Le processeur

Un processeur (aussi appelé microprocesseur ou CPU pour Central Processing Unit) est le coeur de l'ordinateur. Ce composant a été inventé par Intel (avec le modèle 4004) en 1971. Il est chargé de traiter les informations et d'exécuter les instructions. Il ne sait communiquer qu'avec le reste de l'ordinateur via le langage binaire. Le processeur est rythmé par une horloge (quartz) cadencée plus ou moins rapidement (on parle alors de fréquence). A chaque impulsion d'horloge (signal électrique passant du niveau bas au niveau haut en cas de front montant), le processeur lit l'instruction stockée généralement dans un registre d'instruction (un registre est une petite mémoire très rapide située dans le processeur en lui-même) et exécute l'instruction. Dans une même gamme (et donc à architecture comparable) un processeur cadencé plus rapidement est plus efficace car il peut traiter les instructions plus rapidement.

Attention cependant, avec les nouvelles architectures, il y a une baisse de fréquence mais une amélioration substantielle du rendement, ce qui fait qu'un processeur Intel Core 2 duo cadencé à 2.13 GHz est généralement plus performant qu'un Intel Pentium EE cadencé à 3.73 GHz tout en dissipant moins ! Nous reparlerons plus loin des différentes gammes de processeurs, de leurs performances générales et de leur dissipation thermique.

Format d'une instruction :

Pour qu'un processeur puisse exécuter une instruction, encore faut-il qu'il sache de quelle instruction il s'agit et quelles sont les données sur lesquelles agir. C'est pourquoi une instruction sera stockée selon une méthode bien précise. On divise ainsi une instructions en deux codes :

- Le code opération, qui représente le type d'instruction (si il faut déplacer des données d'un registre à l'autre, faire une addition...)
- Le code opérande, qui représente les paramètres de l'instruction (adresse mémoire, constantes utilisées, registres...)

Types principaux d'instructions :

Il existe différents types d'instructions. Les plus courants sont ceux-ci :

- Instructions d'opérations arithmétiques (addition, soustraction, division, multiplication)
- Instructions d'opérations logiques (OU, ET, OU EXCLUSIF, NON, etc...)
- Instructions de transferts (entre différents registres, entre la mémoire et un registre, etc...
- Instructions ayant rapport aux entrées et sorties.
- Instructions diverses ne rentrant pas dans les autres catégories (principalement des opérations sur les bits).

Etapes d'exécution :

Lorsqu'un processeur a besoin d'exécuter des instructions, il le fait toujours dans l'ordre suivant :

- Recherche de l'instruction (fetch)
- Lecture de l'instruction
- Décodage de l'instruction
- Exécution de l'instruction

Les registres :

Un registre est une petite mémoire de taille raisonnable (variant généralement de 32 à 128 bit). Les registres sont utilisés tout le temps, ils sont donc très importants. Nous avons vu plus haut qu'une instruction pouvait faire appel aux registres. L'avantage est que ce type d'opérations est beaucoup plus rapide que de faire appel à la mémoire vive, les registres étant internes au processeur, contrairement à la mémoire vive. Il existe différents types de registres, voici les principaux :

- Le registre d'instruction (RI) qui permet de stocker l'instruction qui va être exécutée.
- Le registre d'état qui permet de stocker des indicateurs sur l'état du système après l'exécution d'une instruction. Voici quelques indicateurs (qui peuvent changer d'appellation, le principe restant le même :
  - C (pour carry) : vaudra 1 si une retenue est présente.
  - V (pour oVerflow) : vaudra 1 en cas de dépassement de capacité (addition de deux chiffres positifs donnant un résultat négatif par exemple).
  - N (pour Negative) : vaudra 1 si le résultat est négatif.
- Le registre PC (Program counter) qui stocke l'adresse de la prochaine instruction à exécuter.

La mémoire cache :

Chaque processeur intègre une quantité variable de mémoire cache. Cette mémoire très rapide est indispensable pour bénéficier de bonnes performances dans les applications. Elle permet de stocker les données les plus fréquemment demandées par le processeur.

