本节为复习课。

1 存储层次（Memory hierarchy）：register cache memory

Register, cache (L1 and L2), memory，从左到右速度减慢、容量增大、相同存储空间价格递减。还可以加上硬盘。

左边一层可以看成右边一层的 cache。

2 AMAT

AMAT 全称为 average memory access time，计算公式十分经典：
$\mathrm{hit\_time} + \mathrm{miss\_rate} \times \mathrm{miss\_penalty}$。

分析内存访问开销时，可以分别从这三个角度来分析。

3 4C misses

Compulsory, Capacity, Conflict, Coherence.

减少 compulsory miss：预热

减少 capacity miss：增大 cache size

减少 conflict miss：增大关联数

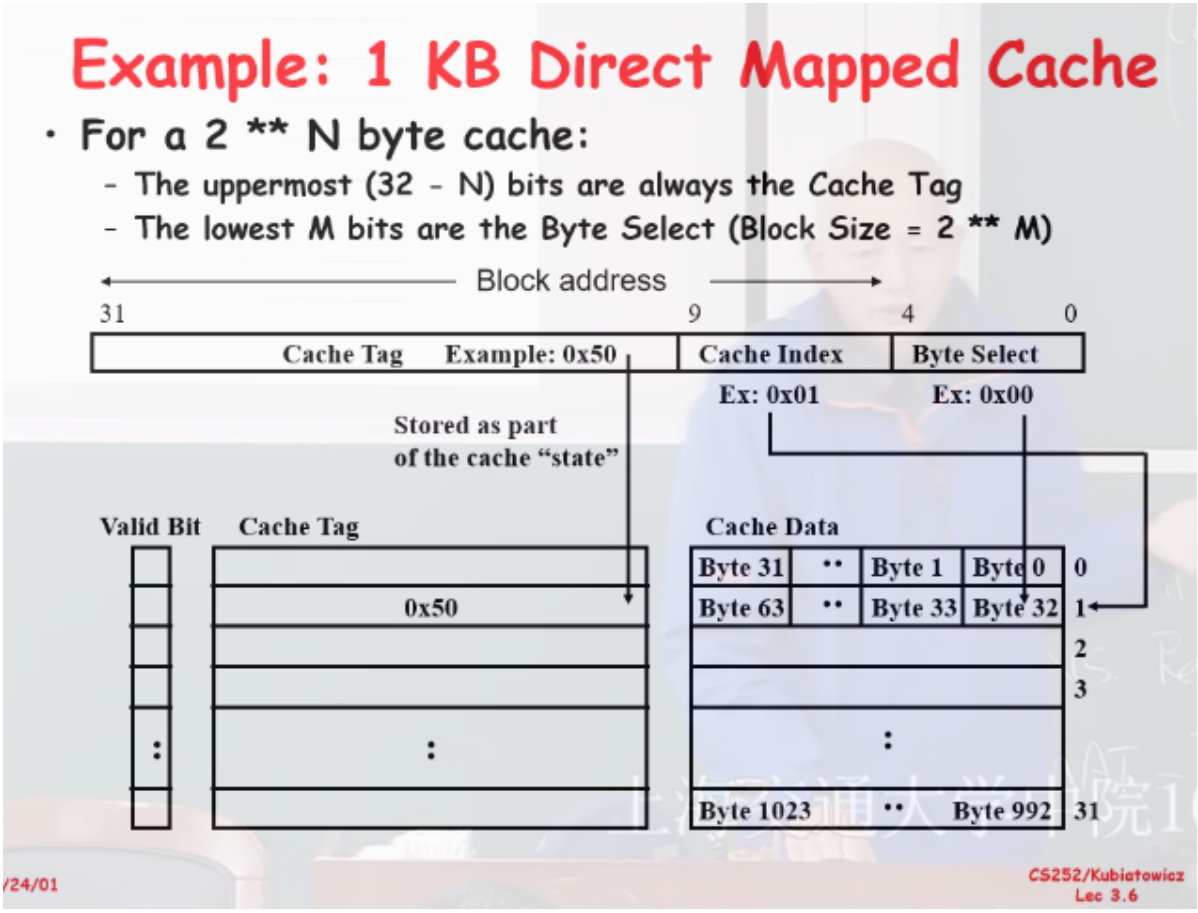
内存性能提升比 CPU 性能提升要慢很多，一定要设法减少内存上的开销。

4 三种常见 cache 类型

4.1 直接映射

将 address 分为三部分：tag、index 和 offset。查询 / 加入时用 index 直接映射到一个位置，查询时检查 valid 和 tag 的匹配性，加入时直接顶掉原来的数据。offset 用于 cache 块内的索引。

