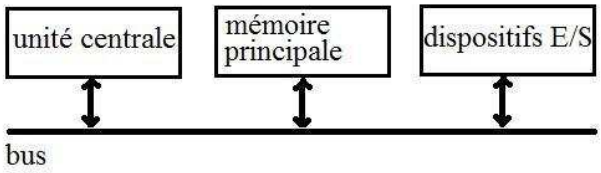
III- L'ordinateur et son architecture

Un ordinateur est formé de trois composants :

- le processeur (ou UC unité centrale, ou *CPU* pour *central processing unit*),
- les mémoires
- les dispositifs d'entrée-sortie,

ces éléments étant interconnectés entre eux par des bus. Cela constitue l'architecture de base de tout ordinateur (dite architecture de Von Neumann). Nous allons étudier chacun de ces éléments. Cela ne suffira pas pour bien comprendre comment marche un ordinateur, où tout se passe à un niveau microscopique aussi bien dans l'espace que dans le temps, mais cela donne du moins la «géographie» de l'ensemble.



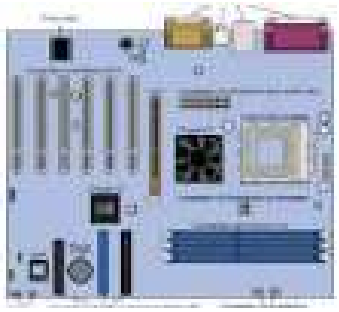
1) L'aménagement global

1-1) L'intérieur de la machine

Ces trois types d'éléments se retrouvent lorsque l'on ouvre un micro-ordinateur. Dans une sorte de boîte (le boîtier pour un ordinateur fixe) se trouve, placée au fond, une plaque de circuits imprimés, appelée la **carte-mère**. Sur cette carte sont placés plusieurs composants électroniques : le microprocesseur, des circuits mémoires, des circuits d'entrée-sortie. Il y a des circuits intégrés et d'autres composants comme des résistances et de condensateurs. La carte-mère comporte aussi des connecteurs (*slots*) permettant d'ajouter des cartes, à savoir les barrettes de mémoire RAM, et aussi des cartes d'extension : carte son, carte vidéo, modem, ainsi que d'autres composants électroniques. Il arrive aussi que les puces électroniques qui constituent ces éléments additionnels soient déjà soudés sur la carte-mère.

Dans le boîtier se trouve aussi l'alimentation, qui reçoit le courant électrique sinusoïdal du secteur (220 volts) et le transforme en courant continu de quelques volts (5, 12 volts). Au niveau des composants électroniques, c'est la présence ou l'absence de courant qui les fait réagir d'une façon ou d'une autre. Car un ordinateur n'est autre qu'un ensemble de circuits électroniques manipulant des données sous forme binaire (avec des bits 0 ou 1 correspondant à une tension électrique proche de 0 volt ou proche de 5 volts par exemple).

Le boîtier contient également divers emplacements réservés à la mémoire de masse (le disque dur, qui est un périphérique) et aux lecteurs de disques. Ces éléments sont associés aux dispositifs d'entrée-sortie. On note aussi la présence de fils, de câbles et de rubans formés de dizaines de fils parallèles accolés qui sont des éléments des bus d'interconnexion. Une pile permet d'assurer la continuité de certaines activités comme l'heure. A l'extérieur du boîtier sont placées des prises pour les connexions externes vers des périphériques comme l'écran, les haut-parleurs, … Il existe pour cela des ports série, des ports parallèles, des ports USB, etc.




A gauche la carte mère (microprocesseur, connecteurs,…), *à droite* le radiateur et le ventilateur qui sont placés au-dessus du microprocesseur

1-2) Les périphériques extérieurs

L'ordinateur traite les instructions d'un programme, au rythme de l'horloge interne elle-même alimentée par une pile. Mais pour cela il faut les lui fournir, d'où la nécessité de communiquer avec des unités d'échange. L'acquisition et la récupération de données se font essentiellement par les périphériques. On distingue :

- les périphériques d'entrée : clavier pour la saisie des textes, micro pour saisie des sons, scanner pour la saisie des images, etc.
- les périphériques de sortie : écran, imprimante, haut-parleur,…
- le stockage des données : mémoires de masse (disque dur, clés USB)
- l'échange de données entre ordinateurs : modem.

Exercices

1) Combien de touches comporte un clavier d'ordinateur ?

Plus de 100 touches.

2) Que se passe-t-il lorsque l'on appuie sur une touche du clavier ?

A chaque pression sur une touche, un signal est transmis dans l'ordinateur. Le clavier utilise un réseau matriciel permettant d'avoir les coordonnées (ligne-colonne) de la touche. Un contact électrique s'établit, qui est transmis à un micro-contrôleur. Le symbole associé à la touche est converti en code binaire (code ASCII par exemple) et envoyé en direction du processeur.

3) Comment fonctionne une imprimante à jet d'encre ?

L'encre arrive par une multitude de tuyaux extrêmement fins. Sous l'effet de la chaleur (300°) des bulles minuscules se forment dans ces tuyaux, provoquant l'éjection de micro-gouttelettes d'encre.

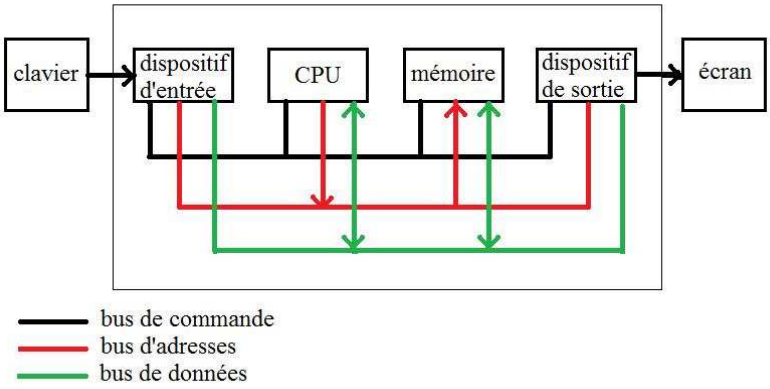
1-3) L'ordinateur minimal

L'ordinateur a besoin d'un minimum de deux choses pour travailler : le processeur et la mémoire. La mémoire contient des nombres qui y sont stockés. Ceux-ci concernent les instructions des programmes et les données à traiter, et le processeur va faire des calculs sur ces nombres. Ces nombres vont de la mémoire vers le processeur, puis retournent vers la mémoire, transformés ou non selon les cas. Le transit des données se fait par le bus des données.

