Answers of Sheet 5 – Cache Memory

a. True or False:

1) Cache memory on Hard Disk is considered as internal memory of computer. TRUE
2) External memory is sometimes faster than internal memory's as in the case of the SSD Hard Disk. FALSE
3) One block can contain more than a word. TRUE
4) Hard Disk and Flash Memory are addressed word by word. FALSE
5) Word size is used as a metric to measure the performance of any memory. FALSE
6) Memory cycle time must be less than memory access time. FALSE
7) Random Access Memory (RAM) apply the concept of memory cycle time. TRUE
8) The time required for the memory to "recover" before next access is Recovery time. TRUE
9) The memory cycle time is concerned with the processor. FALSE
10) Sequential access method starts at the beginning of memory and read in order till it finds the required location. TRUE
11) Cache memory utilizes the Direct access method when dealing with CPU. FALSE
12) Both RAM and Cache memory are independent on the previous location of data. TRUE
13) A comparison with contents of a portion of the stored data is made to achieve sequential access. TRUE
14) According to memory hierarchy, magnetic tapes lie at the bottom of the pyramid with the lowest speed in dealing with data among other memory devices. TRUE
15) Static RAM is slower than Cache. FALSE
16) In a short period of time, the CPU is primarily working with fixed blocks of memory references. TRUE
17) The Main Memory consists of words, these words are grouped into blocks for mapping purposes and each block can contain a different number of words. FALSE
18) Cache includes tags to identify which block of Main Memory we need to retrieve. FALSE
19) Cache hit is when data is present in cache and we transfer it to CPU (fast) while Cache miss is when data isn't present and we read the required block from Main Memory to Cache then we transfer it to CPU (slow). TRUE

20) Mapping function is an algorithm for determining which Main Memory block currently occupying a cache line. TRUE

21) It is possible to arrive at a single “optimum” Cache size for all architectures. FALSE

22) Cache is organized based on the mapping function used in the architecture. TRUE

23) In Direct Mapping, the number of locations in the Main Memory is the number of blocks that can be found in one line of Cache. TRUE

24) In Direct Mapping, Cache searching gets expensive. FALSE

25) In Associative Mapping, a Main Memory block can be loaded into any line of Cache. TRUE

26) Replacement algorithm of Direct Mapping states that if Cache misses, one line must be replaced by the required block data from Main Memory. TRUE

27) Replacement algorithms of Associative Mapping are implemented in software to achieve high speed. FALSE

28) The most effective/popular replacement algorithm for Associative Mapping is LFU. FALSE

29) In write back policy, update/use bit is set when updating Main Memory. FALSE

30) As the block size increases from very small to larger sizes, the hit ratio will at first increase because of the principle of locality. TRUE

b. MCQ:

1) All of the following are examples of internal memory except …….

a) RAM b) Cache on hard disk c) **Hard disk** d) Cache on CPU

2) Internal memory deals with ……., while external memory deals with ……...

a) Byte, word b) **Word, block** c) Word, byte d) Block, byte

3) The word size of any internal memory is determined by the size of ……

a) **Data bus** b) Address bus c) Control size d) DR

4) …….. is a parameter used to measure the how good a memory is w.r.t. capacity.

a) Access time b) Memory cycle time c) Transfer rate d) **Word size**

5) The time required before a second access can commence.is called ……...

a) Memory cycle time b) Access time c) Transfer time d) **Recovery time**

6) The number of words of any internal memory is determined by the size of …….

a) Data bus b) **Address bus** c) Control size d) DR

7) ……. is an access method where individual blocks have unique address and the access time depends on the required location and the previous location of data.

a) Random access method b) **Direct access method** c) Associative access method

d) Sequential access method

8) In random access method, the access time required to access the next location is ……. the time required to access the last location.

a) More than b) Less than c) Multiple d) **Equal to**

9) ……. is an example for a memory device that utilizes the sequential access method.

a) Tape **b) RAM** **c) Hard disk** **d) Cache**

10) RAM is based on ……. technology, Disk & Tape are based on ……. technology and CD & DVD are based on …….. technology.

a) **Semiconductor, Magnetic and Optical** b) Magnetic, Semiconductor and Optical

c) Optical, Magnetic and Semiconductor d) Magnetic, Optical and Semiconductor

11) For a memory with 5 lines address bus and a Cache where each line can contain 4 words, the number of lines of Cache needed to store all the words of the memory is ….

a) 6 b) 7 c) **8** d) 9

12) The size of Cache has to be …….. for the cost purpose, and also has to be ……. so that the overall average access time is close to that of the Cache alone.

a) small, small b) large, small c) **small, large** d) large, large

13) For a 32 Mbyte RAM,64 Kbyte Cache and each block in Cache is 8 bytes, how many locations in Main Memory that can occupy the whole Cache?

