

# Computer Organization & Architecture Design

## Memory Hierarchy

### Chapter 7

saeed\_zafar@comsats.edu.pk  
Comsats Institute of Information Technology

1

## Memory Hierarchy

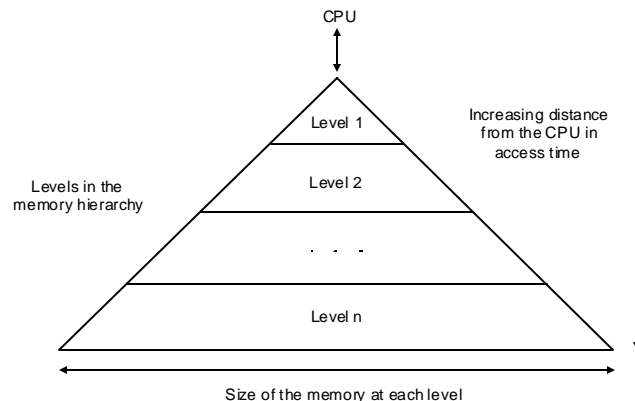
- Principle of *locality*:
- **Temporal locality** (locality in time): If an item is referenced, it will tend to be referenced again soon.
- **Spatial locality** (locality in space): If an item is referenced, items whose addresses are close by will tend to be referenced soon.
- **Method**: The level closer to processor (the fastest) is a subset of any level further away, and all the data is stored at the lowest level (the slowest).

Memory Technology	Typical access time	\$ per Mbyte in 1997
SRAM	5-25 ns	\$100-\$250
DRAM	60-120 ns	\$5-\$10
Magnetic disk	10-20 million ns	\$0.10-\$0.20

DRAM(Dynamic Random Access memory)  
SRAM(Static Random Access memory)

2

## Memory Hierarchy



3

## The Basics of Caches

- The caches are organized on basis of **blocks**, the smallest amount of data which can be copied between two adjacent levels at a time.
- If data requested by the processor is present in some block in the upper level, it is called a **hit**.
- If data is not found in the upper level, the request is called a **miss** and the data is retrieved from the lower level in the hierarchy.
- The fraction of memory accesses found in the upper level is called a **hit ratio**.
- .

4

## Continued...

- The fraction of memory accesses not found in a level of memory hierarchy is called as a **miss rate**.
- The storage, which takes advantage of locality of accesses is called a **cache**.
- **Miss Penalty**: The time required to fetch a block into a level of the memory hierarchy from the lower level, including the time to access the block, transmit it from one level to the other, and insert it in the level that experienced the miss.

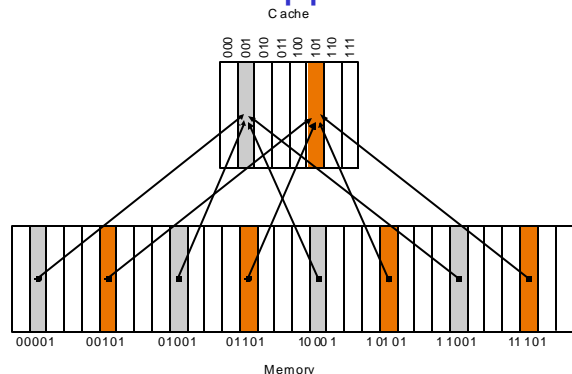
5

## Concept of memory hierarchy



6

## Direct Mapped Cache



**(block address) modulo (No of cache blocks)**

Direct mapped cache with eight entries showing address of memory words b/w 0 and 31 which map the same cache location.

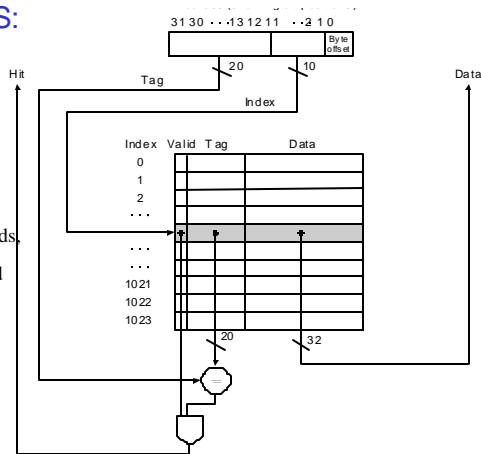
## Continued...

- Each memory location can contain the contents of different memory location.
- How do we know the requested data in the cache corresponds to a requested word?
- Solution is to add tags to the cache
- **Tag**: A field in a table used for a memory hierarchy that contains the address information required to identify whether the associated block in the hierarchy corresponds to a requested word.
- **Valid bit**: A field in the tables of a memory hierarchy that indicates that the associated block in the hierarchy contains valid data.