En fait c'est un peu plus complexe. La mémoire est formée d'une succession de cases (cellules) numérotées. Chaque élément de la mémoire est formé de son adresse (le numéro de la case) et d'un contenu numérique (instruction ou (et) données). Si le processeur veut accéder à une case mémoire, il demande le contenu de son adresse en mémoire. Il est essentiel de ne pas confondre l'adresse d'une cellule mémoire et son contenu.

Il existe trois types de bus :

* le bus des données, où les données circulent à double sens, entre le processeur et la mémoire notamment.
* le bus d'adresses, c'est là que passe le numéro de la case dont le processeur demande le contenu, il est à sens unique, du processeur vers la mémoire
* le bus de commande (ou de contrôle), c'est là que passent les actions à réaliser, le tout dans une certaine synchronisation.



Un exemple minimal d'ordinateur, sans périphériques, est constitué par un circuit intégré tel qu'on le trouve dans une machine à laver par exemple. Il possède une mémoire ROM non volatile où sont gravés une fois pour toutes les quelques programmes nécessaires à son fonctionnement, ainsi qu'un processeur qui traite ces programmes selon le choix de l'utilisateur. C'est là un ordinateur minimal. Un véritable ordinateur n'est qu'une extension de ce modèle et il devient une machine universelle, permettant de recevoir n'importe quel programme conçu par un utilisateur extérieur.

2) Le processeur ou unité centrale

Le processeur ou CPU (*Central Processing Unit*) est aussi appelé unité centrale (UC). C'est le chef d'orchestre de l'ordinateur. Certains spécialistes l'appellent aussi le « cerveau ». Mais c'est un cerveau vide, puisqu'il ne fait qu'obéir aux ordres qu'il reçoit. Il fonctionne séquentiellement au rythme d'un signal de synchronisation - le signal d'horloge en forme de signal carré périodique-. Il a pour charge d'exécuter les programmes qui sont stockés dans la mémoire principale, instruction après instruction, au rythme de l'horloge.

L'unité centrale est composée de deux unités distinctes :

- une unité de commande (ou unitéde contrôle ou encore séquenceur) qui organise l'enchaînement des étapes de la tâche à effectuer, ainsi que le transfert des données. Elle va chercher les instructions situées dans la mémoire, les charge, les décode. C'est elle qui distribue les signaux de commande aux circuits concernés.
- une unité arithmétique et logique (UAL ou ALU pour *Arithmetic and Logical Unit*) qui exécute les opérations indiquées dans les instructions, comme par exemple une addition ou un OU logique.

L'unité centrale possède aussi sa propre mémoire de travail privée, de petite taille (quelques octets) où la lecture et l'écriture se font très rapidement. Il s'agit des registres au nombre d'une dizaine, voire de quelques centaines, chargés de stocker des résultats temporaires ou des informations de commande. Un des registres est le compteur ordinal ou compteur de programme (*PC* pour *Program Counter*) qui pointe sur la prochaine instruction à charger dans l'unité centrale. Un autre est le registre instruction (*RI*) qui contient l'instruction en cours d'exécution. Un autre est l'accumulateur, qui stocke des données en cours de traitement.

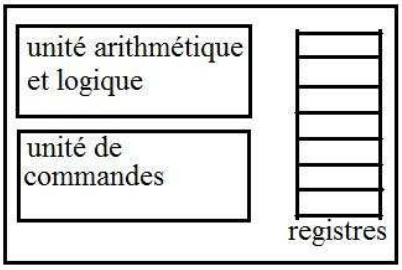


Schéma simplifié de l'unité centrale de l'ordinateur

Qu'est-ce qu'une instruction ? C'est une opération élémentaire exécutée par le processeur, cette opération étant associée à des données sur lesquelles agit l'opération (on les appelle les opérandes). L'instruction est un mot binaire composé de deux champs : le code de l'opération et le code des opérandes. Le jeu d'instructions comporte :

* des opérations arithmétiques, comme l'addition
* des opérations logiques (et, ou, non,…)
* des opérations de lecture ou écriture dans la mémoire
* des opérations de test ou de branchement (pour faire des boucles dans un programme).

Un microprocesseur est caractérisé par

* sa fréquence d'horloge (en GHz),
* le nombre d'instructions qu'il peut exécuter par seconde (en MIPS)
* la taille des données qu'il est capable de traiter, en bits. Un processeur 32 bits traite des mots de 32 bits, soit 4 octets.

L'exécution d'une instruction est cadencée par l'horloge interne, pouvant demander plusieurs tops d'horloge, et elle se fait en répétant les étapes suivantes :

- Envoi par le processeur de l'adresse de l'instruction à traiter (adresse située dans le registre PC) en direction de la mémoire, puis renvoi de l'instruction dans le registre instruction du processeur.

- Analyse et décodage de l'instruction qui vient d'être chargée, par l'unité de contrôle.

- Localisation en mémoire d'éventuelles données nécessaires à l'instruction (les opérandes), et dans ce cas chargement de ces données dans les registres.

- Exécution de l'instruction dans l'UAL, et stockage du résultat par le biais de l'unité de contrôle.

- Modification du compteur ordinal (PC) pour qu'il pointe sur l'instruction suivante.

C'est là un premier pas pour comprendre comment fonctionne un ordinateur. Nous y reviendrons dans le chapitre suivant.

Imaginons un interprète [1] qui fait une traduction en simultané dans une conférence. Le processeur joue ce rôle de traducteur en temps réel. Il est en quelque sorte programmé pour prendre, décoder et exécuter les instructions d'un autre programme, venu de l'extérieur, celui que lui propose l'utilisateur. Autrement dit, il existe une certaine interchangeabilité entre un processeur matériel et un interpréteur logiciel, entre le hard et le soft. Selon la part relative de l'un et de l'autre, il existe deux types d'architecture : l'architecture *CISC* ou l'architecture *RISC*.[2]

Le fonctionnement séquentiel de l'unité centrale, une instruction après l'autre, présente des limites en termes de performances. Si l'on veut arriver à traiter 500 millions d'instructions par seconde (500 MIPS), il est nécessaire d'introduire du « parallélisme », afin de pouvoir traiter plusieurs instructions simultanément. En faisant plusieurs choses en même temps, la vitesse d'exécution des programmes augmente. Il existe deux sortes de parallélisme, soit au niveau des instructions, notamment avec la technique du pipeline [3], soit au niveau de l'unité centrale, avec plusieurs processeurs fonctionnant en parallèle [4].

---

[1] « interprété » signifie traduction en temps réel par opposition à « compilé » qui signifie traduction préalable du programme en entier, comme on le fait pour un livre. Cette traduction aboutit à ce que l'on appelle l'« exécutable » (le *.exe* en langage *C*).

