Gestion de mémoire secondaire

F. Boyer, Laboratoire Lig
Fabienne.Boyer@imag.fr

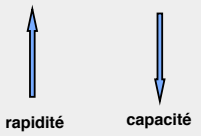
1- Structure d'un disque
2- Ordonnancement des requêtes
3- Gestion du disque
   - formatage
   - bloc d'amorçage
   - récupération d'un bloc défectueux
4- Fiabilité des disques
5- Gestion de l'espace de swap

Ce cours a été conçu à partir de…

- Cours en ligne (free download)
  - http://pages.cs.wisc.edu/~remzi/OSTEP/
- Livre d'Andrew Sylberschatz
  - A. Silberschatz, P. Galvin and G. Gagne, "Operating System Concepts", 9th International student edition, John Wiley, 2013

Supports de stockage

- Mémoires flash
- Disques durs
- Disques optiques
- Bandes magnétiques



- Unité de transfert
  - Flôt de données (caractère)
  - Bloc
- Connexion des périphériques
  - Par bus
  - Par bus avec accès direct à la mémoire
  - Architectures à plusieurs bus

Bandes magnétiques

- Différentes technologies
  - Serpentine, parallèles, hélicoidales
- Capacités
  - Stockage jusqu'à 100Go (NCPT/Next Compatible Tape Product de Philips)
  - Vitesse de 4 à 10 Mo/s
  - → moins cher mais accès aléatoires très lents par rapport aux disques optiques
- Fonctions
  - Append-only

→ Utilisées pour les sauvegardes (systèmes information, bases de données, serveurs Web, etc)

Disques optiques

- **Plusieurs technologies (write-once / réinscriptible)**
- **CD-ROM**
  - 500 à 700 Mo
  - 150 Ko/s (initialement)
- **DVD**
  - 4 à 18 Go (→ 50 Go pour DVD HD / BlueRay)
  - DVD lecture seule
  - DVD-R (inscriptible une fois)
  - DVD-RW (lect/ecr)
  - DVD-RAM (lect/ecr, technologie optique-magnétique)

Mémoires Flash

- **2 technologies**
  - Cartes mémoires
  - Solid State Drive (*lecteur à l'état solide – sans pièces mobiles*)
    - Remplaçant potentiel des disques durs
    - Meilleure résistance aux chocs
    - Faible consommation électrique
- **Caractéristiques**
  - Vitesse de lect/ecr de 96M/s à 1Go/s
  - Temps d'accès moyen de 0,1ms
  - Coût: environ 0,05 €/Gio
  - Capacité jusqu'à 8 To
- **Technologies NOR (cartes)**
  - Accès aléatoires rapides (12 microsec)
  - Temps écriture lent (750 msec)
  - Pas utilisée pour le stockage de masse

Mémoires Flash

- **Technologies NAND (SSD)**
  - Temps écriture plus rapide (2 msec)
  - Accès séquentiels (mode bloc)
  - Cache RAM
- **Nombre de cycles d'écritures limité (100 000 – 300 000)**
  - Algorithmes d'étalement des écritures

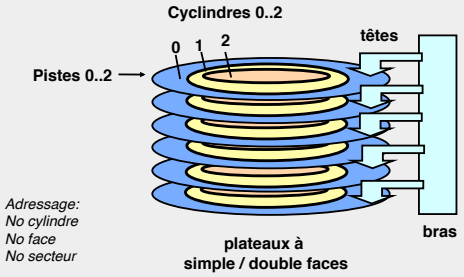
Disques durs

- **Caractéristiques**
  - Accès aléatoire plus lent que les supports de stockage précédents
  - Temps d'accès moyen de 3 à 12 ms
  - Vitesse de lect/ecr de 40 à 260 Mo/s
  - Capacité jusqu'à 8 To
  - Coût: environ 0,5 €/Gio
- **Utilisés pour**
  - Le stockage de données persistantes
  - La mise en œuvre d'une mémoire virtuelle
- **Technologies**
  - IDE → S-ATA (Serial ATA)
  - SCSI (stations de travail et serveurs puissants)
  - RAID

Structure d' un disque dur

- **Logiquement :**
  - Suite de blocs contigüs
  - Le bloc est la plus petite unité de transfert vers la mémoire centrale
- **Physiquement :**
  - Ensemble de plateaux (double faces)
  - Chaque plateau est composé de pistes (circonférences sur le plateau)
  - Chaque piste est composée de secteurs
- **Un bloc = un secteur**

