Cache Design

Lecture Objectives:

1) Define set associative cache and fully associative cache.
2) Compare and contrast the performance of set associative caches, direct mapped caches, and fully associative caches.
3) Explain the operation of the LRU replacement scheme.
4) Explain the concept of a multi-level cache.
5) Explain the three C model for cache.

Least Recently Used

Definitions

- Fully associative cache
  - A cache structure in which a block can be placed in any location in the cache.

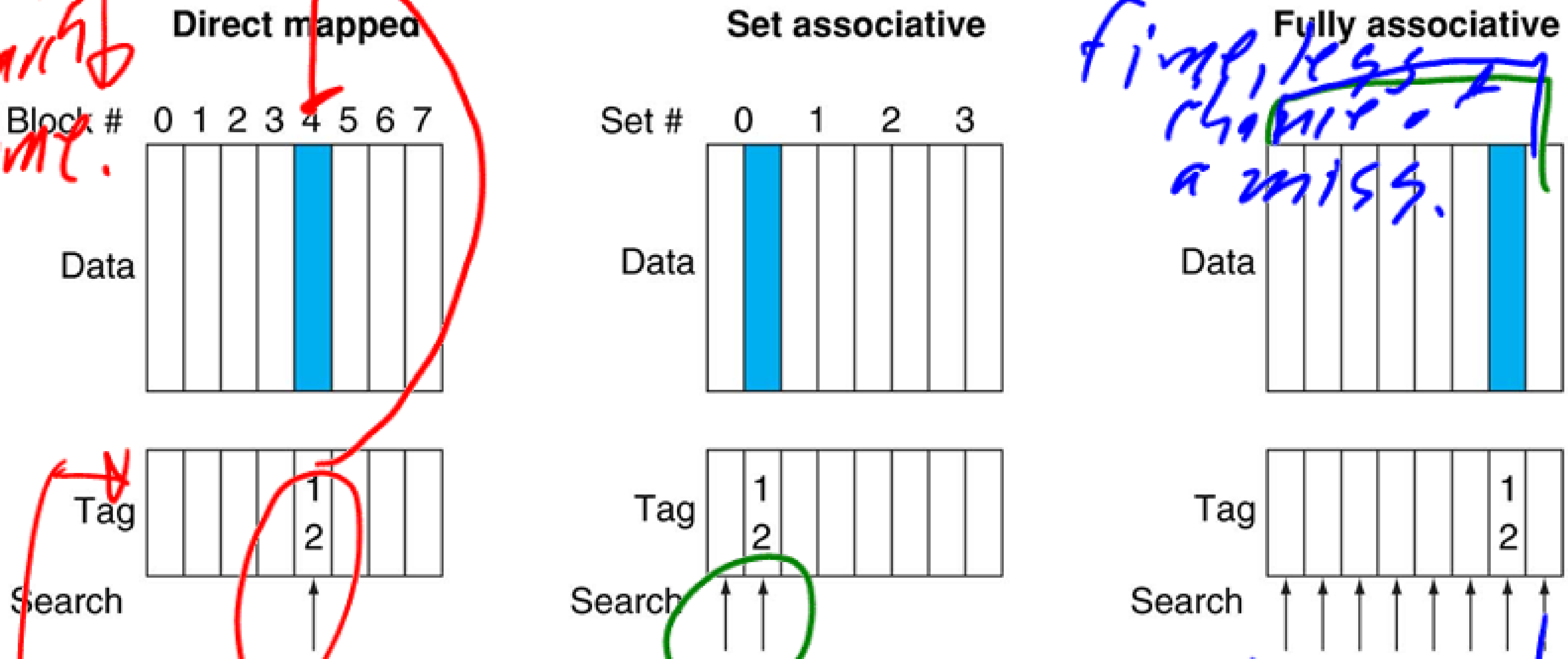
Any Location for a page of data

- Set-associative cache
  - A cache that has a fixed number of locations (at least two) where a given block can be placed

Comparison of cache designs

Less search time.

More search time, less chance of a miss.



Address 12

Direct mapping of where the data is at.

Address 12

2 places.

Can be anywhere

Problem

- Working with your partner, solve the following problem
  - The following block accesses occur in memory over three different cache structures. Calculate the miss rate for each cache structure
    - Structure 1: A direct mapped cache of 4 one word blocks
    - Structure 2: A 2 way set associative cache of 4 words
    - Structure 3: A fully associative cache of 4 words.
- Access trace: 1, 9, 1, 7, 9, 7, 9, 1, 7, 3

Address 0 if miss
1 if hit

Direct Mapped Example

Address of memory Access	Hit or miss	0	1	2	3
1%4 1	0		~~1~~		
9%4 9	0		~~9~~		
1%4 1	0		1		
7%4 7	0		~~1~~		7
7%4 9	0		~~9~~		
7	1				
9	1				
1	0		1		
7	1				
3	0				3

Yuck!

