Cours : *Technologie des processeurs modernes*
Public : *IUT Informatique*
Auteur : *P. Pinault*

Technologies des processeurs modernes

« *J'ai d'abord noté, en 1965, le doublement de la densité des transistors dans les composants produits chaque année et ce 4 ans après que le premier circuit intégré soit créé. La presse appela cela la Loi de Moore. Le nom est resté. Pour être honnête, je ne pensais pas que cette loi reste valable 30 ans après, mais, je suis maintenant confiant qu'elle le restera pour une nouvelle vingtaine d'années. En 2012, Intel devrait être en mesure d'intégrer 1 Milliard de transistors opérant à 10GHz. Il pourra en résulter une performance de 100,000 MIP. Soit là une augmentation de puissance, par rapport à un Pentium II, comparable à celle qu'il y a entre un 386 et un Pentium II. Nous ne voyons pas de barrières fondamentales d'ici 2012 et ce n'est que vers 2017 que nous atteindrons les limitations physiques liées aux technologies de fabrication des Wafers.* » Dr. Gordon E. Moore (fin des années 1990)

L'évolution de la puissance des ordinateurs est poussée par un besoin réel de performances accrues pour des applications de plus en plus complètes et donc gourmandes en ressources. Les évolutions réalisées ne passent pas seulement par l'augmentation de fréquence des processeurs, elles font appel à l'utilisation de technologies fonctionnelles, c'est à dire à la création d'architectures matérielles optimisées. A titre d'analogie avec le monde du logiciel, il s'agit de choisir un algorithme plus performant plutôt que d'exécuter le programme sur un ordinateur plus puissant.

Dans ce chapitre, nous allons aborder une partie des solutions architecturales utilisées dans les processeurs actuels.

1. Augmentation de fréquence des processeurs

Le micro-ordinateur familial fonctionnait entre 75 et 90 Mhz en 1995, il utilise maintenant une horloge à plusieurs Giga Hertz. Cette augmentation est considérable, ce n'est toutefois pas pour autant que les programmes exécutés sur les ordinateurs actuels fonctionnent 10 à 20 fois plus rapidement.

L'augmentation de la fréquence d'horloge des circuits électroniques est principalement liée à la réduction d'échelle opérée lors de la gravure des composants et l'utilisation de substrats autres que le Silicium et plus performants. Les avancées technologiques dans ces domaines ont permis de réduire considérablement la taille des transistors, éléments de base des circuits logiques. Un transistor plus petit permet une commutation plus rapide et donc une utilisation à fréquence plus élevée. Il offre en outre une consommation plus faible et donc une dissipation thermique moindre permettant à la fois la réalisation de circuits plus complexes et une utilisation à plus haute fréquence sans surchauffe. Dans les processeurs actuels, la réduction d'échelle ne permet pas une diminution de consommation suffisante, c'est pourquoi le coeur des processeurs utilise des tensions d'alimentation de seulement 1 à 2 Volts. La tension d'alimentation étant le facteur dominant de la consommation énergétique.

Cours : *Technologie des processeurs modernes*
Public : *IUT Informatique*
Auteur : *P. Pinault*

2. Les freins à la performance

La fréquence d'utilisation d'un processeur n'est pas le seul critère à prendre en compte lorsque l'on parle de ses performances. En effet, la fréquence caractérise seulement le nombre de cycles qu'un processeur peut exécuter par seconde. Or, ce cycle peut correspondre seulement à une partie d'une instruction exécutée, il peut aussi correspondre à une ou plusieurs instructions.

Selon l'architecture du processeur, les instructions peuvent être micro-codées, dans ce cas, une instruction est le résultat de l'enchaînement de plusieurs sous-instructions, chacune prenant un cycle pour s'exécuter. Dans une architecture de type RISC, chaque instruction ne demande qu'un seul cycle. Toutefois, les instructions utilisées sont plus simples. Dans une architecture dite super-scalaire, plusieurs instructions sont exécutées à chaque cycle. De ce fait, pour caractériser la puissance d'un processeur, plutôt que de parler de Méga Hertz, on utilise une unité : le MIPS. Celle-ci indique le nombre de millions instructions exécutées par seconde.