[2] Dans les premiers temps de l'informatique, avec l'architecture CISC (*Complex Instruction Set Computer*) les fabricants ont plutôt conçu des ordinateurs utilisant largement les interpréteurs sous forme logicielle. Les microprocesseurs étaient réduits à leur plus simple expression, et la complexité était reléguée dans la mémoire contenant l'interpréteur. La conception d'un composant électronique complexe était contournée au profit de la conception d'un logiciel complexe.

Plus tard est arrivée la technologie RISC (*Reduced Instruction Set Computer*) où l'aspect interpréteur logiciel était supprimé. De tels systèmes possèdent un petit nombre d'instructions simples qui s'exécutent en un seul cycle d'horloge, et plusieurs instructions simples permettent de produire une instruction complexe, comme l'addition de nombres en flottants. Aujourd'hui les principes de la conception RISC sont assez largement adoptés. Toute instruction est traitée directement par des composants matériels, sans aucune interprétation logicielle.

[3] Le pipeline comporte plusieurs étages, chaque étage ayant pour office de faire l'une des tâches inhérente à l'exécution des instructions. Plusieurs registres font office de mémoire tampon (*buffer*). Prenons un modèle simplifié. Le premier étage du pipeline a pour spécificité de chercher les instructions en mémoire principale et de les ranger dans le buffer. Le deuxième étage est chargé de décoder les instructions. Le troisième étage a pour mission d'exécuter les instructions. Dans une première étape de temps (un cycle d'horloge), une première instruction passe dans le premier étage du pipeline. Lors de la deuxième étape de temps, elle est traitée dans le deuxième étage du pipeline, tandis qu'une deuxième instruction commence à être traitée au premier étage. Et ainsi de suite. Le pipeline fait du travail à la chaîne, au sens propre du terme. La vitesse de traitement est multipliée par le nombre des étages. L'architecture en pipeline permet de multiplier les performances par dix, au maximum.

[4] Pour multiplier les performances par 100, il faut faire travailler plusieurs processeurs ou plusieurs ordinateurs en parallèle. La société d'ordinateurs *Cray* est à l'origine du premier ordinateur vectoriel, où le même type d'opérations est fait en simultané sur une colonne de processeurs, qui lâchent tous leur réponse en même temps. Puis sont apparus des systèmes dits multiprocesseurs, où tous les processeurs sont autonomes et indépendants. Ils ont chacun leur UAL et leur unité de commande. Le lien qui les unit se fait par le biais de la mémoire principale qu'ils se partagent. C'est le système d'exploitation qui doit gérer ce partage pour éviter qu'un processeur n'écrive dans une zone qui n'est pas la sienne. Un groupe de 64 processeurs reste relativement simple à concevoir. Pour aller plus loin, les concepteurs se sont orientés vers des systèmes parallèles basés sur des ordinateurs autonomes interconnectés en réseau. Il s'agit de systèmes distribués (*multicomputer*).

3) La mémoire principale de l'ordinateur

Il existe deux catégories de mémoires présentes dans les ordinateurs : la mémoire principale et la mémoire secondaire qui fait partie des périphériques. A elle seule, la mémoire principale ne suffit pas pour traiter toutes les données et programmes. Elle doit être secondée par des mémoires secondaires. La mémoire principale, qui contient le programme en cours d'exécution, a un temps d'accès assez faible, quelques dizaines de nanosecondes, et ce temps est indépendant de la position du mot cherché (par son adresse) dans la mémoire. Par contre, les mémoires secondaires ont des temps d'accès nettement plus longs, qui se comptent en millisecondes.

Dans la mémoire principale sont rangés les programmes et les données que va mouliner la machine. C'est là que le processeur lit et écrit. Cette mémoire travaille comme le reste en arithmétique binaire. Elle ne contient que des 0 et des 1. Ces *bits* -éléments atomiques d'information-, sont les plus petits que l'on puisse envisager. Une telle arithmétique binaire est considérée comme la plus efficace, la moins susceptible d'erreurs. Il suffit en effet de savoir distinguer deux états associés à un phénomène physique continu comme une tension électrique. Le 0 correspond à une tension proche de 0, le 1 à la tension proche du maximum (5 volts par exemple). Avec 16 bits, on peut déjà représenter 2^{16} = 65 536 nombres différents. Avec 32 bits, on a 2^{32} nombres, soit quelques milliards. Bien sûr, si l'on savait découper précisément en dix parties une tension de 0 à 10 volts, on pourrait utiliser l'arithmétique décimale, mais on n'en est pas là.[5]

La mémoire est formée d'un grand nombre de cellules, contenant toutes le même nombre de bits, et possédant chacune un numéro qui constitue son adresse, donnant par là-même le moyen d'y accéder précisément. La mémoire est comme une longue rue où les maisons (les cellules) se succèdent avec leurs adresses qui augmentent de un en un. Ces adresses sont aussi en binaire. Avec *k* bits par adresse, on peut obtenir 2^k adresses.

La cellule ou case mémoire constitue la plus petite quantité d'information à laquelle on peut s'adresser. Les fabricants d'ordinateur se sont accordés à utiliser des cellules de 8 bits, soit un octet (ou *byte*). Les cellules peuvent aussi être regroupées en mots. Si ces mots sont formés de 4 cellules, on aura un ordinateur 32 bits. Les registres auront alors 32 bits, et les instructions, comme celles d'addition ou soustraction, manipuleront des mots de 32 bits. Là où les fabricants ne se sont pas mis d'accord, c'est sur le sens de la numérotation des octets dans un mot, de gauche à droite pour les grosses machines IBM par exemple, de droite à gauche pour les microprocesseurs Intel. Cela peut constituer un inconvénient majeur pour l'échange de données entre machines différentes.

Malgré le découpage de la tension en deux valeurs extrêmes, les mémoires sont sensibles à divers phénomènes comme les surtensions ou les parasites sur les lignes d'alimentation. Pour se prémunir contre ces interférences, elles utilisent des techniques de détection et de correction d'erreurs. A chaque mot mémoire s'intègrent quelques bits supplémentaires pour contrôler et vérifier la bonne qualité de l'information. [6]

---

[5] Remarque : sur certains grands ordinateurs, il arrive que l'on utilise quatre bits pour représenter un chiffre décimal (de 0 à 9). Chaque chiffre d'un nombre en base 10 est écrit avec 4 bits. Par exemple le nombre 3945 s'écrit 0011 1001 0100 0101. Avec 16 bits au total on peut obtenir 10 000 nombres, de 0 à 9999. Avec le même nombre de bits, mais directement en binaire, on obtient 2^{16} = 65 536 nombres, beaucoup plus.

