Cache Basics

Lecture Objectives:

1) Define temporal and spatial locality.
2) Define hit rate and miss rate.
3) Define the term cache
4) Explain how a direct mapped cache determines the location in the cache.
5) Given a memory address and cache size information, perform the calculations to convert an address to a cache block number.

Exceptions and Interrupts

- Unexpected" events requiring change in flow of control
  - Different ISAs use the terms differently
- Exception
  - Arises within the CPU
    - e.g., undefined opcode, overflow, syscall, …
- Interrupt
  - From an external I/O controller
- Dealing with them without sacrificing performance is hard

Handling Exceptions

- In MIPS, exceptions managed by a System Control Coprocessor (CP0)
- Save PC of offending (or interrupted) instruction
  - In MIPS: Exception Program Counter (EPC)
- Save indication of the problem
  - In MIPS: Cause register
  - We'll assume 1-bit
    - 0 for undefined opcode, 1 for overflow
- Jump to handler at 8000 00180

An Alternate Mechanism

- Vectored Interrupts
  - Handler address determined by the cause
- Example:
  - Undefined opcode: C000 0000
  - Overflow: C000 0020
  - …: C000 0040
- Instructions either
  - Deal with the interrupt, or
  - Jump to real handler

- An Exception Demo and example…

Activity with your neighbor

- I will post two grocery lists
- Determine how to quantify how many aisles you will visit to obtain the items in order
- Determine how many times you need to change aisles in order to find the items
- Determine which grocery list will be faster to purchase

In the grocery store, which list will be faster to purchase?

	List 1	List 2
1.	Tomatoes	Cherry Jelly
2.	Corn	Corn
3.	Beans	Ice Cream
4.	Carrots	Rib Eye Steaks
5.	Bread	Beans
6.	Peanut Butter	Mozzarella Cheese
7.	Cherry Jelly	Carrots
8.	Fresh Chicken Breasts	Bread
9.	Rib Eye steaks	Fresh Chicken Breasts
10.	Spaghetti Sauce	Spaghetti Sauce
11.	Pasta Noodles	Milk
12.	Milk	Tomatoes
13.	Yogurt	Peanut Butter
14.	Mozzarella Cheese	Yogurt
15.	Ice Cream	Pasta Noodles
16.	Cherry Lemon Sherbet	Cherry Lemon Sherbet

Fresh

Meat

Dairy

Ordered

Freezer.

MSOE

Lets assume you could obtain anything from the same aisle in 0 time, but to change aisles took 10 units of time.

1. Tomatoes
2. Corn
3. Beans
4. Carrots
5. Bread
6. Peanut Butter
7. Cherry Jelly
8. Fresh Chicken Breasts
9. Rib Eye steaks
10. Spaghetti Sauce
11. Pasta Noodles
12. Milk
13. Yogurt
14. Mozzarella Cheese
15. Ice Cream
16. Cherry Lemon Sherbet

1. Cherry Jelly
2. Corn
3. Ice Cream
4. Rib Eye Steaks
5. Beans
6. Mozzarella Cheese
7. Carrots
8. Bread
9. Fresh Chicken Breasts
10. Spaghetti Sauce
11. Milk
12. Tomatoes
13. Peanut Butter
14. Yogurt
15. Pasta Noodles
16. Cherry Lemon Sherbet

Lets assume you could obtain anything from the same aisle in 0 time, but to change aisles took 10 units of time.