Le nombre de MIPS donné par le fabriquant n'étant généralement obtenu que de façon théorique, alors que les meilleures conditions sont réunies, il existe d'autres tests reconnus. Les tests SPEC (Standard Performance Evaluation Corporation) permettent alors une comparaison réelle de la puissance des micro-ordinateurs. Il s'agit de tests basés sur l'exécution de programmes et remis à jour régulièrement.

L'architecture du système informatique dans son ensemble a un impact très important sur sa performance. En effet si d'un point de vue théorique le processeur peut exécuter des milliards d'instructions par seconde, celles-ci sont extraites de la mémoire de laquelle on ne peut en lire qu'une centaine de millions par seconde. Un facteur dix, qui bien que compensé par les mémoires caches, est très pénalisant.

Les co-processeurs intégrés dans le système sont eux aussi des éléments augmentant ou diminuant ses performances. En effet, un périphérique spécialisé peut venir en aide au processeur pour une tâche spécifique. Celle-ci est alors réalisée bien plus rapidement car câblée dans le co-processeur et exécutée en parallèle d'autres traitements. Les cartes accélératrices 3D ou les DMA sont des exemples de ce qu'il est possible de faire. Grâce à de telles co-processeurs, les performances obtenues sont bien plus grandes que le simple gain lié à l'augmentation de fréquence.

Enfin, le jeu d'instruction d'un processeur, et par là leur architecture interne, peut aussi permettre un gain de performance. Les DSP par exemple forment une famille de processeurs spécialisés dans le traitement du signal. Ils sont capables en un seul cycle d'exécuter des traitements en nécessitant plusieurs sur d'autres plate-formes. Par exemple, leur jeu d'instructions permet de réaliser des opérations de multiplication/accumulation (MAC) (de types v+=a*x) en un seul cycle. Les calculs polynomiaux de degré n sont alors effectués en n cycles. Les micro-contrôleurs télécoms intègrent généralement des unités spécialisées dans la réception de trames qui simplifient et optimisent grandement ce type de traitements.

3. Les pipelines

L'utilisation de pipeline permet d'augmenter la performance des processeurs en exécutant de façon parallèle plusieurs sous-parties d'instructions. En effet, une instruction est souvent réalisée en plusieurs étapes : sa lecture, le chargement de ses opérandes, la réalisation du calcul et enfin le stockage des résultats. Dans une architecture sans pipeline, l'instruction suivante ne débute que lorsque l'instruction précédente a terminée toutes ces étapes. L'architecture pipeline permet de commencer le traitement d'une instruction dès lors que la première partie de la précédente est terminée :

Cours : *Technologie des processeurs modernes*
Public : *IUT Informatique*
Auteur : *P. Pinault*

Programme : Mov A,#1
Mov B,#2

temps

Lecture	Chargement opérandes	traitement	Sauvegarde
Mov A,#1			
	Mov A,#1		
		Mov A,#1	
			Mov A,#1
Mov B,#2			
	Mov B,#2		
		Mov B,#2	
			Mov B,#2

Exécution sans pipeline sur 8 cycles

temps

Lecture	Chargement opérandes	traitement	Sauvegarde
Mov A,#1			
Mov B,#2	Mov A,#1		
	Mov B,#2	Mov A,#1	
		Mov B,#2	Mov A,#1
			Mov B,#2

Exécution avec pipeline sur 5 cycles

La mise en oeuvre d'une architecture avec pipeline demande l'ajout d'électronique dédiée aux contrôles et à la mémorisation du résultat de l'étape précédente. Le pipeline, une fois plein permet l'exécution des instructions, en moyenne, en un seul cycle. Il permet en outre la réduction de la taille des chemins critiques et ainsi, il permet un fonctionnement des processeurs à fréquence plus élevée.

Toutefois, cette technique a ses limites : dans l'exemple suivant, la troisième instruction ne peut être exécutée tant que le résultat de la première n'est pas connu, puisqu'elle l'utilise. Une bulle est alors insérée dans la pipeline, le débit en est alors amoindri.