a) 256 b) 1024 c) **512** d) 128

14) In Direct Mapping, we only compare the tag field of Main Memory address with the tag of exactly one line in Cache because ……

a) Each byte in Cache must be loaded into one byte in Main Memory.
b) **Each block from Main Memory must be found in only one line in Cache.**
c) Each block in Main Memory has unique tag.
d) Each block in Cache must be loaded into one byte in Main Memory

15)……. is a mapping function where the tag of each line of Cache is compared to the tag in the Main Memory address.

a) **Associative** b) Direct c) Indirect d) Set associative

16) Write policy is used to ensure ……..

a) **The consistency/validation of data in Cache and Main Memory.**

b) The efficiency of data transfer between Cache and Main Memory.

c) The redundancy of data between Cache and Main Memory.

d) The capability of the system to store data.

c. Problems:

1. What are the differences among sequential access, direct access, associative access and random access? (Reference)
   Answer:
   Sequential access: Memory is organized into units of data, called records. Access must be made in a specific linear sequence.
   Direct access: Individual blocks or records have a unique address based on physical location. Access is accomplished by direct access to reach a general vicinity plus sequential searching, counting, or waiting to reach the final location.
   Random access: Each addressable location in memory has a unique, physically wired-in addressing mechanism. The time to access a given location is independent of the sequence of prior accesses and is constant.

2. What is the general relationship among access time, memory cost, and capacity? (Reference)
   Answer:
   Faster access time, greater cost per bit;
   greater capacity, smaller cost per bit;
   greater capacity, slower access time

3. What is an addressable unit? (Slides)
   Answer:
   Smallest location which can be uniquely addressed, it can be:
   i. Word location on internal memory
   ii. Block location on external memory

4. What are the 2 most important characteristics of a memory? (Old sheet 6)
   Answer:
   The two most important characteristics of memory are capacity and performance.

5. Mention and explain the parameters used to measure how good a memory can be w.r.t performance and capacity. (Slides and old sheet 6)

   Answer:

   w.r.t performance:

   i. **Access time:** For Random- Access Memory (RAM), this is the time it takes to perform a read or write operation, that is, the time from the instant that an address is presented to the memory to the instant that data have been stored or made available for use.
      For non- random- access memory, access time is the time it takes to position the read-write mechanism at the desired location.
   ii. **Memory Cycle time:** Time may be required for the memory to "recover" before next access. It is the time required before a second access can commence.
      Cycle time = access time + recovery time
   iii. **Transfer Rate:** Rate at which data can be moved

   w.r.t capacity:

   i. **Word size:** The natural unit of organisation (Depending on the size of Data Bus)
   ii. **Number of words (or Bytes):** Depending on the size of Address Bus

6. Explain and give 1 example for the 4 memory access methods then state if each one of them is dependent or independent on current or previous location of data. (Slides)

   Answer:

   i. **Sequential**: Start at the beginning and read through in order, Example: Tape (Dependent)
   ii. **Direct**: Individual blocks have unique address, we search inside the block for our word., Example: Hard disk (Dependent)
   iii. **Random**: Individual addresses identify locations exactly (i.e. the access time required to access the next location is the same time required to access the last location) Example: RAM (Independent)
   iv. **Associative**: Data is located by a comparison with contents of a portion of the store, Example: Cache memory (Independent)

7. What is meant by volatility, erasability and power consumption of different memory technologies. (Slides)

   Answer:

   volatility: is data stored permanently or get erased with the power switch.
   Erasability: How to erase the contents of a memory (using what technology and erase byte by byte or bit by bit or block by block).
   Power Consumption: How much power does the memory use during normal activities and idle state.

8. For a memory with 4-bit address lines, 8-bit word and a cache with 4 lines, how many words are there in each line of cache? (To store the whole memory) (Slides)

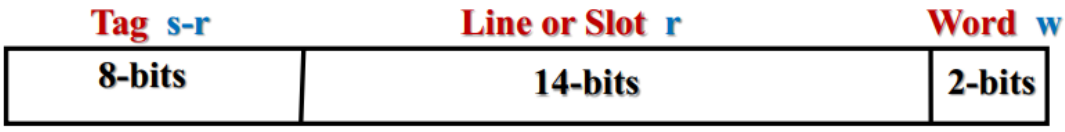
Answer:

Total number of words in memory = 2^4 = 16
Number of words per line in cache = 16/4 = 4 words per line

9. For a Direct-mapped Cache, a Main Memory address is viewed as consisting of three fields. List and define the three fields. (Reference)

Answer:

One field identifies a unique word or byte within a block of main memory. The remaining two fields specify one of the blocks of main memory. These two fields are a line field, which identifies one of the lines of the cache, and a tag field, which identifies one of the blocks that can fit into that line.