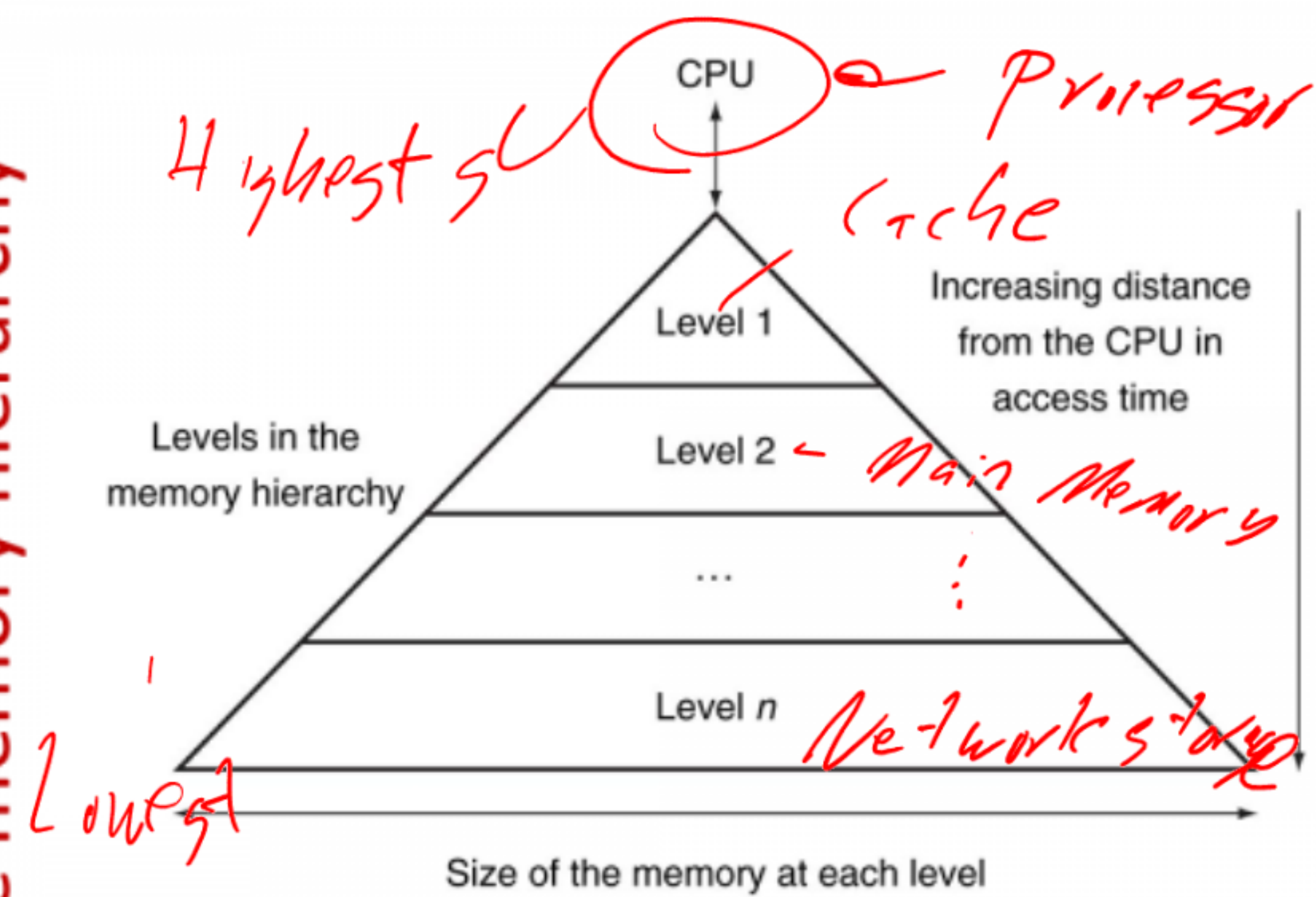
Definitions

- Temporal Locality = Programming
  - The principle stating that if a data location is referenced then it will again be referenced soon. "loose"
- Spatial Locality
  - The locality principle stating that if a data location is referenced, data locations with nearby addresses will tend to be referenced "soon."

```
for ( x = 0; x<100; x++ )
{
    system.out.println(data[x]);
};
```

