

高等计算机体系结构

第九讲: 分层存储体系结构及Cache基础

栾钟治
北京航空航天大学 计算机学院 中德联合软件研究所

1

阅读材料

- 分层存储体系结构
- Patterson & Hennessy's *Computer Organization and Design: The Hardware/Software Interface* (计算机组成与设计: 软硬件接口)
 - 第五章: 5.1-5.3
- Maurice Wilkes早期关于cache的论文
 - Wilkes, "Slave Memories and Dynamic Storage Allocation," IEEE Trans. On Electronic Computers, 1965.

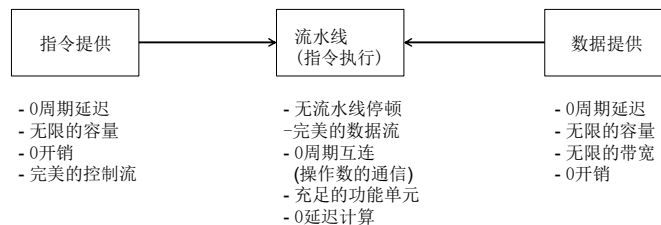
2

2

理想化

到目前为止, 我们想象

- 程序看到连续的4GB内存
- 在一个处理周期中可以访问内存的任意位置



4.1. Ideally one would desire an indefinitely large memory capacity such that any particular aggregate of 40 binary digits. word (cf. 2.3), would be immediately available—i.e. in a tin

---- Burks, Goldstein, von Neumann, 1946

3

3

真实的世界

- 无法负担也不需要像用户地址空间那样大的内存(想想64位的ISAs)
- 大多数机器在若干程序之间是“多任务”执行的
- 找不到既能支持千兆字节 (GB), 又能在千兆赫 (GHz) 主频下使用的存储技术
 - “魔法”内存
- “魔法”内存仍然是非常接近现实的“可用”抽象, 因为:
 - 分层存储: 又大又快
 - 虚拟内存: 连续且私有

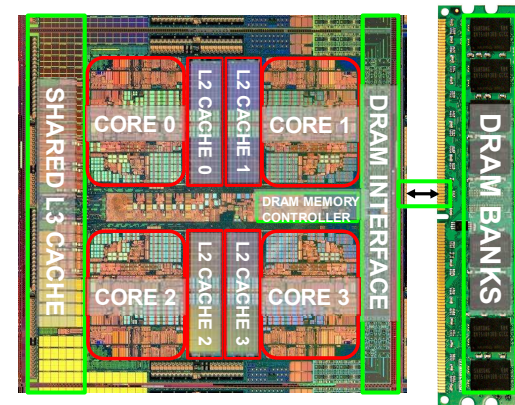
4

4

分层的存储体系结构

5

现代系统中的存储



6

理想存储器

- 0访问时间(延迟)
- 无限的容量
- 0开销
- 无限的带宽 (支持多路并行访存)

7

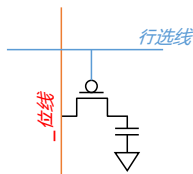
问题

- 理想存储器的需求之间互相制约
- 越大则越慢
 - 越大 → 确定位置花费的时间越长
- 越快则越贵
 - 存储器技术: SRAM vs. DRAM
- 带宽越高越贵
 - 需要更多的bank, 更多的端口, 更高的频率, 或更快的技术

8

存储器技术: DRAM

- 动态随机存取存储器(Dynamic random access memory)
- 电容器充电状态表示了存储的值
 - 电容器充电或者未充电代表存储1或0
 - 1 个电容器
 - 1 个存取晶体管
- 电容器向行选线方向漏电
 - DRAM单元随时间的推移损失电荷
 - DRAM单元需要刷新

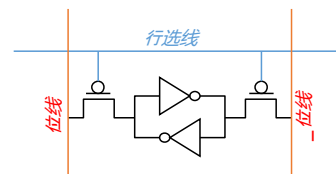


9

9

存储器技术: SRAM

- 静态随机存取存储器(Static random access memory)
- 两个交叉耦合的反相器存储1bit
 - 反馈路径使被存储的值保持在SRAM单元中
 - 4个晶体管用于存储
 - 2个晶体管用于存取



10

10

存储器Bank的组织 and 操作

-
- 读访问过程:
 1. 译码行地址并驱动字线
 - 读整行
 2. 选择位驱动位线
 3. 放大行数据
 4. 译码列地址并选择行的子集
 - 发送至输出
 5. 预充电位线
 - 为下次访问做准备

