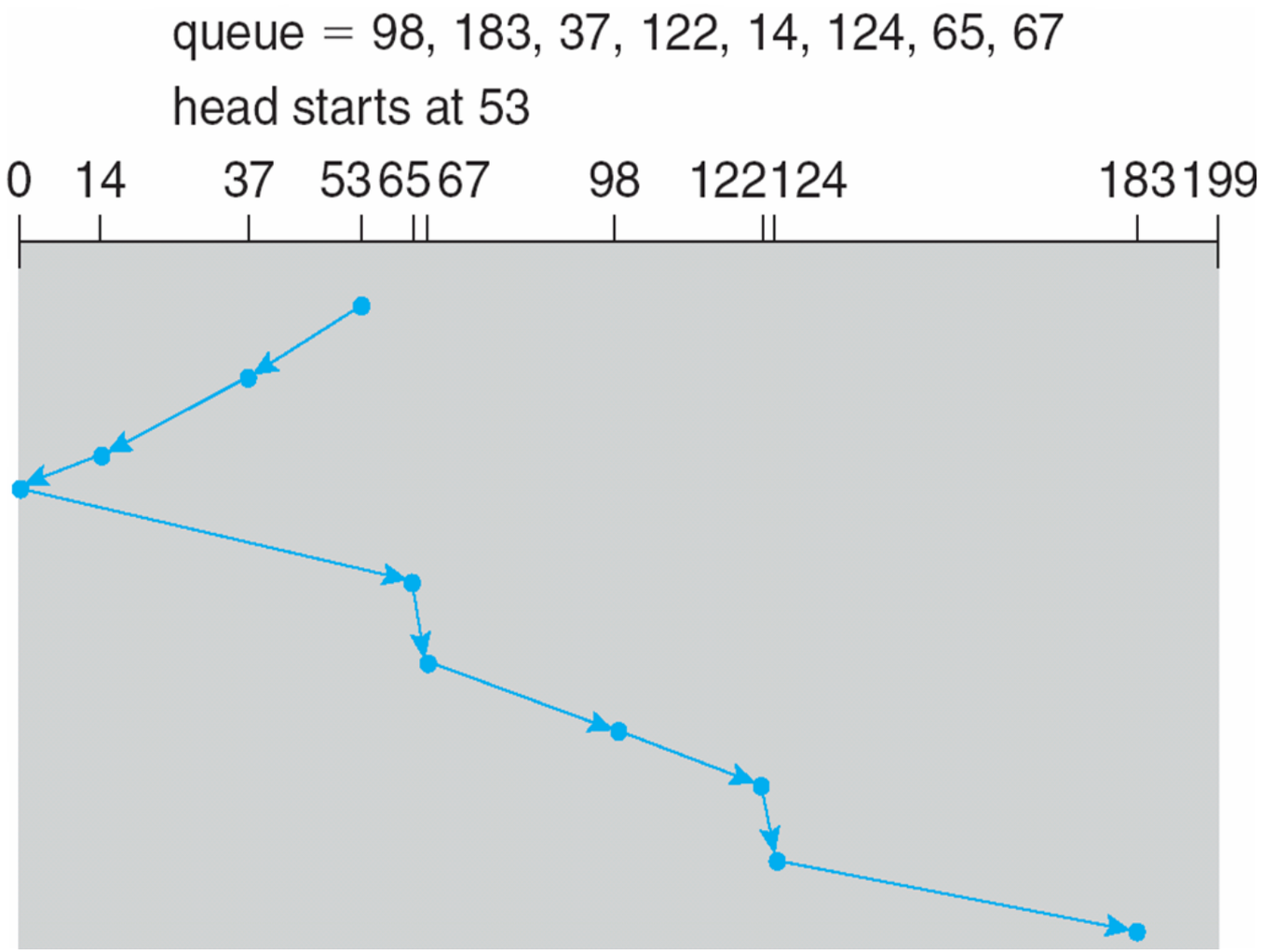
head starts at 53



Totale = 236

SCAN

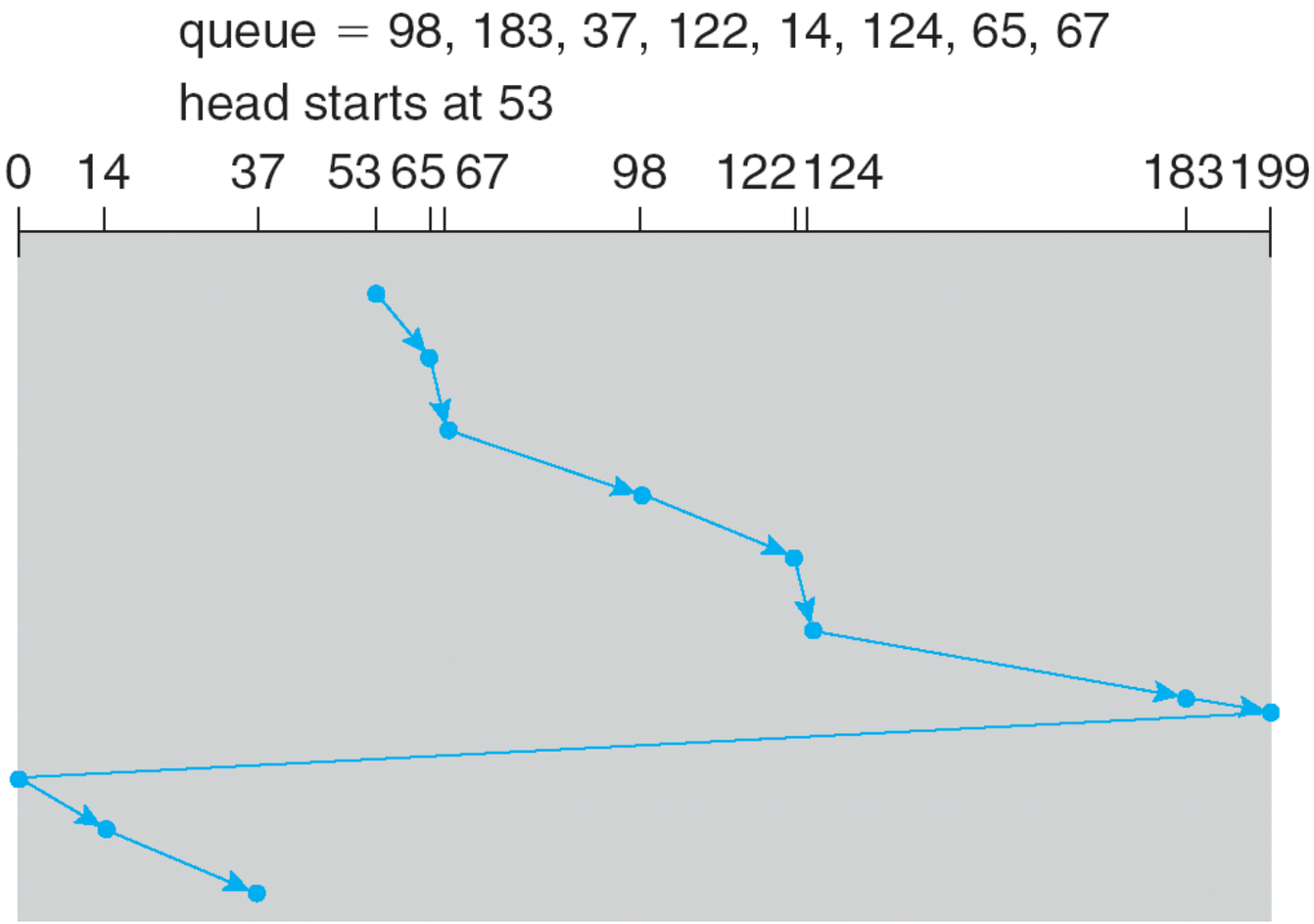
- La tête parcourt toute la surface dans un sens puis dans l'autre
- Aussi appelé algorithme de l'ascenseur