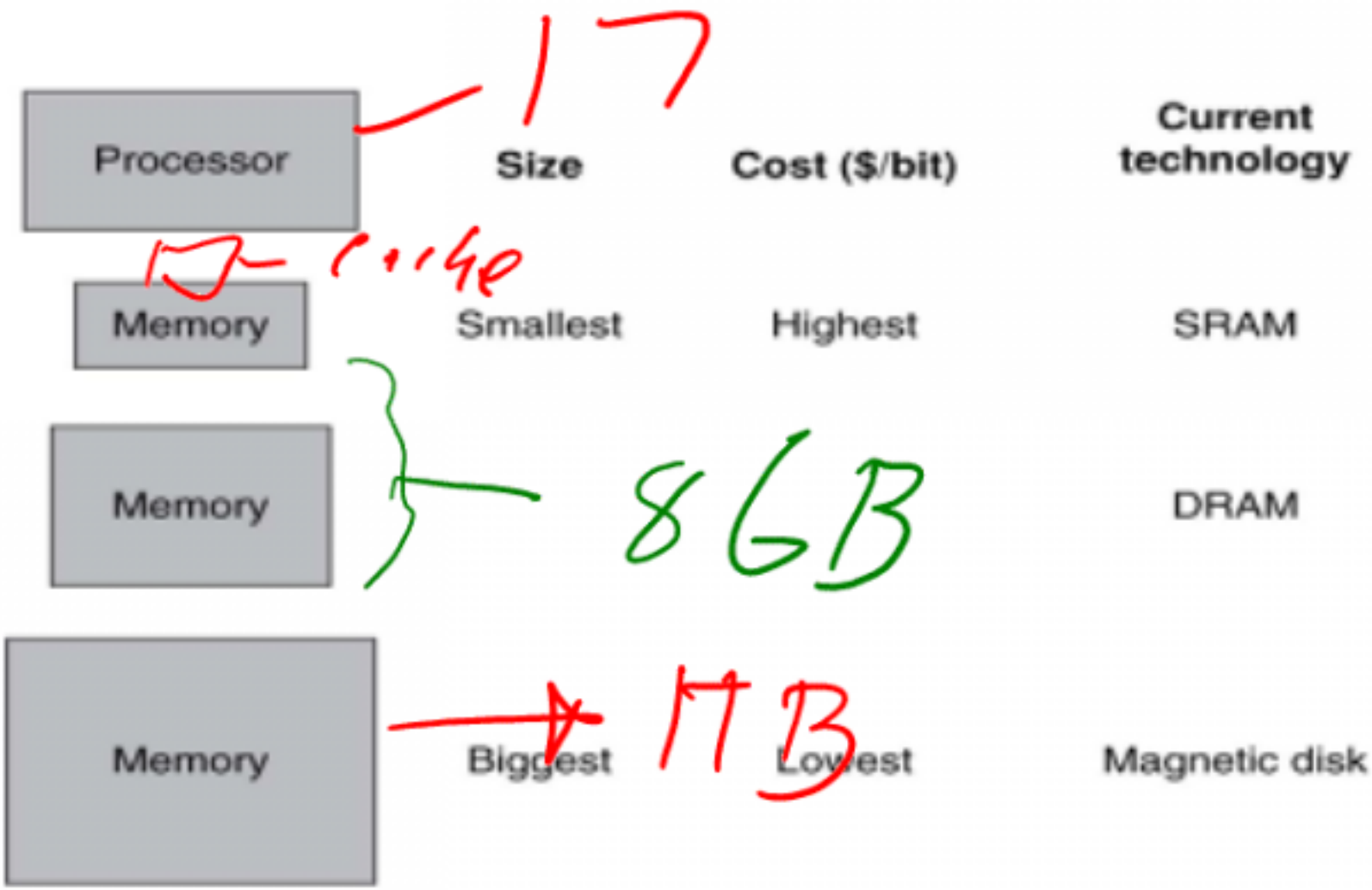

MSOE

The memory hierarchy



SYX SG-125 Gaming PC

Intel Core i7 2600 3.4 GHz,
Genuine Windows® 7
Professional 64 Bit, 1GB
NVIDIA GeForce GTX 550 Ti,
8GB DDR3, 1TB 7200rpm
HDD, Blu-Ray, USB 3.0