10. Mentions the pros and cons of Direct Mapping and Associative Mapping. (Slides and old sheet 7)

Answer:

	Pros	Cons
Direct mapping	• its simplicity and ease of implementation • compare the tag once to decide it the required data is in Cache or not.	• it has a greater access time than any other method. • its performance is degraded if two or more blocks that map to the same location are used alternately.
Associative mapping	• Associative mapping is fast. • It has short access time.	• Cache Memory implementing associative mapping is expensive as it requires to store address along with the data. (Not just tag) • Compare with all the entries of the cache.

11. Why do we have to check that Main Memory is up to date before replacing a line/block in Cache. (Slides)

Answer:

When a block resident in Cache is to be replaced:
If at least one writes operation has been performed on a word in that block of the Cache, then Main Memory must be updated by writing the block of Cache out to the block of Main Memory before bringing in the new block. More than one device may have access to Main Memory. For example, an I/O module may be able to read-write directly to memory. If a word has been altered only in the Cache, then the corresponding Main Memory word is invalid.

12. Discuss each of the following: (Slides)
    - "Write through" policy
    - "Write back" policy

    Answer:

    Write through: all write operations are made to Main Memory as well as to the Cache, ensuring that Main Memory is always valid. (Lots of traffic, slows down writes) Write back: updates are made only in the Cache, when an update occurs, a use bit, associated with the Block is set. Then, when a block is replaced, it is written back to Main Memory if and only if the use bit is set i.e. (use bit =1). (uses extra bit per block)

13. Consider a machine with a byte addressable main memory of 2^{16} bytes and block size of 8 bytes. Assume that a direct mapped cache consisting of 32 lines is used with this machine. (Reference)
    a. How is a 16-bit memory address divided into tag, line number, and byte number?
    b. How many total bytes of memory can be stored in the cache?
    c. Why is the tag also stored in the cache?

    Answer:

    Number of blocks in memory = 2^{16} / 8 = 2^{16} / 2^3 = 2^{13}
    Number of lines in cache = 32 = 2^5 , then line size is 5 bits

    a. 8 leftmost bits = tag; 5 middle bits = line number; 3 rightmost bits = byte number
    b. Total number of byte stored in cache = 32 * 8 = 256 bytes.
    c. Because items with different memory addresses can be stored in the same place in the cache. The tag is used to distinguish between them.

d. Assignment:

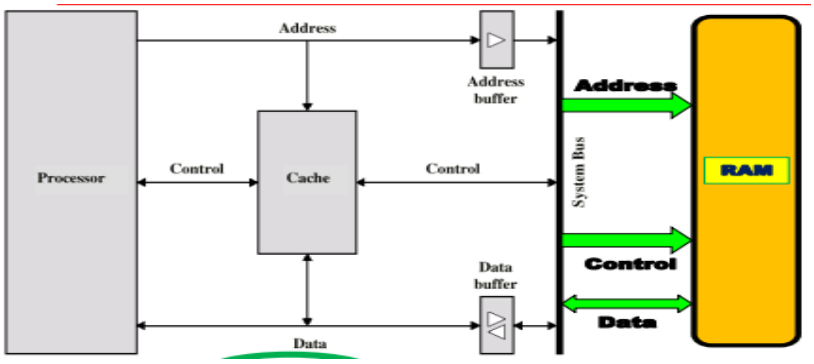
1. Mention the full hierarchy list of memory devices and the technology used to make each one of them.

   Answer:

   • Registers: Semiconductor
   • L1 Cache: Semiconductor
   • L2 Cache: Semiconductor
   • Main memory: Semiconductor + capacitors
   • Disk cache: Semiconductor
   • Disk: Magnetic
   • CD & DVD: Optical
   • Tape: Magnetic

2. Explain the cache operation (cache hit and cache miss).
   Answer:
   A cache hit occurs (is block containing RA in cache), the Data and Address buffers are disabled and communication is only between processor and cache
   A cache miss occurs (is block containing RA in cache), the desired Address is loaded onto the system bus and the Data are returned through the data buffer to both the cache and the processor.

3. In the typical cache organization, how does the Cache connect to the processor?
   Answer:
   In this organization, the cache connects to the processor via Data, Control, and Address lines.
   The Data and Address lines also attach to Data and Address buffers, which attach to a system bus from which main memory is reached.



4. Mention and explain the techniques used in mapping function.
   Answer:
   Direct mapping: Each block of Main Memory maps to only one cache line.
   Associative mapping: A Main Memory block can load into any line of cache.
   Set associative mapping: Search the internet.

5. What is meant by LRU, FIFO and LFU replacement algorithms.
   Answer:
   LRU: Replace that block in the set that has been used in the cache since longest time.
   FIFO: Replace that block in the set that has been in the cache longest time.
   LFU: Replace block which had fewest hits