Totale = 236

SCAN circulaire

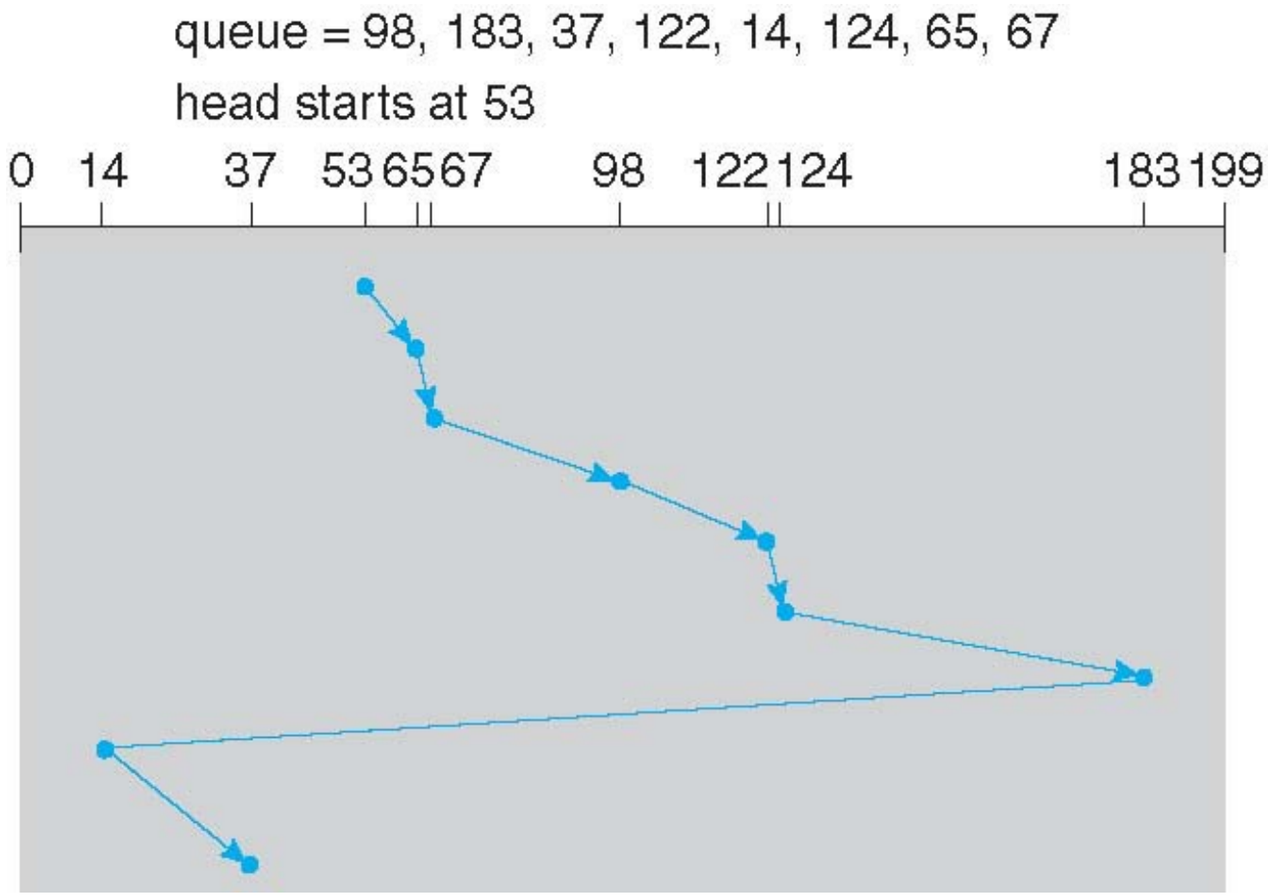
- La tête parcourt la surface toujours dans le même sens
- Diminue l'attente maximum par rapport à SCAN



Déplacement total plus grand (382) mais 1x200 < 2x100

LOOK circulaire (C-LOOK)

- LOOK = SCAN sans aller vraiment jusqu'au bout
- C-LOOK = C-SCAN en évitant aussi les déplacements inutiles



Total = 322

Fonctionnement de l’ordonnanceur

- Les performances dépendent du nombre et des types de demandes
- Les demandes de service de disque peuvent être influencées par la méthode d'allocation de fichiers

- SSTF ou LOOK est un choix raisonnable pour l'algorithme par défaut

- Qu'en est-il de la latence rotatoire? (souvent aussi grand que le seek time)
  - Difficile pour l'OS de calculer

- Le contrôleur de disque peut avoir son propre algorithme d’ordonnacement, ce qui soulagerait le système d'exploitation de cette tâche
  - Mais s'en remettre à cela retarderait toutes les connaissances du système d'exploitation sur les priorités (crash, écriture vs lecture, petit espace restant pour la mémoire paginée, etc.)

- Comment la mise en file d'attente sur disque affecte-t-elle les efforts du système d'exploitation?