$$\text{Hit rate} : \frac{\text{Hits}}{\text{total Accesses}} \Rightarrow \frac{3}{10} \Rightarrow 30\%$$

2 Way Set Associative Example

Address of memory Access	Hit or miss	Set 0	Set 0	Set 1	Set 1
1%2 1	M			1	
9%2 9	M			1	9
1%2 1	H				
%2 → 7	M			1	7
9%2 → 9	m			9	7
→ 7	H				
→ 9	H				
1%2 1	m			9	1
7%2 7	m			7	1
3%2 3	m			7	3

Yuck!

$$HR = \frac{\text{Hits}}{\text{total Accesses}} \Rightarrow \frac{3}{10}$$

Any data Anywhere

Fully Associative Example

Address of memory Access	Hit or miss	Block 0	Block 1	Block 2	Block 3
1	M	1			
9	M	1	9		
1	H				
7	M	1	9	7	
9	H				
7	H				
9	H				
1	H				
7	H				
3	M	1	9	7	3

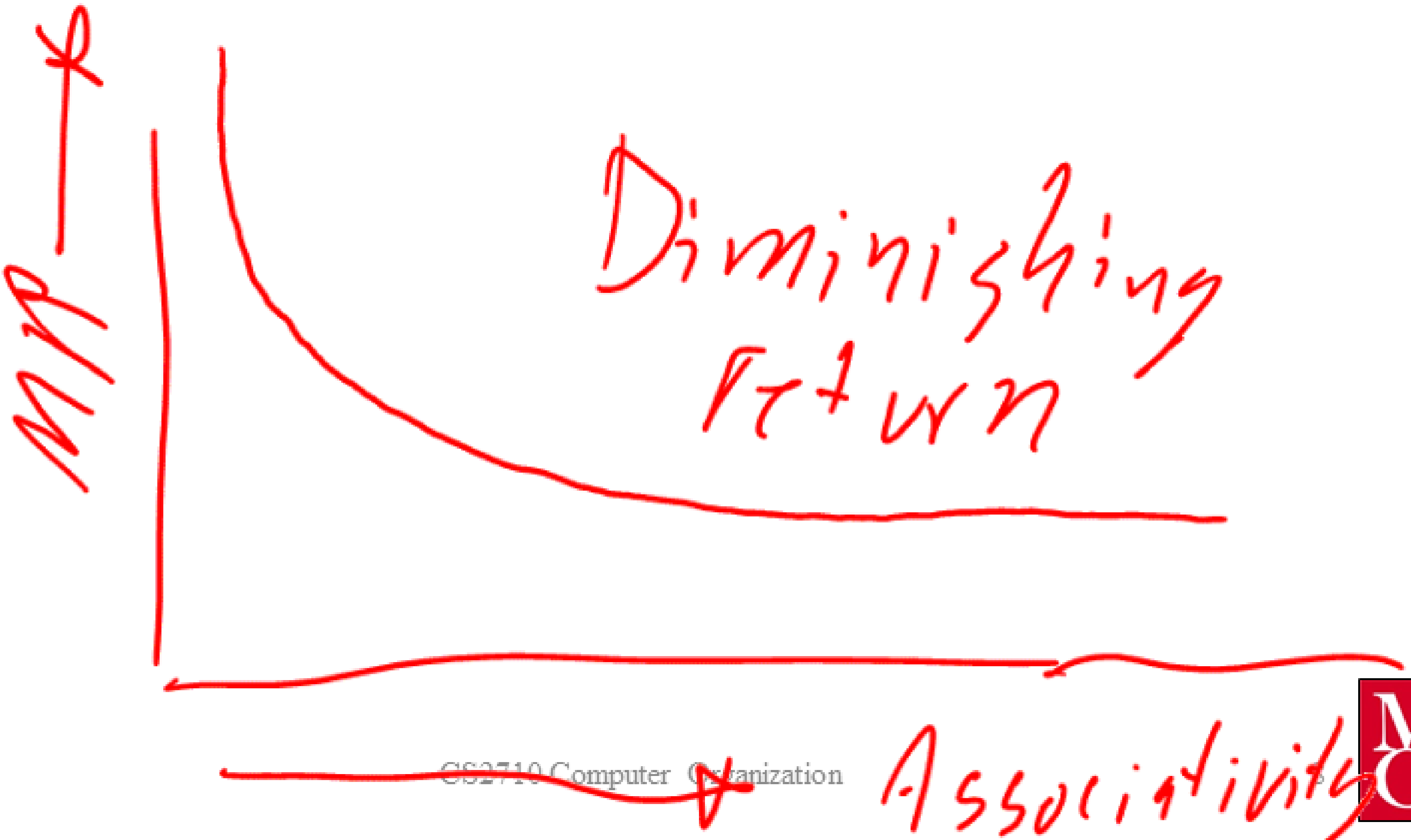
Better

$$HR = \frac{Hits}{Total} \Rightarrow \frac{6}{10} \Rightarrow 60\%$$

Data Miss Rates for various associativities

Direct mapped cache

Associativity	Data miss rate
1	10.3%
2	8.6%
4	8.3%
8	8.1%



How to determine which cache item to replace?