11

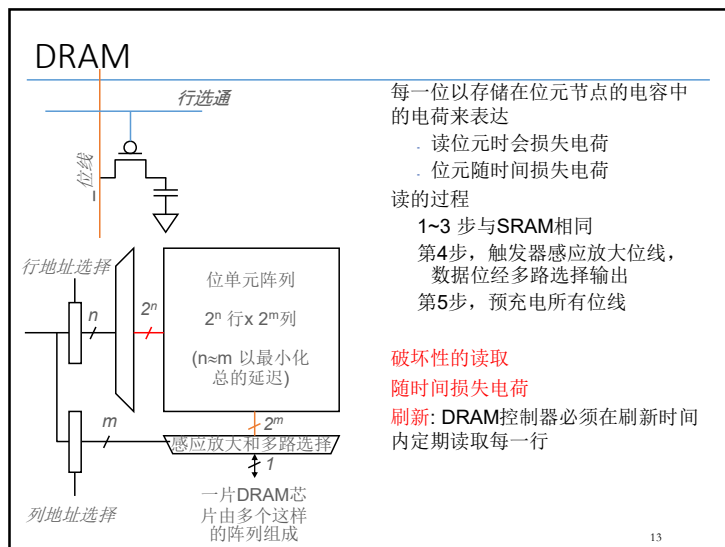
11

SRAM

-
- 读过程
1. 地址译码
 2. 驱动行选通
 3. 被选的位单元驱动位线 (整行一起读)
 4. 差分感应并选通列 (数据准备好)
 5. 预充电所有位线 (为下一次读或者写)
- 第2、3两步是产生访问延迟的主要阶段
第2、3和5步是整个循环中最耗时的
- 第2步与 2^m 成正比
 - 第3和5步与 2^n 成正比

12

12



13

DRAM vs. SRAM

- DRAM
 - 存取更慢 (电容)
 - 密度更高 (每单元1个晶体管、1个电容)
 - 成本更低
 - 需要刷新 (功耗, 性能, 电路)
 - 需要把电容器和逻辑电路加工到一起
- SRAM
 - 存取更快 (无需电容)
 - 密度更低 (每单元6个晶体管)
 - 成本更高
 - 不需要刷新
 - 与加工逻辑电路过程一致 (没有电容)

14

问题

- 越大越慢
 - SRAM, 512 B, 亚纳秒
 - SRAM, KB~MB, ~纳秒
 - DRAM, GB, ~50 纳秒
 - 硬盘, TB, ~10 毫秒
- 越快越贵(\$ 以及芯片面积)
 - SRAM, < 10\$ 每MB
 - DRAM, < 1\$ 每MB
 - 硬盘, < 1\$ 每GB
 - 这些数据随时间变化很快
- 其它技术也在发展
 - 闪存, 相变存储器 (技术还没有成熟)

15

为什么要有分层存储体系结构?

- 我们想要既快又大
- 但是我们无法仅靠一层存储达到目的
- 思路: 采用多层的存储 (越大并且越慢的离处理器越远) 并且确保处理器需要的大多数数据在更快的层中

16

存储器分层

向这里移动需要用的数据

利用良好的局部性，
存储系统看起来似乎
又大又快

在这里备份
所有的数据

大但是慢

17

17

存储器分层

• 基本的tradeoff

- 快存储: 小
- 大存储: 慢

• 思路: 存储器分层



• 延迟, 成本, 尺寸, 带宽

18

18

解决方案背后的基本原则——#1

局部性原则

- 最近的过去是不久将来的最佳预测器
- 时间局部性(Temporal Locality):如果你刚刚做了某事，很可能你很快会再次做同样的事情
 - 你今天教室里，很有可能今后你会不断地定期来教室
 - 反之亦然
- 空间局部性(Spatial Locality):如果你刚刚做了某事，很可能你会做类似/相关的事情
 - 每次在这个教室里，你可能都坐在同一个座位上(或附近)
 - 你可能坐在同一个人旁边

程序往往比人更容易预测

19

19

解决方案背后的基本原则——#1

内存局部性

- “典型”程序在内存引用中具有很强的局部性
 - 典型的程序通常由“循环”导致
- 时间:
 - 程序倾向于在很短的时间内多次引用(读和写)相同的存储位置
- 空间:
 - 程序倾向于引用一组邻近的内存位置
 - 最值得注意的例子：1，指令存储器的引用；2，数组/数据结构引用)
- 推论: 一个程序可能在其生命周期中引用大量不同的内存位置，但不是在同一时间