Programme : Inc A,#1
Mov B,#2
Add A,R0
Sub B,#3

temps

Lecture	Chargement opérandes	traitement	Sauvegarde
Inc A,#1			
Mov B,#2	Inc A,#1		
Add A,R0	Mov B,#2	Inc A,#1	
NOP	NOP	Mov B,#2	Inc A,#1
Sub B,#3	Add A,R0	NOP	Mov B,#2
	Sub B,#3	Add A,R0	NOP
		Sub B,#3	Add A,R0
			Sub B,#3

cas d'échec de pipeline

Le problème des échecs de pipeline se produit aussi au moment des sauts. En effet, lors d'un saut conditionnel, il y a deux suites d'instructions à exécuter. Celle qui le sera vraiment ne pourra être déterminée qu'au dernier moment. Les processeurs actuels utilisent des techniques de prédiction de sauts et tâchent ainsi d'exécuter la bonne suite d'instructions. Le Pentium considère en guise de prédiction que le test échouera, du fait que les boucles utilisent principalement des conditions de maintien.

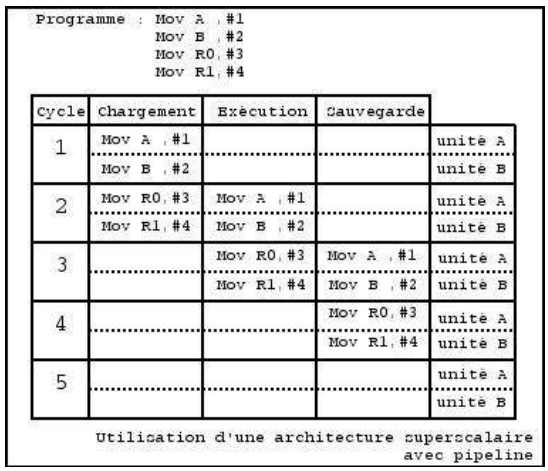
La technologie pipeline ne permet donc qu'une exécution théorique des instructions en un seul cycle. Pour qu'elle soit au mieux utilisée, le compilateur (ou le programmeur) doit organiser les instructions de sorte à limiter les cas d'échec.

Cours : *Technologie des processeurs modernes*
Public : *IUT Informatique*
Auteur : *P. Pinault*

4. Les architectures super-scalaires :

L'utilisation de techniques super-scalaire consiste à exécuter non plus une instruction à chaque cycle mais plusieurs en même temps. Il s'agit donc d'exécuter les instructions en parallèle. Elles nécessitent la multiplication des unités de traitement et par conséquent une augmentation importante du nombre de transistors utilisés.

L'exemple suivant illustre l'utilisation d'une architecture super-scalaire possédant deux unités d'exécution :



Utilisation d'une architecture superscalaire avec pipeline

Tout comme pour l'utilisation de pipeline, il peut y avoir des échecs de parallélisation des instructions lorsque celles-ci font appel aux mêmes registres ou zones mémoire. Le programme doit être là encore organisé au mieux pour profiter pleinement de cette architecture.

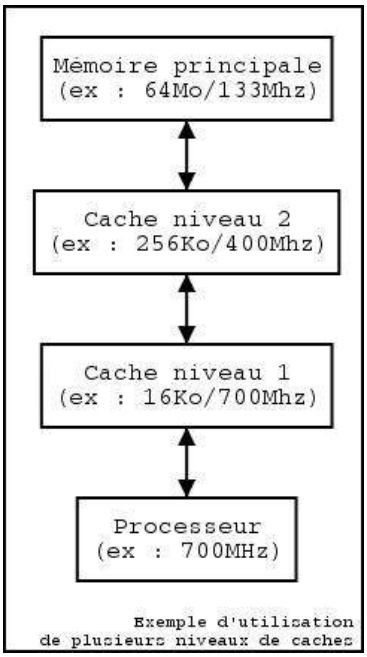
Certains processeurs sont d'ailleurs eux-mêmes capables de réorganiser les instructions à exécuter au fur et à mesure qu'elles se présentent.