On distingue trois niveaux de cache :

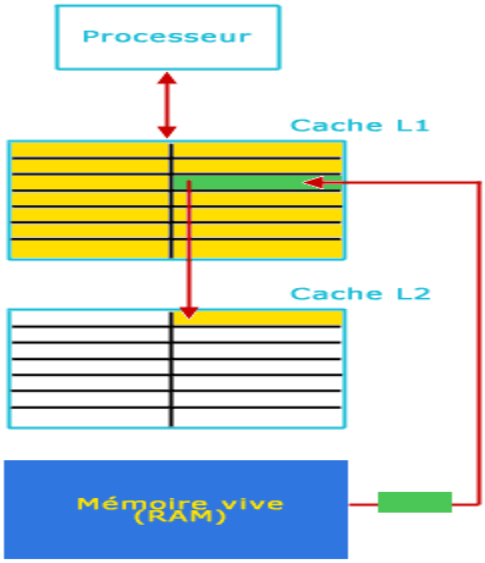
- Le cache L1 (cache de premier niveau) : La quantité intégrée est généralement faible (de 8 à 64 Ko généralement)
- Le cache L2 (cache de second niveau) : Cette quantité varie de 128 Ko à 1 Mo, ce cache est légèrement moins rapide que le cache L1
- Le cache L3 (cache de troisième niveau) : Ce cache est disponible seulement sur certains processeurs et peut vous permettre de gagner 10% de performances en fonction des applications, en réalité les gains sont très moyens voire inexistants.

Lorsqu'un processeur a besoin de lire des données, il va d'abord regarder si celles-ci se trouvent dans la mémoire cache. Si elles s'y trouvent, on parle de succès du cache (cache hit), dans le cas contraire d'échec du cache (cache miss) les données étant placées ensuite en cache à partir de la mémoire vive. Il y a perte de temps en cas de cache miss car le processeur a regardé dans le cache pour rien. Le taux de réussite s'appelle le hit rate, le taux d'échec miss rate. Afin d'augmenter les performances et donc de diminuer le miss rate, il existe différentes techniques ayant chacune leurs avantages et inconvénients. Outre différents algorithmes pouvant être implantés dans le programme, on parlera ici d'une technique matérielle, celle consistant à adopter une mémoire cache de type inclusive ou exclusive. La différence

principale entre ces deux types de gestion du cache se trouve principalement dans leur manière de stocker les données. Lorsque le cache L1 est plein, il faut libérer de la place pour pouvoir placer dans le L1 la donnée que l'on vient de lire en mémoire vive.

Avec le cache Exclusif :

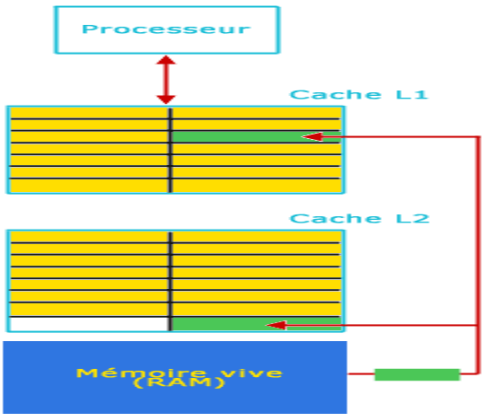
Il faut faire de la place dans le cache L1, on déplace donc une donnée (celle la moins récemment utilisée) et on la place dans le cache L2, puis ensuite on place la donnée provenant de la mémoire vive dans le cache L1.



Lorsqu'une donnée est présente dans le cache L2 mais pas dans le L1, on la fait remonter dans le L1 en permutant la donnée la moins récente du L1 avec la donnée du L2. Les caches L1 et L2 ne contiennent jamais les mêmes données (elles passent de l'un à l'autre des caches), on parle donc de cache exclusif. L'avantage est de pouvoir avoir une liberté sur la taille des caches (la taille efficace étant l'addition des tailles des caches des différents niveaux). Le désavantage de cette technique est que les performances du cache L2 sont réduites étant donné qu'il faut écrire une donnée dans le cache L1 à chaque fois qu'on récupère une donnée dans le L2.

