W6D1-Notes

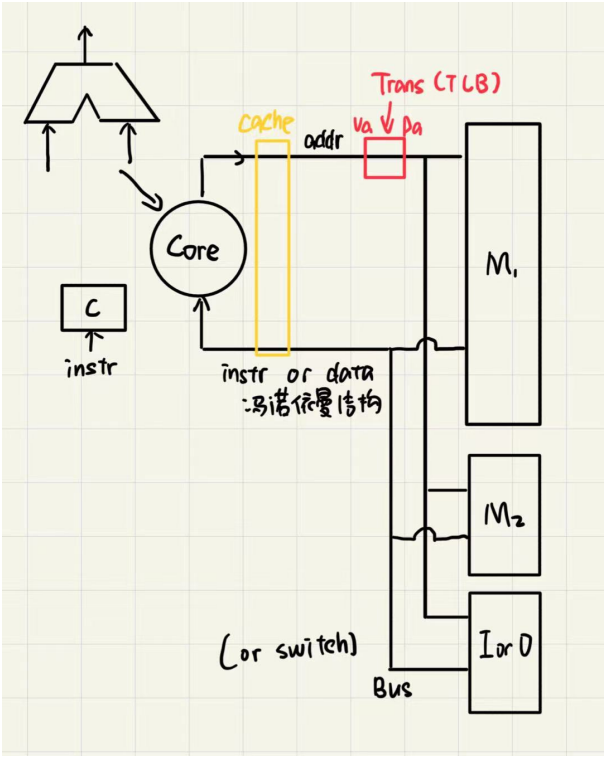
Caches and Memory Systems I

overview

提高cache效率需减少平均内存访问时间AMAT
$AMAT_{\$} = T_{hit} + \eta_{miss-rate} \times T_{penalty}$
improving Cache Performance

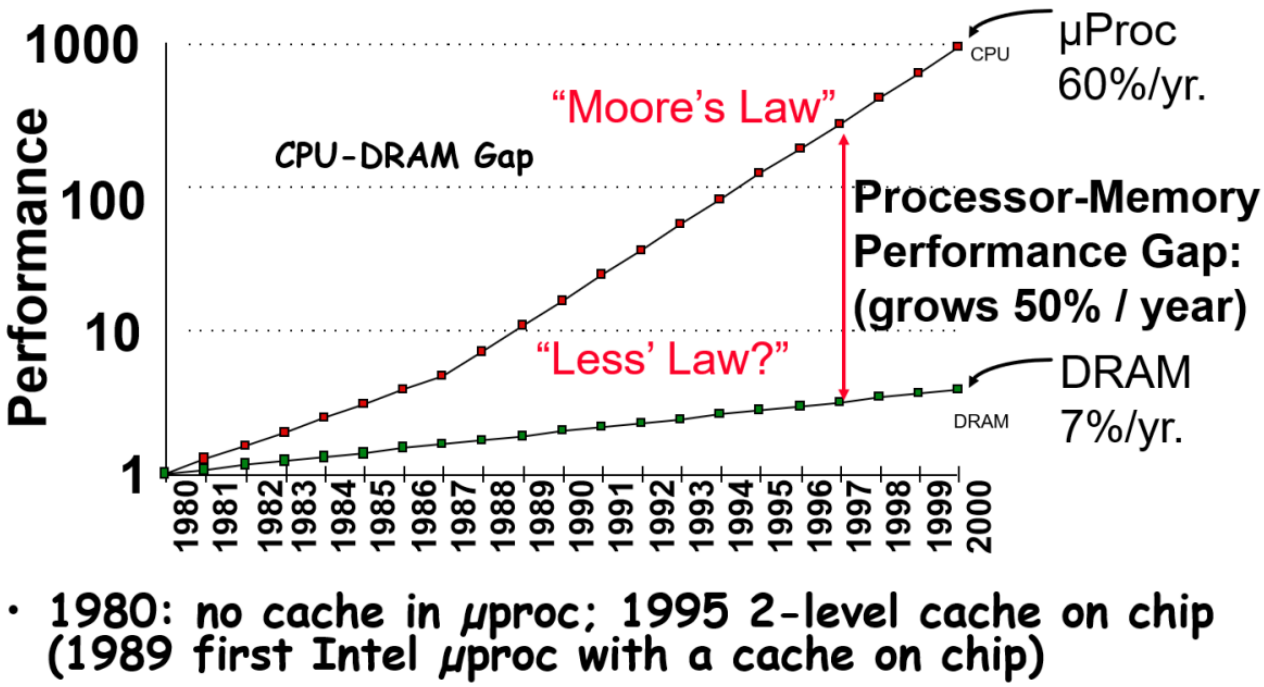
1. Reduce miss rate
2. Reduce $T_{penalty}$
3. Reduce T_{hit}
   本节课主要讨论Reduce miss rate

cache

- cache（*locality*）$\neq$ buffer（*elastic*）
  - buffer 缓冲 两边速度或力量有差异 匹配两边速度或力量
  - cache 为下一级备份
- 整体架构
  - 