20

20

局部性

- 最近的过去是预测不久的将来的最好参考
- **时间局部性(Temporal Locality)**: 如果你刚刚做了什么事, 很有可能你马上还会做相同的事
 - 今天你在这里, 有很大的机会你将会有规律的一次又一次出现在这里
- **空间局部性(Spatial Locality)**: 如果你做过什么事, 很有可能你将会做类似或相关的(空间上)其它事
 - 每次我在这间屋里看到你, 你可能都和某个相同的人坐的很近

21

21

存储局部性

- 一个“典型”的程序在引用存储器方面有很多的局部性
 - 比如, 很多典型的程序是由“循环”组成的
- **时间局部性**: 一个程序往往会在一个小的时间窗口内多次引用相同的存储位置
- **空间局部性**: 一个程序倾向于一次引用一串存储位置
 - 最引人关注的例子:
 - 1. 指令对存储的引用
 - 2. 数组或类似数据结构的引用

22

22

解决方案背后的基本原则——#2

避免重复运算

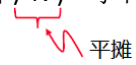
- 如果计算某个东西计算代价很高, 那么可以记住一会儿答案, 以防不久之后又需要它
- 需要局部性才能有效
- 缺乏局部性
 - 存储大量不同的答案(其中许多答案从未重复使用)
 - 从大量存储的答案中查找答案可能比重计算更贵
- 有局部性
 - 少量答案会一直重复使用!
 - 存储少量频繁用到的答案可以避免大多数重计算

23

23

解决方案背后的基本原则——#3

成本平摊

- 间接成本: 构建某些东西的一次性成本
- 每单元成本: 操作的每个单元的成本
- 总成本 = 间接费用 + 每单元成本 $\times N$
- 平均成本 = 总成本 / $N = (\text{间接成本} / N) + \text{每单元成本}$

- 如果成本可以平摊到大量的单元上, 那么高昂的间接成本通常是可以接受的
⇒ 降低了平均成本

24

24

cache的基本要素:利用时间局部性

- 思路: 将最近访问过的数据保存在自动管理的快速存储中 (称为 **cache**)
- 预期: 这些数据将会很快被再次访问
- 时间局部性原理
 - 最近访问的数据将会在不久的将来被再次访问
 - Maurice Wilkes:
 - Wilkes, "Slave Memories and Dynamic Storage Allocation," IEEE Trans. On Electronic Computers, 1965.
 - "The use is discussed of a fast core memory of, say 32000 words as a slave to a slower core memory of, say, one million words in such a way that in practical cases the effective access time is nearer that of the fast memory than that of the slow memory."

25

25

cache的基本要素:利用空间局部性

- 思路: 将与最近访问过的数据地址相邻的数据保存到自动管理的快速存储中
 - 逻辑上将存储器划分为大小相等的块
 - 以整块访问的方式向 **cache** 取数据
- 期待: 附近的数据将很快被访问
- 空间局部性原理
 - 存储器中相邻的数据将会在不久的将来被访问
 - 比如, 顺序的指令存取, 数组的遍历
 - IBM 360/85的实现
 - 16 KB 的 cache, 64B 的数据块
 - Liptay, "Structural aspects of the System/360 Model 85 II: the cache," IBM Systems Journal, 1968.

26

26

以书架类比

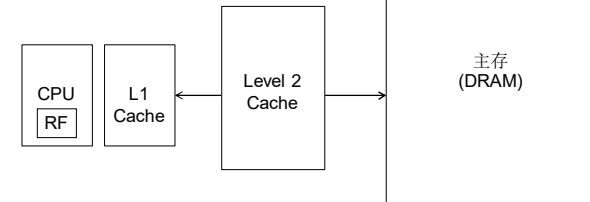
- 手里的书
- 书桌
- 书架
- 家里的储物盒
- 仓库里的储物盒
- 最近使用的书更可能被放在桌上
 - 计算机体系结构相关的书
 - 直到书桌被堆满
- 书架上相邻的书在同一时间段可能都是需要的
 - 如果书架整理的很好的话

27

27

流水线设计中的cache

- Cache需要和流水线紧密的集成
 - 理想情况下, 只用1个周期存取, 可以使相关的操作不用停顿
- 高频流水线 → 不能让cache太大
 - 但是, 我们又想要一个大的cache 并且 是流水线设计
- 思路: **Cache分层**



