Les mémoires

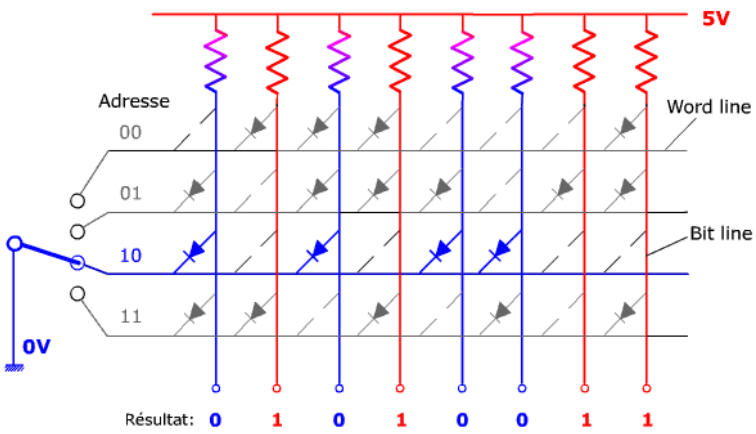
La ROM

La **ROM** ***"Read Only Memory"*** (mémoire à lecture seule) est aussi appelée mémoire morte. Il est impossible d'y écrire, leur contenu est en principe immuable. Leur principal avantage par rapport aux mémoires vives (RAM) est que les mémoires mortes n'ont pas besoins d'être alimentées électriquement pour conserver l'information. Les ROM sont programmées par leurs fabricants pour contenir des informations permanentes telles que les fonctions de certains BIOS.

Nous en verrons quatre variantes : PROM, EPROM, EEPROM et Flash EPROM

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

Le principe de fonctionnement d'une PROM est relativement simple. Cette mémoire contient une matrice de diodes. L'adresse du mot à lire agit sur un décodeur qui dans le schéma ci-dessous est représenté symboliquement par un commutateur à quatre positions. Ce schéma représente donc une PROM de 4 octets. Le code en sortie de la mémoire est une combinaison de bits à 1 et à 0. Les niveaux '1' sont fournis au travers de résistances électriques reliées à la tension d'alimentation du circuit. Par endroits, des diodes forcent les bits les bits de la ligne sélectionnée vers une tension qui correspond au niveau logique 0.



Dans une PROM vierge, les diodes sont en série avec de petits fusibles.

La programmation se fait en brûlant certains fusibles pour les positions des bits devant être mis à 1.

Cette opération est irréversible.

L'**EPROM** ***"Erasable PROM"*** est effaçable.
On efface ces mémoires en les laissant 10 à 20 minutes sous des rayons ultraviolets. Le composant possède une petite fenêtre qui permet le passage des UV. Une fois effacée, l'EPROM peut être reprogrammée.

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 **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, les GSM, les PDA et l'informatique embarquée.

La caractéristique essentielle de toutes ces « mémoires mortes », n'est donc pas qu'elles peuvent uniquement être lues mais plutôt qu'elles ne s'effacent pas quand l'alimentation est coupée.

La RAM

La **mémoire vive** est généralement appelée **RAM** pour *Random Access Memory* ce qui signifie mémoire à accès aléatoire, entendez "accès direct".

> *Elles ont été dénommées mémoires à accès aléatoire pour des raisons historiques. En effet pour les premiers types de mémoire, les cartes perforées ou les bandes magnétiques par exemple, les temps d'accès dépendaient des positions des informations sur ces supports. Avec ces mémoires à accès séquentiel, il fallait faire défiler une kyrielle d'informations avant de trouver celle que l'on cherchait.*

La RAM du PC contient tous les programmes en cours d'exécution ainsi que leurs données. Les performances de l'ordinateur sont fonction de la quantité de mémoire disponible. Aujourd'hui une capacité de plusieurs milliards d'octets (Go) est nécessaire pour pouvoir faire tourner les logiciels de plus en plus gourmands. Quand la quantité de mémoire ne suffit plus, le système d'exploitation a recours à la mémoire virtuelle, il mobilise une partie du disque pour y entreposer les données qu'il estime devoir utiliser moins souvent.

RAM statiques / RAM dynamiques

Il y a deux technologies de fabrication des RAM : statiques et dynamique, elles ont chacune leur domaine d'application.

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 à mémoriser. Les informations y restent mémorisées tant que le composant est sous tension. Les cartes mères utilisent une SRAM construite en technologie CMOS et munie d'une pile pour conserver de manière non volatile les données de configuration (*setup*) du BIOS. Le circuit de cette RAM CMOS est associé au circuit d'horloge qui lui aussi a besoin de la pile pour fonctionner en permanence même quand l'ordinateur est éteint.