Avec le cache Inclusif :

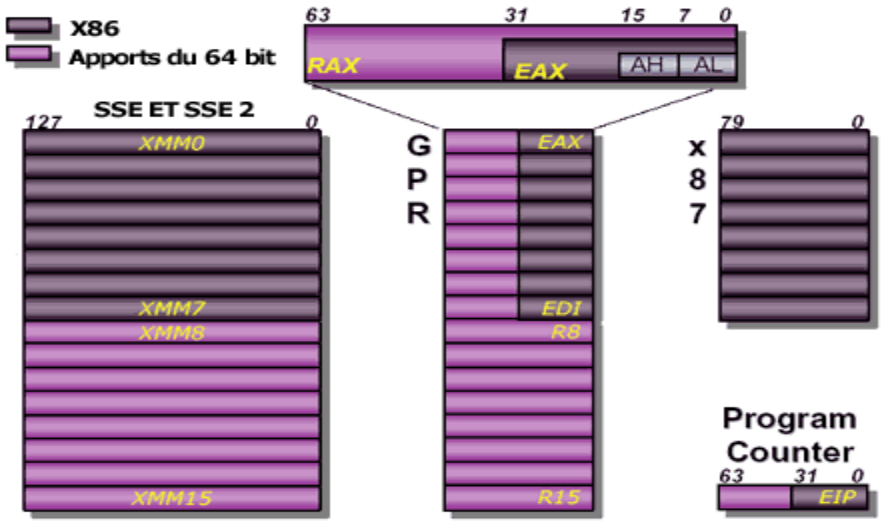
Lorsque la donnée ne se trouve ni dans le L1 ni dans le L2, on copie la donnée de la mémoire vive dans les deux niveaux de cache (L1 et L2). La ligne écrasée du cache L1 n'est pas déplacée dans le L2 car elle y est déjà.



A ce stade, le cache L2 contient des données supplémentaires par rapport au cache L1, cependant toutes les données qui sont présentes dans le cache L1 sont présentes dans le cache L2. On parle d'un cache inclusif. L'avantage majeur de ce type de cache est de ne pas avoir à réécrire la donnée dans le cache L1 en cas de cache hit en lecture dans le cache L2. Les performances du cache L2 sont donc supérieures. L'inconvénient majeur est la taille totale du cache efficace ainsi que la contrainte de taille des caches L1 et L2. Pour être efficace, cette technique doit être mise en oeuvre avec un cache L2 très grand devant la taille du cache L1.

Le 64 bit :

Avec les nouveaux processeurs 64 bits (X86-64), la taille de différents registres est passée de 32 à 64 bit, avec plusieurs avantages à la clé : un adressage maximal de la mémoire qui n'est plus limité à 4 Go comme c'était le cas en 32 bit, et une rapidité généralement accrue des applications en tirant partie (car plus de registres disponibles signifie qu'on aura généralement plus de place pour stocker des données au lieu d'utiliser la RAM avec un code optimisé). Les processeurs disposant du 64 bit portent en général la mention EMT64 (Enhanced Memory 64 Technology) chez intel et "64" chez AMD (bien qu'il y ait des exceptions chez AMD).



Tous les processeurs sont composés de ces éléments :

- L'UAL (unité arithmétique et logique, aussi appelée ALU) : c'est l'unité de calcul qui gère ce qui porte sur des nombres entiers.

- La FPU (Floating Point Unit) est l'unité de traitement des nombres à virgules (aussi appelés nombres flottants).

- Le décaleur : il est le spécialiste des divisions et multiplications par deux. Son rôle est de décaler les bits vers la gauche ou vers la droite.

- Les registres

- Le circuit de données : son rôle est d'acheminer les données provenant de l'UAL vers les registres.

- La MMI (Mémoire de micro instructions) : cette zone du processeur contient toutes les instructions nécessaires à celui-ci pour comprendre les instructions du langage machine.

- Le SEQ (séquenceur) : cet organe traduit les instructions compliquées en instructions plus simples pour permettre au processeur de les traiter.

- L'unité de gestion des instructions : elle recueille les instructions demandées, les décode puis les envoie à l'unité d'éxécution.

- L'unité d'exécution : son rôle est d'exécuter les tâches que lui a envoyé l'unité d'instruction.

- L'unité de gestion des bus : elle permet de gérer les informations entrantes et sortantes.