28

28

手动 vs. 自动管理

- **手动:** 程序员管理跨层的数据迁移
 - 对于大的程序而言, 程序员会非常痛苦
 - 上世纪50年代的“磁芯” vs “磁鼓” 存储器
 - 仍在某些嵌入式处理器中使用
- **自动:** 硬件管理跨层的数据迁移, **对程序员透明**
 - ++ 程序员的人生更美好
 - 简单的启发式方法: 将最近使用的内容保存在cache中
 - 一般的程序员不需要了解这一点
 - 不需要知道cache有多大, 也不需要了解它是如何工作的, 就能够写出“正确”的程序! (如果你想要一个“快”的程序呢?)

29

29

分层存储结构中的自动管理

- Wilkes, “**Slave Memories and Dynamic Storage Allocation**,” IEEE Trans. On Electronic Computers, 1965.

Slave Memories and Dynamic Storage Allocation

M. V. WILKES

SUMMARY

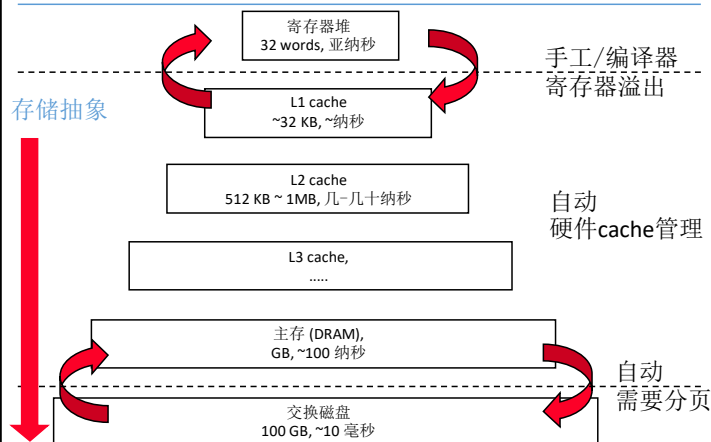
The use is discussed of a fast core memory of, say, 32 000 words as a slave to a slower core memory of, say, one million words in such a way that in practical cases the effective access time is nearer that of the fast memory than that of the slow memory.

- “By a slave memory I mean one which **automatically accumulates to itself words** that come from a slower main memory, and keeps them available for subsequent use without it being necessary for the penalty of main memory access to be incurred again.”

30

30

现代的分层存储体系结构



31

31

分层的延迟分析

- 对于给定的存储层次 i , 它有一个技术上固有的访问时间 t_i , 我们感知到的访问时间 T_i 比 t_i 要长
- 除了最外层, 每当要寻找一个给定地址时
 - 有一定的“命中”机会(命中率 h_i), 访问时间是 t_i
 - 有一定的“缺失”机会(缺失率 m_i), 访问时间是 $t_i + T_{i+1}$
 - $h_i + m_i = 1$
- 因此

$$T_i = h_i \cdot t_i + m_i \cdot (t_i + T_{i+1})$$

$$= t_i + m_i \cdot T_{i+1}$$

记住, 这里的 h_i 和 m_i 定义的命中率和缺失率针对的是在 L_{i+1} 缺失的引用

32

32

层次设计注意事项

- 递归的延迟方程

$$T_i = t_i + m_i \cdot T_{i+1}$$

- 目标: 在可以接受的开销范围内获得满意的 T_1
- $T_i \approx t_i$ 将是令人满意的
- 保持低的缺失率 m_i
 - 增加容量 C_i 以降低缺失率 m_i , 但是要注意会增加 t_i
 - 通过更好的管理降低缺失率 m_i (替换::预测你不需要什么, 预取::预测你需要什么)
- 保持低的 T_{i+1}
 - 让更低的层次更快, 但是要注意会增加成本
 - 引入中间层做折衷

33

33

层次设计注意事项

- DRAM
 - 针对容量/美元进行优化
 - 无论容量如何, T_{DRAM} 基本相同
- SRAM
 - 首先针对容量/延迟优化, 再针对容量/美元优化
 - 容量和延迟之间存在不同的折衷可能
- 分层结构弥合了CPU速度和DRAM速度之间的差异
 - $T_{\text{pclk}} \approx T_{\text{DRAM}} \Rightarrow$ 无需分层结构
 - $T_{\text{pclk}} \ll T_{\text{DRAM}} \Rightarrow$ 通过一级或多级SRAM, 在保持成本可控的情况下最小化 T_1