Minute Quiz

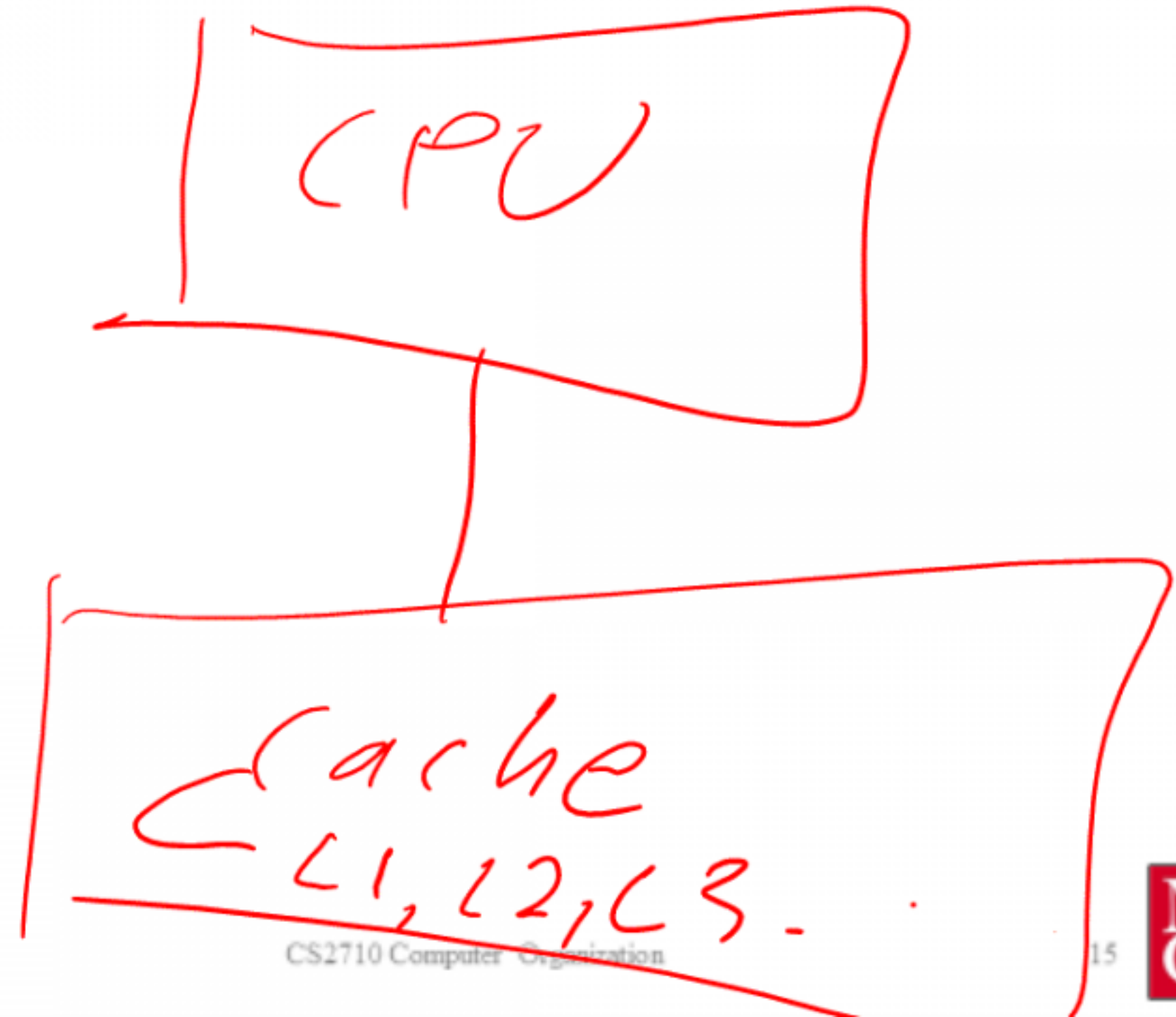
- Which of the following statements is generally false?
  - A. Caches take advantage of temporal locality
  - B. On a read, the value returned depends upon which blocks are in the cache
  - C. Most of the capacity of the memory hierarchy is at the lowest level
  - D. The most expensive portion of the memory hierarchy is at the highest level

