Architecture des ordinateurs

Mémoire Cache

Florent Devin, Matthias Colin

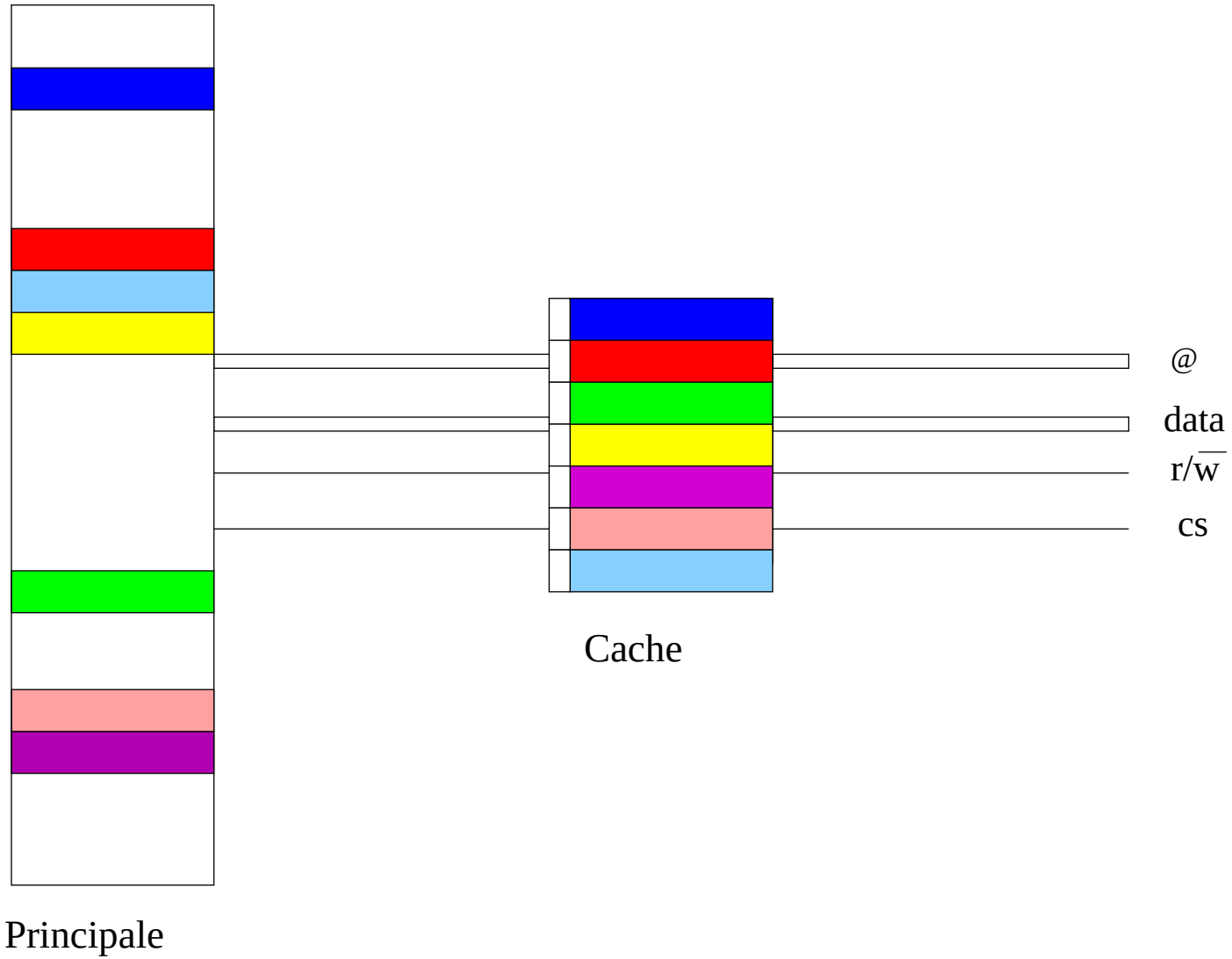
Ecole Internationale des Sciences du Traitement de l'Information

Mémoire cache

Présentation

- niveau de mémorisation intermédiaire
- très rapide : plusieurs dizaines de fois que la mémoire principale
- de petite capacité
- mémorise les données ou instructions les plus récentes
- situé
  - entre le processeur et la mémoire
  - entre le processeur et un autre cache

Mémoire cache



Mémoire cache

Principe

- Recherche d'une donnée dans le cache
  - succès cache : la donnée est présente dans le cache
  - défaut de cache : la donnée est absente du cache $\Rightarrow$ recherche dans la mémoire suivante