Menu

- Structure du disque et performance
- Ordonnancement de disques
- **Gestion de disques**
- RAID
- Autres périphériques de stockage

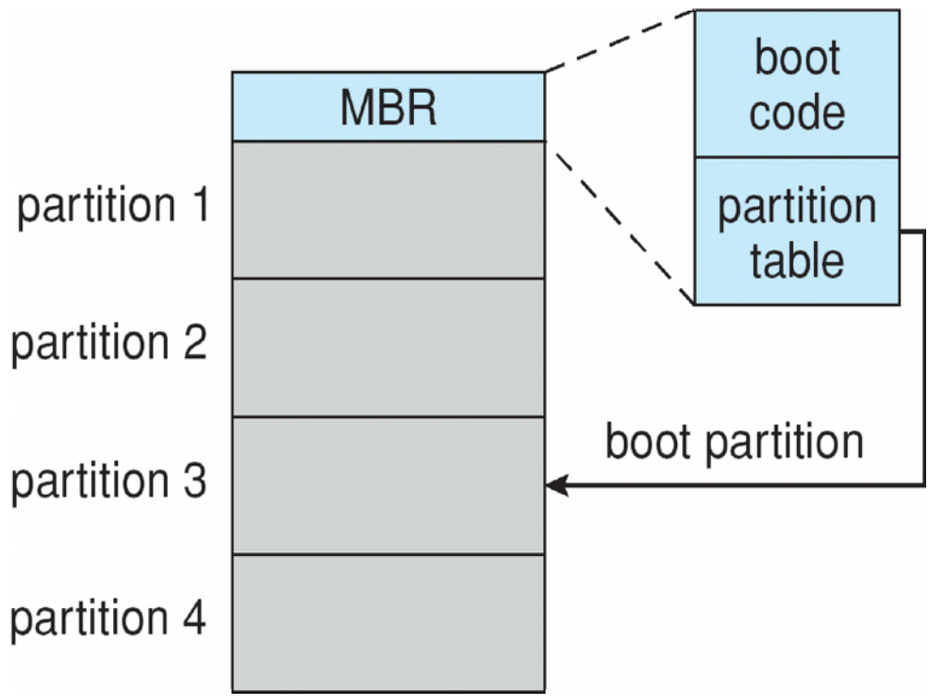
Gestion des disques

- **Low-level formatting** — Division d'un disque en secteurs que le contrôleur de disque peut lire et écrire
  - Chaque secteur peut contenir des informations (header), des données, plus un code de correction d'erreur (**ECC**)
  - Normallement 512 bytes
  - Le contrôleur teste l'ECC après lecture et indique s'il peut corriger une erreur

- Pour utiliser un disque pour stocker des fichiers, le système d'exploitation doit enregistrer ses propres structures de données sur le disque
  - **Partitionner** le disque en un ou plusieurs groupes de cylindres, chacun traité comme un disque logique
  - **Logical formatting** pour créer un système de fichiers

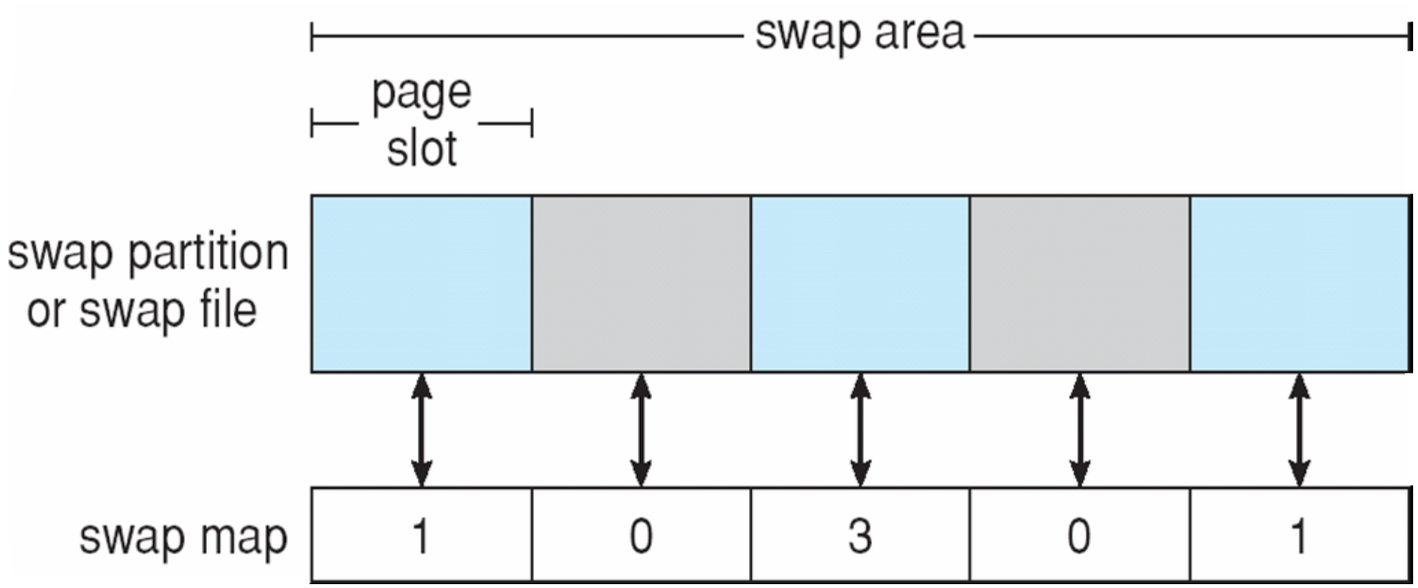
Boot loader

- Boot block initialise le système
  - The bootstrap est stocké en ROM
  - **Bootstrap loader** programme stocké dans les boot blocks de la boot partition



Gestion du swap space

- Swap-space — La mémoire virtuelle utilise l'espace disque comme une extension de la mémoire principale
- Swap space peut être dans le système de fichiers normal (mais moins efficace en raison de la traversée du système de fichiers, de la fragmentation supplémentaire, des accès disque supplémentaires, etc.) ou, plus généralement, sur une partition de disque séparée (raw) (optimisé pour l'accès plutôt que le stockage)
- Noyau utilise **swap maps** pour tracker l'utilisation de l'espace de swap



Menu

- Structure du disque et performance
- Ordonnancement de disques
- Gestion de disques
- **RAID**
- Autres périphériques de stockage

Error correcting codes (ECC)

- Parity code: peut détecter les erreurs mais ne les corrige pas

1 1 0 0 0 1 1 0 parity bit: 0 (even parity)

- Hamming codes: peut détecter les erreurs de deux bits et corriger les erreurs d'un bit

