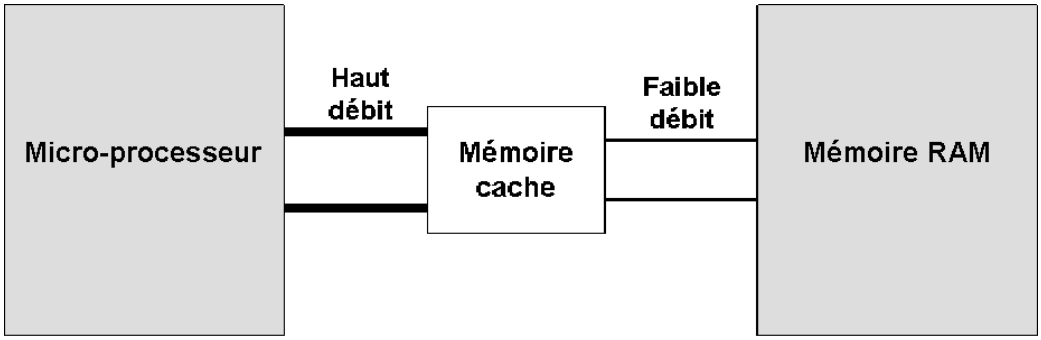
La mémoire cache
UE USI

La mémoire cache ou mémoire tampon est devenue en quelques années le compagnon indispensable des microprocesseurs ou disques durs. Quand on dit indispensable c´est un euphémisme, elle peut carrément transformer un processeur en fin de carrière en un jeune premier fringant.

Mais voilà, il n´existe pas une seule mémoire cache, mais plusieurs qui se différencient par leur architecture, leur taille ou encore leur vitesse de fonctionnement. C´est pour vous aider à faire un peu mieux la part des choses que nous allons voir ces différents aspects.

I. C´EST QUOI LA MÉMOIRE CACHE ?

La mémoire cache a pour fonction d´accélérer les communications entre un microprocesseur (Pentium, Athlon, ARM, ...) et un composant servant à stocker les données (RAM, disque dur). Attention, la mémoire cache n´accélère pas la vitesse de la mémoire ou du disque dur (la vitesse est une caractéristique physique). C´est plus subtil et on va éclaircir tout ça dans la suite de ce document.



1. Pourquoi a-t-on besoin de mémoire cache ?

Un microprocesseur a besoin d´informations pour exécuter ses instructions. Celles-ci sont situées dans une unité de stockage. Or les microprocesseurs sont devenus si performants qu´aucune unité de stockage n´est capable de fournir autant d´informations que le microprocesseur peut en traiter.

En effet, les microprocesseurs équipant les ordinateurs personnels actuels ont une fréquence de fonctionnement d'environ 3 Ghz, alors que celles des mémoires RAM est de 400 Mhz. Pour exécuter une instruction simple du genre (A := B + C – donner à A le résultat de l'addition de B et C), le microprocesseur doit effectuer 3 accès mémoire (2 accès en lecture pour lire les valeurs de B et C, 1 accès en écriture pour écrire la valeur de A). Ceci signifie que le processeur, au lieu de fonctionner à sa pleine vitesse d'environ 750 millions d'instructions par secondes, devra à chaque fois attendre pour lire et écrire des informations dans la mémoire. Ceci le ralenti d'environ 10 fois.

La mémoire cache permet en partie de palier à cette insuffisance.

2. Comment ça marche ?

Lorsque le microprocesseur a besoin d´une donnée, il regarde si elle est disponible dans la mémoire cache, si ce n´est pas le cas, il va la chercher dans l´unité de stockage et en même temps la dépose dans la mémoire cache. Ainsi la prochaine fois qu´il aura besoin de cette information, il y accédera directement par la mémoire cache et donc plus rapidement.