Structure d'un disque dur



Adressage:
No cylindre
No face
No secteur

Adressage de l'espace disque

- **Linéarisation de l' espace**
- **Un exemple de parcours**
  - Par secteur
    - Par face
      - Par cylindre

Adressage de l'espace disque (version 1)

- Soit N un numéro de secteur logique
- N doit être décomposé comme suit :

No cylindre nc	No face nf	No secteur ns

x = N div nbsf
ns = x div nbsc
nf = N mod nbsf
nc = x mod nbsc

nbf = nombre de faces du disque
nbc = nombre de cylindres du disque
nbsc = nombre de secteur par cylindre
*nbsf = nombre de secteurs par face = nbf * nbc*

Adressage de l'espace disque (version 2)

- Soit N un numéro de secteur logique
- N doit être décomposé comme suit :

No face nf	No piste np	No secteur ns

nf = N mod nbf
np = (N div nbf) div nbsp
ns = (N div nbf) mod nbsp

nbf = nombre de faces du disque
nbsp = nombre de secteur par piste

Opérations élémentaires

- Chargement d'un bloc
- Déchargement d'un bloc

- Un bloc = 1 secteur
(ou un nombre fixe de secteurs)

Quantum d'allocation

- Taille des blocs
  - → grande diversité
  - un disque de 20 Mo structuré en secteurs de 512 octets en contient 40 000
  - un disque de 5Go structuré en secteurs de 4Ko en contient 1.2 millions

En général, blocs de 4Ko à 64Ko

→ Définir le quantum d'allocation

GRAND → Perte d'espace

PETIT → Coût des algorithmes (Nombre d'unités allouables grand)

Gestion des requêtes disque

- Une requête spécifie :
  - type d'opération
    - entrée / sortie
  - adresse disque
    - numéro de bloc, traduit par le gestionnaire de disque en adresse disque composée
  - adresse mémoire
    - où (ou vers où) copier
  - nb octets à transférer

Ordonnancement des requêtes disque

- Objectif: minimiser les temps d'accès
- Temps d'accès
  - *Temps de positionnement du bras* = temps de déplacement de la tête de lecture/écriture sur la bonne piste
  - *Temps de positionnement rotationnel* = temps d'attente pour que le bloc désiré passe sous la tête
  - Temps de transfert
- Bande passante = nombre total de bits transférés divisé par le temps total entre l'émission de la requête et sa terminaison.

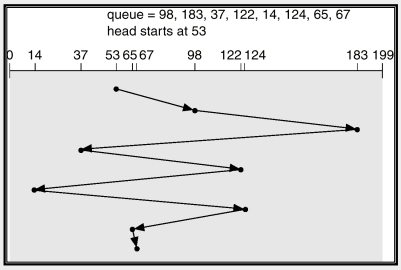
Algorithmes d'ordonnancement disque

- Objectif : minimiser les déplacements de la tête de lecture sur les pistes
- Différents algorithmes (FCFS, SSTF, SCAN, ...)
- Illustration avec une liste de requêtes portant sur des pistes allant de 0 à 199 :
  - 98, 183, 37, 122, 14, 124, 65, 67
  - Hypothèse : initialement, la tête pointe sur 53

FCFS

- Pros
  - Equitable
  - Ordonné selon l'ordre applicatif
- Cons
  - Mouvements non optimisés de la tête de lecture
  - Balayages incessants possibles

FCFS(First Come - First Served)

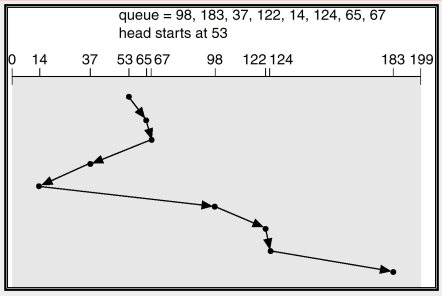


L'illustration montre un mouvement total de la tête de 640 pistes.