Definitions

- Block
  - The minimum unit of information that can be either present or not present in the cache
- Hit Rate
  - The fraction of memory accesses found in the level of the memory hierarchy
- Miss Rate
  - The fraction of memory accesses not found in a level of the memory hierarchy
- Miss penalty
  - The time required to fetch a block into a level of memory hierarchy from the lower level.

- The level of memory closest to the Processor

Cache Memory



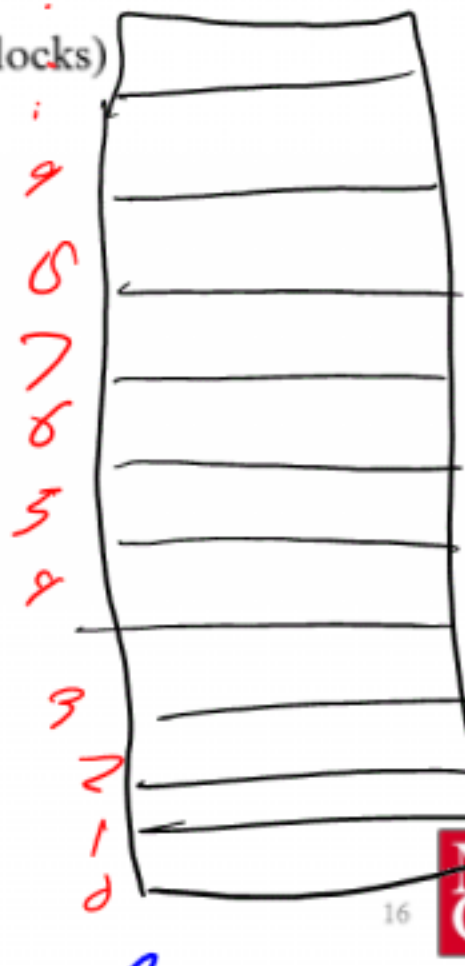
Direct Map

16

(Block Address) mod(Number of Blocks In Cache)

Determining where to look

Block to access	Cache Location (16 blocks)
0x01	location 1
0x05	Location 5
0x15	Also location 5
0x22	
0x37	
0x3F	
0x3D	



