Chapter 5

Cache and Main Memory, Secondary Storage

LEARNING OBJECTIVES

- *Characteristics of memory system*
- *Memory hierarchy*
- *Locality of reference*
- *Cache memory*
- *Basic operation of cache*
- *Elements of cache design*
- *Replacement algorithm*
- *Secondary storage*
- *Disk*
- *Diskette*
- *Magnetic tape*
- *Optimal memory*

Characteristics of Memory System

1. **Location:** The term refers to whether memory is internal or external to the computer. The location of memory may be
   - Processor
   - Internal (main)
   - External (secondary)
2. **Capacity:** The capacity of internal memory is expressed in terms of bytes. The capacity specified using
   - Word size
   - Number of words
3. **Unit of transfer**
   - For internal memory, the unit of transfer is equal to the number of data lines into and out of the memory module. The unit of transfer need not equal a word or an addressable unit.
   - For external memory, data are often transferred in much larger units than a word, and these are referred to as blocks.
4. **Access method:** The various methods of accessing units of data are

   (i) Sequential access: Memory is organized into units of data, called records.

   Example: Magnetic tapes

   (ii) Direct access: Individual blocks or records have a unique address based on physical location.
   Example: Magnetic disks

   (iii) 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.
   Example: Main memory

   (iv) Associative: This is a random access type of memory that enables one to make a comparison of desired bit locations within a word for a specified match.

5. **Performance:** Three performance parameters are:

   (i) Access time (latency):
   - For random access memory, this is the time it takes to perform a read or write operation.
   - 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: For a random access memory it consists of the access time plus any additional time required before a second access can commence.

   (iii) Transfer rate: This is rate at which data can be transferred into or out of memory unit.

For Random access memory,

$$\text{Transfer rate} = \frac{1}{\text{Cycle Time}}$$

For non-random access memory, $T_N = T_A + \frac{N}{R}$

Where, T_N = Average time to read or write N-bits.
T_A = Average access time
N = Number of bits
R = Transfer rate in bits per second

6. **Physical type:** The physical type of a memory will be
   i. Semiconductor
   ii. Magnetic
   iii. Optical
   iv. Magneto-optimal
7. **Physical characteristics:** The memory may be
   - Volatile/non-volatile
   - Erasable/non-erasable
8. **Organization:** There is a trade–off among the three key characteristic of memory.
   i. Cost
   ii. Capacity
   iii. Access time

Memory Hierarchy

Consider the following memory hierarchy, which shows the various memory components.

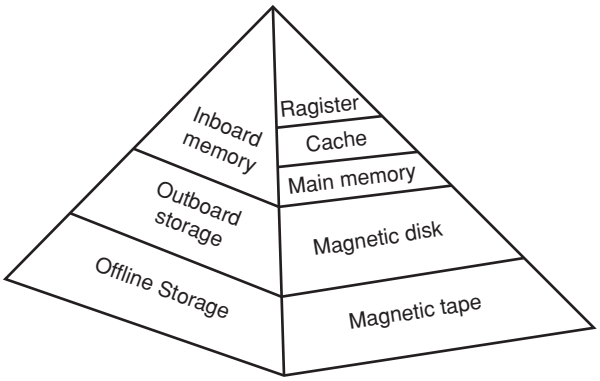


Figure 1 Memory Hierarchy

As one goes down the hierarchy, the following occur:

1. Decreasing cost per bit
2. Increasing capacity
3. Increasing access time
4. Decreasing frequency of access of the memory by the processor.

Locality of Reference

During the course of execution of a program, memory references by the processor, for both instructions and data, tend to cluster. This is referred as principal of locality.

(i) Registers: The fastest, smallest and most expensive type of memory consists of the registers internal to the processor.

(ii) Main memory: The principal internal memory system of the computer is main memory. Each location in main memory has a unique address.

(iii) Cache: Main memory is usually extended with a higher speed, smaller cache. The cache is not visible to the programmer or, indeed, to the processor. It is a device for staging the movement of data between main memory and processor registers to improve performance.