0 1 0 0 1 1 0 1

P1 P2 0 P4 1 0 0 P8 1 1 0 1

P1 0 1 0 1 0 P1 = 0

P2 0 0 0 1 0 P2 = 1

P4 1 0 0 1 P4 = 0

P8 1 1 0 1 P8 = 1

- pour 2^N bits besoin de N + 1 bits de parité

RAID: Redundant Array of Inexpensive Disks

- RAID – Copies redondantes sur plusieurs disques, pour la fiabilité
- **Temps moyen jusqu'à l'échec**
  - 100,000 heures pour un disque
  - 100,000/100 = 1000 heures (41 jours): un disque dans un array de 100 disques
  - Avec mirroring, avoir deux disques spécifiques échouent en même temps
    - e.g.100,000x100,000 / 2 x 10 heures pour réparer = $500x10^6$ (57k years)
  - Cela suppose que les défaillances de disque sont indépendantes

- En plus: Accès parallèles à plusieurs disques, pour la performance
- "Disk **striping**" utilise un groupe de disques comme une unité de stockage
  - E.g., avec 8 bits de données, stocker un bit par disque, 8 disques peuvent transférer 8x plus de données (bit-level striping)
  - Peut "stripe" par byte, secteur, bloc (le plus commun)
  - Striping augmente le throughput
  - Striping n'améliore pas la fiabilité

RAID-0: Striping

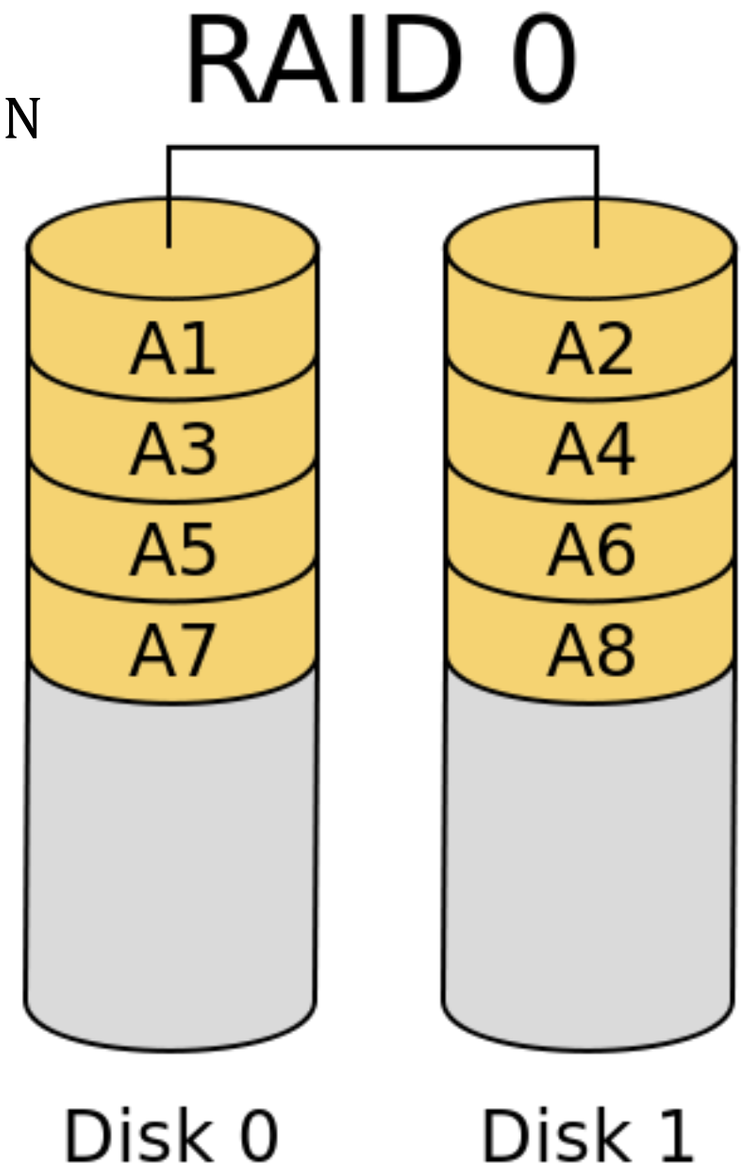
- Combiner **N** disques **PD_i** en un grand disque **LD**
- Données réparties finement sur tous les disques
- Divisé en “stripes”. Stripe **S** placée sur disque $S \bmod N$

$$\text{taille (LD)} = \Sigma \text{ taille } (PD_i)$$

$$\text{throughput (LD)} = \Sigma \text{ throughput } (PD_i)$$

$$\text{IOPS (LD)} = \Sigma \text{ IOPS}(PD_i)$$

$$\text{fiabilité (LD)} \cong 1/N * \text{fiabilité}(PD_i)$$

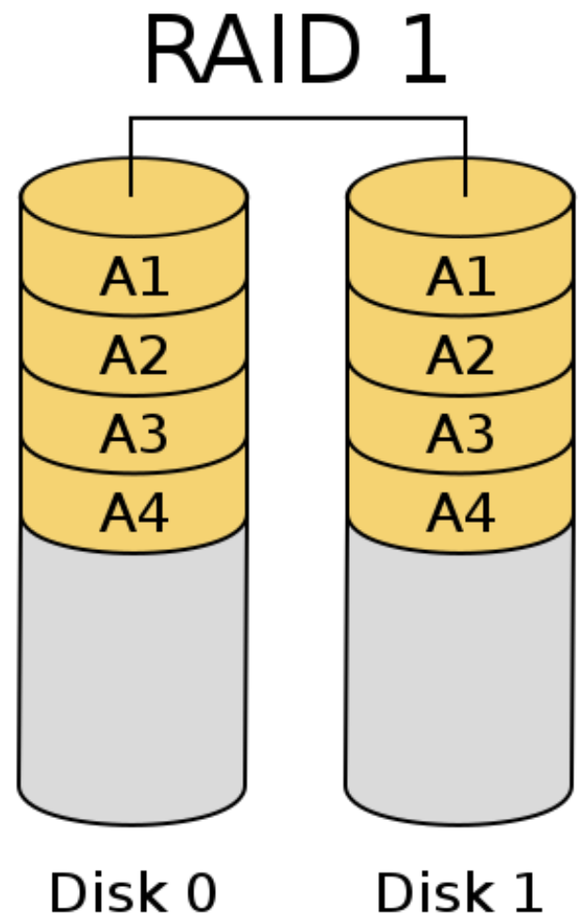


RAID-1: Mirroring

- Combiner N disque PD_i en un disque LD de même taille
- Données copiées N fois: chaque disque est une copie des autres

$$\text{fiabilité}(LD) \cong N \times \text{fiabilité}(PD_i)$$
$$\text{IOPS\_read}(LD) \cong \Sigma\ \text{IOPS\_read}(PD_i)$$

$$\text{IOPS\_write}(LD) \cong \text{IOPS\_write}(PD_i)$$
$$\text{taille}(LD) \cong \text{taille}(PD_i)$$