[6] Comme exemple de code détecteur d'erreur, il existe le bit de parité. A chaque mot est ajouté un bit, qui est choisi de façon que la somme des 1 du mot et du bit supplémentaire soit paire. Ainsi le mot 0100110 sera suivi du bit de parité 1. Si un bit du mot se trouve modifié, il y aura détection d'erreurs. Ce sera aussi le cas si c'est le bit de parité qui se trouve modifié, alors que le mot reste valable. Il en est de même si un nombre impair de 1 se trouvent transformés en 0. Par contre, si deux 1 sont transformés, la détection d'erreurs ne se fera pas. Autrement dit, ce code détecteur d'erreurs ne peut vraiment être utilisé que pour détecter une erreur simple et unique. Cette indication d'erreur sera transmise au processeur pour qu'une action éventuelle soit

La mémoire principale se divise en deux types de mémoire : la mémoire vive, accessible en lecture-écriture, mais qui perd ses données lors de la mise hors tension, et la mémoire morte, à lecture seule, mais qui conserve ses données.

3-1) La mémoire vive ou RAM

Les mémoires qui permettent la lecture et l'écriture d'informations sont en général les mémoires dites RAM pour *Random Access Memory*, mémoire à accès aléatoire, ce qui signifie simplement que le temps d'accès à l'information est le même quel que soit le mot sollicité (cela par opposition à un accès séquentiel). Autrement dit, le temps mis pour accéder à une information ne dépend pas de son adresse. Ces mémoires garantissent la mémorisation de l'information aussi longtemps que l'alimentation électrique est maintenue. On distingue les RAM statiques (*SRAM*) et les RAM dynamiques (*DRAM*).

• Les SRAM sont très rapides : le temps d'accès est de l'ordre de quelques nanosecondes. Mais elles sont d'un coût élevé, chaque bit nécessitant quatre transistors formant une bascule électronique, plus deux transistors d'accès. Ces mémoires servent souvent comme mémoires caches.

• les DRAM ont, elles, une structure interne à base de transistors et de condensateurs. Comme la charge d'un condensateur a tendance à diminuer dans le temps, les RAM dynamiques doivent être régulièrement rafraîchies, environ une fois toutes les quelques millisecondes, afin d'éviter la disparition de l'information. Il faut prévoir des interfaces permettant de donner ces instructions de rafraîchissement, ce qui est un handicap. Mais les DRAM ont comme atout d'avoir une capacité de mémoire importante. Elles ont aussi la simplicité d'utiliser un seul transistor et un condensateur pour chaque bit, ce qui facilite leur intégration dans un système. D'où un coût moins élevé que les SRAM. Aussi les mémoires principales d'ordinateur sont-elles presque toutes des mémoires DRAM, et la tendance va en faveur des SDRAM synchrones (*Synchronous DRAM*) pilotées par une horloge commune.

Barrette de mémoire RAM à placer sur un support SIMM. Une mémoire de 32 Mo est formée de 8 circuits mémoires de 4 Mo.

Les mémoires RAM ne suffisent pas pour répondre à tous les problèmes. Dans de nombreuses applications, des bouts de programmes et des données internes doivent obligatoirement être mémorisées de façon permanente, même lorsque l'alimentation électrique est fermée. On utilise alors des mémoires ROM (*Read Only Memory*), qui ne permettent que la lecture.

3-2) La mémoire morte ou ROM

Il y a aussi accès arbitraire aux données, comme pour les RAM, mais l'information écrite dans ces mémoires, faite par le fabricant de la machine, ne peut plus être ni modifiée, ni effacée, intentionnellement ou pas, par l'utilisateur. Elle peut seulement être lue. Ce type de mémoire est indispensable, surtout au démarrage de l'ordinateur, car il faut lui donner un programme initial, une amorce, notamment ce que l'on appelle le BIOS (*Basic Input Output System*). Dès que l'on allume

---

entreprise. Pour la correction d'erreurs, on utilise souvent le code dit de Hamming. Le fait de faire un XOR (ou exclusif) entre deux mots permet de savoir le nombre de chiffres qui diffèrent dans un mot (puisque 0+0=0, 1+1=0, 0+1=1 et 1+0=1). Le nombre de 1 obtenus dans le XOR de deux mots est appelée la distance de Hamming. On applique alors l'algorithme dit de Hamming pour déceler où se trouve l'erreur et la corriger.

l'ordinateur, le BIOS prend le contrôle, il vérifie le fonctionnement du matériel et teste les composants. Il passe ensuite la main au système d'exploitation.[7]

Les informations gravées dans une ROM sont enregistrées lors de sa fabrication. Cela se fait par une technique d'insolation sur le support photosensible, en passant à travers un masque reflétant les configurations binaires désirées. Le prix d'une ROM est pour cette raison plus élevé que celui d'une RAM.

D'autres types de mémoires mortes existent, qui permettent non seulement la lecture des informations, mais aussi l'écriture en mémoire morte, ainsi que leur effacement et leur modification. On distingue :

- Les mémoires PROM (*Programmable ROM*), programmables, mais une fois pour toutes.

- Les mémoires EPROM (*Erasable PROM*) : ce sont des PROM que l'on peut programmer, mais encore effacer et reprogrammer. Pour effacer, on doit utiliser un matériel spécialisé. On soumet la fenêtre qui se trouve au centre de la mémoire à un rayonnement ultraviolet, pendant une quinzaine de minutes. Après cela, tous les bits sont mis à 1. Ces mémoires sont pratiques dès qu'il s'agit de mettre au point une application, où l'on est amené à pratiquer des modifications et rectifications. Cela rend les EPROM plus économiques que les simples PROM.

- Les mémoires EEPROM (*Electrically Erasable Memory*). L'effacement se fait dans le cas présent en appliquant une impulsion électrique. L'avantage sur les modèles précédents est la possibilité d'effacer et de reprogrammer une EEPROM sans la déplacer de son support. Mais ces EEPROM son dix fois plus lentes que les RAM et ont des capacités 100 fois plus faibles. Aussi les utilise-t-on surtout dans les situations où la non-volatilité est cruciale.

- Les mémoires *flash*. Elles ressemblent aux EEPROM, dans la mesure où l'on efface totalement ou partiellement une mémoire flash par une impulsion électrique.[8] Ces mémoires flash sont à l'œuvre dans de nombreux équipements, comme les appareils photos ou baladeurs numériques ainsi que les clés USB. Peut-être seront-elles amenées à concurrencer les disques durs, grâce à leurs meilleures performances. Le temps d'accès à une mémoire flash est en effet d'une centaine de nanosecondes, contre une dizaine de millisecondes pour les disques durs. Mais en l'état actuel des techniques elles ont une légère tendance à se dégrader, devenant inutilisables après une centaine de milliers de réécritures, d'où une certaine prédominance des disques durs.

