
Slide Set #16: Exploiting Memory Hierarchy

1

Memory, Cost, and Performance

- Ideal World: we want a memory that is
 - Fast,
 - Big, &
 - Cheap!
- Real World:
 - SRAM access times are .5 – 5ns at cost of \$4000 to \$10,000 per GB.
 - DRAM access times are 50-70ns at cost of \$100 to \$200 per GB. (2004)
 - Disk access times are 5 to 20 million ns at cost of \$.50 to \$2 per GB.
- Solution?

3

ADMIN

- Chapter 7 Reading
 - 7.1-7.3

2

Locality

- A principle that makes caching work
- If an item is referenced,
 1. it will tend to be referenced again soon
why?
 2. nearby items will tend to be referenced soon.
why?

4

Caching Basics

- Definitions

1. Minimum unit of data: "block" or "cache line"

For now assume, block is 1 byte

2. Data requested is in the cache:
3. Data requested is not in the cache:

- Cache has a given number of blocks (N)

- Challenge: How to locate an item in the cache?

- Simplest way:

Cache index = (Data address) mod N

e.g., N = 10, Address = 1024, Index =

e.g., N = 16, Address = 33, Index =

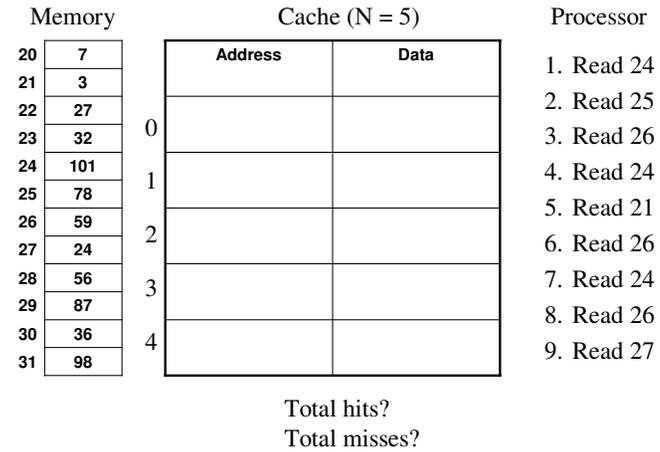
- Implications

For a given data address, there is _____ possible cache index

But for a given cache index there are _____ possible data items that could go there

5

Example – (Simplified) Direct Mapped Cache



6

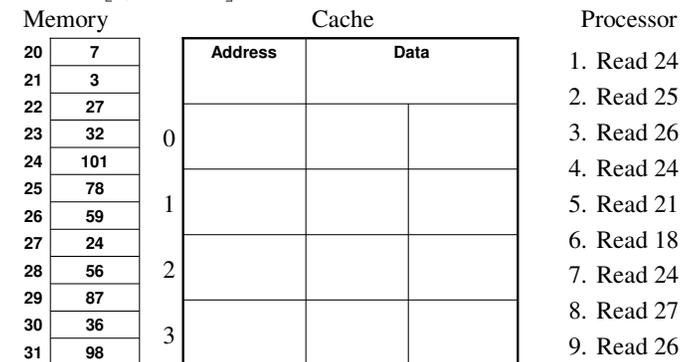
Improving our basic cache

- Why did we miss? How can we fix it?

7

Approach #1 – Increase Block Size

$$\text{Index} = \left\lfloor \frac{\text{ByteAddress}}{\text{BytesPerBlock}} \right\rfloor \bmod N$$



8

Approach #2 – Add Associativity

$$\text{Index} = \left\lfloor \frac{\text{ByteAddress}}{\text{BytesPerBlock}} \right\rfloor \bmod \frac{N}{\text{Associativity}}$$

Memory

20	7
21	3
22	27
23	32
24	101
25	78
26	59
27	24
28	56
29	87
30	36
31	98

Cache

	Address	Data
0		
1		

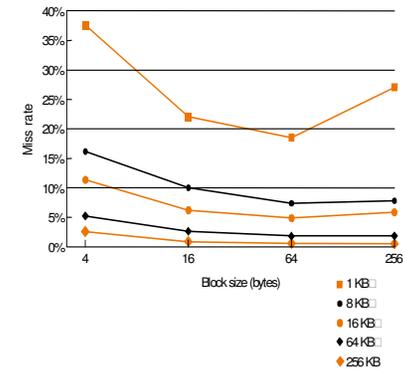
Processor

1. Read 24
2. Read 25
3. Read 26
4. Read 24
5. Read 21
6. Read 18
7. Read 24
8. Read 27
9. Read 26

9

Performance Impact – Part 1

- To be fair, want to compare cache organizations with same data size
 - E.g., increasing block size must decrease number blocks (N)
- Overall, increasing block size tends to decrease miss rate:



10

Performance Impact – Part 2

- Increasing block size...
 - May help by exploiting _____ locality
 - But, may hurt by increasing _____ (due to smaller _____)
 - Lesson – want block size > 1, but not too large
- Increasing associativity
 - Overall N stays the same, but smaller number of sets
 - May help by exploiting _____ locality (due to fewer _____)
 - May hurt because cache gets slower
 - Do we want associativity?

11

How to handle a miss?

- Things we need to do:
 1. _____ the CPU until miss completes
 2. _____ old data from the cache
Which data?
 3. _____ the needed data from memory
Pay the _____
How long does this take?
 4. _____ the CPU

What about a write miss?

12