These three forms of memory are volatile and employ semi conductor technology.

(iv) Magnetic tapes and disks: Data are stored more permanently on external mass storage devices, of which the most common are hard disk and removable media.

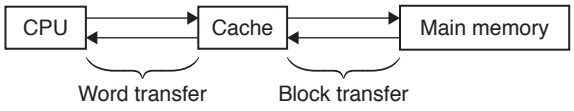
- External, non-volatile memory is also referred to as secondary or auxiliary memory.
- Used to store program and data files, which are visible to the programmer in the form of files and records.

Cache Memory

The locality of reference property states that over a short interval of time, the address generated by a typical program refer to a few localized areas of memory repeatedly, while the remainder of memory is accessed relatively infrequently (Because of frequent loops and subroutine calls).

If the active portions of the program and data are placed in a fast small memory, the average memory access time can be reduced, thus reducing the total execution time of the program. Such a fast small memory is referred to as a cache memory.

Cache memory is intended to give memory speed approaching that of the fastest memories available and at the same time provide a large memory size at the price of less expensive types of semiconductor memories. The following figure shows the structure of cache/main memory system.



The fundamental idea of cache organization is that by keeping the most frequently accessed instructions and data in the fast memory, the average memory access time will approach the access time of cache.

Basic Operation of Cache

- When the CPU need to access memory, the cache is examined. If the word is found in cache, it is read otherwise main memory is accessed to read the word.

- The performance of cache memory is measured in terms of hit ratio.
- When the CPU refers to memory and find the word in cache, it is called hit.
- If the word is not found in cache and is in Main Memory, it is called miss.

$$\text{Hit ratio} = \frac{\text{hits}}{\text{hits} + \text{misses}}$$

Average access time = $hc + (1 - h)(c + m)$m

Where, $c \rightarrow$ Cache access time

$m \rightarrow$ Main memory access time

$h \rightarrow$ hit ratio

- Let main memory consists of up to 2^n addressable words, with each word having a unique n-bit address.
- For mapping purposes, this memory is considered to consist of a number of fixed length blocks of K words each.

$$\therefore \text{ Number of blocks } (M) = \frac{2^n}{K}$$

- The cache consists of C lines.
- Each line contains K words, plus a tag of a few bits.
- The number of words in a line is referred to as the line size.
- The number of lines is considerably less than the number of main memory blocks i.e., $C << M$.
- Each line includes a tag that identifies which particular block is currently being stored.

  The tag is usually a portion of the main memory address.

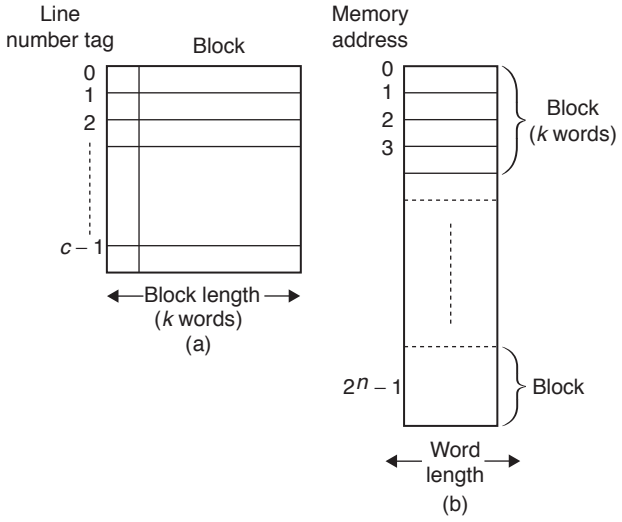


Figure 2 (a) Cache, (b) Main memory

Elements of Cache Design

1. Cache size
2. Mapping function
   - Direct
   - Associative
   - Set-associative
3. Replacement algorithm
4. Write policy
   - Write through
   - Write back
   - Write once
5. Line size
6. Number of caches
   - Single or two level
   - Unified or split

Cache size

The size of the cache to be small enough so that the overall average cost per bit is close to that of main memory alone and large enough so that the overall average access time is close to that of the cache alone.