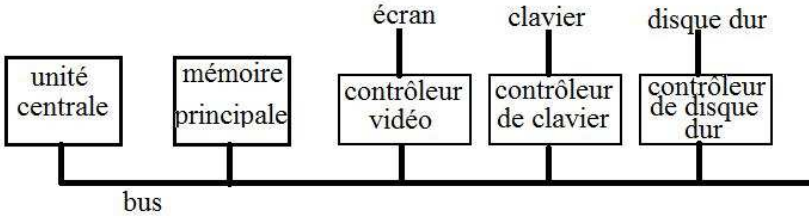
4) Le bus et les entrées-sorties

Les dispositifs d'entrée-sortie permettent d'assurer la communication entre l'ordinateur proprement dit et le monde extérieur. C'est là que viennent s'accrocher les terminaux et périphériques indispensables que sont le clavier, l'écran, la souris, etc., ainsi que le modem et les canaux d'échanges de données comme le réseau Internet. En informatique industrielle, il y a aussi divers capteurs, senseurs, sondes, etc. Chaque périphérique possède sa zone d'adresses spécifique dans la mémoire principale, et un décodeur d'adresses lié au bus d'adresses se charge d'aiguiller les données concernées, lorsqu'elles circulent sur le bus des données.

---

[7] Lorsque l'on allume l'ordinateur, on a la possibilité d'accéder au BIOS, possède un *setup* comme on le constate au bas de l'écran, *press DEL to enter SETUP* (le programme de configuration). Mais mieux vaut ne pas trop jouer avec.

[8] Mais dans le cas de la mémoire flash, l'effacement minimal est à l'échelle d'un secteur (un bloc) de données, à la différence de l'EEPROM où il peut se pratiquer au niveau d'un octet.



Chaque périphérique a son propre contrôleur, installé entre lui et le bus. Ce contrôleur est une carte additionnelle insérée dans le châssis du micro-ordinateur et raccordée à un des *slots* du bus. Le rôle du contrôleur est de piloter son périphérique en gérant ses accès au bus. Par exemple, lorsqu'un programme souhaite obtenir des données situées sur un disque externe, il en fait la demande au contrôleur du disque. Le contrôleur transmet au disque les commandes permettant d'accéder à l'information demandée, notamment le positionnement de la tête. Lorsque la piste et le secteur sont localisés, le disque envoie au contrôleur les informations lues par la tête, où elles arrivent sous forme d'un flot de bits en série. A partir de ce flot continu, le contrôleur construit les mots destinés au système, et les transmet à la mémoire principale dès qu'ils sont assemblés.[9] Lorsque le transfert est terminé, le contrôleur génère une interruption. Le système d'exploitation est alors prévenu qu'une opération d'entrée-sortie est terminée.

Comment fonctionne le bus ? Le bus sert à double sens, aussi bien du contrôleur vers le processeur qu'en sens inverse, lorsque ces deux éléments dialoguent avec la mémoire. Le contrôleur et le processeur ne peuvent pas utiliser le bus simultanément, car ce dernier n'accepte qu'une seule communication à la fois. En cas de conflit, un arbitrage a lieu, et la préférence est en général donnée au contrôleur, sinon il faudrait arrêter le disque externe alors qu'il est à pleine vitesse.

Si l'évolution technologique a provoqué un gain énorme des performances des processeurs, mémoires, entrées-sorties, et périphériques, les bus en ont moins profité. Les fabricants de cartes additionnelles, suivant en cela les vœux des acheteurs, préfèrent continuer à diffuser leurs produits, en garantissant la compatibilité entre les anciens et les nouveaux modèles d'ordinateur, alors qu'une modification du bus obligerait à tout changer. [10]

Le bus, qui constitue la structure d'interconnexion entre diverses unités fonctionnelles d'un ordinateur, est en fait un groupe de lignes qui ont chacune leur propre fonction. Comme on l'a déjà vu précédemment, on distingue le bus (ou la ligne) d'adresses, le bus de données et le bus de commandes. Pour obtenir une instruction du programme qu'il est en train d'exécuter, le microprocesseur commence par placer l'adresse de l'instruction sur le bus d'adresses. Il active ensuite un signal du bus de commande pour indiquer à la mémoire qu'il demande une opération de lecture. En réponse, la mémoire place le mot demandé, l'instruction, sur le bus de données et prévient le microprocesseur que l'opération est terminée, grâce à un signal sur le bus de commande. Le microprocesseur peut alors lire l'instruction présente sur le bus de données et la range dans l'un des registres mémoires.

[9] Si le contrôleur lit ou écrit des mots en mémoire sans intervention du processeur, on parle d'accès direct à la mémoire (DMA pour *Direct Memory Access*).

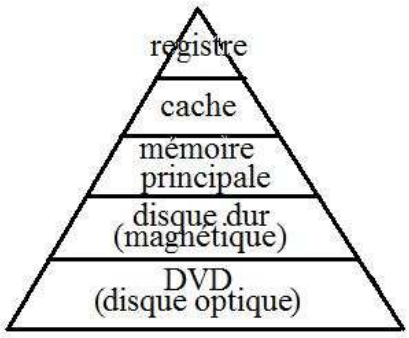
[10] Le problème s'est posé à IBM, lorsqu'il a sorti son appareil PS/2 en y mettant un nouveau bus, bien plus performant, mais différent de celui du PC. Les fabricants de clones de PS/2 n'ont pas suivi IBM et ont continué à utiliser le bus classique. Confronté à des problèmes d'incompatibilité, IBM a dû jeter l'éponge et remettre les bus traditionnels. Les bus ont cependant évolué peu à peu au fil des ans.

En fonction du type de communication à travers le bus, une hiérarchie s'établit entre les diverses unités ou circuits concernés. On distingue celle qui a l'initiative de la communication sur le bus, on l'appelle maître, et celle qui reste passive, à savoir l'esclave. Par exemple, lorsque le microprocesseur demande au contrôleur de gestion du disque dur de lire un bloc d'information, il est du type maître, tandis que le contrôleur est du type esclave. Mais si c'est le contrôleur qui s'adresse à la mémoire principale pour lui dire d'enregistrer une information en provenance de son disque, il devient maître à son tour.

