

I- Introduction aux structures de fichiers

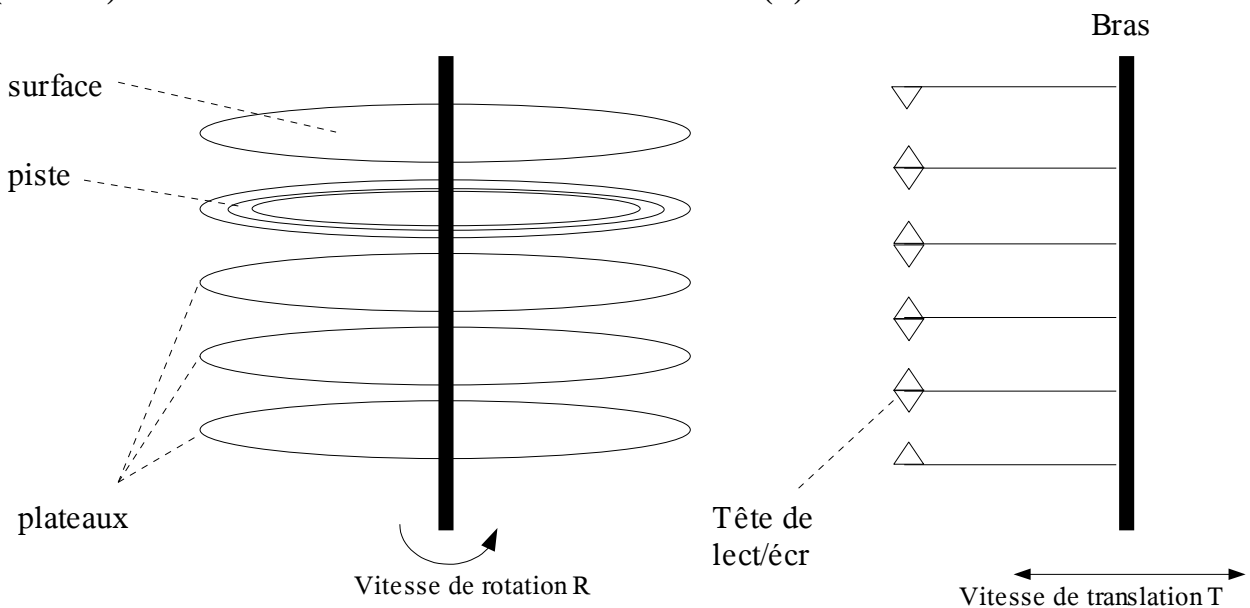
1) Entrées/Sorties (E/S) sur disque

La mémoire secondaire (MS) est généralement de grande taille mais le temps d'accès est très grand relativement à la mémoire centrale (MC).

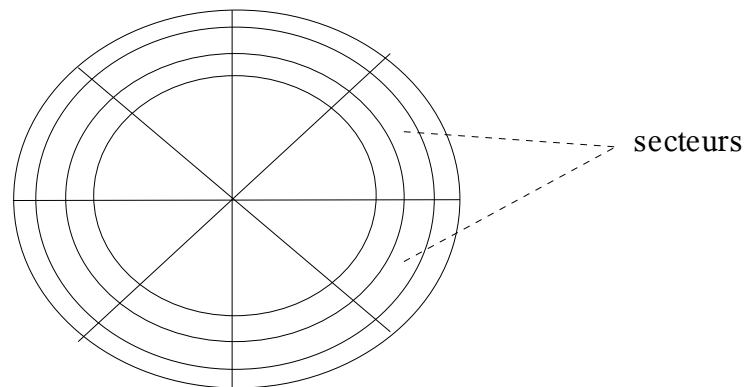
De plus, la MS est généralement non volatile et le coût à l'unité de stockage est moindre que celui de la MC.

Le type de MS le plus utilisé de nos jours reste le disque dur (DD).

Un DD est un ensemble de plateaux superposés, chacun formé de deux faces magnétiques (surfaces) et tournant avec une vitesse de rotation constante (R).



Chaque surface est composée d'un nombre fixe de pistes concentriques, pouvant stocker la même capacité en octets. Chaque piste est découpée en un nombre (généralement fixe) de secteurs, constituant l'unité de transfert entre la MS et la MC (le secteur est donc la plus petite quantité de données pouvant être lue ou écrite sur le disque à l'aide d'une opération E/S).



Le DD dispose d'une tête de lecture/écriture par piste. Toutes les têtes de lecture/écriture sont attachées à un même bras qui se déplace latéralement pour se positionner sur une piste donnée au niveau de chaque surface. L'ensemble des pistes sur lesquels les têtes de lecture/écriture sont positionnés à un moment donné, représentent un cylindre.