它在三种方式中实现最为简单。



4.2 组关联

每个 index 可以对应多个组，每个组中存一个块。查询时检查同一 index 对应的多个组有没有匹配的信息，加入时可能涉及替换，可以采用 LRU (Least Recently Used) 策略。

优点是减少 conflict miss，缺点有：电路比较复杂，比较器多；multiplexer 多，增加延迟；miss 与否、数据在哪一组内确定后，数据才能被传输出去。

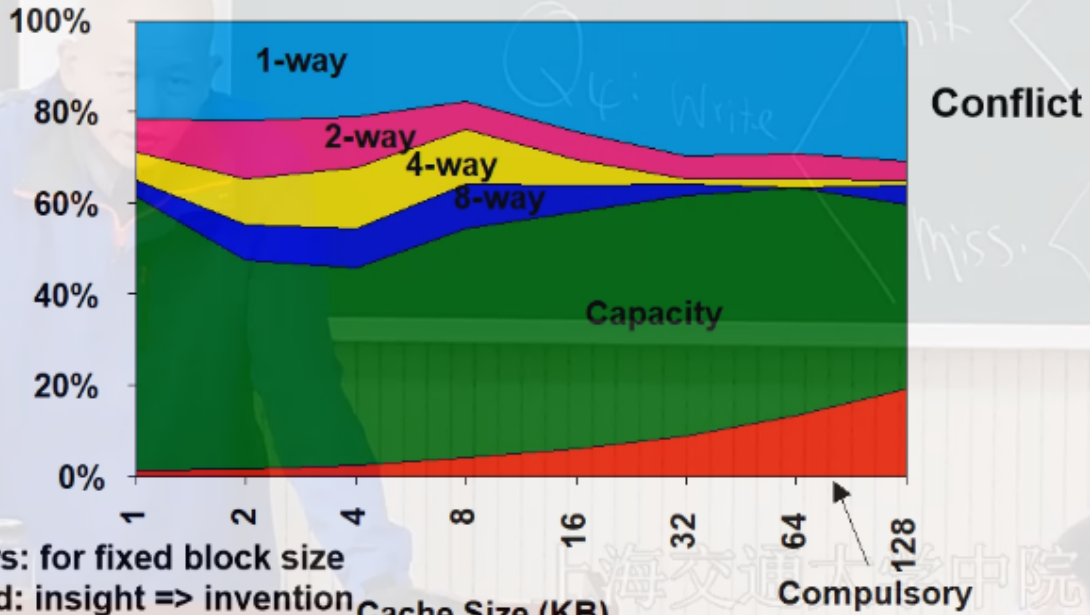
4.3 全关联

组关联的极端情况，没有 index。会将组关联的缺点放大。

Review: Four Questions for Memory Hierarchy Designers

- Q1: Where can a block be placed in the upper level? *(Block placement)*
  - Fully Associative, Set Associative, Direct Mapped
- Q2: How is a block found if it is in the upper level? *(Block identification)*
  - Tag/Block
- Q3: Which block should be replaced on a miss? *(Block replacement)*
  - Random, LRU
- Q4: What happens on a write? *(Write strategy)*
  - Write Back or Write Through (with Write Buffer)

3Cs Relative Miss Rate



Flaws: for fixed block size
Good: insight => invention

5 降低 miss rate 的方法

5.1 Victim cache

Miss 的数据放进另外一个 cache 当中。设想这样一个场景：两套数据映射到同样的位置，然后它们还要被交替访问。这时 victim cache 的出现就可以显著降低 miss rate，否则换一套数据就要 miss 一次。

5.2 Pseudo-Associativity

把 cache 拆成两部分（可以是一部分直接映射，一部分组关联），直接映射 miss 了，去组关联那里看看在不在。如果在，可以看作 pseudo-hit。

缺点：hit time 和 pseudo hit time 不一致，给 CPU 的流水线设计带来挑战。