



BROWN

# Computer System Design Lecture 18: Improving Memory Performance

Prof. R. Iris Bahar  
EN164  
March 9, 2007

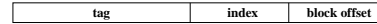
Reading: Appendix C, Sections C.3



BROWN

## Locating a Block

- Address portions



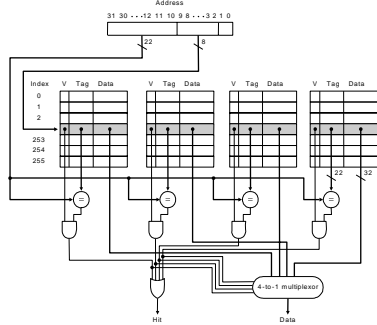
- Index selects the set.
  - Tag chooses the the block by comparison.
  - Block offset is the address of the data within the block.
- The costs of an associative cache
    - comparators and multiplexers
    - time for comparison and selection

EN164  
Lecture 17-2



BROWN

## 4-Way Set-Associative Cache



EN164  
Lecture 17-3



BROWN

## 4-way Cache Efficiency

- Cache memory size
  - 256 sets X 4 words/set X 32 b = 32 kb
- Tag memory size
  - 256 X 4 X 22 b = 22 kb
- Valid information
  - 1024 · 1 b = 1 kb
- Efficiency
  - 32/55 = 58.2 %
- What else do we need to account for here?*

EN164  
Lecture 17-4



BROWN

## Replacement Strategy

- For a direct mapped cache, there is no choice as to which block to replace on a miss.
  - What about an associative caches?*
- Several different replacement strategies are possible:
  - Random
  - First-in-first-out
    - oldest block is replaced
  - LRU (Least Recently Used)
    - the block having been unused for the longest time is replaced
- What are the advantages/disadvantages of each of these?*

EN164  
Lecture 17-5



BROWN

## Other Issues for Set-Associative Caches

- Set-associative caches have a significant HW overhead
- Tag lookup is more complicated
- The CPU would like the data as soon as possible
  - For direct mapped caches, there is only one choice of which data to send
  - What about a set-associative cache?
- Can you send the data to the CPU before the tag has been checked?*
- What about power concerns?*

EN164  
Lecture 17-8

BROWN

## Types of Cache Misses

- **Compulsory misses:** happens the first time a memory word is accessed
  - the misses for an infinite cache
- **Capacity misses:** happens because the program touched many other words before re-touching the same word
  - the misses for a fully-associative cache
- **Conflict misses:** happens because two words map to the same location in the cache
  - the misses generated while moving from a fully-associative to a direct-mapped cache
- *Can a fully-associative cache have more misses than a direct-mapped cache of the same size?*

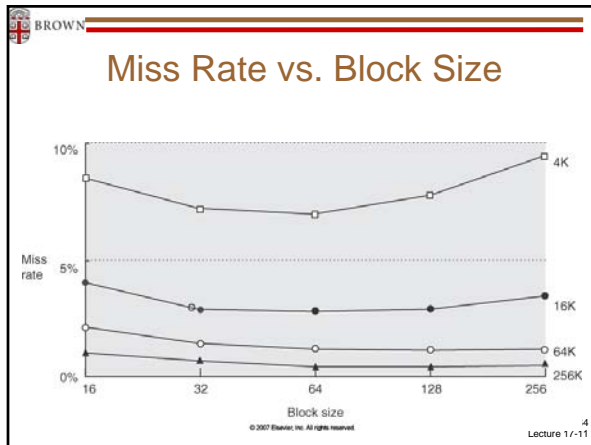
EN164  
Lecture 17-9

BROWN

## Reducing Miss Rate

- **Large block size:**
  - Reduces compulsory misses, reduces miss penalty in case of spatial locality
  - Increases traffic between different levels, wastes space, and can increase conflict misses
- **Large caches:**
  - Reduces capacity/conflict misses
  - Access time penalty
- **High associativity:**
  - Reduces conflict misses (Rule of thumb: 2-way cache of capacity N/2 has the same miss rate as direct mapped cache of capacity N)
  - Access time penalty
- **Way prediction:**
  - by predicting the way, access time is effectively like a direct-mapped cache
  - can also reduce power consumption

EN164  
Lecture 17-10



BROWN

## What Influences Cache Misses?

	Compulsory	Capacity	Conflict
Increasing cache capacity			
Increasing number of sets			
Increasing block size			
Increasing associativity			

EN164  
Lecture 17-13