La mémoire

Rôle de la mémoire

Le but de l'informatique est de traiter des données. Il faut pour cela pouvoir les ranger ainsi que les programmes qui les manipulent dans une mémoire ou plus exactement dans une variété de composants mémoire. Ceux-ci se caractérisent par leur vitesse, leur capacité, leur volatilité, leur prix et leurs dimensions physiques.

On distingue :

1. la **mémoire de masse**, dont le rôle est d'être une zone de stockage permanent. Ce rôle est assuré par les disques (disques durs, SSD, CD-ROM ou DVD) ou encore par des bandes magnétiques. Les données y sont enregistrées par des procédés magnétiques, électroniques ou optiques, elles subsistent même quand ces équipements sont hors tension. Ce sont par contre des systèmes relativement lents.
2. la **mémoire centrale** dont le rôle est d'être une zone de travail et de stockage temporaire. Les programmes que l'on veut exécuter et les données que l'on veut traiter doivent d'abord être chargés en mémoire centrale pour y être à la disposition du processeur.

La mémoire centrale est un organe passif qui reçoit des ordres de lecture/écriture du CPU.

- Les ordres de lecture/écriture lui sont envoyés par le bus de commande (*Control bus*)
- Les emplacements à lire ou écrire sont signifiés par le bus d'adressage (*Address bus*)
- Les données transitent par le bus des données (*Data Bus*)

La ROM

La **ROM *"Read Only Memory"*** (mémoire à lecture seule) est aussi appelée mémoire morte. Il est impossible d'y écrire. Les ROM sont programmées par leurs fabricants pour contenir des informations immuables telles que les fonctions du BIOS.

Il existe d'autres variantes:

- La **PROM *"Programmable ROM"*** est une ROM qui peut être programmées à l'aide d'un graveur de PROM. Une fois écrite, il est impossible d'en modifier le contenu.

- L'**EPROM *"Erasable PROM"*** était la première mémoire morte à pouvoir être reprogrammée. On l'effaçait en la laissant 10 à 20 minutes sous des rayons ultra violet. La puce du composant était visible sous une petite fenêtre qui permettait le passage de cette lumière. Pour les programmer il fallait les placer dans un programmateur d'EPROM. La lenteur de ces opération fait que les EPROM sont maintenant avantageusement remplacées par les mémoires Flash.

- L'**EEPROM *"Electricaly Erasable PROM"*** est une EPROM qui s'efface par des impulsions électriques. Elle peut donc être effacée sans être retirée de son support.

- La **FEPROM *"Flash EPROM"*** plus souvent appelée mémoire Flash est un modèle de mémoire effaçable électriquement. Les opérations d'effacement et d'écriture sont plus rapides qu'avec les anciennes EEPROM. C'est ce qui justifie l'appellation "Flash". Cette mémoire, comme les autres ROM, conserve les données même quand elle n'est plus sous tension. Ce qui en fait le composant mémoire amovible idéal pour les appareils photos numériques.

La RAM

La mémoire vive est généralement appelée RAM pour *Random Access Memory* ce qu'on traduit habituellement par "mémoire à accès aléatoire" car on peut arbitrairement accéder à n'importe laquelle de ses adresses. La traduction "mémoires à accès direct " est sans doute plus appropriée.

> *Ces mémoires ont été dénommées ainsi pour des raisons historiques. Les premières mémoires telles que les cartes perforées et les bandes magnétiques, étaient des mémoires à accès séquentiel car il fallait faire défiler une kyrielle de données avant d'atteindre celle qui est recherchée.*

La RAM du PC contient l'ensemble des programmes en cours d'exécution ainsi que leurs données. Les performances de l'ordinateur dépendent donc de la quantité de mémoire disponible. Quand l'espace mémoire ne suffit plus, le système d'exploitation a recours à la mémoire virtuelle, il mobilise pour ce faire une partie du disque et y entrepose les données utilisées le moins souvent.