Mapping function

Because there are fewer cache lines than main memory blocks, an algorithm is needed for mapping main memory blocks into cache lines. Three techniques can be used for mapping.

(i) Direct
(ii) Associative
(iii) Set-associative

Direct mapping Maps each block of main memory into only one possible cache line. Figure 2 illustrates the general mechanism. The mapping is expressed as

$i = j$ modulo m, where
i = cache line number
j = main memory block number
m = number of lines in the cache

For purpose of cache access, each main memory address can be viewed as consisting of three fields.

- The least significand w bits identify a unique word or byte within a block of main memory.
- The remaining s-bits specify one of the 2^s blocks of main memory. The cache logic interpret these s-bits as a tag of s-r bits. (most significand portion)
- A line field of r-bits, to identify one of 2^r lines.

To summarize,

Address length = $(s + w)$ bits

Number of Addressable units = 2^{s+w} words or bytes

Block size = line size = 2^w words or bytes

Number of blocks in main memory = $\frac{2^{s+w}}{2^w} = 2^s$ Number of lines in cache = $M = 2^r$.

Size of Tag = $(s - r)$ bits

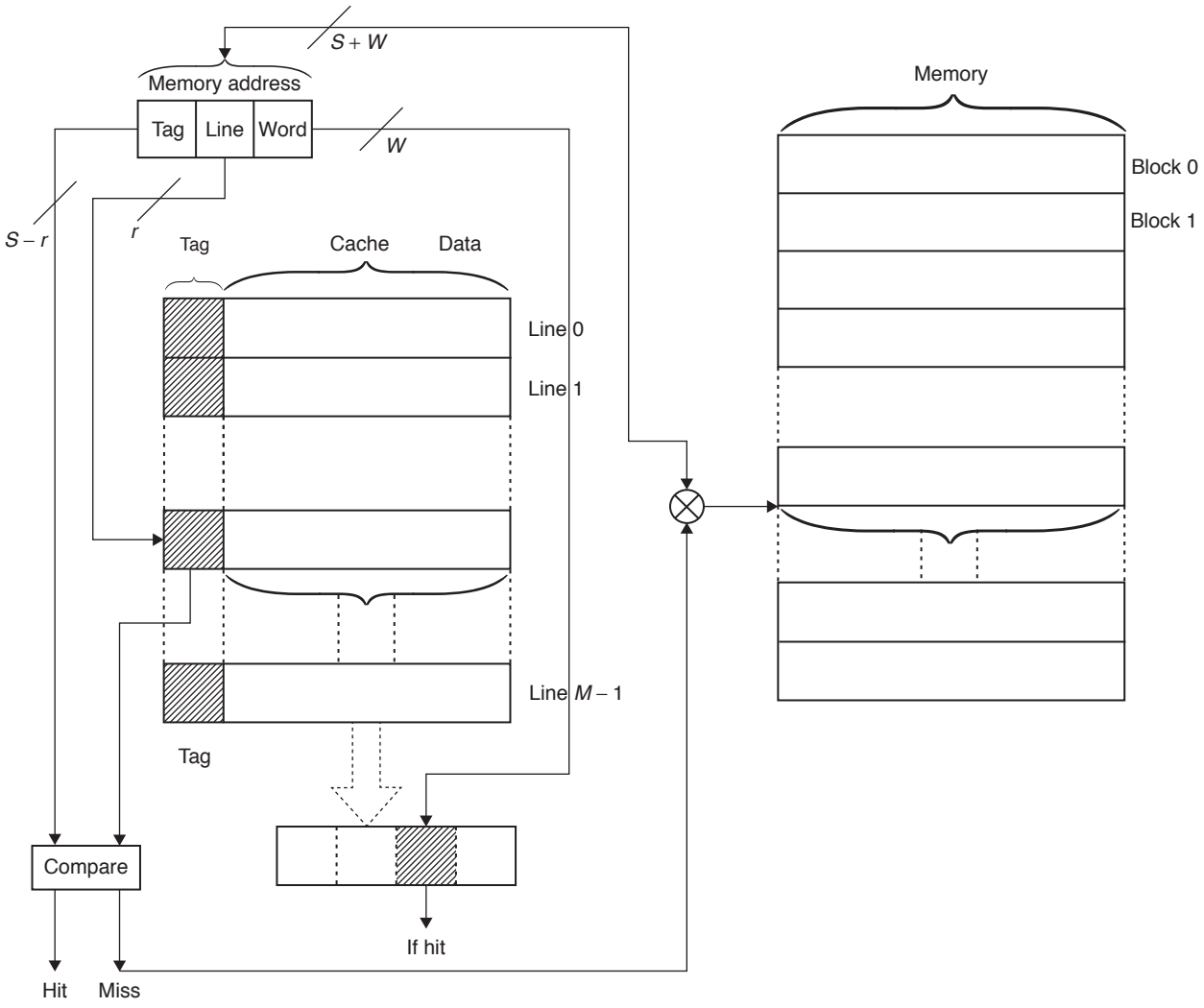


Figure 3 Direct Mapping

The effect of this mapping is that blocks of main memory are assigned to lines of the cache as follows:

Cache Line	Main Memory Blocks Assigned
0	0, m, $2m$, … $2^s - m$
1	1, $m+1$, $2m+1$, … $2^s - m + 1$
.	
.	
.	
.	