On différencie les microprocesseurs de part :

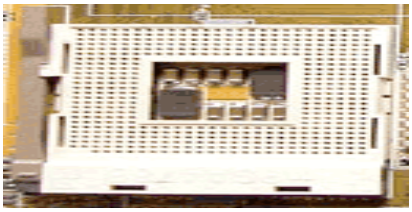
- Leur fréquence (vitesse de traitement maximale si vous préférez). Cette fréquence s'exprime en MHz (Méga-hertz) ou GHz (Giga-hertz). La fréquance s'obtient en multipliant la fréquence du FSB (Front Side Bus aussi appelé Bus système) par un coefficient multiplicateur.

- La fréquence de leur FSB : plus cette fréquence est élevée, meilleures sont les performances (à familles de processeurs égales).

- Leur architecture interne : Nombre d'ALU, de FPU pour un même processeur, contrôleur mémoire interne ou non, tout ça peut changer radicalement les performances et il est donc ridicule de comparer les performances d'un Athlon 64 avec celles d'un Pentium 4 à fréquence égale car les deux architectures n'ont pas été conçues pour les mêmes fréquences (même si les objectifs en matière de fréquence n'ont pas pu être atteints par Intel qui tablait sur 10 GHz, alors que le pentium 4 plafonne à 3.8 GHz).

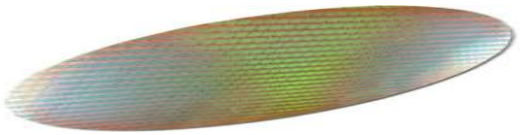
- Leur quantité de mémoire cache (répartie sur 1, 2 ou 3 niveaux).

- Leur mode de connexion à la carte-mère (appelé socket, dont le nombre de trous et l'appellation varient. En général on appelle un socket par son nombre de trous, on parle alors de "socket 478", "socket 775", etc...) :



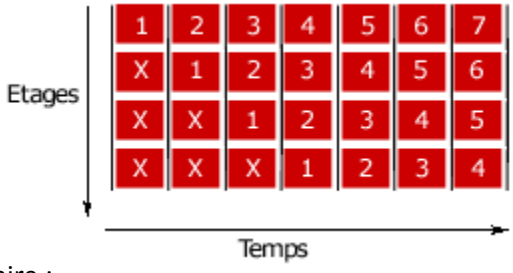
Fabrication des processeurs :

Les processeurs sont tous gravés sur des plaques appelées Wafers. Les différentes séries de processeurs n'ont pas forcément la même finesse de gravure (mesurée en micromètres (µm) ou nanomètres (nm)). Les processeurs actuels sont gravés en 0.09µ et 0.065µ (soit respectivement 90 et 65 nanomètres). Diminuer la finesse de gravure permet de produire plus de processeurs à la fois sur un Wafer et permet donc d'abaisser leur coût de fabrication. Cette technique permet également de diminuer la consommation du processeur et donc la quantité de chaleur produite ce qui permet d'abaisser la consommation d'énergie et de monter plus haut en fréquence. Une finesse de gravure accrue permet également de loger plus de transistors dans le core (aussi appelé die) du processeur, et donc d'ajouter des fonctionnalités supplémentaires tout en gardant une surface aussi compacte que les générations précédentes.



La technique du pipeline :

Afin d'optimiser le rendement, la technique du pipeline est apparue sur les 386 d'Intel. Le pipeline permet de commencer à traiter l'instruction suivante avant d'avoir terminé la précédente via un mécanisme de "travail à la chaîne". L'inconvénient de cette technique est que plus le pipeline est profond (contient d'étapes) plus la perte de performances est importante si une erreur de prédiction survient. L'avantage de cette technique est qu'elle permet d'augmenter la fréquence du processeur plus facilement. L'inconvénient majeur de cette technique est qu'elle entraîne une baisse des performances à fréquence égale. Il y a également une augmentation du dégagement thermique et donc de la température du processeur. Pourquoi cette hausse de température ? Lorsque l'instruction doit être exécutée en un temps donné, quelle que soit la profondeur du pipeline, cette instruction sera toujours exécutée aussi rapidement. Plus il y a d'étages au pipeline, plus l'instruction doit être "découpée" en une quantité de "micro-instructions" qui seront exécutées en un temps très court, bien plus court que le temps nécessaire pour traiter l'instruction. Or, plus le pipeline comporte d'étages et plus le délai de traitement d'une "micro-instruction" doit être faible, ce qui nécessite généralement plus de transistors, ces transistors chauffent, ont besoin généralement de plus de courant pour fonctionner plus rapidement. Voilà une des causes du dégagement thermique supérieur des processeurs comportant un nombre important d'étages de pipeline.