La SRAM est très rapide et est pour cette raison le type de mémoire qui sert également aux mémoires cache.

La **DRAM** pour RAM dynamique est de réalisation beaucoup plus simple que la SRAM. Ce qui permet de faire des composants de plus haute densité et dont le coût est moindre.

Chaque bit d'une DRAM est mémorisé par une charge électrique stockée dans un petit condensateur. Ce dispositif offre l'avantage d'être très peu encombrant mais a l'inconvénient de ne pas pouvoir garder l'information longtemps. Le condensateur se décharge au bout de quelques millisecondes (ms). Aussi pour ne pas perdre le bit d'information qu'il contient, il faut un dispositif qui lit la mémoire et qui la réécrit de suite pour recharger les condensateurs. On appelle ces RAM des RAM dynamiques car cette opération de rafraîchissement doit être répétée régulièrement.

Structure de la RAM

L'adressage des cellules à l'intérieur des composants mémoire nécessite un certain nombre de broches pour l'interconnexion des composants au bus d'adressage et un nombre bien plus important de portes logiques pour la sélection des cellules. Le nombre de cellules adressables avec n lignes d'adresse est de 2^n. Il faut par exemple 20 lignes d'adresses pour former 2^{20} soit 1024*1024 adresses distinctes.

L'organisation la plus simple est semblable à ce que nous avions déjà vu pour l'adressage des octets d'une ROM (Figure1 page **Erreur ! Signet non défini.**) :

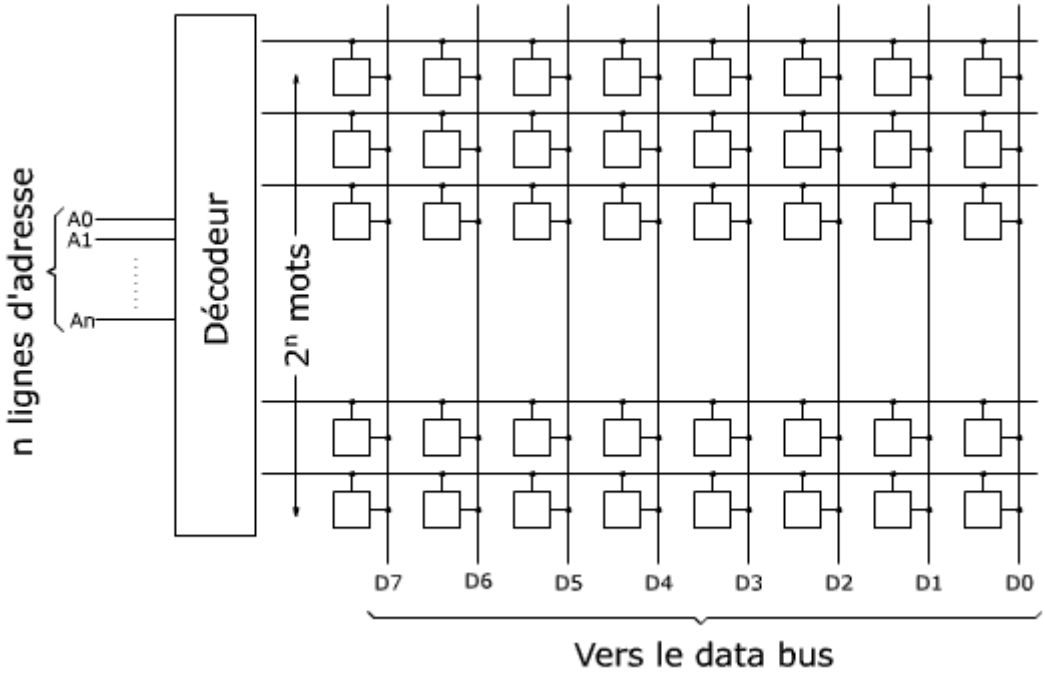


Figure 1 – Adressage linéaire de 2^n bytes

Les lignes d'adresses sont connectées aux n entrées d'un décodeur qui sélectionne une seule des 2^n lignes du composant mémoire. Les bits qui appartiennent à la ligne sélectionnée sont connectés au bus des données.

Le composant représenté ci-dessus produit 8 bits de données. Dans la pratique les puces disposées sur les barrettes RAM fournissent en général 4, 8 ou 16 bits. Elles y sont en nombre suffisant pour donner autant de bits que nécessaire pour la largeur du bus des données (64 bits pour les barrettes DIMM). (8 puces de 8 bits, ou 16 puces de 4 bits).