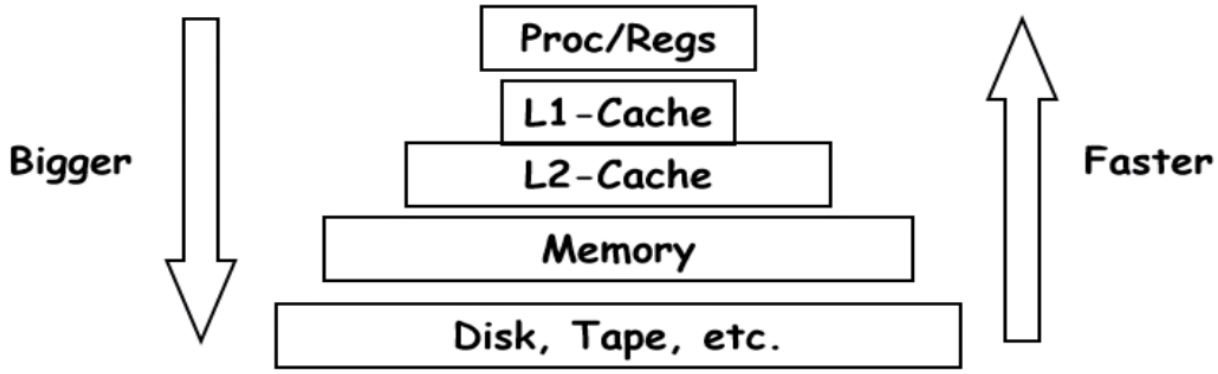
cache的必要性

- memory 的速度跟不上 cpu 的速度；
- 图上 gap 的存在导致 cache 的引入；



什么是cache

- 用于提高对慢速存储器的平均访问时间的小而快的存储器。
- 利用空间和时间局部性
- 在计算机体系结构中，几乎所有东西都是缓存



量化 Cache Performance

- AMAT = Average Memory Access Time
  $CPUtime = IC \times (\frac{ALUOps}{Inst} \times CPI_{AluOps} + \frac{MemAccess}{Inst} \times AMAT) \times CycleTime$
  $AMAT = HitTime + MissRate \times MissPenalty$
- 冯诺依曼结构：
  - **指令和数据**是不加区别地混合存储在同一个存储器中的，共享数据总线，根据**行为**判断类型。

- 哈佛结构：
  - 一级缓存中指令放在指令 cache ，数据放在数据 cache
  - 相比之下哈佛结构会有更小的 AMAT，因为虽然 cache 容量减少，miss 率增大，但不会有 structure hazard

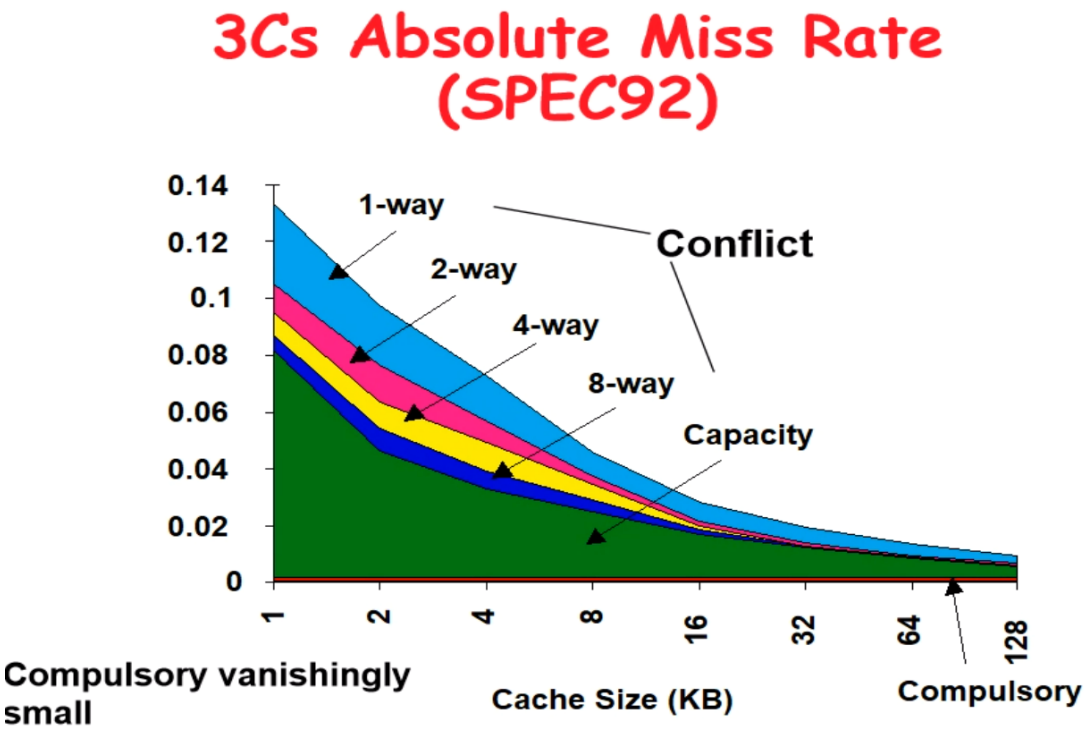
improving Cache Performance

1. Reduce miss rate
2. Reduce $T_{penalty}$
3. Reduce T_{hit}
   本节课主要讨论Reduce miss rate

如何 reduce misses

- miss 原因
  - Compulsory：对块的第一次访问不在缓存中，因此必须将块放入缓存中。也称为 cold start misses 或 first reference misses 。(即使 cache 无限大也会 miss)
  - Capacity：如果缓存不能包含程序执行期间所需的所有块，则由于块被丢弃并稍后被检索而导致 miss 。
  - Conflict：如果块放置策略设置为关联的或直接映射的，冲突缺失(除了上两者)将会发生，因为如果有太多块映射到它的集合，块可以被丢弃，然后再被检索。也称为碰撞失误或干扰失误。
  - Coherence：缓存一致性导致的丢失。多核 CPU 中每个核都有自己的 cache，如果一个核改变了内存中一处的 值，为保证数据正确性，一般采用强制将其他核的 cache invalid，从而导致 cache miss。
- 一些优化方法
  - 预热可减少 Compulsory miss
  - 增大 cache 容量可减少 Capacity miss

- 增大 way 数会显著减小 conflict miss，但 cache size 足够大时效果减弱



- miss rate of 1-way associative cache size X = miss rate 2-way associative cache size X/2