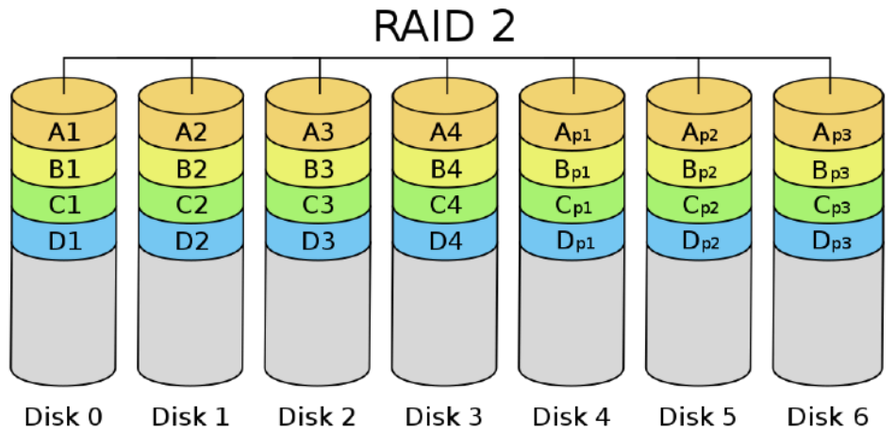


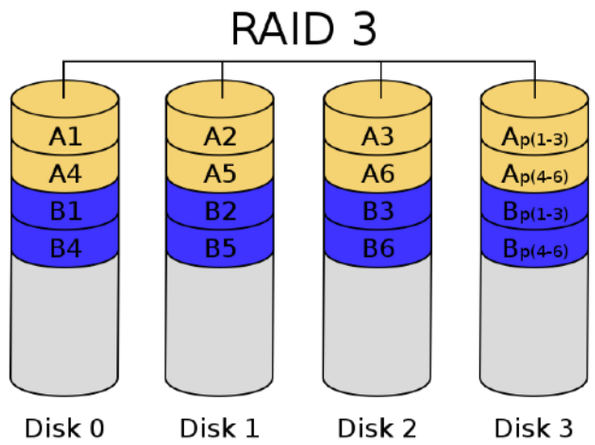
RAID 2-4: Parity

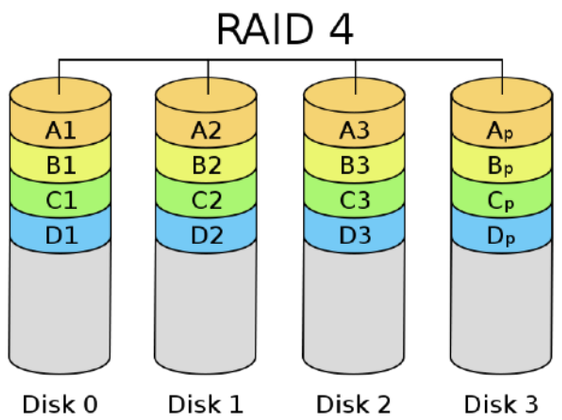
- Striping sur N disques plus un disque supplémentaire de parité
- En lecture: comme RAID-0 avec le disque de parité inutilisé
- Fiabilité meilleure que RAID-0: un disque peut mourir sans perte
- Mais en écriture: peut être pire que RAID-1

- Chaque écriture touche au disque de parité
- Le calcul de la parité peut nécessiter des lectures supplémentaires

- RAID 2 - Hamming code (bit level striping)
- RAID 3 - byte level striping
- RAID 4 - block level striping
  - une lecture d'E / S n'accède qu'à un seul disque - nous pouvons faire beaucoup de lectures en parallèle