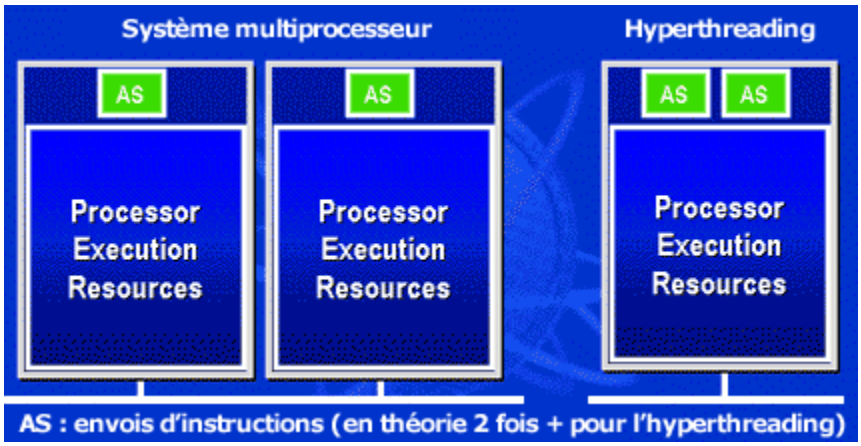


L'architecture super scalaire :

Cette astuce consiste à doubler le nombre d'unités de traitement pour traiter plusieurs instructions par cycle.

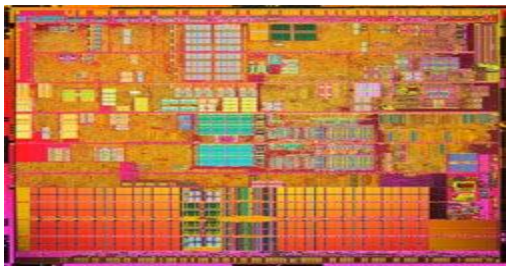
L'hyperthreading :

sous ce terme un peu barbare se cache une optimisation d'Intel pour ses processeurs Pentiums 4 (à partir du core B avec le processeur à 3.06 GHz). L'hyperthreading consiste à émuler au sein d'un seul processeur physique deux processeurs logiques, ce qui permet de gaver le processeur de plus d'instructions et améliorer son rendement :

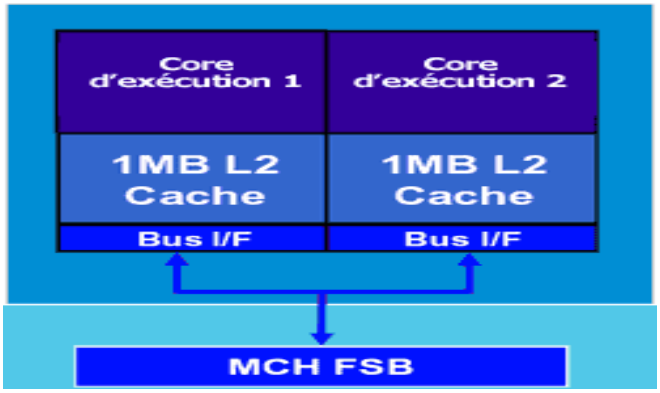


Le dual core (et plus généralement le multicore) :

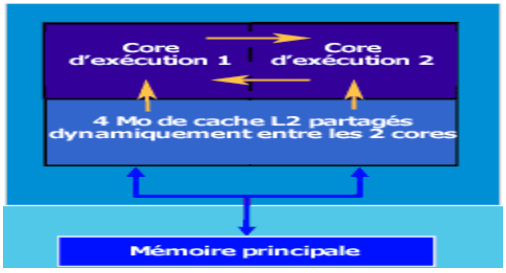
Le core (ou die) d'un processeur est toute sa partie centrale qui ressemble à ça (agrandie des milliers de fois) :



