計算機方式論

第13章 キャッシュ

–性能の指標–

1

キャッシュ性能の指標

①ヒット率(hit ratio)
②アクセス時間
③ミスペナルティ時間
④主記憶更新時間
⑤キャッシュ容量
⑥ラインサイズ

2

①キャッシュ性能の指標–ヒット率p(hit ratio)

- CPUが要求するアクセス対象(番地)のキャッシュ中に存在する確率。
- 高いほど良い。
- 対象への実効アクセス時間(平均アクセス時間)T_e
  ヒット率 p
  キャッシュへのアクセス時間 T_c
  主記憶にアクセスしたとき要する時間 T_{miss}としたとき、
  $T_e = pT_c + (1-p)T_{miss}$

3

ヒット率pと実効アクセス時間Teの例

- $T_e = pT_c + (1-p)T_{miss}$

  T_c = 5ns、$T_{miss} = 31T_c$ = 155ns とすると、

  $T_e=(31-30p)T_c$

  p = 100%　$T_e = T_c$= 5ns
  キャッシュへのアクセス時間(最良)
  p = 90%　T_e= 4 T_c= 20ns
  p = 50%　T_e= 16 T_c= 80ns
  p = 0%　$T_e = T_{miss}$= 31T_c= 155ns(最悪)

4

②キャッシュ性能の指標–アクセス時間Tc

- Tc = キャッシュメモリ自身のアクセス時間
  ＋ キャッシュラインへのマッピング時間
- キャッシュメモリ自身のアクセス時間は、構成する記憶素子で決まる…数nsのSRAMが使われる。
- マッピング時間とは、アクセス対象の主記憶番地から、キャッシュラインを決定する時間。連想検索方式のマッピング（アソシアティブ方式）のときは、キャッシュライン数が増えると、マッピング時間が増える。
- ヒット率pが高い程、Tcは実効アクセス時間Teに影響する。

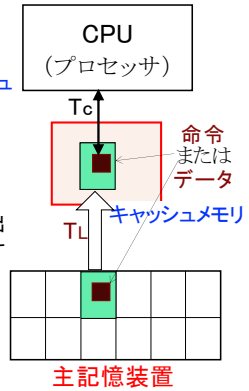
$Te = pTc + (1-p)Tmiss$

③キャッシュ性能の指標–ミスペナルティ時間Tmiss

- ミスヒット時の処理時間で、主として、ライン置換時間。
- ライン置換アルゴリズム。
- 割り込み処理性能。
- 主記憶–キャッシュ間のデータ転送速度に依存。

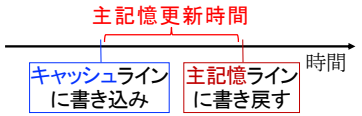
ミスヒット時の実効アクセス時間Te

- 対象がミスヒット時、
  - 空のキャッシュラインが在れば、主記憶ラインを時間TLでコピーし、キャッシュライン中の対象を時間Tcでアクセス。
    Tmiss=TL+Tc
  - 空のキャッシュラインがなければ、アソシアティブマッピング方式では、ライン置換アルゴリズムが、時間TRで、追出すキャッシュラインを選び主記憶に書き戻すことで空きキャッシュラインをつくって、そこに主記憶ラインを時間TLでコピーし、対象を時間Tcでアクセス。
    Tmiss=TR+TL+Tc
- Tmiss=TR+TL+Tc とすると、
  $Te = pTc + (1-p)Tmiss = Tc + (1-p)(TL+TR)$



④キャッシュ性能の指標–主記憶更新時間

- 更新されたキャッシュラインを主記憶ラインに書き戻すまでの時間。
- 短い程有効
- 書き込みアクセスが多いときほど、主記憶を更新することも多くなる。
- 主記憶の素子、ラインサイズに依存。