SSTF (Shortest Seek Time First)

- **Sélectionner la requête la plus proche de la position courante de la tête**
- **Pros**
  - Minimise les temps de positionnement
- **Cons**
  - SSTF est une forme de scheduling pouvant causer la famine de certaines requêtes

SSTF

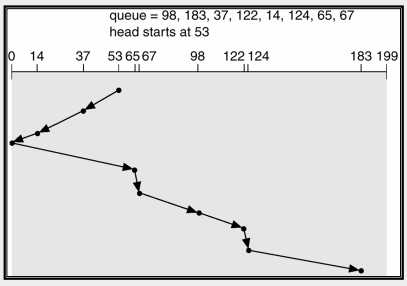


Mouvement de la tête de 236 pistes (65, 67, 37, 14, 98, 122, 124, 183)

SCAN

- **SCAN = balayage**
  - Parfois appelé *algorithme de l'ascenceur (ou chasse-neige)*
  - La tête démarre à une extrémité et se déplace jusqu'à l'autre extrémité en servant toutes les requêtes au passage (piste par piste)
  - La tête reprend le déplacement inversé
- **Pros**
  - Temps borné pour chaque requête
- **Cons**
  - Une requête en bout de piste attend longtemps

SCAN

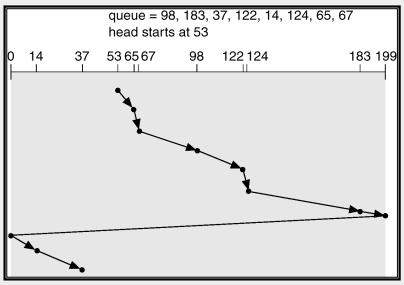


Mouvement de la tête de 208 pistes

C-SCAN

- **Circular-SCAN**
  - Quand la tête arrive à une extrémité du disque, elle retourne immédiatement au début du disque
- **Pros**
  - Fournit un temps d'attente encore plus uniforme que SCAN
- **Cons**
  - Temps de déplacement de la tête sur le retour non exploité

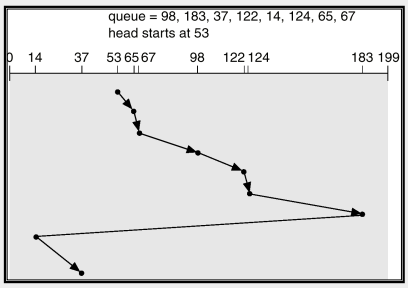
C-SCAN



C-LOOK

- **Version de C-SCAN**
  - Recherche une requête avant de traverser le disque
  - Après la dernière requête, le bras retraverse toute la surface du disque et reprend le service dans le même sens
- **Version réaliste de C-SCAN**

C-LOOK



Sélection d'un algorithme

- SSTF largement utilisé
- SCAN et C-SCAN plus performants pour les systèmes qui font une utilisation intensive des disques

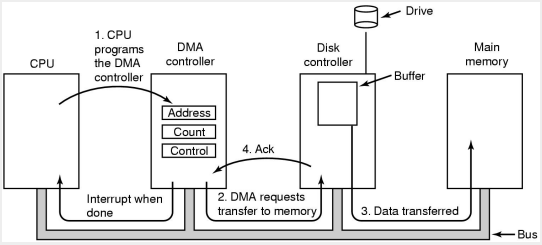
Dans tous les cas :

- Performance dépend du nombre et type des requêtes
- Requêtes peuvent dépendre de l'implémentation du SGF (fichiers contigus, chainés, indexé)

Pilotage des entrées-sorties depuis le noyau

- **E/S synchrones**
  - Le processeur est bloqué tant que l'opération d'entrée-sortie n'est pas terminée
  - Utilisé dans des configurations spécialisées uniquement
- **E/S asynchrones par interruption**
  - Le processeur continue d'exécuter des instructions mais est interrompu à chaque échange de données entre la mémoire et le support externe
- **E/S asynchrones par DMA**
  - Le DMA contrôle les échanges de données entre la mémoire et le support externe
  - Le processeur n'est interrompu qu'après le transfert du bloc entier

Principe du DMA



Source: www.cs.uiuc.edu/class/fa07/cs423/Lectures/IO01-concepts.ppt

Entrées-Sorties Tamponnées

- **Principe**
  - Transférer les données par petits bouts coûte cher
  - Transférer un bloc ne coûte pas plus cher que transférer quelques octets
  - Donc il vaut mieux vaut factoriser les transferts
- **Bufferisation des écritures**
  - Les données écrites sont placées dans un buffer
  - Elles seront transférées lorsque le buffer sera plein
  - Implique de synchroniser les processus
- **Bufferisation des lectures**
  - Lecture en avance de données voisines (pre-fetching)

Entrées-Sorties Spoolées

- Cas des périphériques très lents (imprimantes)
- Risque de monopoliser les buffers en raison de la lenteur du transfert
- Les données à imprimer sont transférées dans un fichier sur disque
- De façon asynchrone, ces données sont lues par un processus qui les envoie à l'imprimante sans passer par les buffers système