$m-1$	$m-1$, $2m-1$, $3m-1$,… $2^s - 1$

Example 1: Let cache capacity = 64 KB
Line size = 4 B
Main memory capacity = 16 MB = 2^{24} B
Using direct mapping, Address length = $s + w$ = 24-bits
Block size = 2^2 B

$$\text{Number of blocks in main memory} = \frac{2^{24}}{2^2} = 2^{22}$$

$$\text{Number of lines in cache} = m = 2^r = \frac{2^{16}}{2^2} = 2^{14}$$

∴ Size of tag = $s - r = 22 - 14 = 8$
∴ Main memory address =

Tag	Line	Word
8	14	2

The mapping becomes

Cache Line	Starting Memory Address of Blocks (Hexa)
0	00000, 010000, … FF0000
1	000004, 010004, … FF0004
.	.
.	.
.	.
.	.
$2^{14} - 1$	00FFFC, 01FFFC,… FFFFFC

Note: No two blocks that map into the same line number have the same tag number.

Advantages:
- Simple and cheap
- The tag field is short; only those bits have to be stored which are not used to address the cache.
- Access is very fast.

Disadvantages: A given block fits into a fixed cache location, i.e., a given cache line will be replaced whenever there is a reference to another memory block which fits to the same line, regardless what the status of the other cache line is.

This can produce a low hit ratio, even if only a very small part of the cache is effectively used.

Associative mapping This technique overcomes the disadvantage of direct mapping by permitting each main memory block to be loaded into any line of the cache. Here the cache control logic interprets a memory address as two fields.

1. Tag
2. Word

Figure shows associative mapping technique:

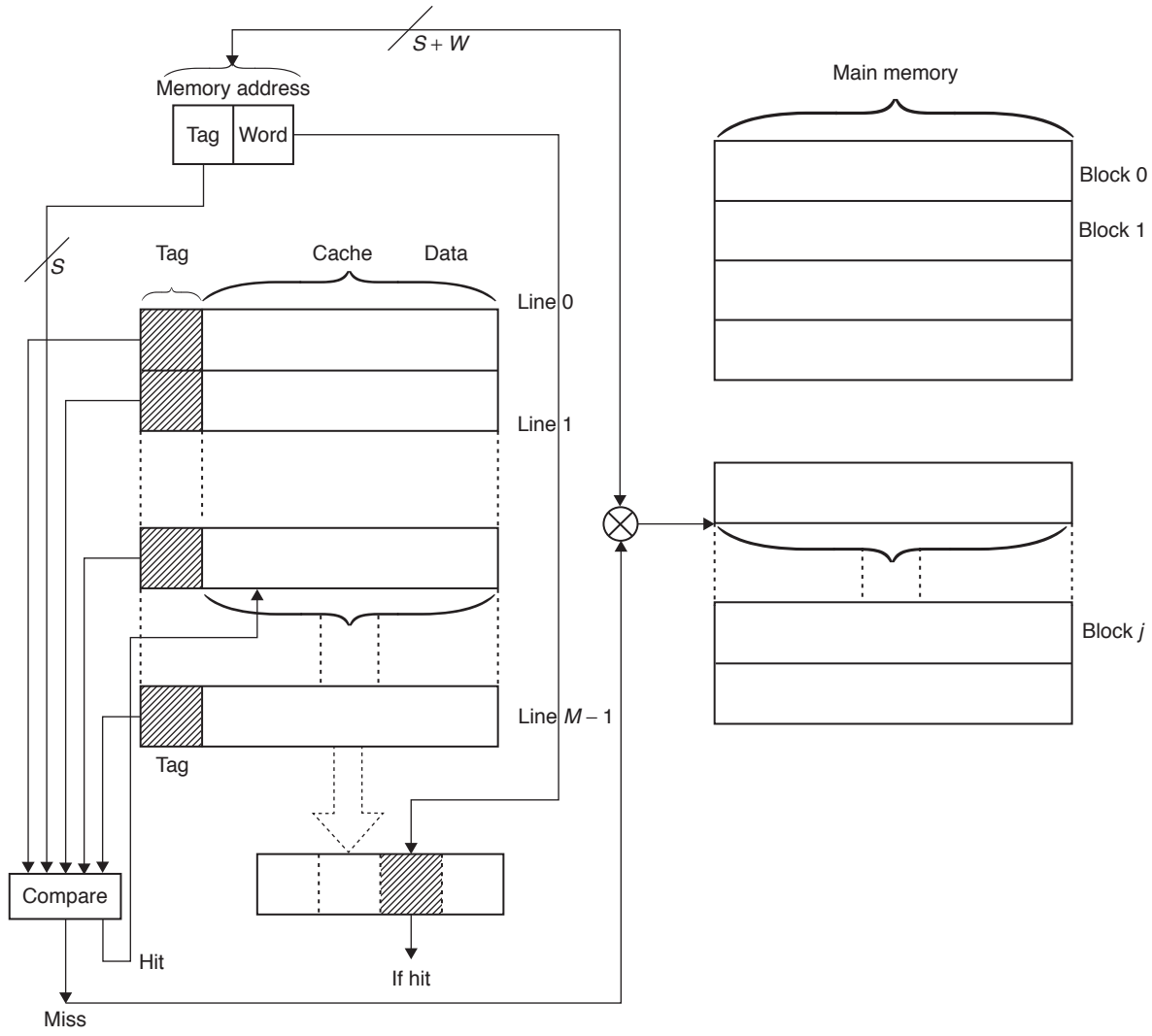


Figure 4 Associative mapping

To determine whether a block is in the cache, the cache control logic must simultaneously examine every line tag for a match. No field in the address corresponds to line number, so that the number of lines in the cache is not determined by the address format.

To summarize,

Address length = $(s + w)$ bits
Number of addressable units = 2^{s+w} words or bytes
Block size = line size = 2^w words or bytes
Number of blocks in main memory = $\frac{2^{s+w}}{2^w} = 2^S$
Number of lines in cache = undetermined
Size of tag = s-bits

Example 2: Cache size = 64 KB
Line size = 4 B
Main memory capacity = 16 MB
Number of blocks in main memory $= \frac{2^{24}}{2^2}$
$= 2^{22}$.
∴ Size of tag = 24 – 2 = 22-bits
For example, the tag of the hexadecimal main memory address 16339C is 058CE7
Main memory address =

Tag	Word
22	2

Advantages: Associative mapping provides the highest flexibility concerning the line to be replaced when a new block is read into a cache.