5. Les mémoires caches

La mémoire principale des systèmes informatiques est d'un temps d'accès relativement lent. Par conséquent, elles ne sont utilisées qu'à quelques centaines de mégahertz alors que le processeur est capable de les solliciter plusieurs milliards de fois chaque seconde. De cette différence de débit provient un des principaux freins à la puissance de nos systèmes. En effet, le processeur peut alors se retrouver en famine, c'est à dire en attente de données ou d'instructions. Cette attente peut représenter un temps important lors de l'exécution d'un programme.
Les technologies nécessaires à l'utilisation de mémoire plus rapide en grande quantité se révélant trop coûteuses, les systèmes actuels utilisent pour compenser le phénomène de famine plusieurs niveaux de mémoires caches.

La mémoire cache est une mémoire de type statique, d'accès beaucoup plus rapide. Elle est présente en petite quantité dans le système. Elle contient des blocs données, copies de la mémoire principale, accédés dernièrement par le processeur. Ce système repose sur le principe que le processeur accède de façon successive à des données ou instructions proches les unes des autres.

Un système comprend généralement plusieurs niveaux de caches dont le temps de réponse est de plus en plus bref. Les caches sont gérés de façon transparente pour le processeur, seuls, ils synchronisent leurs données, chargent de nouveaux blocs.
Toutefois, le processeur peut souhaiter accéder à un bloc non chargé, alors celui-ci sera copié dans les caches, entraînant une famine le temps de cette opération. Une nouvelle fois, les compilateurs viennent en aide aux programmeurs en optimisant au mieux les accès mémoire pour limiter les échecs d'accès aux caches.



Exemple d'utilisation de plusieurs niveaux de caches

Le cache de niveau 1 est en général intégré dans le processeur, celui de niveau 2 est lui plutôt intégré à la carte mère ou placé sur la carte processeur dans le cas échéant.

La mémoire cache a un impact très important sur les performances d'un système, toutefois, le gain apporté suit une courbe logarithmique. Il n'est donc pas nécessaire d'utiliser de grandes quantités de cette mémoire dans un système. Un compromis est généralement trouvé entre les performances gagnées et le surcoût engendré.

6. L'utilisation de co-processeurs

Le processeur peut être aidé dans ses tâches par un composant externe spécialisé. Ce composant, alors appelé co-processeur, est capable de réaliser en parallèle du fonctionnement du micro-processeur un traitement spécifique évolué. Ce peut être par exemple un calcul 3D comme le font les cartes vidéo actuelles ou un transfert de données entre la mémoire et un périphérique comme le font les DMA.

Les DMA sont très utiles par exemple lorsqu'il est nécessaire de transférer des données par bloc de façon périodique : le processeur traitera un lot de données alors que le suivant, sera en cours de transfert. Ce traitement parallèle permet de réaliser d'importantes économies de charge CPU.

Certains co-processeurs sont intégrés au coeur du micro-processeur. Les Pentium intègrent un co-processeur dédié au calcul flottant et des co-processeurs dédiés aux traitements multimédia : MMX et 3DNow.

7. Evolutions récentes et futures

Du point de vue physique, les recherches actuelles montrent qu'il est possible de passer outre les barrières évoquées par Moore. En effet, l'utilisation de nano-technologies, voire de technologies atomiques devraient permettre de gagner encore et toujours en puissance de calcul et capacité de stockage. L'utilisation de technologies optiques permettront sans doute de passer outre les problèmes liés à l'utilisation de bus à très haute fréquence sur les cartes mères et aideront ainsi à l'utilisation de mémoires plus rapides.

Du point de vue structurel, le passage de 32 à 64 bits permettra entre autre d'augmenter le débit des données échangées entre le processeur et ses périphériques. De nombreuses améliorations des techniques actuelles, permises par l'augmentation de densité des composants, permettra l'utilisation au mieux des capacité des processeurs. Des technologies comme l'hyper-threading sont déjà intégrées aux processeurs de dernière génération. Celle-ci permet de confier au niveau matériel la gestion de thread auparavant géré au niveau système. Il est ainsi possible de remplacer les bulles insérées dans les pipelines, lors des conflits, par les instructions d'un autre thread et d'utiliser alors au mieux les capacités de traitement parallèle de nos processeurs.

Il est évident que de nombreuses technologies vont encore voir le jour. La bataille que se livrent les fondeurs pour produire le processeur le plus puissant du marché garanti une innovation continue et soutenue pour les années à venir.