L'organisation des cellules adressables des RAM dynamiques (DRAM) ne peut cependant pas être aussi simple que ce que la figure 2 pourrait nous laisser croire. Imaginez par exemple que cette puce comporte 1 Go. Le décodeur possèderait donc 30 lignes d'entrées pour l'adresse sur 30 bits ($2^{30}\approx10^9$) mais il devrait aussi comporter plus d'un milliard de portes logique pour un milliard de sorties !

A cette structure linéaire, on préfère une organisation matricielle des cellules mémoire avec une matrice aussi carrée que possible.
Prenons un exemple plus simple, une RAM de 1Ko, les 10 lignes d'adresse de cet exemple (2^{10}=1024) pourraient être réparties comme suit : les 7 bits d'adresse les plus

significatifs (A9 à A3) sont connectés à un décodeur qui n'a plus que 2^7 = 128 sorties (au lieu de 1024) tandis que les 3 bits les moins significatifs de l'adresses (A2, A1 et A0) commandent un multiplexeur qui sélectionne 8 signaux (1 octet) hors de 64 colonnes.

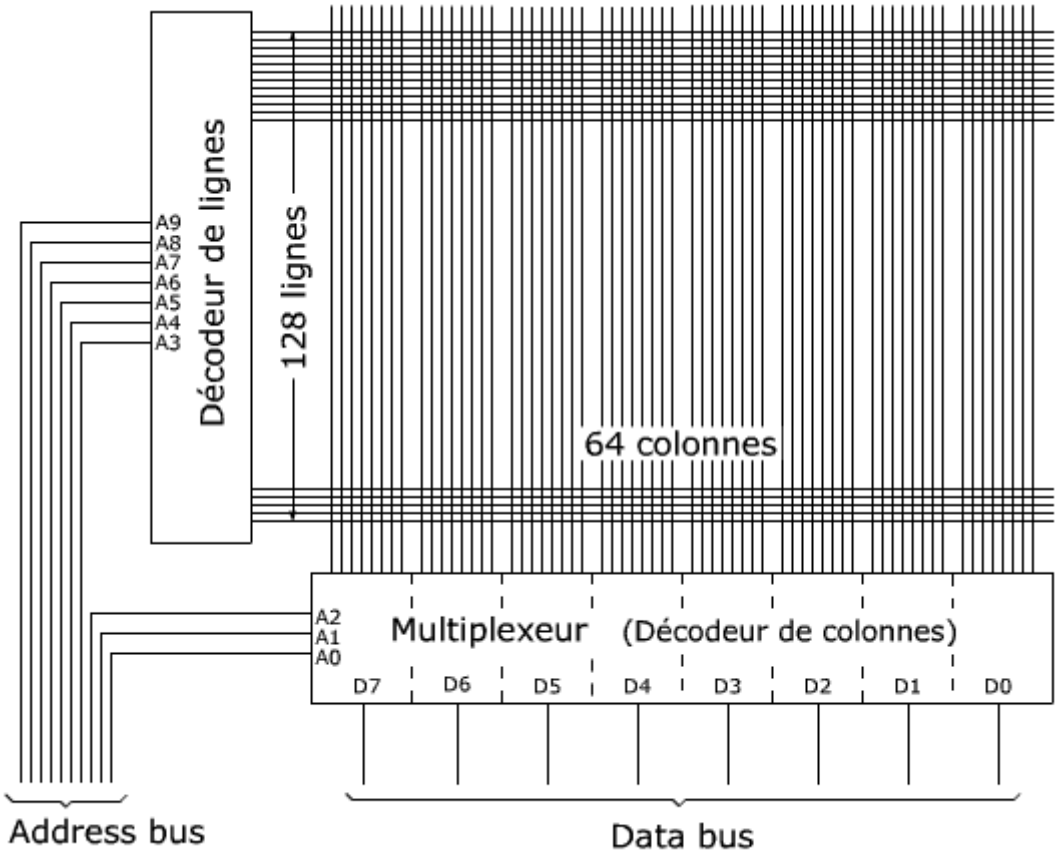


Figure 2 – Multiplexage des adresses de lignes et colonnes

Le nombre de portes logiques qui constituent le décodeur et le multiplexeur est considérablement réduit par rapport au schéma précédent mais le mode d'adressage de la RAM s'en trouve modifié. L'adressage se fait en deux dimensions : lignes et colonnes (*rows and columns*).

L'adressage se fait aussi en deux temps 1° sélection d'une ligne, 2° sélection du numéro de colonne.