Mémoire cache

Principe du défaut de cache

- bloc mémoire : ensemble de mots d'adresses contigües
- la mémoire est découpée en bloc
  - 32 octets pour un processeur Alpha AXP 21064
- accès à une adresse
  - défaut de cache : le bloc entier est copié dans le cache
  - succès de cache : Rien

Mémoire cache

Caractéristiques

- utilisation de multiple caches, organisés en niveau (level)
- Un cache peut
  - être situé sur la même puce que le processeur (on-chip/internal cache)
  - n'être qu'accessible via un bus externe au processeur (external cache).
- L'utilisation d'un cache interne permet
  - d'augmenter les performances
  - de laisser le bus externe disponible

Mémoire cache

Organisation typique

- un cache interne (de niveau 1)
- un cache externe (de niveau 2)
  - doit être de 10 à 100 fois plus grand que le/les caches de niveau 1
  - sinon AUCUN intérêt

Mémoire cache

Spécialisation

- L'utilisation de plusieurs caches
  - stockage des données
  - stockage des instructions
- L'intérêt permet de dissocier le mécanisme
  - d'exécution des instructions
  - recherche et décodage des instructions

Mémoire cache

Taille du cache

- suffisamment petit pour que
  - son coût soit proche de celui d'une mémoire principale
  - temps d'accès soit le plus intéressant possible
- suffisamment grand pour ne pas avoir à trop accéder à la mémoire principale
- Des études ont montré que les caches les plus efficaces on une taille inférieure à 512 K mots

Mémoire cache

Correspondance

- taille du cache plus petite que la taille de la mémoire
- définition d' une stratégie de copie des blocs de données
- 3 stratégies possibles
- Aujourd'hui la grande majorité des caches sont à correspondance directe ou à correspondance associative par ensemble de 2 ou 4 blocs.

Mémoire cache

Correspondance

- **correspondance directe** : chaque bloc mémoire ne peut être placé que dans un seul bloc du cache,
- **correspondance totalement associative** : chaque bloc mémoire peut être placé dans n'importe quel bloc du cache
- **correspondance associative par ensemble** : chaque bloc mémoire peut être placé dans n'importe quel bloc du cache parmi un ensemble de n blocs.

Mémoire cache

Algorithme de remplacement

- plusieurs manières de déterminer quel bloc du cache doit être remplacé, dont les principales (du - au + efficace)
  - choisir le plus ancien bloc du cache (FIFO, First In First Out)
  - choisir un bloc candidat de manière aléatoire
  - choisir le bloc le moins récemment utilisé (LRU Least Recently Used)
  - choisir le bloc le moins fréquemment utilisé (LFU Least Frequently Used)

Mémoire cache

Politique d'écriture

- Deux situations possibles selon la présence dans le cache
- Donnée présente dans le cache :
  - écrire à la fois dans le bloc du cache et dans le bloc de la mémoire (écriture simultanée, ou *write through*)
  - écrire uniquement dans le bloc du cache, et différer l'écriture de ce bloc en mémoire lorsque l'emplacement qu'il occupe sera désigné pour recevoir un nouveau bloc mémoire (réécriture ou *write back*).

Mémoire cache

Politique d'écriture

- Deux situations possibles selon la présence dans le cache
- Donnée non présente dans le cache, alors on peut
  - de charger le bloc de la mémoire dans le cache puis effectuer l'opération d'écriture (écriture allouée)
  - d'effectuer l'écriture directement dans la mémoire (écriture non allouée).
- optimisation classique : utiliser un tampon d'écriture permettant au processeur de continuer à travailler dès que la donnée est écrite dans le tampon

Mémoire cache

Tags à ajouter aux données dans le cache

- dépendent du mode de gestion choisi
- clé : une partie du numéro de bloc (sous-partie @)
- gestion FIFO, LRU
- bit de modification

Mémoire cache

Performance

- Évaluation de la performance par le calcul du temps d'accès mémoire moyen
  - temps d'accès mémoire moyen = temps d'accès succès + taux d'échec * pénalité d'échec
  - temps d'accès succès = temps d'accès à une donnée résidant dans le cache
  - taux d'échec = nombre de défaut de cache / nombre d'accès cache

Mémoire cache

Exemple

- durée d'un cycle horloge : τ
- pénalité d'échec : 10 cycles
- durée d'une instruction (sans référence mémoire) : 2 cycles
- nombre de références mémoire par instruction : 1.33
- taux d'échec : 2%
- temps d'accès succès : négligeable

Mémoire cache

Exemple

- Avec cache
  - temps d'exécution moyen d'une instruction : $(2 + 1.33 * 2\% * 10)\tau = 2.27\tau$
- Sans cache
  - temps d'exécution moyen d'une instruction : $(2 + 1.33 * 10)\tau = 15,3\tau$