En général, les signaux électriques qui transitent dans le bus ne sont pas suffisamment puissants pour ne pas subir d'affaiblissement ou de dégradations gênantes. Pour y remédier, les circuits à l'état maître disposent de composants capables de redonner de la puissance aux signaux. On appelle *drivers* ces transmetteurs amplificateurs de puissance. Pour les circuits de type esclave, des récepteurs de bus se chargent de recevoir et d'interpréter des signaux quelque peu dégradés. Enfin, les circuits qui peuvent être tantôt maîtres, tantôt esclaves, utilisent des composants mixtes appelés *transceivers*.

Selon le cadencement de leurs échanges, les bus sont soit synchrones, soit asynchrones. Un bus synchrone dispose d'une ligne d'horloge spécifique pilotée par un oscillateur à quartz. Le cycle du bus est alors la période de l'horloge. Une horloge de fréquence 40 MHz donne un cycle de bus de 25 nanosecondes. Même s'il existe des bus bien plus rapides, cela reste lent par rapport à la vitesse d'horloge des microprocesseurs (des centaines de MHz). Par contre un bus asynchrone ne dispose pas d'horloge. Un cycle de bus est plus ou moins long, et il n'est pas nécessairement le même pour toute liaison maître-esclave. Pour satisfaire de façon optimale la connexion d'un ensemble de circuits hétérogènes, il est préférable d'utiliser un bus asynchrone plutôt qu'un bus synchrone, car il n'impose pas un cadencement d'horloge prédéfini pour assurer le fonctionnement du bus. Bien que l'avantage du bus asynchrone soit net, par les libertés supplémentaires qu'il permet, la plupart des bus d'ordinateurs demeurent synchrones. Il est en effet plus simple de concevoir un bus synchrone. Le processeur place ses signaux sur le bus, et la mémoire ou le circuit d'entrée-sortie ne font que répondre au rythme des sollicitations, sans qu'il y ait besoin de concertation ou de synchronisation avec le processeur.

5) Organisation hiérarchique des mémoires



De haut en bas de la pyramide des mémoires, trois paramètres interviennent :

* les temps d'accès aux informations sont de plus en plus longs (de l'ordre de la nanoseconde pour les registres, de quelques dizaines de nanosecondes pour la mémoire principale, de quelques millisecondes pour le disque dur, et quelques secondes pour les disques optiques).

* les capacités de stockage sont de plus en plus grandes (quelques dizaines d'octets pour les registres, quelques Go pour la mémoire principale, quelques dizaines ou centaines de Go pour le

disque dur, mais moins pour les disques optiques : un DVD simple face contient 4,7 Go, mais on peut en posséder autant que l'on veut.

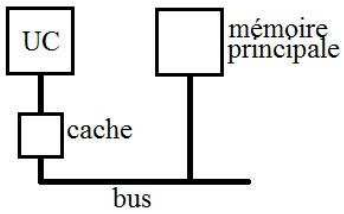
* le prix du stockage d'un octet est de plus en plus faible.

Un problème inhérent à l'ordinateur est la différence du temps de réaction entre le processeur et les mémoires, celles-ci étant beaucoup moins rapides. Si les performances des mémoires s'améliorent au fil des ans, celles des processeurs augmentent aussi, et l'écart de rapidité demeure constant. Sur une puce de silicium, on met toujours plus de composants. Pour les mémoires c'est surtout leur capacité de stockage qui s'est multipliée, et non leur vitesse de fonctionnement. Lorsque le processeur sollicite la mémoire, il passe une bonne partie du temps à attendre que la mémoire réagisse. Il peut s'écouler trois cycles d'horloge avant que le processeur n'obtienne la donnée qu'il a demandée en lecture. Une façon de résoudre ce problème consiste, dès que la demande *Read* de lecture est faite, à bloquer le processeur lorsque l'instruction qui suit a besoin d'utiliser le mot mémoire. Mais ces blocages répétés deviennent vite un handicap.

En fait, il serait possible de construire des mémoires aussi rapides que les processeurs, mais pour cela il faudrait les intégrer sur la puce du processeur. Car c'est le temps de propagation du mot mémoire sur le bus qui ralentit principalement les opérations. Mais l'intégration d'une mémoire de grande dimension sur la puce augmenterait notablement ses dimensions et son prix, ce qui va à l'encontre des intérêts économiques des fabricants d'ordinateur. La règle générale est de disposer d'une faible quantité de mémoire très rapide à proximité du processeur –ce sont les registres, et d'une grosse quantité de mémoire principale nettement plus lente. Nous avons déjà vu ces deux types de mémoires. Il nous reste à voir les autres mémoires.

5-1) La mémoire cache

Une solution intermédiaire, faisant le compromis entre la rapidité et la capacité de stockage pour un prix raisonnable, consiste à employer une mémoire cache. Celle-ci est placée entre le processeur et la mémoire principale. Cette mémoire cache contient les mots mémoires les plus fréquemment utilisés. Lorsque le processeur à besoin d'un mot mémoire, il commence par le chercher dans la mémoire cache, où le temps d'accès est réduit. C'est seulement s'il ne le trouve pas qu'il va le prendre dans la mémoire principale. La mise en place d'une mémoire cache améliore fortement les performances du processeur. Il convient de bien choisir la taille du cache : plus il est grand, meilleure est la performance, mais il est aussi plus coûteux. Il faut définir aussi une politique de gestion du cache, en déterminant quels sont les mots à y placer, et pendant combien de temps.



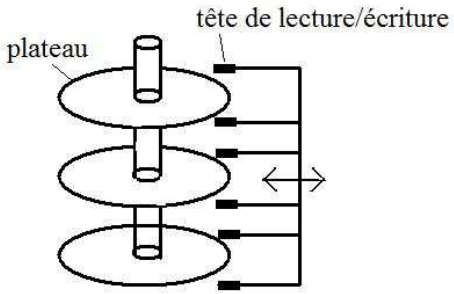
5-2) Les mémoires secondaires (ou mémoires de masse)

Malgré ses capacités de plus en plus grandes, la mémoire principale demeure insuffisante. Avec l'évolution de l'informatique, il y a toujours bien plus d'informations à stocker que la mémoire principale ne peut en contenir. Il suffit de penser aux films enregistrés, aux encyclopédies, à tout ce qui combine les données, le son et les images. D'où la mise en place de mémoires secondaires. Il s'agit principalement des disques magnétiques et des disques optiques, accessoirement des bandes magnétiques. Mais une nouvelle tendance se dessine. Il s'agit du *cloud computing* (informatique en nuage), où le stockage se fait sur des serveurs à distance et à grande échelle, via Internet. Cela peut

poser quelques problèmes de sécurité. Certains voient dans le *cloud computing* un phénomène de mode, voire un piège, les utilisateurs risquant de perdre le contrôle de leurs données.