8

## Direct Mapped Cache

- For MIPS:



\*Cache has 1024 words  
\*Block size of 1-word

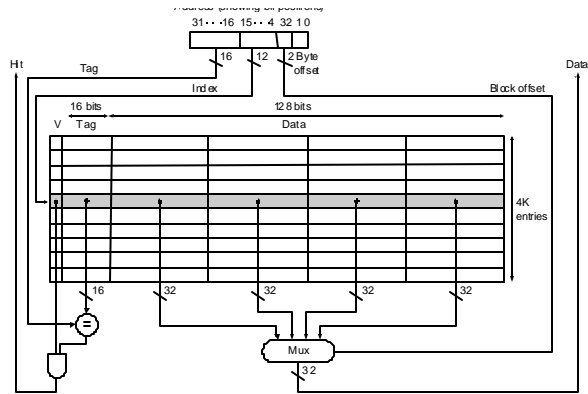
9

## Steps For Instruction Cache Miss

1. Send the PC value(current pc-4) to the memory.
2. Instruct the main memory to perform a read & wait for memory complete access.
3. Write the cache entry, putting data from memory in the data portion, upper bits of address in tag field, and turning the valid bit on.
4. Restart the instruction execution, this time finding it in the Cache.

10

## Direct Mapped Cache



- Taking advantage of spatial locality:

11

## Reducing Cache Miss

- A block can go in exactly one place in the cache called as **direct mapped** cache.
- Block can be placed in any location in cache called as **fully associative** cache.
- The middle range (at least two locations) of the designs b/w direct mapped and fully associative called as **Set associative** cache.

12

## Example of cache

- 8-block cache can be configured as a
  - Direct mapped cache
  - Two-way set associative cache
  - Four-way set associative
  - Fully associative cache

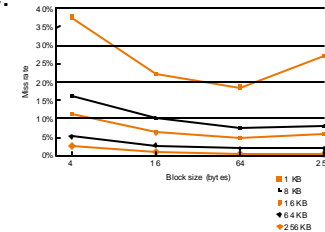
### Problem

There are three small caches, each consisting of four one-word blocks. One is Fully associative cache, 2<sup>nd</sup> is Two-way set associative, & the third is direct mapped. Find the no of the misses & hits for each cache organization given the following sequence of the block address: 0,8,0,6,8

13

## Performance

- Increasing the block size tends to decrease miss rate:

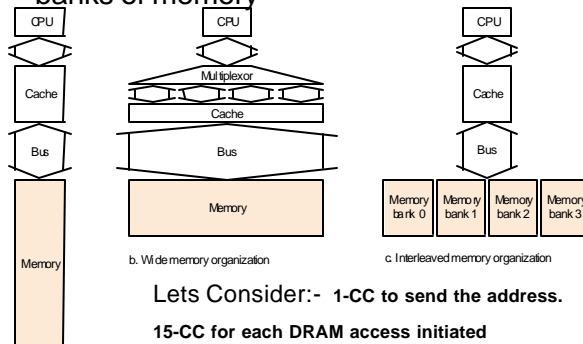


Program	Block size in words	Instruction miss rate	Data miss rate	Effective combined miss rate
gcc	1	6.1%	2.1%	5.4%
	4	2.0%	1.7%	1.9%
spice	1	1.2%	1.3%	1.2%
	4	0.3%	0.6%	0.4%

14

## Hardware Issues

- Make reading multiple words easier by using banks of memory



a. One-word-wide memory organization

b. Wide memory organization

c. Interleaved memory organization

Lets Consider:- 1-CC to send the address.

15-CC for each DRAM access initiated

1-CC to send a word of data.

Find Miss penalty & No of bytes transferred per clk cycle for single miss?

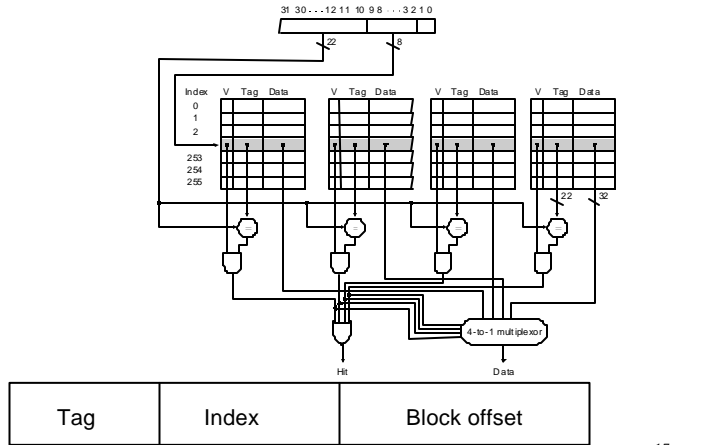
15

## Measuring & Improving cache Performance

- $CPU\ time = (CPU\ Exec\ CC + Mem-stall\ CC) \times CC\ time$
- $Mem-stall\ CC = Read-stall\ cycles + Write-stall\ cycles$
- $Read-stall\ cycles = \frac{Read}{Program} \times Read\ miss\ rate \times Read\ miss\ penalty$
- $Write-stall\ cycles = \left[ \frac{writes}{Program} \times Write\ miss\ rate \times Write\ miss\ penalty \right] + Write\ Buffer\ stall$
- $Mem-stall\ CC = Mem\ accesses/program \times Miss\ rate \times Miss\ penalty$
- $Mem-stall\ CC = Instructions/program \times Misses/Instruction \times Miss\ penalty$

16

## Locating a Block in the cache



17

## Reducing Miss penalty using Multilevel caches

- Caches implemented with same die CPU.
- L1 and L2 caches.
- Miss penalty is reduced due to presence of secondary level cache.

18

## Virtual memory

- Virtual memory block is called **page**.
- V.M miss is called **page fault**.
- **Memory Mapping or address translation**: CPU generates a virtual address which is translated in physical address with the help of sw &Hw.

19