Cette sélection en deux temps a pour principal objectif de réduire le nombre de contacts des puces mémoire. On divise le nombre de ligne d'adresse par deux en se servant de mêmes lignes pour véhiculer tantôt le numéro de ligne tantôt le numéro de colonne. La distinction numéro de ligne/ numéro de colonne (multiplexés sur les mêmes contacts) est rendue possible par l'ajout de signaux sur le bus de contrôle : RAS (*Row address Strobe*) et CAS (*Column address Strobe*)

Les temps d'accès

Les mémoires statiques (SRAM) ont des temps d'accès très courts adaptés aux fréquences des processeurs qui 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 = 10^{-9} s = 1 ns (1 nano seconde)

Pour gagner du temps, les barrettes mémoires sont organisées en ***bancs***, généralement quatre, entre lesquels sont distribuées une à une les adresses successives. Ainsi s'il faut accéder à quatre données contiguës, l'accès à la première requière des périodes d'attente qui ne sont plus nécessaires pour les trois données suivantes puisque les quatre adressages ont pu se faire presque simultanément. On dit que les données sont traitées en ***mode rafale*** (***burst mode***).

Une mémoire SDRAM cadencée à 133 MHz a besoin de 5 cycles de 7,5 ns (1 / 133.10^6 Hz = 7,5 10^{-9} s) 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.

Nous détaillerons le timing des RAM dynamiques un peu plus loin (0 page 8)

Types de RAM dynamiques

La préoccupation des constructeurs de mémoire est triple : ils cherchent à obtenir des mémoires de plus en plus grosses et de plus en plus rapides à un prix toujours moindre.

RAM FPM

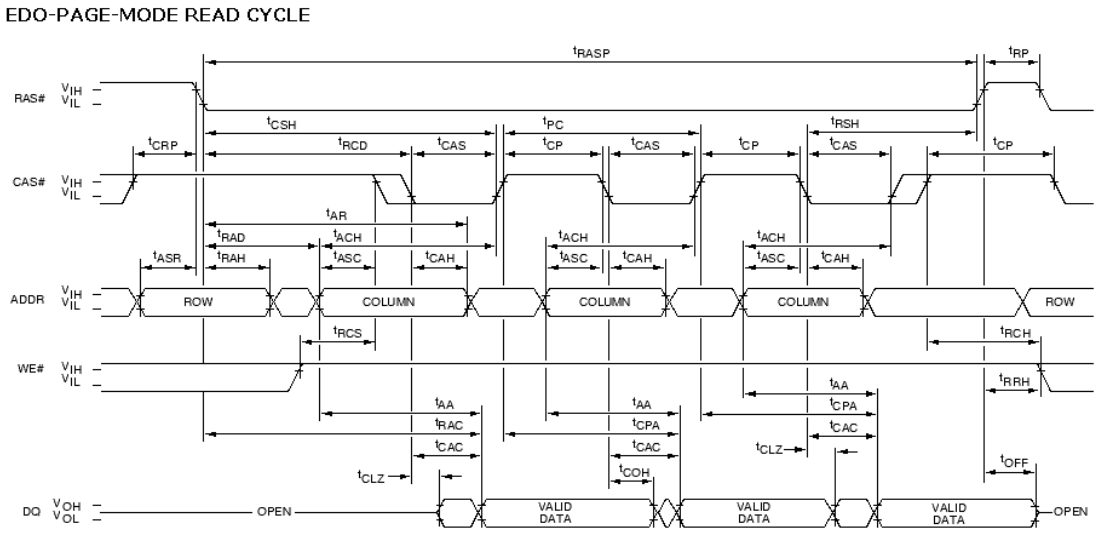
La RAM FPM *"Fast Page Mode"* , dépassée aujourd'hui, utilise un adressage en mode page. Cela consiste à adresser les données en deux temps. La première partie de l'adresse spécifie la page (une ligne) et la seconde y indique l'emplacement mémoire visé (une colonne). La plupart du temps les accès mémoire se font sur des données voisines. Les données peuvent donc être lues en rafale, la première partie de l'adressage n'est nécessaire que pour l'accès à la première donnée mais ne doit plus à être répété pour les données situées aux adresses suivantes. On pouvait lire sur les boîtiers de ces composants des nombres tels que 70 ou 80 ils indiquaient les temps d'accès en nanosecondes (70 ou 80 ns)
(Lecture en mode rafale en 14 cycles : 5-3-3-3)

RAM EDO

En 1995, la RAM EDO *"Extended Data Out"* a amélioré la technique précédente en permettant que les adressages successifs du mode rafale soient faits pendant les lectures des données précédentes. Les temps d'accès sont alors passés à 60 et 50 ns.
(Lecture de 4 données en mode rafale en 11 cycles : 5-2-2-2)