Gestion de caches

- Optimisation par cache de blocs
  - LRU (least recently used) ou LFU (least frequently used)
- Main memory disk caching: cache géré en mémoire principale
- On-Disk caching: RAM intégrée au contrôleur disque
  - ATA Disk : 5MB, IBM Ultra 160 : 16MB
  - Avantageux pour les lectures en cas de localité d'accès
  - Cher, et complexe pour la gestion des écritures fiables

Gestion du disque

- Formatage
  - Formatage physique = découpage en secteurs
    - Secteur : en-tête, contenu
    - En-tete possède un code correcteur d'erreur (ECC), actualisé à chaque écriture
  - Formatage logique
    - Installation des données du système (ex: FAT, inodes, etc)
- Partitions
  - Disque contenant des secteurs est divisé en un ou plusieurs groupes de cylindres
  - Chaque groupe est considéré comme un disque individuel, un disque logique

Gestion des secteurs défectueux

- Secteurs
  - Initialement défectueux
  - Devenant défectueux en cours d'utilisation
- Gestion des secteurs défectueux par le contrôleur
  - Au moment du formatage : les secteurs défectueux ne seront jamais insérés dans la liste des secteurs libres
  - Réserve de secteurs sains pour remplacer des secteurs défectueux
- Information de contrôle
  - ECC (Error Correcting Code) – valeur dépend du contenu
  - Permet de corriger le contenu du secteur si celui-ci a été corrompu

Gestion des pannes

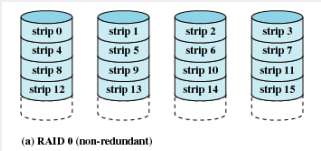
- Disque = composant le moins fiable d'un système
  - Le risque de panne transitoire est limité (les écritures sont atomiques)
  - Mais le risque de panne globale du disque est réél
- Sauvegardes
  - Récupération des données de reprise
- Disques RAID
  - Prévention de la perte de données (duplication)
  - Parallélisation des accès

Disques RAID

- RAID = Redondant Arrays of Inexpensive Disks
  - Redondance des données
  - Parallélisation des E/S
  - 5 niveaux de disques RAID
  - Niveaux 1 et 5 très utilisés
- Objectifs
  - Rapidité
  - Fiabilité
- RAID SCSI Controller

Disque RAID 0

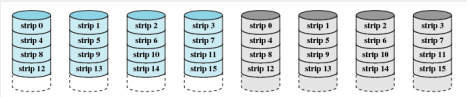
- Découpage d'une partition en bandes de taille égales
- Répartition des partitions sur différents disques
- Ecritures consécutives réparties selon la stratégie round-robin (striping)



Source: I/O Management and Disk Scheduling, engr.smu.edu/~kocan/7343/fall05/slides

(+) Disques RAID (niveau 1)

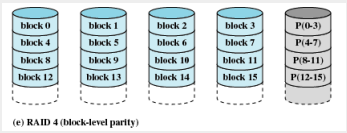
- Striping de RAID-0
- Utilise un disque miroir pour chaque disque
- Duplique chaque écriture
- La lecture peut se faire sur n'importe quel disque miroir



Source: I/O Management and Disk Scheduling, engr.smu.edu/~kocan/7343/fall05/slides

Disques RAID (niveau 4)

- Parité entrelacée
- Utilise un disque pour stocker des blocs de parité (EOR)
- Parité vérifiée à la lecture

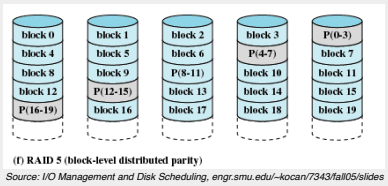


(e) RAID 4 (block-level parity)

Source: I/O Management and Disk Scheduling, engr.smu.edu/~kocan/7343/fall05/slides

Disques RAID (niveau 5)

- Répartit les blocs de parité sur différents disques (pas de bottleneck)
- 100 disques → temps moyen de perte de données = 90 ans (2 à 3 ans avec les gros disques chers)



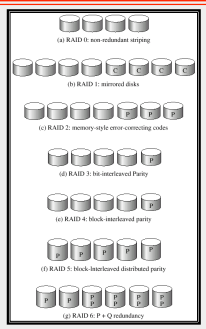
(f) RAID 5 (block-level distributed parity)

Source: I/O Management and Disk Scheduling, engr.smu.edu/~kocan/7343/fall05/slides

Disques RAID

→ Solutions peu chères
→ performantes
→ fiables

RAID Levels



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