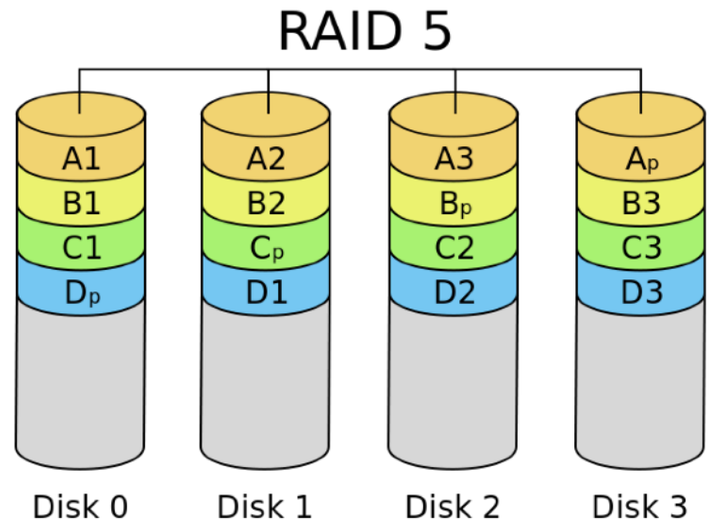




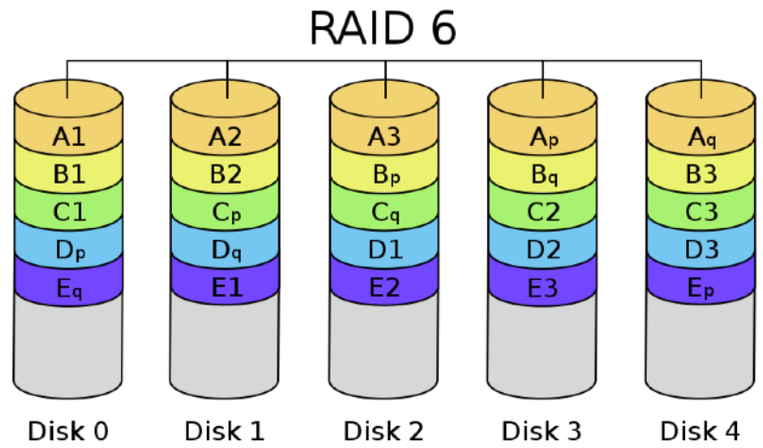


RAID 5-6

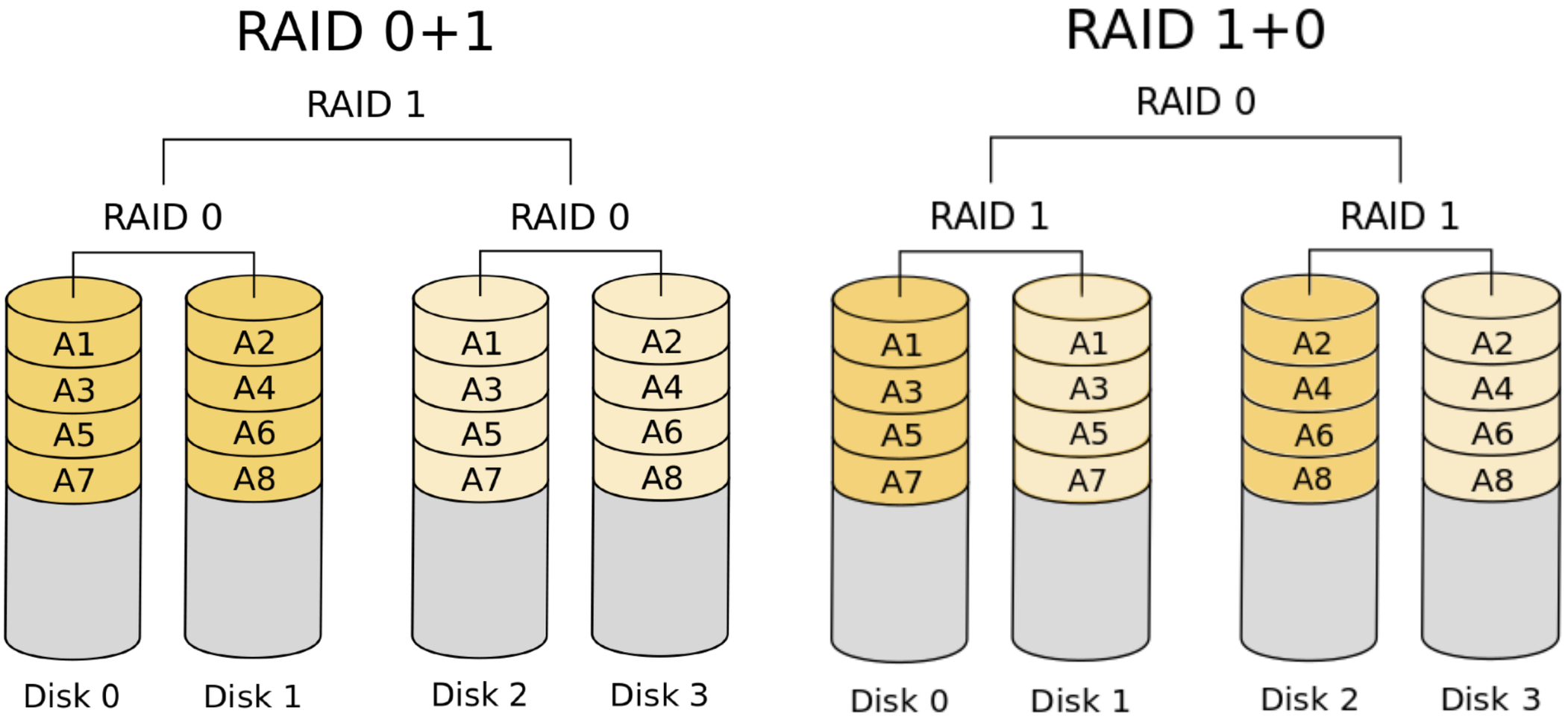
- RAID 5
  - identique à RAID 4, mais distribuez le disque de parité sur tous les disques en round-robin
  - récupération plus complexe en cas de panne de disque



- RAID 6
  - ajoutez deux bits redondants pour permettre la récupération de deux disques



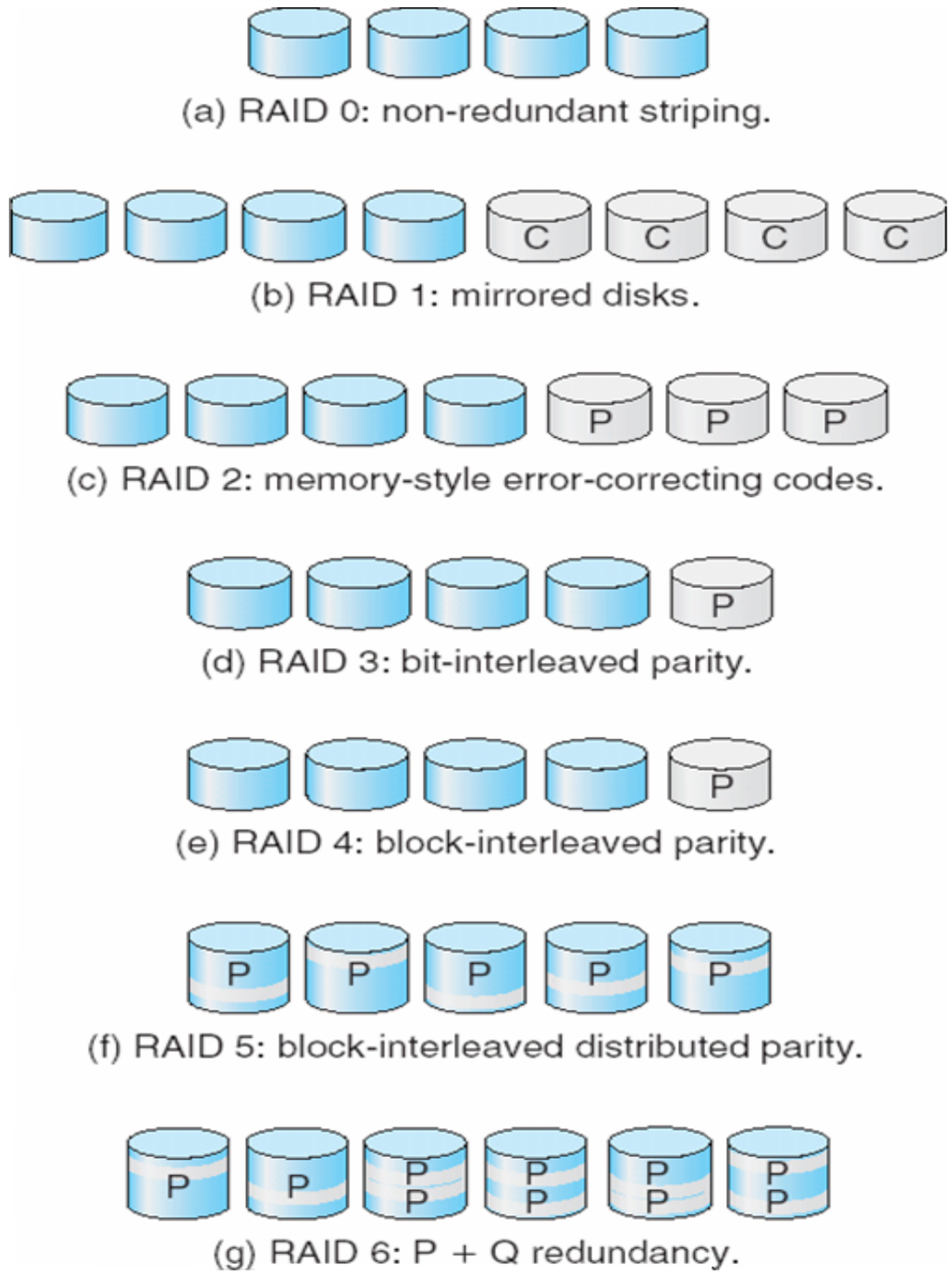
RAID (0 + 1) and (1 + 0)