La figure ci-dessous montre à titre indicatif à quoi ressemble le timing d'une RAM EDO. Le nombre de paramètres est impressionnant. Contentons-nous d'observer que

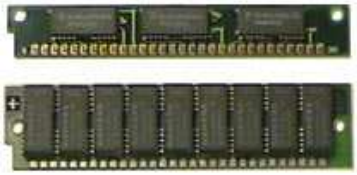
l'adresse ADDR est donnée en deux parties et que pour une même ligne (ROW) on sélectionne ici successivement 3 colonnes (COLUMN)

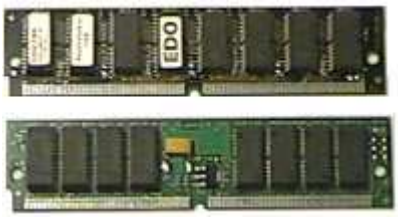


Barrettes SIMM (Modèles très anciens pour RAM dynamiques FPM et EDO)

Les SIMM (*Single Inline Memory Module*) sont des modules mémoire à une rangée de contacts. En fait, il y a des contacts sur les deux faces mais ils sont reliés par des trous métallisés.

- Les SIMM 30 broches mémorisent les données par mots de 8 bits. Elles étaient utilisées à l'époque des 486 et étaient montées par "bancs" de quatre barrettes puisque le bus de données y avait une largeur de 32 bits.





- Les SIMM 72 broches mémorisent les données par mots de 32 bits. Puisque le bus du Pentium communique sur une largeur de 64 bits, il faut monter ces barrettes par bancs de deux.

Barrettes DIMM (Modèles actuels pour SDRAM et DDRSDRAM)

Les barrettes DIMM *"Dual Inline Memory Module"*. Ce sont des modules mémoire à deux rangées de contacts. En augmentant le nombre de contacts par rapport aux barrettes SIMM les données peuvent être échangées par groupes de 64 bits, soit huit octets en une fois.

SDRAM

La SDRAM *"synchronous DRAM"* a supplanté les types de RAM précédentes en 1997. Lors d'un cycle de lecture en mode rafale, les lectures sont synchronisées avec le bus. Dès que la première donnée a pu être lue les suivantes se succèdent en étant synchronisés à la fréquence du bus système. La vitesse s'exprime dès lors en MHZ et non plus en ns.
(Lecture en mode rafale en 8 cycles : 5-1-1-1)

Les DIMM SDRAM ont 168 contacts (84 par face) Deux encoches servent de détrompeurs. Elles sont généralement alimentées en 3,3V mais il existe d'autres tensions d'alimentation. Ces variantes sont signalées par de petites modifications des positions des détrompeurs empêchant ainsi l'insertion d'une barrette prévue pour 3,3V dans un socket prévu pour alimenter des barrettes en 5V.

RDRAM

La RDRAM (Rambus DRAM) est un type de mémoire complètement différent qui a été développé par la société RAMBUS Inc en 2000 et qui était destinée aux Pentium IV. Alors que dans les modèles précédents, on essayait de transférer les données sur des bus aussi larges que possible, la RDRAM utilise un canal étroit de 16 bits seulement mais à des fréquences beaucoup plus élevées.

La RDRAM n'a pas connu le succès attendu et trop chère a vite été supplantée par la DDR SDRAM.

Les RIMM *"RDRAM Inline Memory Module"* sont des modules à deux rangées de contacts. Les détrompeurs sont deux encoches très rapprochées. Ces modules RAMBUS ont 184 contacts comme les DIMM - DDR. Les barrettes RIMM sont recouvertes par un boîtier en aluminium qui facilite la dissipation de la chaleur.

DDR SDRAM

La DDR RAM *" Double Data Rate "* (double taux de transfert) est une variante de la SDRAM dans la quelle on effectue deux transferts par cycle d'horloge.

Les puces d'une DDR200 fonctionnent donc bien à la fréquence de 100 MHz mais puisqu'il y a deux transferts par cycle cela équivaut à une fréquence de 200 MHz.

La conception de cette RAM est assez proche de celle des SDRAM ce qui a permis aux constructeurs de SDRAM de se reconvertir en l'an 2000 pour la fabrication des DDR sans investir autant que pour passer aux RDRAM.

Les barrettes DIMM équipées de DDR RAM ont 184 contacts et une seule encoche vers le milieu des contacts. La tension d'alimentation est plus basse que pour la SDRAM, elle est ramenée à 2,5 ou 2,6V selon les constructeurs.