5-2-1) Les disques durs

Un disque dur est un disque magnétique formé d'un ou plusieurs plateaux métalliques circulaires. Ces plateaux rigides, non magnétiques dans leur masse, et dont le diamètre est de quelques centimètres, sont recouverts d'un enduit magnétisable. Ils tournent à vitesse constante, de l'ordre de 100 ou 200 tours par seconde. La tête magnétique est une bobine électromagnétique fixée à un bras mobile. Elle flotte sur un coussin d'air, à quelques microns au-dessus des plateaux. L'écriture sur le disque se fait par la magnétisation locale de la surface du disque au-dessous de la tête, suivant un phénomène d'induction. Inversement, lorsque la tête passe au-dessus d'une zone magnétisée, cela induit un courant dans le bobinage de la tête. Il est ainsi possible de lire et d'écrire des séquences de bits sur le disque.



Disque dur formé de trois plateaux double-face, avec ses six bras et têtes de lecture

Le disque, ou ses plateaux, sont partagés en pistes en forme de couronnes circulaires, de 5 à 10 microns de large, à raison d'un millier de pistes par centimètre. Celles-ci sont divisées en secteurs de 512 octets. Au début d'un secteur se trouve une zone de préambule qui permet à la tête de se synchroniser avant la lecture ou l'écriture. A la fin se trouve une autre zone permettant d'effectuer un contrôle d'erreur sur les données enregistrées, cette technique (*Error Correcting Code*) étant basée sur le code de Hamming notamment. Deux secteurs voisins sont séparés par un espace vide. Une telle organisation du disque est appelée formatage. Sur les disques à haute densité, on atteint aujourd'hui des densités de 100 000 bits par centimètre.

Comme on l'a vu, le disque est associé à un contrôleur de disque, qui pilote son fonctionnement. Il s'agit d'un circuit intégré, d'une sorte de processeur, qui reçoit les ordres de l'unité centrale, comme *READ, WRITE, FORMAT*. Il est chargé d'exécuter ces ordres qui se traduisent par un certain nombre de tâches à accomplir, comme celle du positionnement du bras, ou encore la conversion des groupes d'octets reçus en parallèle à partir du bus en un flot de bits à écrire en série sur le disque, ou inversement.

5-2-2) Les disques optiques

Avec l'explosion de la micro-informatique, la diffusion massive des logiciels sur le marché a nécessité l'utilisation de petits disques amovibles, pouvant être mis ou enlevés de l'ordinateur. A l'origine, dans les années 1980, il y eut les *floppy disk*s, constitués d'un plateau circulaire en plastique souple enveloppé dans une jaquette plastique plus rigide. Ils ont d'abord été au format de 5,25 pouces, puis sont passées à 3,25 pouces.[11] On est ensuite passé aux CD-ROM, puis aux DVD. Dans ces deux cas, il s'agit de disques optiques, offrant une capacité de stockage très supérieure à

[11] Il s'agit aussi de disques magnétiques, mais à la différence des disques durs, la tête des lecteurs de disquette est en contact avec la disquette. Pour réduire les phénomènes d'usure et les déchirements du plateau, la rotation de plateau s'arrête lorsqu'il n'y a aucune action de lecture ou d'écriture à effectuer, et le bras de lecture se rétracte, ce qui provoque une certaine lenteur de réaction.

celle des disques magnétiques. C'est la principale raison de leur succès commercial. On les rencontre aujourd'hui sur la quasi-totalité des ordinateurs. Ils contiennent des programmes, des jeux, des encyclopédies, des films. Tous les logiciels commerciaux les utilisent.

Le DVD (*digital versatile disk* ou *digital video disk*) est officiellement lancé en 1995, avec des spécifications qui ont été faites par des sociétés d'électronique grand public, notamment japonaises, en collaboration avec les studios du cinéma américain. Les producteurs de films voulaient en effet remplacer leurs bandes vidéo analogiques par des produits numériques à haute performance, moins coûteux, moins encombrants, et d'une plus grande durée de vie. Par contre les industriels des télécommunications et de l'informatique n'ont pas été conviés à ces travaux, ce qui explique que les DVD soient surtout liés au marché de la vidéo.

Comme les CD, les DCD sont des disques de 12 cm de diamètre, possédant une piste en spirale comme c'était le cas pour les antiques microsillons, à la différence des pistes circulaires d'un disque magnétique. Cette piste unique continue est creusée d'alvéoles minuscules de l'ordre d'un demi-micron. Le passage d'une micro-cuvette à une zone plane signifie 1, et celui d'une zone plane à une micro-cuvette donne un 0. La lecture se fait par un rayon laser de faible puissance. Un photo-détecteur mesure l'énergie du rayon après réflexion sur le disque. Pour que la lecture de la piste se fasse à vitesse constante, il faut que la vitesse angulaire de rotation du disque augmente lorsque l'on se rapproche du centre du disque. C'est là encore une différence avec les disques magnétiques. Une autre différence est la faible vitesse de rotation, qui varie entre 200 et 500 tours/minute. On est loin des 6000 tours/ minute d'un disque magnétique.

La capacité de stockage d'un DVD est sept fois plus grande que celle d'un CD-ROM, et atteint 4,7 Go. En utilisant une compression suivant la norme MPEG-2, un DVD peut contenir un film complet en haute résolution. On peut aussi bien y mettre six films compressés en DIVX. Il existe même des DVD double face et double couche ayant des capacités de stockage bien plus grandes.[12]

6) Les modems et le raccordement au réseau téléphonique

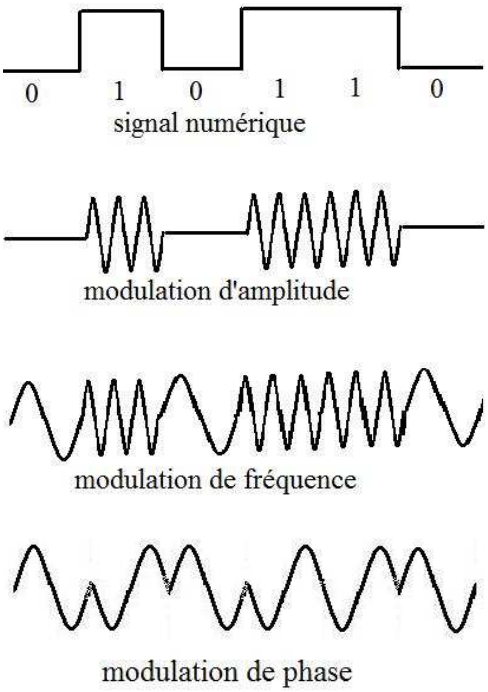
Avec l'essor des réseaux, il est devenu courant qu'un ordinateur entre en relation avec un ordinateur distant. C'est le cas lorsque l'on se connecte au réseau Internet avec un micro-ordinateur. Dans ce cas, le réseau téléphonique sert le plus souvent de moyen de communication.