RAM statiques / RAM dynamiques

Nous distinguons deux technologies de fabrication des RAM

La **SRAM** ou RAM Statique est la plus ancienne. Les bits y sont mémorisés par des bascules électroniques dont la réalisation nécessite six transistors par bit. Les données y restent enregistrées tant que le composant est sous tension. Certaines cartes mères utilisent une SRAM munie d'une pile pour former une mémoire non volatile destinée à conserver les données du setup. Cette technique tend à être remplacée par l'utilisation de mémoire flash.
La SRAM est très rapide et est pour cette raison le type de mémoire qui sert aux mémoires cache.

La **DRAM** pour RAM dynamique est de réalisation beaucoup plus simple que la SRAM. Cela permet de faire des composants de plus haute densité et dont le coût est plus faible. Chaque bit y est mémorisé par une charge électrique stockée dans un minuscule condensateur. Ce dispositif présente l'avantage d'être très peu encombrant mais n'est pas capable de garder l'information longtemps. Le condensateur se décharge au bout de quelques millisecondes. Il faut, pour ne pas perdre cette information, un dispositif qui lit la mémoire et la réécrit aussi tôt en rechargeant les condensateurs avant que leur contenu ne se dissipe. On appelle ces RAM des RAM dynamiques car cette opération de rafraîchissement doit être répétée régulièrement.

Les différents types de barrettes mémoire et leurs technologies sont décrites dans le chapitre Hardware et maintenance de ce site.

Les temps d'accès

Les commandes de lectures ou d'écritures ne se font jamais instantanément.
Le **temps d'accès** est le délai minimum entre l'instant où la commande est envoyée et celui de l'accès réel à la donnée (lecture ou écriture). Mais il faut parfois aussi tenir compte d'un **temps de cycle** supérieur au temps d'accès. Il s'agit dans ce cas du **cycle mémoire**, c'est à dire de l'intervalle minimum de temps entre deux accès successifs.

Les mémoires statiques ont des temps d'accès très courts qui leur permettent de s'adapter aux fréquences des processeurs et en font les candidates idéales pour les mémoires cache.

Les mémoires dynamiques (DRAM) ont des temps d'accès supérieurs. Elles sont organisées en matrices et l'adressage qui y sélectionne successivement les lignes et les colonnes nécessite un **temps de latence** qui vaut plusieurs cycles du processeur.

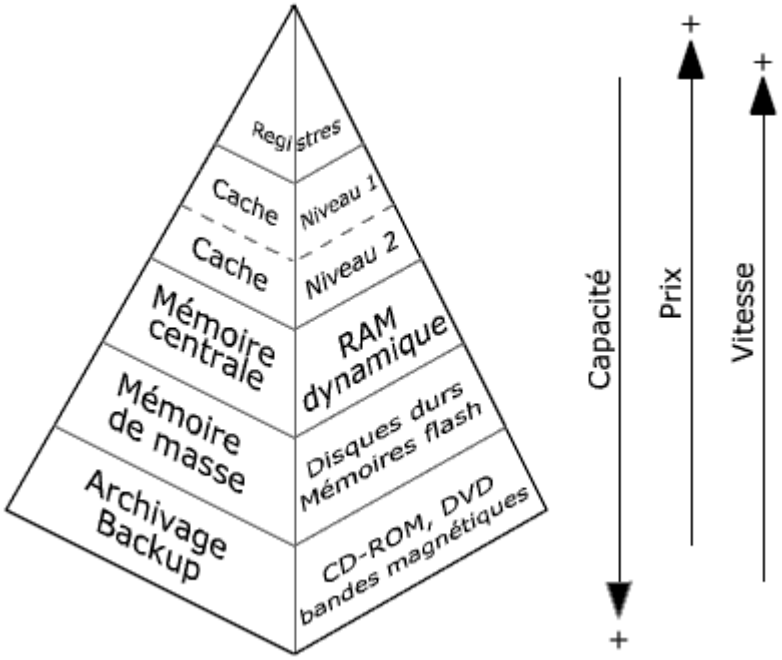
La durée d'un cycle du processeur est égale à l'inverse de la fréquence d'horloge. Si par exemple le CPU tourne à 1 GHz (10^9 Hz) son cycle dure $1/10^9$ s = 1 ns (1 nano seconde)