Temps d'accès

Pour transférer le contenu d'un secteur donné, le contrôleur du disque déplace le bras pour le positionner sur le bon cylindre, attend que le secteur passe sous la tête de lecture/écriture puis active la tête associée à la surface sur laquelle le secteur se trouve pour récupérer l'information stockée.

Le temps nécessaire à cette opération (avec la technologie actuelle) est au voisinage de quelques ms en moyenne. Si on le compare au temps d'accès à la mémoire centrale (quelques ns), la différence est énorme (le temps d'accès au disque est un million de fois plus lent que celui de la MC).

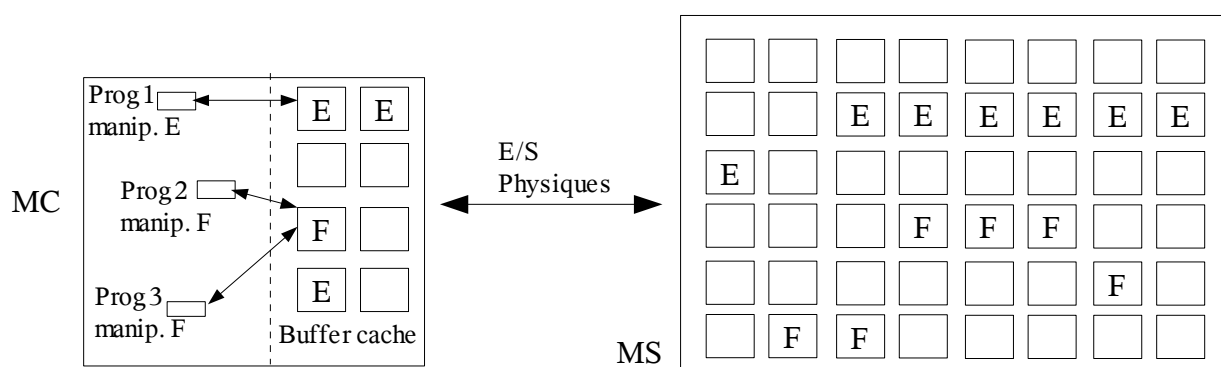
Parmi les méthodes utilisées pour diminuer l'effet de cette grande différence dans le temps d'accès, on retrouve le « **buffering** ». Il s'agit de réserver une zone en mémoire centrale pouvant contenir plusieurs blocs en même temps et utiliser des stratégies de remplacement (comme LRU, FIFO, ...) afin de minimiser le nombre d'accès disque. D'autres techniques, comme la multi-programmation (durant l'opération d'E/S d'un programme, le processeur est alloué à un autre programme) ou la parallélisation des E/S (les données sont réparties sur plusieurs disques – RAID) sont aussi utilisées pour diminuer le goulot d'étranglement sur les E/S.

De plus, comme le temps de déplacement du bras est plus grand que celui nécessaire pour une rotation du disque, l'accès à des secteurs physiquement proches (situés sur la même piste ou le même cylindre) est beaucoup plus rapide que l'accès à des secteurs nécessitant un déplacement du bras. Le « **clustering** » est une technique de stockage rassemblant les données susceptibles d'être traitées ensemble dans des secteurs physiquement proches.

2) Notion de fichier

Au niveau applicatif (logique), un fichier est un ensemble de données (enregistrements) stockées en mémoire secondaire (ex en Pascal : File of Tenreg). On peut aussi définir un fichier comme étant une suite d'octets, on parle alors de flux (ex en C : FILE *). C'est des fichiers « non typés ».

Au niveau physique, un fichier est un ensemble de blocs, organisés d'une certaine manière en MS. Les données (enregistrements) du fichier sont stockés à l'intérieur des blocs.



Les programmes d'applications manipulent les fichiers à travers des appels systèmes (read, write, ...) pour accéder aux données en mémoire cache. Le système synchronise périodiquement (et quand c'est nécessaire) les buffers en cache avec le disque à l'aide d'E/S physiques (lireBloc/écrireBloc)

Caractéristiques

Pour pouvoir manipuler correctement un fichier, on doit donc connaître sa structure et son organisation. Certaines de ces informations (caractéristiques) sont sauvegardées en MS et chargées en MC lors de la manipulation du fichier.