Dans le cas du réseau téléphonique, conçu pour transmettre la voix humaine, comment passer à la transmission de signaux codés avec des 0 et des 1 créés par deux niveaux de tension électrique? Si l'on transmettait directement ces signaux à deux niveaux tels quels sur une ligne téléphonique, ils subiraient de grosses distorsions et modifications qui entraîneraient des erreurs de transmission irrémédiables. Par contre un signal sinusoïdal d'une fréquence de l'ordre de 2000 Hertz passe dans de bonnes conditions dans une ligne téléphonique. C'est ce genre de signal qui est utilisé dans la plupart des communications de type analogique. On prend ce type de signal comme base, il constitue le courant porteur ou porteuse. Puis on le soumet à des modulations pour faire ressortir les signaux binaires avec leurs 0 et les 1. Il suffit pour cela de faire varier l'un des trois paramètres d'un signal sinusoïdal : l'amplitude du signal, sa fréquence ou sa phase, en suivant le rythme d'apparition des bits 0 ou 1.

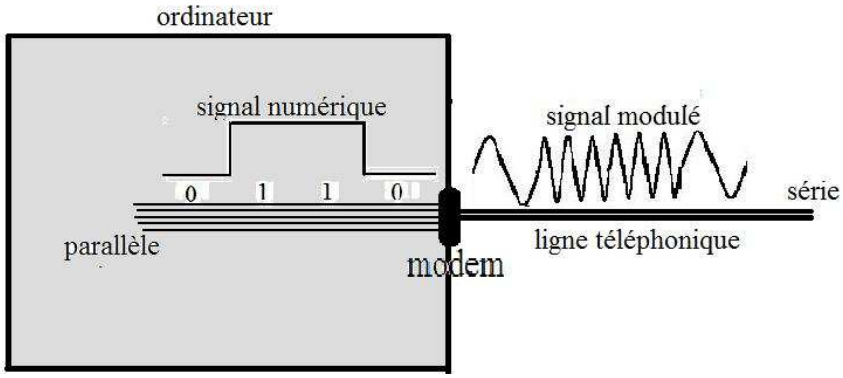
En modulation d'amplitude, on crée deux niveaux d'amplitude, associés à 0 et à 1. Dans le cas d'un signal audible, cela se traduit par deux niveaux sonores : un faible et un fort. En modulation de fréquence, l'amplitude du signal reste constante, et l'on fait varier la fréquence suivant deux niveaux de période. Dans le cas d'un signal audible, cela donne deux tonalités différentes, l'une

[12] Un DVD double couche est constitué d'une couche supérieure translucide semi-réfléchissante et d'une couche inférieure réfléchissante. Pour lire l'une ou l'autre de ces deux couches, le rayon laser venant du dessus doit modifier son intensité : avec une intensité faible, le rayon est réfléchi sur la couche supérieure, avec une intensité plus élevée le rayon passe à travers la première couche et se réfléchit sur la couche inférieure, dont la spirale est inversée par rapport à celle du haut.

dans le grave et l'autre dans l'aigu. C'est cette modulation de fréquence que l'on préfère utiliser. Il reste la modulation de phase, où l'amplitude et la fréquence du signal ne bougent pas. Seule la phase (un certain décalage du signal dans le temps) est modifiée lorsque l'on passe de 0 à 1 ou de 1 à 0.



La rapidité de modulation est exprimée dans une unité de mesure appelée baud, correspondant au nombre de modulations effectuées chaque seconde. Par exemple, un modem de 9600 bauds indique que l'intervalle de temps de modulation est de 104 microsecondes (il suffit de diviser 1 000 000 par 9600). Si le rythme se fait bit après bit, le baud est aussi le nombre de bits transmis par seconde. Mais lorsque l'on transmet deux bits ou plus par état de modulation, ce qui peut arriver avec la modulation de phase dite quaternaire,[13] le débit en bits peut être supérieur à la rapidité de modulation en bauds.



Prenons le cas de caractères codés sur 8 bits. Dans l'ordinateur, chaque groupe de 8 bits circule en parallèle sur les lignes du bus. Mais comme la ligne téléphonique ne comporte qu'une paire de fils comme canal de transmission, il est nécessaire de transmettre les caractères codés sur 8 bits en série, bit après bit. Le caractère analogique du réseau impose de moduler les signaux numériques

[13] Au lieu de faire un seul changement de phase pour distinguer le 0 du 1, par exemple l'opposition de phase (une demi-période) comme dans le dessin ci-dessus, on peut en faire quatre différents, ce qui permet de traiter des groupes de deux bits au lieu d'un. D'où la modulation de phase dite quaternaire.

avec un équipement spécialisé. Cet appareil qui effectue la modulation est le modem. Il a le double rôle d'émettre et de recevoir des signaux analogiques modulés sur la ligne téléphonique. D'un côté de la ligne, le modem réalise une modulation du signal au rythme des bits de données que lui fournit l'ordinateur, et à l'autre extrémité de la ligne, c'est aussi un modem qui reconstitue les bits de données à partir du signal modulé. Dès que le modem reçoit un début de signal, son horloge interne se synchronise avec celle du modem émetteur.

Aujourd'hui, il existe des techniques de modulation très performantes. La quasi-totalité des modems actuels fonctionnent en *duplex* ou *full duplex*. Cela signifie qu'ils peuvent transmettre les données dans un sens et dans l'autre simultanément, dans le sens ordinateur-réseau en même temps que dans le sens réseau-ordinateur. Il existe aussi des systèmes où la communication se fait dans les deux sens, mais pas simultanément. Elle se fait alternativement dans un sens puis dans l'autre. Cela s'appelle le *half duplex*. Il en est ainsi notamment pour les radios CB ou les talkies-walkies. Enfin, lorsque les systèmes communiquent dans un seul sens, comme la radio ou la télévision, on appelle *simplex* un tel mécanisme de communication.

Les modems ne sont qu'un aspect de la technologie des réseaux. Ils en sont l'entrée en matière. Comme on pourra le constater, les réseaux reprennent, mais à grande échelle, les principales fonctionnalités mises en œuvre à l'intérieur d'un ordinateur.