34

34

Intel Pentium 4

- 90nm P4, 3.6 GHz
- L1 D-cache
 - $C_1 = 16\text{KB}$
 - $t_1 = 4$ 周期 整型 / 9 周期 浮点
- L2 D-cache
 - $C_2 = 1024\text{KB}$
 - $t_2 = 18$ 周期 整型 / 18 周期 浮点
- 主存
 - $t_3 = \sim 50\text{ns}$ 或 180 周期
- 注意
 - 最好情况的延迟不再是 1 个时钟周期
 - 最坏情况的访问延迟视情况不同可达到 300+ 时钟周期

35

为什么DRAM慢?

- DRAM的制造是超大规模集成电路技术的前沿, 在容量和成本上与摩尔定律同步扩展, 但不是速度
- 1980 ~ 2004年间的DRAM
 - 64K bit \rightarrow 1024M bit (约55%/年指数级)
 - 250ns \rightarrow 50ns (线性)
- 这是一个非常慎重的选择
- 如果需要, 我们可以“设计”更快的DRAM
- 内存容量需要随着CPU速度线性增长, 以保持系统平衡——Amdahl的另一定律
- DRAM/处理器速度的差异通过内存分层结构进行协调(L1, L2, L3,)
- L2在1990年代开始普及
- L3在2000年初开始普及

36

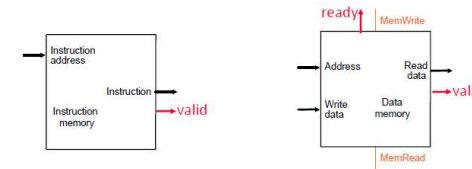
36

Cache 基础和操作

37

高速缓存(Cache)

- 通常，任何可以通过“记忆”频繁操作的结果，以避免从头重复执行长延迟操作的结构，都可以叫做cache，比如web cache
- 最常见的是，能够自动管理的基于SRAM的分层存储
 - 将DRAM存储中被最频繁访问的内容记忆在SRAM中以避免重复出现DRAM的访问延迟



38

cache基础

■ Block (line): cache中的存储单元

- 内存存在逻辑上按照cache block划分并映射到cache的相应位置

■ 当数据被引用

- 命中HIT: 如果在cache中, 使用被缓存的数据, 不再访存
- 缺失MISS: 如果不在cache中, 将相应的block调入cache
 - 可能不得不将某些别的block踢出cache

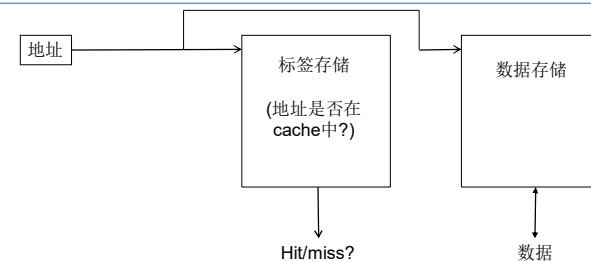
■ 一些重要的cache设计决策 (C<<M)

- 放置: 在哪儿以及如何能在cache中放置/寻找一个block?
- 替换: cache中哪些数据应该被移除?
- 管理的粒度: 大的, 小的还是统一的block?
- 写策略: 写cache的时候应该怎么做?
- 指令/数据: 应该分别对待吗?

39

39

Cache的抽象和指标



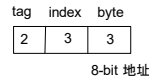
- Cache命中率: $\text{Cache hit rate} = (\# \text{命中}) / (\# \text{命中} + \# \text{缺失}) = (\# \text{命中}) / (\# \text{访问})$
- 平均内存访问时间: Average memory access time (AMAT)
 $= (\text{命中率} * \text{命中延迟}) + (\text{缺失率} * \text{缺失延迟})$
减小AMAT对性能有什么影响?