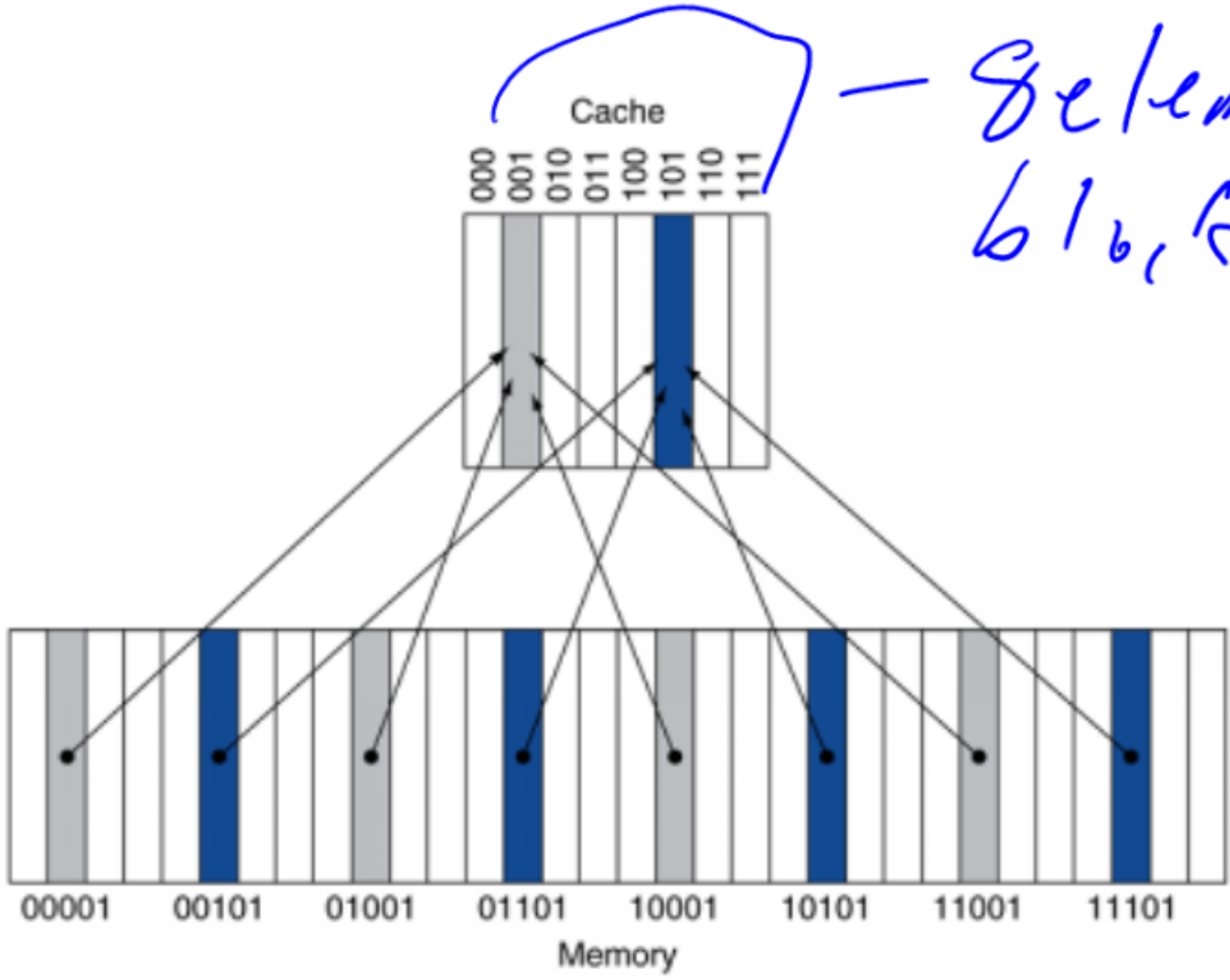
MSOE

cache

Direct relationship between address and data in the cache

Direct Mapped Cache

- #Blocks is a power of 2
- Use low-order address bits



- If we search the cache, how do we know if the data is valid or not?

Valid bits

tells us if the data can be used or not.

- How do we know which specific block in the cache?

Tags and

Tells us which address in specific is in the cache.

MSOE

Example Address stream

Index	V	Tag	Data
000	N		
001	N		
010	N		
011	N		
100	N		
101	N		
110	N		
111	N		

- Address stream
- 10110, 11010, 10000, 10110, 00011, 10010,

Address (showing bit positions)

31 30 · · · 13 12 11 · · · 2 1 0

Byte offset

20

Tag

10

Index

Hit

Data

Index | Valid | Tag | Data

0
1
2
…
…
…
1021
1022
1023

20

32

=

Address Subdivision

Determining Sizes

$$TagFieldSize = AddressBusSize - (n + m + 2)$$

$$CacheSize = 2^{n}\ Blocks$$

$$BlockSize = 2^{m}\ Words = (2^{m+2}\ bytes)$$