3. Quel composant se sert de mémoire cache ?

Toute unité de stockage peut servir de mémoire cache. Il suffit juste qu´elle soit plus rapide que l´unité de stockage principale. Voici trois exemples :

1 - Microprocesseur -> mémoire cache -> RAM
2 - Microprocesseur -> mémoire cache -> Disque dur
3 - Microprocesseur -> mémoire cache -> Internet

Dans le premier exemple, la mémoire cache sera une mémoire RAM ultra rapide (intégrée au microprocesseur ou mémoire SRAM par exemple), dans le second exemple, de la mémoire RAM traditionnelle fera très bien l´affaire, dans le troisième cas, c´est votre disque dur qui fera office de mémoire cache.

4. Pourquoi ça marche ?

Les capacités de stockage de la mémoire cache peuvent être 100 fois (ou plus) plus petites que celles de l´unité de stockage principale. On ne peut donc pas tout mettre en mémoire cache.

> Question : Alors comment ce fait-il que cela permette d´accélérer les communications entre la mémoire principale et le microprocesseur et donc la vitesse de traitement du microprocesseur.

Réponse : Un microprocesseur exécute un programme fonction par fonction. Chacune de ces fonctions correspond à une ou deux tâches, rarement plus. Or en général chacune de ces tâches n´utilise qu´une petite quantité de données qui peut être contenue en grande partie en mémoire cache.

> Question : Mais quand le microprocesseur a besoin d´une nouvelle donnée, elle ne se trouve pas en mémoire cache, il faut bien aller la chercher dans l´unité de stockage et cela prend le même temps avec ou sans mémoire cache.

Réponse : C´est vrai, mais l´intérêt même de l´informatique c´est d´être capable d´exécuter des tâches répétitives, c´est d´ailleurs à ça qu´un microprocesseur passe le plus clair de son temps. C´est donc lorsqu´il voudra accéder une deuxième, troisième, ... fois que le traitement sera accéléré.

II. INDICATEURS DE PERFORMANCE DE LA MÉMOIRE CACHE

Il existe 3 méthodes d'organisation des caches :

- Direct Mapped
- N-way associative
- Fully associative

Pour pouvoir comparer les performances respectives de chaque méthode d´organisation, on utilise deux informations :

1. Ratio de réussite :

C´est le rapport entre le nombre total d´accès au cache sur le nombre d´accès ayant permis de trouver l´information dans le cache. Cette valeur s´exprime en pourcentage. Autrement dit, c´est la chance qu´a le microprocesseur de trouver une information dans la mémoire cache.

Plus cette chance est grande, moins le microprocesseur fait appel à la mémoire RAM et donc plus le traitement d´un programme sera fait rapidement. Ce ratio mesure l´efficacité du cache. Sans tenter d´entrer dans des détails de calcul statistique, on considère les ratios suivants pour les différentes méthodes d´organisation :

2. Temps de latence :

C´est le temps moyen que le microprocesseur met pour consulter les lignes de cache. Autrement dit, c´est le temps qu´il faut au microprocesseur pour savoir si l´information qu´il cherche est ou n´est pas dans la mémoire cache. En fait, il s´agit plutôt d´une durée relative car ce qui nous intéresse, c´est la différence de temps entre chaque méthode d'organisation du cache.

Le temps de latence est fortement défavorable à la méthode fully associative et cette différence avec les autres méthodes n´est pas compensée par un meilleur taux de réussite. Cette méthode est à réserver pour cacher de petites quantités de mémoire, afin que le nombre de lignes de cache soit le plus petit possible. La méthode direct mapped apparaît par contre comme un cas idéal. Si la taille des blocs de mémoire RAM que couvre chaque ligne de cache n´est pas trop importante, elle se révèle assez efficace, d´autant plus que sa conception (du point de vue électronique) reste simple et donc moins coûteuse à fabriquer.

Pour conclure ce paragraphe, on peut constater que la mémoire N-way associative récupère le meilleur des deux autres méthodes. Ceci explique pourquoi c´est la plus employée actuellement.

3. Bande passante :

Tout comme la bande passante d'une mémoire, la bande passante du cache est conditionnée par deux valeurs, à savoir la taille et la fréquence de son bus externe !