Disadvantages:
- Complex
- The tag field is long
- Fast access can be achieved only using high performance associative memories for the cache, which is difficult and expensive.

Set-associative mapping: It exhibits strengths of both the direct and associative approaches and reduces their disadvantages.
Here the cache is divided into V sets, each of which consists of K lines

i.e., $m = V \times K$
$i = j$ modulo V
Where i = cache set number
j = main memory block number
m = number of lines in cache

As there are K lines in each set, this is referred as K-way set associative mapping. The cache control logic interprets a memory address simply as three fields.

1. Tag
2. Set
3. Word

The d set bits specify one of $V = 2^d$ sets. The S-bits of the tag and the set fields specify one of the 2^S blocks of main memory.
Figure 3 shows Set-associative mapping.

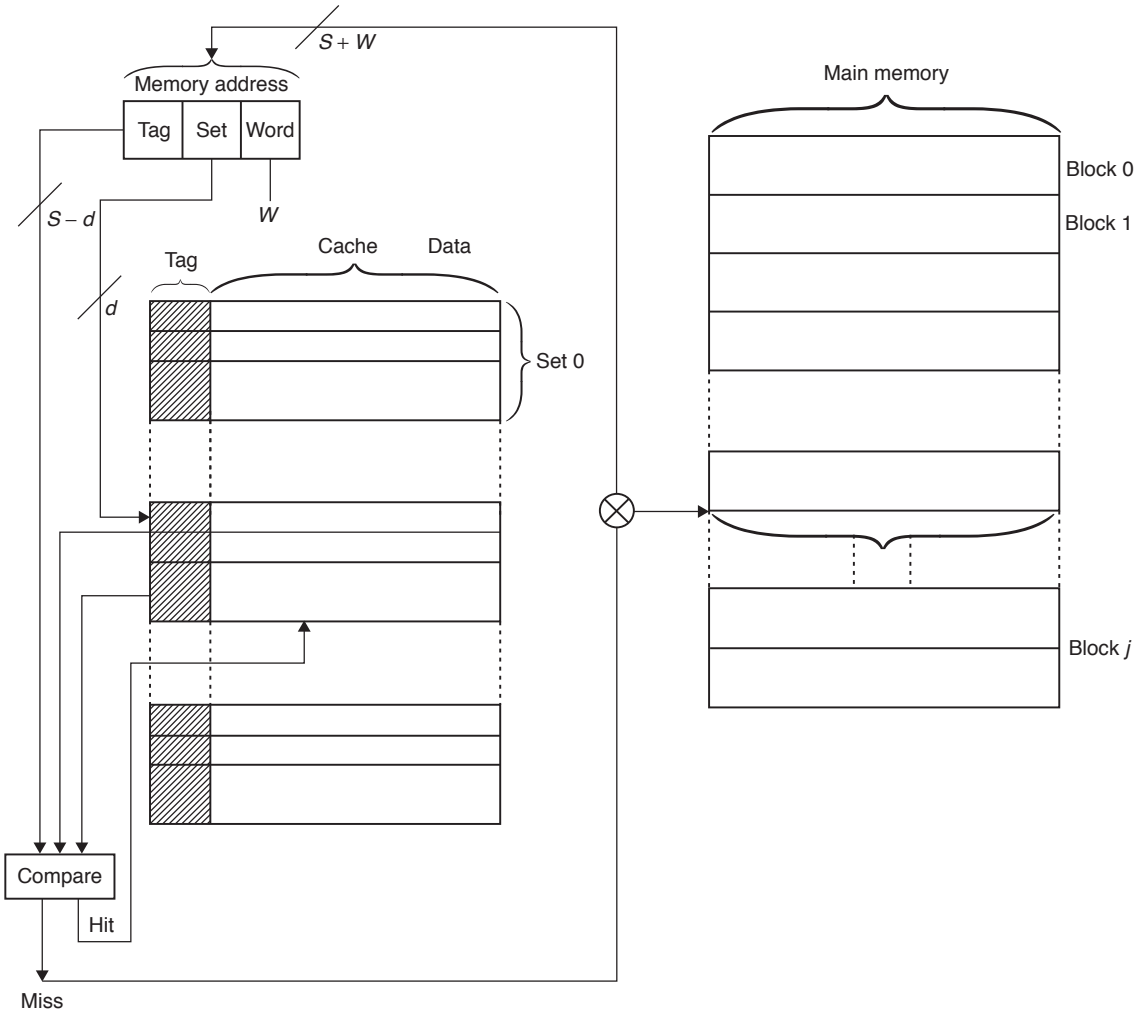


Figure 5 *K*-way set associative cache

Here the tag in a memory address is much smaller and is only compared to the K tags within a single set. To summarize,

Address length = $(s + w)$ bits

Number of Addressable units = 2^{s+w} words or bytes.

Block size = line size = 2^w words or bytes.

Number of blocks in main memory = $\frac{2^{s+w}}{2^w} = 2^s$

Number of lines in set = K

Number of sets $V = 2^d$

Number of lines in cache = $KV = K \times 2^d$

$\therefore$ Size of tag = $(s - d)$ bits

Example 3: Cache capacity = 64 KB
Block size = 4 B
Main memory capacity = 16 MB

Number of blocks in main memory $= \frac{2^{24}}{2^2} = 2^{22}$

For 2-way set associative mapping,
Number of lines in a set $K = 2$
Number of sets = $V = 2^d$

Number of lines in cache = $K \times 2^d = \frac{2^{16}}{2^2}$

$= 2^{14}$

$\Rightarrow \quad 2 \times 2^d = 2^{14}$

$\Rightarrow \quad 2^d = 2^{13}$

$\Rightarrow \quad d = 13$

$\therefore$ Size of Tag = 22 − 13 = 9

Hence main memory address =

Tag	Set	Word
9	13	2

In practice, 2 and 4-way set associative mapping are used with very good results. Larger sets do not produce further significant performance improvement.

If a set consist of a single line, i.e., it is direct mapping; If there is one single set consisting of all lines i.e., it is associative mapping.