Parmi les caractéristiques courantes, on retrouve :

- le nombre de blocs utilisés
- le nombre d'enregistrements
- la date de création du fichier
- la date du dernier accès au fichier
- l'identité du propriétaire du fichier (cas des systèmes multi-utilisateurs)
- les droits d'accès au fichier (cas des systèmes multi-utilisateurs)
- ... etc

Méthode d'accès

Une structure de fichier (méthode d'accès) consiste à définir :

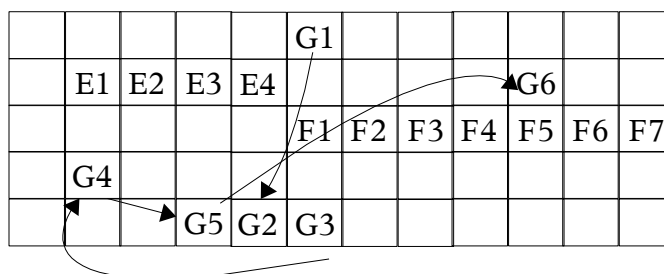
- une manière d'organiser les blocs d'un fichier sur MS
- le placement des enregistrements à l'intérieur des blocs
- l'implémentation des opérations d'accès (recherche, insertion, suppression, ...).

3) Modèle de fichiers

Afin d'étudier les méthodes d'accès aux fichiers dans un cadre général, on modélise la MS par une zone contigüe de blocs numérotés séquentiellement (adresses de blocs).

Les blocs sont des zones contigües d'octets de même taille, renfermant les données (enregistrements) des fichiers.

Dans le schéma suivant, la MS contient 3 fichiers E, F et G



Le fichier E est formé de 4 blocs contigus, le fichier F est formé de 7 blocs contigus et le fichier G est une liste de blocs.

Pour écrire des algorithmes sur les structures de fichiers on utilisera la machine abstraite définie par le modèle suivant:

{ouvrir, fermer, lireDir, écrireDir, lireSeq, écrireSeq, aff_entete, entete, allocbloc }

Dans ce modèle, on manipule des numéros de blocs relatifs au début de chaque fichier (adresses logiques). L'utilisation des adresses physiques n'est pas d'une utilité particulière à ce niveau.

un fichier est donc un ensemble de blocs **numérotés logiquement** (1, 2, 3, ... n)

Déclaration d'un fichier 'f' et de ses zones tampons 'buf1', 'buf2', ... :

var

f : FICHER de TypeBloc BUFFER buf1, buf2, ... ENTETE (type1, type2, ...typem);

dans cette ligne de déclaration, il y a la définition de :

- la variable f de type FICHER
- les variables buf1, buf2, ...de type 'TypeBloc' qui servent comme zones tampons pour lire et écrire les données du fichier.
- la structure de l'entête du fichier, formée par m caractéristiques dont les types sont spécifiés

Ouvrir(f, nomfichier, mode)

pour ouvrir ou créer un fichier de nom 'nomfichier' selon la valeur de 'mode'

« a » : ouvrir un ancien fichier en lecture/écriture et charge ses caractéristiques en MC

« n » : créer un nouveau fichier en lecture/écriture et alloue une zone en MC pour ses caractéristiques

Fermer(f)

ferme le fichier f, en sauvegardant ses caractéristiques sur disque

LireDir(F, i, buf)

lire dans 'buf' le contenu du bloc numéro 'i' du fichier 'f' (num logique de blocs)

EcrireDir(F, i, buf)

écrit le contenu de la variable 'buf' dans le bloc numéro 'i' du fichier (num logique)

LireSeq(F, buf)

lire dans 'buf' le contenu du bloc courant du fichier 'f'. (lecture séquentielle)

EcrireSeq(F, buf)

écrit le contenu de la variable 'buf' dans le bloc courant du fichier (écriture séquentielle)

Entete(f, i)

retourne la valeur de la 'i'ème caractéristique du fichier 'f' (en MC)

Aff_entete(f, i, val)

affecte 'val' dans la 'i'ème caractéristique du fichier 'f' (en MC)

AllocBloc(f)

alloue un nouveau bloc au fichier et retourne son numéro

4) Objectifs du cours

Etudier les méthodes d'accès aux fichiers (structures et algorithmes) ainsi que leurs performances (temps d'exécution et occupation mémoire)

Pour la manipulation d'un fichier d'enregistrements donné, on étudie les différentes possibilités de stockage sur MS. Un enregistrement est un ensemble de champs dont un ou plus jouant le rôle de clés de recherche (champs utilisés dans la recherche et l'identification des enregistrements).

Ex:

```
Tenreg = structure { matricule : entier, nom : chaine[20], adr : chaine[80] };  
// matricule est (par exemple) la clé de l'enregistrement
```

Les principaux paramètres à prendre en considération, lors de l'élaboration d'une méthode d'accès sont:

- Le temps d'accès des opérations de base (recherche, insertion, ...)

- Le facteur de chargement (pourcentage de remplissage de l'espace alloué au fichier)

- La taille du fichier en MS

- La taille des structures auxiliaires de la méthode en MC et en MS

- Le type d'accès:

 - fichier statique – peu ou pas d'insertions/suppressions

 - fichier dynamique – en constante évolution