- Direct Mapped?

No choice

- Set Associative

  - Prefer non valid entry if there is one. Otherwise, choose among entries in the set

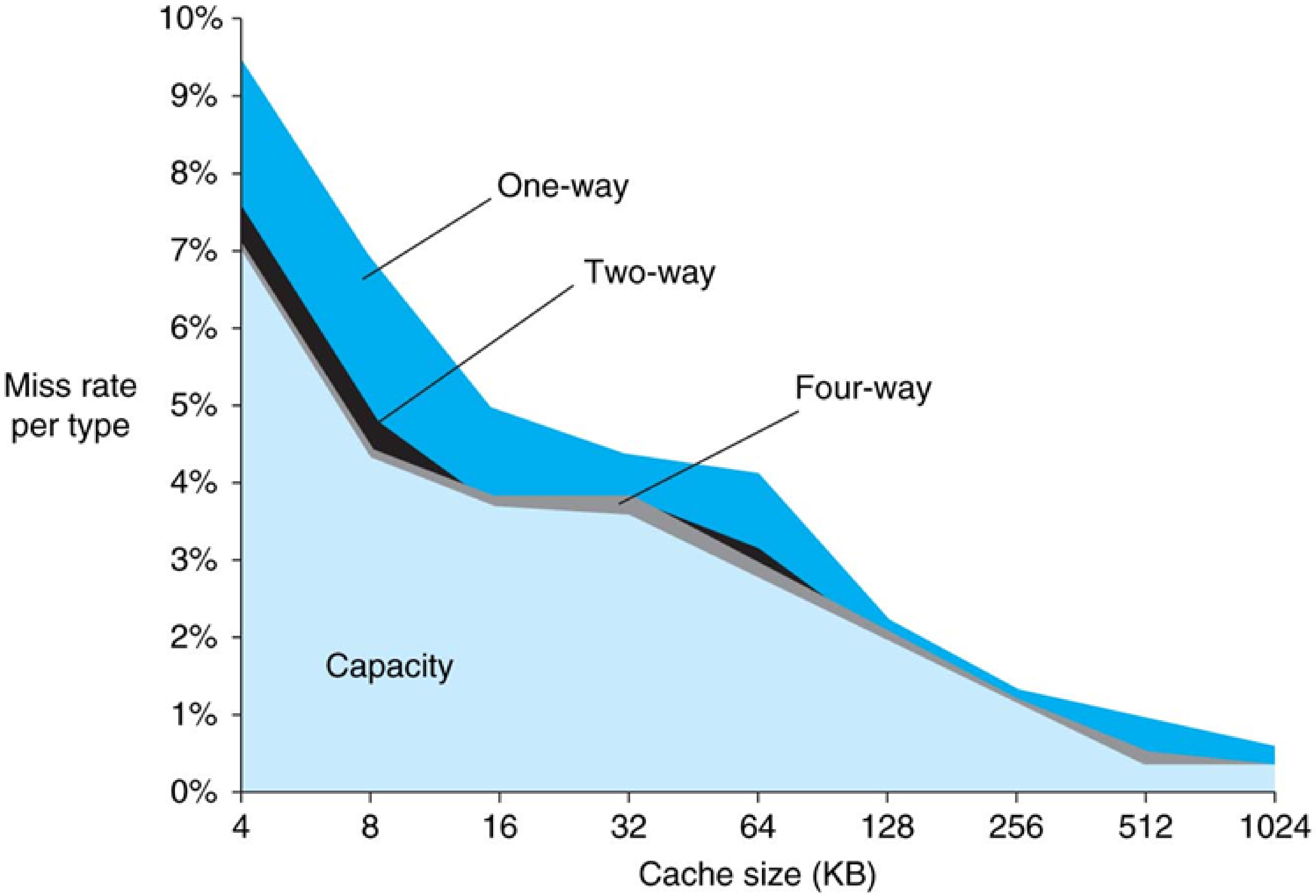
LRU (Least recently used)
  - choose the unused one for the longest time.

- Random
  - Approximately same performance as LRU at high associativity.

The C Cs model

- A Cache model in which all cache misses are classified into one of three categories
  - Compulsory Misses
  - Capacity Misses
  - Conflict Miss

Source of misses



- Compulsory misses not visible (0.006%)

Basic Design challenges

Design Change	Effect on miss rate	Possible negative performance impact
Increase the cache size	Decreases capacity misses	May increase access time
Increase Associativity	Decreases miss rate due to conflict misses	May increase access time
Increase Block Size	Decreases miss rate for a wide range of block sizes due to spatial locality	Increases miss penalty. Very large blocks could increase miss rate.