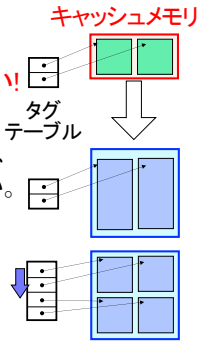
⑤キャッシュ性能の指標–キャッシュ容量

- キャッシュ容量を**増やした**とき、
  ヒット率pは増えるが、
  実効アクセス時間Teが**減るとは限らない!**

(1)**ラインサイズ**を**大きく**すると、
主記憶–キャッシュ間転送時間TLが増え、
実効アクセス時間Teが**減るとは限らない。**

$$Te = Tc + (1-p)(TL + TR)$$

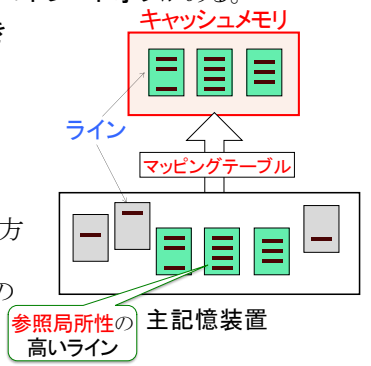
(2)**ラインサイズ**を**同じか小さく**すると、
ライン数が**増え**、
連想検索を行う**マッピング方式**では、
タグ検索を行うため、Tcが**増え**、
実効アクセス時間Teが**減るとは限らない。**



⑥キャッシュ性能の指標–ラインサイズ

- **ライン**が主記憶–キャッシュ間の**転送単位**のため、
  そのラインの**大きさ**でつぎの**トレードオフ**がある。
- **ラインサイズ**が**大きい**とき
  (**キャッシュ容量**は同じ)、

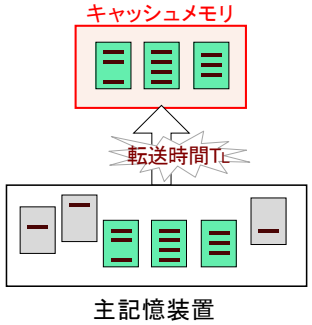
(1)**参照局所性**が**高ければ**、
ヒット率が高い。
(2)キャッシュ内の**ライン数**が
少なくなり、
連想検索を行うマッピング方式では、**検索時間**が**減る**。
また、**マッピングテーブル**の
構成コストも**安く**なる。



⑥キャッシュ性能の指標–ラインサイズ

- **ライン**が主記憶–キャッシュ間の**転送単位**のため、
  その大きさでつぎの**トレードオフ**がある。
- **ラインサイズ**が**大きい**とき
  (**キャッシュ容量**は同じ)、

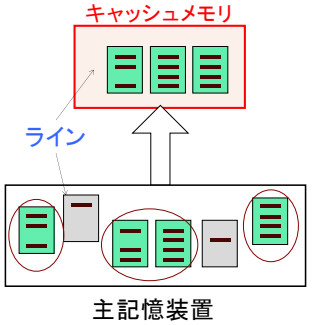
(3)主記憶–キャッシュ間の
データ転送時間TLがかかる。



⑥キャッシュ性能の指標–ラインサイズ

- **ライン**が主記憶–キャッシュ間の**転送単位**のため、
  その大きさでつぎの**トレードオフ**がある。
- **ラインサイズ**が**大きい**とき
  (**キャッシュ容量**は同じ)、

(4)**キャッシュ内ライン数**が**少なすぎる**と、
参照局所性を**活用できなくなり**、
ヒット率が**低下**する。



⑥キャッシュ性能の指標−ラインサイズ（まとめ）

- ラインが主記憶−キャッシュ間の転送単位のため、ラインの大きさでつぎのトレードオフがある。
- ラインサイズが大きいとき（キャッシュ容量一定）は、
  (1)参照局所性が高ければ、ヒット率pは高い $T_e = T_c + (1-p)(T_L + T_R)$
  (2)キャッシュ内のライン数が少なくなり、マッピングテーブルの構成コストが安くなる。連想検索を行うマッピング方式では、検索時間（キャッシュアクセス時間T_c）が減る。一方、
  (3)主記憶−キャッシュ間のデータ転送時間T_Lがかかる。
  (4)キャッシュ内ライン数が少なすぎると、参照局所性を活用できなくなり、ヒット率pが低下する。
- ラインサイズが小さいときは、この長短が逆になる。
- マイクロコンピュータのラインサイズは、32～128Bである。
- キャッシュ容量は、32K～256KB。

13

Intel Core i7−4770 Q2'13 specifications

- Core i7, 4core, 3.4GHz, 32GB

L1キャッシュ
命令キャッシュ4x32KB, 8way set associative
データキャッシュ4x32KB, 8way set associative

L2キャッシュ 4x256KB, 8way set associative

L3キャッシュ 8MB, 16way set associative

ラインの大きさ 64B

14

Intel Core i9−12900KS Q1'22 specifications

- Core i9, 16core, 3.4GHz, 128GB

L1キャッシュ
命令キャッシュ16x32KB, 8way set associative
データキャッシュ16x48KB, 8way set associative

L2キャッシュ 10x1.25MB, 8way set associative

L3キャッシュ 30MB, 16way set associative

ラインの大きさ 64B

15