Technologie récente et à la mode, le dual core consiste à mettre "deux processeurs en un" même si en pratique ce n'est pas tout à fait ça. Suite à l'impossibilité ou presque de monter en fréquence pour les deux fondeurs principaux (Intel et AMD), l'optimisation des performances passe désormais par l'ajout de cores et par l'optimisation du rendement. Avec le dual core on peut en théorie doubler les performances sans changer la fréquence. Cette architecture est très efficace en multitâche, en monotâche elle ne vous fera rien gagner à moins d'augmenter les performances des cores (ajout de cache, hausse de fréquence...). Ainsi, les fréquences plus basses et l'efficacité par cycle sont redevenues à la mode (avec l'Intel pentium M et le core 2 duo notamment). Il existe plusieurs architectures dual core. Nous allons en regarder deux et montrer pourquoi l'une est plus performante que l'autre. Commençons avec l'architecture des premiers processeurs intel dual Core, à savoir les processeurs Pentium D :



On constate tout d'abord que les deux core ont leur propre mémoire cache. Il y a 2 Mo de mémoire cache L2 au total mais chaque core ne peut utiliser qu'1 Mo de mémoire. Lorsqu'une application a besoin de beaucoup de mémoire cache, cette solution n'est pas optimale. Ensuite, si les deux cores veulent dialoguer entre eux (et c'est là où se situe l'énorme goulôt d'étranglement), les deux cores n'ont pas un bus les reliant en interne, ils doivent passer par le bus principal (FSB) pour dialoguer. Les performances en multitâche sont donc bridées, même si Intel a augmenté la fréquence du FSB pour limiter les pertes de puissance. Intel a cependant innové avec l'architecture "Core" (celle qu'utilise le Core 2 duo) :



Les modifications sont conséquentes : outre le dialogue possible entre les deux cores sans passer par le FSB, intel utilise désormais un cache L2 partagé entre les deux cores (cette technologie est appelée SMART Cache). Si un core a besoin de 4 Mo de cache L2 et que l'autre core n'en a pas besoin, tout le cache sera utilisable par un seul des deux cores. En pratique, cette architecture offre d'excellentes performances.

Les instructions spécialisées :

Avec les Pentiums MMX (MultiMedia Extensions) sont apparues les instructions spécialisées dans le traitement de tâches comme la 3D ou le traitement audio et vidéo. Les processeurs Intel les plus récents sont dotés des instructions SSE 3 et SSE 4 qui permettent d'accélérer l'encodage vidéo notamment. Les processeurs AMD les plus récents disposent pour leur part des instructions SSE 2, SSE 3 et du 3D now! qui leur permettent d'avoir de bonnes performances dans les jeux 3D et le multimédia. Si un programme est spécialement optimisé pour une instruction donnée, un processeur sans instructions devra être très puissant pour compenser l'absence de ces instructions et aller aussi vite que son concurrent dans le même programme.

Voici quelques domaines d'utilisation d'un PC pour juger du type de processeur qu'il vous faudra :

- PC de bureautique légère : un processeur monocore suffit amplement quelle que soit sa fréquence de fonctionnement.

- PC de bureautique avancée : un processeur monocore suffit également parfaitement à la tâche.

- PC destiné à faire du multitâches intensivement ou/et du rendu 3D : Un processeur multicores est plus que recommandé, même si vous prenez de l'entrée de gamme.

- PC destiné aux jeux intensifs et à toutes sortes d'applications gourmandes : Un processeur multicore hautement cadencé en fréquence devrait vous ravir.

Il n'est plus vraiment nécessaire aujourd'hui de raisonner en terme de fréquence car les derniers processeurs sont suffisamment performants. Chez intel, les processeurs à partir du core 2 duo 6300 sont largement capables de faire tourner les applications les plus exigeantes. Chez AMD les processeurs athlons 64 X2 4600+ et supérieurs en seront également capables. Les entrées de gamme de ces deux fondeurs sauront combler les amateurs de performances élevées tout en préservant le budget, ils seront en outre parfaitement adaptés pour une utilisation bureautique, multimédia, ou ludique.

Technologies :

- Le 64 bit n'est pas encore utile, bien que tous les processeurs l'intègrent désormais en standard.

- Les instructions SSE 2, SSE 3 et SSE 4 seront très utiles si vous faites beaucoup d'encodage vidéo