Pour gagner du temps, on profite du fait que le plus souvent les accès mémoire se font sur des données consécutives. On les traite en **mode rafale (*burst mode*)**

Si par exemple on accède à quatre données consécutives, seul l'adressage de la première donnée sera vraiment long, car pendant ce temps l'adressage des données suivantes s'organisent déjà. Prenons l'exemple d'une mémoire SDRAM cadencée à 133 MHz. Elle a besoin de 5 cycles de 7,5 ns pour obtenir le transfert de la première donnée mais chacun des trois accès suivants ne prend qu'un seul cycle. Ce qui fait un total de 8 cycles pour quatre accès (5+1+1+1) soit une moyenne de deux cycles par transfert.

La hiérarchie des mémoires

Le but de l'informatique étant de traiter des informations, il nous faut stocker ces informations et les programmes qui les manipulent sur divers supports : les mémoires. Il en existe de plusieurs types qui se distinguent par leur mode d'enregistrement (électronique, magnétique, optique), leur capacité, leur rapidité, le fait quelles soient volatiles ou non, leur prix, la densité d'information, la manière d'y accéder, la fiabilité etc.
Il est d'usage pour classifier les mémoires de les hiérarchiser en les situant dans une représentation appelée pyramide des mémoires.



Au sommet de la pyramide, se trouvent les registres qui font partie du processeur. Ils sont extrêmement rapides et fonctionnent à la vitesse du CPU mais ils ne peuvent contenir que quelques mots, les instructions et les données, qui y sont traités en quelques milliardièmes de secondes.

Les données que le processeur traite, doivent être facilement accessibles à proximité immédiate du CPU dans les mémoires caches disposées sur la même puce que le processeur. Ce sont des mémoires très rapides. Elles mettent à la disposition du processeur les copies de quelques ensembles de données et d'instructions prises dans la mémoire centrale trop lente par rapport au processeur. Il y a généralement deux niveaux de cache. La cache de niveau 1 étant plus rapide mais de taille plus restreinte que la cache de niveau 2.

La mémoire centrale se présente sous forme de barrettes disposées à proximité du processeur et reliées à lui par ce qu'on appelle le bus système. Les informations y sont stockées sous forme électronique comme pour les mémoires cache et les registres mais avec une technologie différente qui permet d'avoir une quantité plus importante de données pour un coût moindre. Elle est en contrepartie plus lente. La vitesse de réaction des barrettes RAM et la celle du bus système sont insuffisantes pour pouvoir répondre rapidement aux commandes du processeur. C'est la raison pour laquelle les données y sont lues non pas une par une mais par blocs mis à portée de main du processeur par l'entremise de la mémoire cache.
Sous la mémoire centrale, dans cette représentation hiérarchisée, se trouvent les mémoires de masse. Les disques durs conservent les données sous forme magnétique. Les temps d'accès sont plus lents que pour la mémoire de masse mais le coût du byte mémorisé ainsi que le volume physique pour le mémoriser sont bien moindre. Les mémoires de type flash (disques SSD) remplacent parfois avantageusement les disques magnétiques mais coûtent plus cher.

A la base de cette pyramide se trouvent les disques optiques CD-ROM ou DVD qui offrent des volumes de données plus importants encore pour un prix moindre. Leurs faibles coûts en font des supports idéaux pour l'archivage d'informations auxquelles il n'est pas nécessaire d'accéder souvent. Les bandes magnétiques ont les temps d'accès encore plus longs puisqu'elles sont à accès séquentiel. On les utilise toujours actuellement pour les backups.