RAID Sommaire

Content for 4 disks of data
P: error-correcting bits
C: second copy

- **Mirroring** or **shadowing** (**RAID 1**) conserve un copie de chaque disque
- **Block interleaved parity** (**RAID 4, 5, 6**) utilise beaucoup moins de redondance
- **RAID 2** ECC; **RAID 3** bit level parity
- **RAID 4** lit un bloc sur un disque, mais les petits accès sont chers;
- **RAID 5** stocke la parité et bloque sur n'importe quel disque;
- **RAID 6** utilise plus de bits pour la correction d'erreurs (permettre 2 échecs de disque)
- Striped mirrors (**RAID 1+0**) ou mirrored stripes (**RAID 0+1**) fournit de hautes performances et une grande fiabilité



Menu

- Structure du disque et performance
- Ordonnancement de disques
- Gestion de disques
- RAID
- **Autres périphériques de stockage**

Solid state drives (SSDs)

- Une mémoire non volatile
- Plus cher par MB
- Performance beaucoup plus élevée qu'un HDD:
  - Bande passante: ~500MB/s
  - Latence: < 0.1ms
  - Meilleure fiabilité mécanique
  - Plus basse consommation d'énergie

Périphériques de stockage tertiaires

- Moins cher
- Peut être enlevé - exige généralement que les données soient copiées dans le stockage secondaire à utiliser
- E.g., diquettes, CD-ROM

Disquettes

- Disquettes — Disque flexible mince recouvert d'un matériau magnétique, enfermé dans un boîtier en plastique protecteur
  - La plupart des disquettes contiennent environ 1 MB
  - Les disques magnétiques amovibles peuvent être presque aussi rapides que les disques durs, mais ils présentent un risque plus élevé de dommages dus à l'exposition

Disques optiques

- Les disques optiques n'utilisent pas de magnétisme; ils utilisent des matériaux spéciaux qui sont modifiés par la lumière laser
- **WORM** (“Write Once, Read Many Times”) ou “read only”
- Film mince en aluminium pris en sandwich entre deux plateaux en verre ou en plastique
- Pour écrire un “bit”, le lecteur utilise une lumière laser pour brûler un petit trou à travers l'aluminium; l'information peut être détruite en ne changeant pas
- Très durable et fiable
- e.g. CD-ROM, DVD

Bande magnétique

- Était un support de stockage secondaire
- Relativement permanent et contenant de grandes quantités de données
- Temps d'accès très lent
  - Accès aléatoire environ 1000 fois plus lent que le disque
- Principalement utilisé pour la sauvegarde, le stockage de données rarement utilisées, le transfert de support entre les systèmes
- Une fois les données sous la tête, les taux de transfert comparables au disque (~ 1 GB / sec = 125 MB / sec)
- 200 GB à 1,5 TB

Vitesse

- Deux aspects de la vitesse dans le stockage tertiaire sont la bande passante et la latence.
- La bande passante est mesurée en bytes par seconde.
  - Bande passante soutenue - débit de données moyen lors d'un transfert long;
  - Bande passante effective - moyenne sur toute la durée d'E / S, y compris seek () ou locate (),
- Latence d'accès - temps nécessaire pour localiser les données
  - Temps d'accès pour un disque - déplacez le bras vers le cylindre sélectionné et attendez la latence de rotation; <35 millisecondes
  - L'accès sur bande nécessite l'enroulement des bobines de bande jusqu'à ce que le bloc sélectionné atteigne la tête de bande; des dizaines ou des centaines de secondes
- Une bibliothèque amovible est le mieux dédiée au stockage de données rarement utilisées, car la bibliothèque ne peut satisfaire qu'un nombre relativement faible de requêtes d'E / S par heure