DIMM DDR2

La DDR2 tout comme la DDR classique fait deux échanges sur le bus par cycle d'horloge. En interne par contre, elles possèdent deux canaux vers des puces. La fréquence du bus est donc double de celle des composants mémoire ce qui double une fois de plus la bande passante. Cette fois avec des puces cadencées à 100 MHz on parlera de DDR2-400. (Deux échange par cycle sur un bus cadencé à 200 MHz)

Contrairement aux DDR, premières du nom, les DDR2 possèdent 240 contacts. La tension d'alimentation est réduite à 1,8V pour limiter la quantité de chaleur produite.

Ces barrettes sont apparue en 2003 avec des fréquences de 200 et 266 MHz sous les appellations DDR2-400 / PC3200 ou DDR2-533 / PC4200.

DIMM DDR3

La DDR3 (*Double Data Rate 3rd generation)* a succédé à la DDR2 en 2007 en doublant une fois de plus le taux de transfert par rapport à la génération précédente.
En partant de cellules mémoire cadencée à 100 MHz on obtient donc des barrettes DDR3-800.

Caractéristiques des mémoires

La fréquence

Les mémoires communiquent actuellement à des fréquences de l'ordre de 133, 166 ou 200 MHz. Ces nombres représentent la fréquence du FSB, le bus système. Les désignations des barrettes DDR font référence à la fréquence de la RAM qui vaut le double de la fréquence du FSB puisque pour les DDR il y a 2 transferts par cycle d'horloge. On parle donc de DDR266, DDR333 ou DDR400.

La bande passante

Les échanges étant faits sur un bus de 64 bits, la bande passante se calcule en multipliant la fréquence de la RAM par 8 (64 bits = 8 octets). Ainsi l'appellation DDR400 est tout à fait équivalente à PC3200. Le "3200" fait référence à la bande passante alors que la valeur 400 faisait allusion à la fréquence.

Exemples :

Fréquence FSB	Désignation qui tient compte de la fréquence de la RAM	Désignation qui se réfère à la bande passante
133 MHz	DDR266	PC2100
166 MHz	DDR333	PC2700
200 MHz	DDR400	PC3200
ou	DDR2-400	PC2-3200
266 MHz	DDR2-533	PC2-4200
333 MHz	DDR2-667	PC2-5300
400 MHz	DDR2-800	PC2-6400
	DDR3-800	PC3-6400
533 MHz	DDR3-1067	PC3-8500
667 MHz	DDR3-1333	PC3-10600
800 MHz	DDR3-1600	PC3-12800
1000 MHz	DDR3-2000	PC3-16000

Le timing des RAM dynamiques

Les RAM actuelles sont caractérisées par la fréquence d'une part mais aussi par 4 nombres qui résument leur timing et que nous nommerons ici CL, TRCD, TRP et TRAS. Le premier de ces paramètres, **CL** (***CAS Latency***) est le plus important. Il est parfois directement inscrit sur les barrettes SDRAM à côté de la fréquence.

L'indication **DDR 133 CL3** signifie par exemple que la RAM est cadencée à 133 MHz, la durée d'un cycle est donc de 1/133 MHz = 7,5 ns et le « CAS Latency Time » est de 3 cycles (donc de 3 x 7,5ns = 22,5 ns)

Les constructeurs donnent des indications telles que DDR266 133MHz 2.5-3-3-6
Voyons pour comprendre ce que signifient ces 4 nombres ce qu'en disent les spécifications du JEDEC *Joint Electron Device Engineering Council*, l'organisme de normalisation des composants à semi-conducteur et en particulier des RAM. Les spécifications de toutes ces normalisations sont publiées dans le catalogue du JEDEC.
Voici en quelques mots ce qui ressort de la lecture de leur spécifications en ce qui concerne les DDR SDRAM :