40

40

Cache的块和寻址

- 内存逻辑上按照cache块划分
- 每一块映射到cache中的某个位置, 由地址中的索引(index) 位决定
 - 用来索引标签和数据存储
- Cache的访问: 基于地址中的索引位索引到标签和数据存储, 检查标签存储中的有效位, 比较地址中的标签位和标签存储中存储的标签
- 如果一个块在cache中 (cache 命中), 标签存储中就应该有相应块的标签

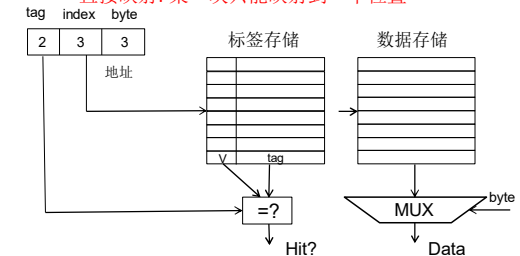


41

41

直接映射 Cache: 放置和访问

- 假设内存是字节寻址:
256 字节, 8-字节的块 → 32 块
- 假设cache: 64 字节, 8 块
 - 直接映射: 某一块只能映射到一个位置



- 有相同索引的地址会争用同一个位置
- 产生冲突缺失

42

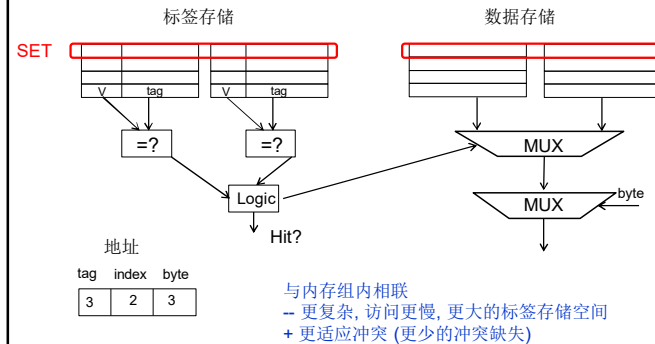
42

直接映射 Cache

- 直接映射 cache: 当内存中的两个不同块映射到cache中相同的索引上时, 这两块不能同时出现在cache中
 - 一个索引 → 一个表项
- 当多个映射到相同索引上的块交替被访问时可能导致0%的命中率
 - 假设地址A和B有相同的索引位和不同的标签位
 - A, B, A, B, A, B, A, B, ... → cache索引冲突
 - 所有的访问都发生冲突缺失(conflict miss)

组(set)相联

- 直接映射的cache中, 地址0和8总是产生冲突(刚才的例子中)
- 采用两组各4块代替1组8块



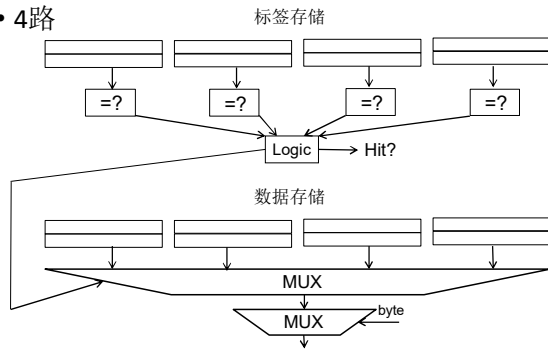
44

44

43

更高的相联度

• 4路



-- 更多的标签比较器，更大的数据多路选择器，更大的标签存储
+ 冲突缺失的可能性更低

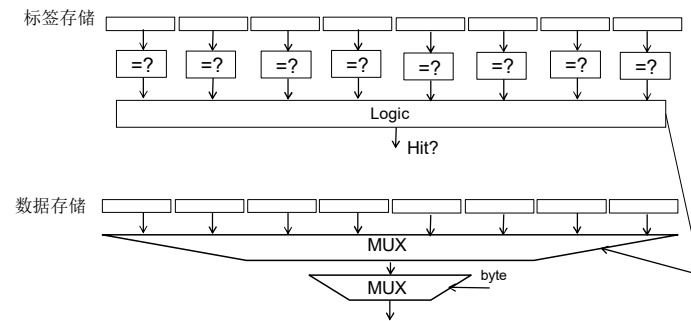
45

45

全相联

• 全相联cache

- 某一块可以放在cache的任何位置



46

46

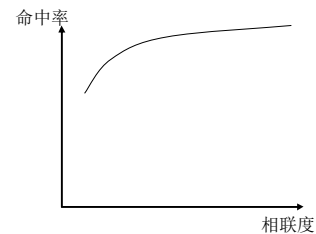
相联 (折衷)

• 多少块可以映射到一个相同的索引(或组)?

• 更高的相联度

- ++ 更高的命中率
- 更长的cache访问时间(命中延迟和数据访问延迟)
- 更昂贵的硬件(更多的比较器)

• 边际收益递减



47

47