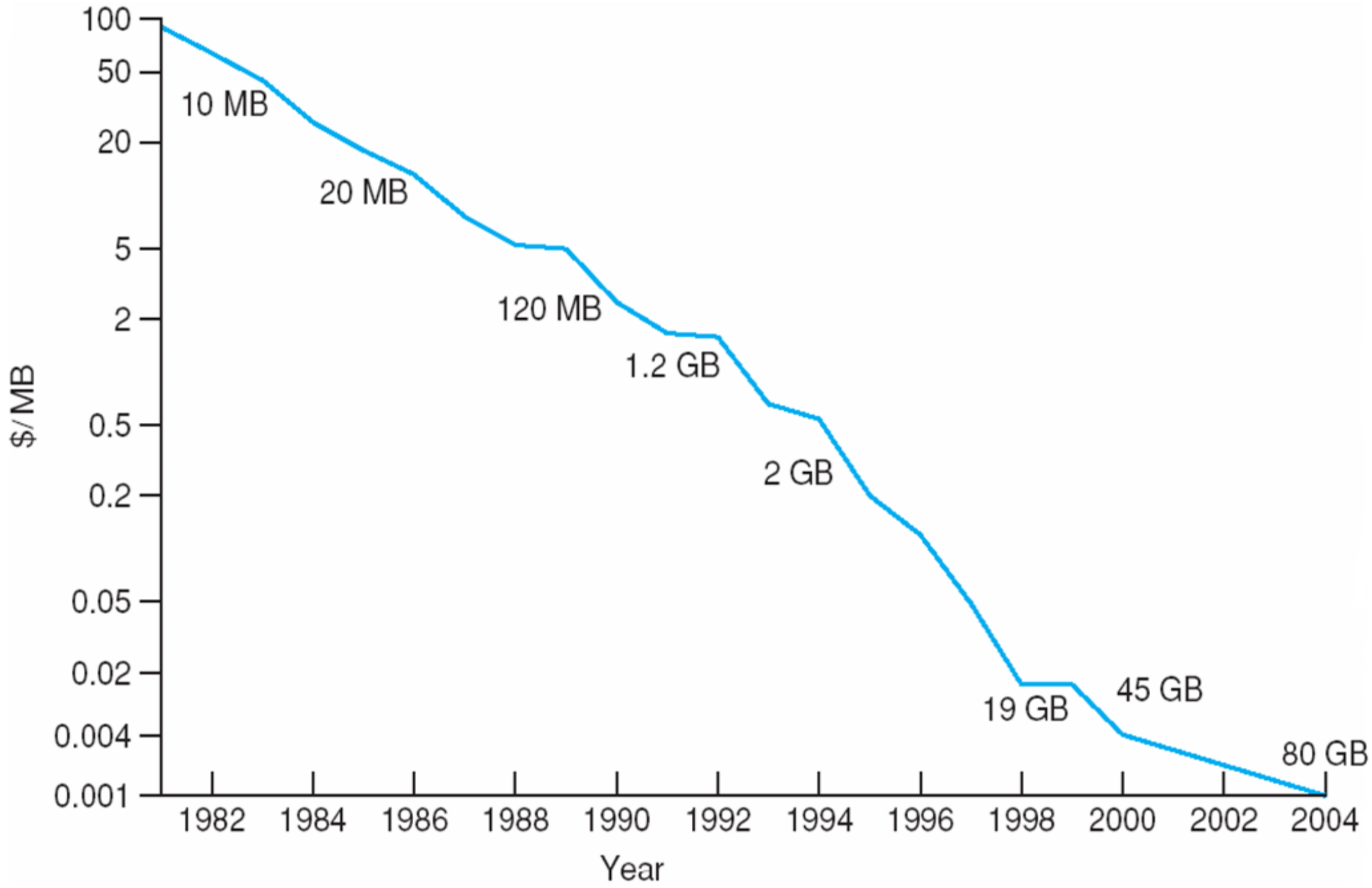
Fiabilité

- Un lecteur de disque fixe est probablement plus fiable qu'un disque amovible ou un lecteur de bande
- Une cartouche optique est probablement plus fiable qu'un disque magnétique ou une bande
- Un “head crash” dans un disque dur fixe détruit généralement les données, alors que la défaillance d'un lecteur de bande ou d'un lecteur de disque optique laisse souvent la cartouche de données indemne

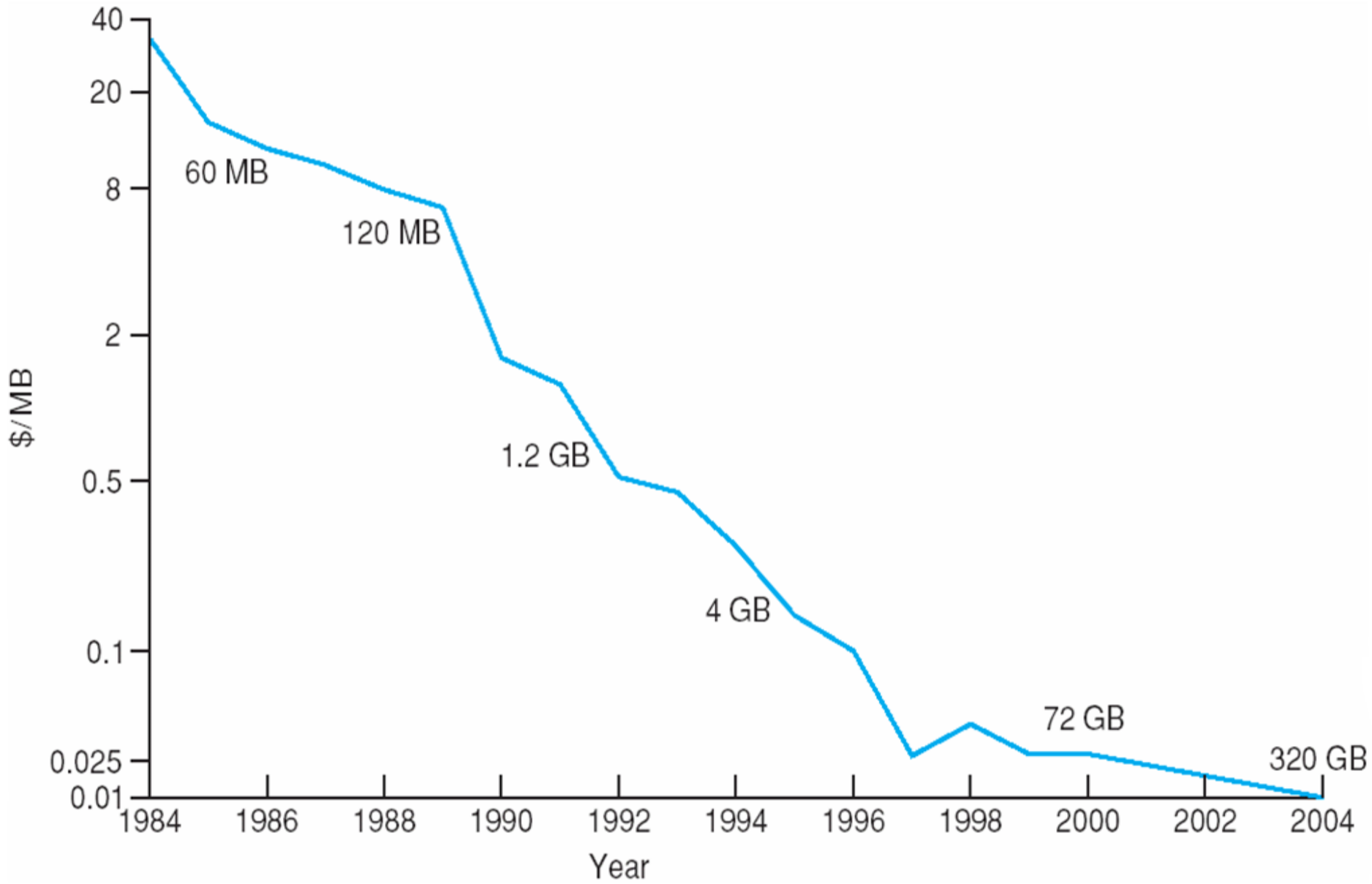
Coût

- La mémoire principale est beaucoup plus chère que le stockage sur disque
- Le coût par MB de stockage sur disque dur est compétitif avec la bande magnétique
- Les lecteurs de bande les moins chers et les lecteurs de disques les moins chers ont eu à peu près la même capacité de stockage

Prix par MB de disque dur magnétique de 1981 à 2004



Prix par MB d'un lecteur de bandeDe 1984 à 2004



Sommaire

- Sur la plupart des ordinateurs, les disques durs sont utilisés pour le stockage secondaire
- Les performances d'un disque sont spécifiées par des mesures telles que les IOPS, le temps de réponse et le throughput
- Un algorithme de planification est utilisé pour gérer les demandes de disque
- Les systèmes RAID assurent l'efficacité et la redondance grâce à la “striping”, “mirroring” et à l'utilisation de bits de parité
- Les autres types de stockage incluent les disques SSD, les bandes et le stockage tertiaire