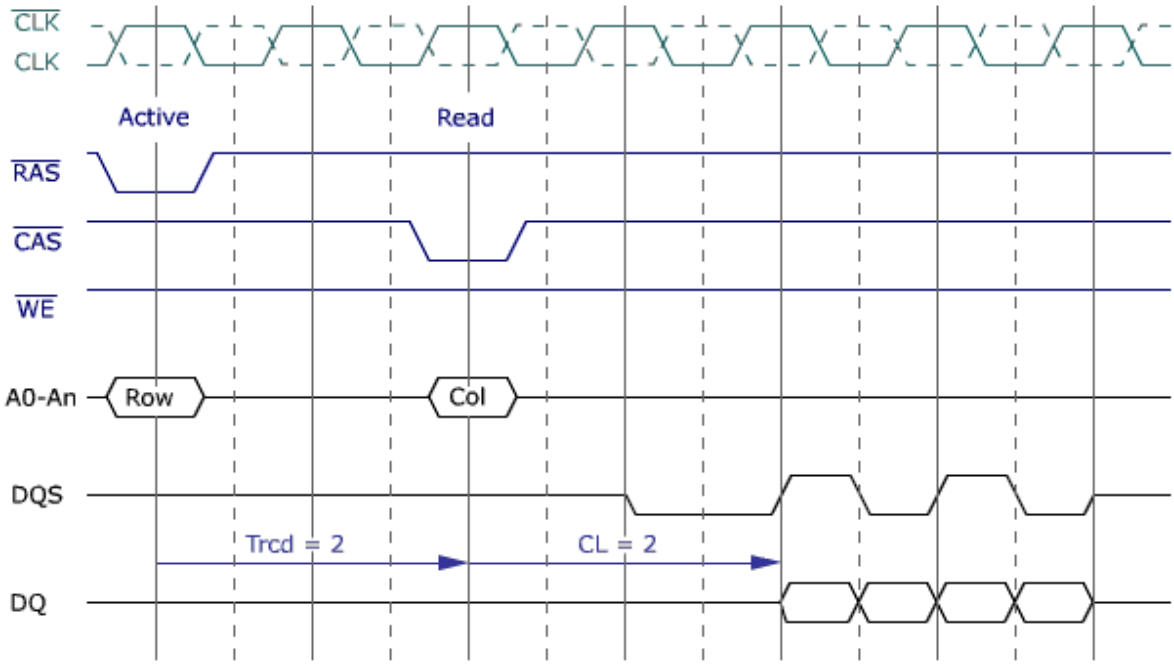
La DDR SDRAM contient 4 bancs de RAM dynamique. L'ensemble est raccordé à des tampons d'entrées sorties capables de transmettre aux broches du circuit deux mots de données par cycle d'horloge.

Les accès en lecture/écriture sont prévus pour être la plus souvent en mode rafale.
Ils commencent par une commande d'activation d'une ligne suivie par une ou des commandes de lecture ou d'écriture de colonnes. Les bits d'adresse envoyés pendant la commande d'activation sélectionnent le banc et la ligne. Ils désignent la première donnée de la "rafale".

La DDR est un composant dont le mode de fonctionnement est programmable.
On y configure entre autre la longueur (2, 4 ou 8 locations) et le type de mode rafale (séquentiel ou entrelacé), la latence en lecture (*CAS latency*) et le mode de fonctionnement.

La figure suivante montre la chronologie des principaux signaux :
/RAS (*Row Address Strobe*) passe à 0 pour indiquer que le code présent sur les bits d'adresse représente le numéro de la colonne.
/CAS (*Col Address Strobe*) fait de même pour indiquer que les bits d'adresse désignent une colonne.
La combinaison des 3 signaux /RAS, /CAS et /WE (*Write Enable*) forme des codes de commandes telles que « Activer une ligne », « Lire » =activer une colonne avec /WE=1, « Ecrire » = activer une colonne avec /WE=0, « Précharger » = désactiver une colonne etc.

Les broches DQ servent à la connexion du bus des données et DQS (*Data Strobe*) est un signal pour la synchronisation et capture des données.

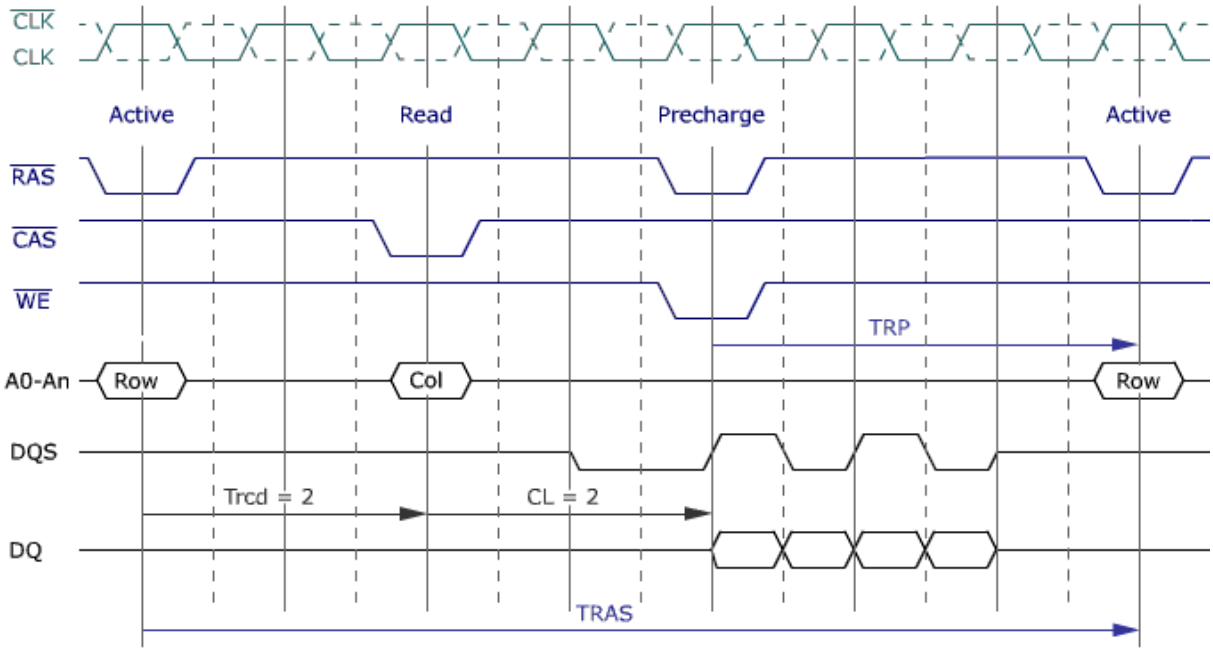


La latence en lecture notée ici « **CL** » est le nombre de cycles d'horloge entre la commande READ et l'instant où la première donnée est transmise.

TRCD (*RAS to CAS delay*) est le second nombre souvent renseigné dans les timings. Il est moins important que CL car ce temps ne concerne que les accès mémoires qui ne sont pas consécutifs sur une même ligne. Ce qui, vu le nombre de bits sur une ligne de la matrice est relativement beaucoup moins fréquent.

Il faut pour passer d'une *row* à la suivante, désactiver la première ligne avant de pouvoir en activer une autre. Le tout doit se faire en respectant deux temps minimum :

- TRP est le temps minimum entre la commande de désactivation (PRECHARGE) et l'activation d'une autre ligne (ACTIVE)
- TRAS est le temps minimum entre deux activations. C'est le quatrième et le plus grand des nombres qui caractérisent le timing.

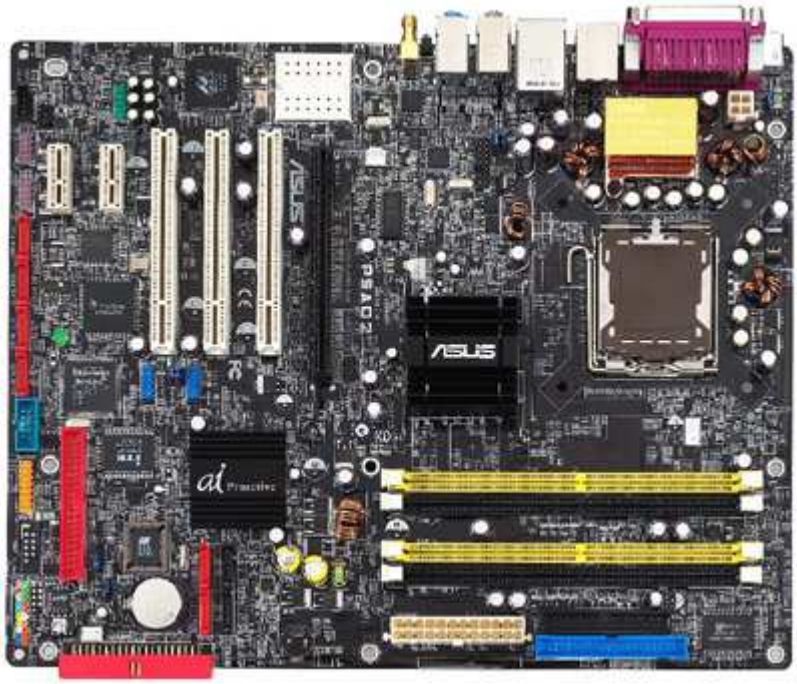


Le « Dual channel »

La dernière astuce pour augmenter la bande passante consiste à commander les barrettes DDR de sorte à faire des accès simultanés sur deux barrettes.

Initialement cette technique avait pour but de transformer le bus qui relie le contrôleur mémoire aux barrettes d'une largeur 64 bits en un bus deux fois plus large (128 bits). Actuellement le double canal est configuré pour permettre des accès simultanés sur deux bus de 64 bits indépendants que se partages les cœurs du processeur.

L'utilisateur a le choix de monter les barrettes en mode simple ou double canal. Ces cartes mère qui offrent ce choix sont identifiables par les couleurs alternées des sockets mémoire.



Pour être gérées en mode « Dual Channel », les barrettes identiques (même capacité et même fréquence) doivent être montées dans des supports appariés.

Le canal triple (*triple channel*) a fait son apparition avec le processeur Core i7 900