Replacement algorithms

Once the cache has been filled, when a new block is brought into the cache, one of the existing blocks must be replaced. For direct mapping, there is only possible line for any particular block, and no choice is possible.

For associative and set associative techniques, a replacement algorithm is needed. Four of the most common replacement algorithms are

(i) LRU (Least recently used): Replaces the block in the set that has been in the cache longest with no reference to it.

(ii) FIFO (First-in-first-out): Replace the block in the set that has been in the cache longest.

(iii) LFU (Least frequently used): Replace the block in the set that has experienced the fewest references.

(iv) Random

Write policy

When a block that is resident in the cache is to be replaced, there are two cases to consider.

(i) If the old block in the cache has not been altered, then it may be over-written with a new block without first writing out the old block.

(ii) If at least one write operation has been performed on a word in that line of the cache, then main memory must be updated by writing the line of cache out of the block of memory before bringing in the new block.

The write policies are

(a) Write through: All write operations are made to main memory as well as to the cache, ensuring that main memory is always valid.
Drawback: Creates substantial memory traffic

(b) Write back: This technique minimizes memory writes. It updates are made only in the cache. When a block is replaced it is written back to main memory if and only if it is updated.
Drawback: There some portions of main memory are invalid and hence accesses by I/O modules can be allowed only through the cache.

Line size

Larger blocks reduce the number of blocks that fit into a cache. Because each block fetch overwrites older cache contents, a small number of blocks results in data being over written shortly after they are fetched.

As a block becomes larger, each additional word is farther from the requested word, therefore less likely to be needed in the near future.

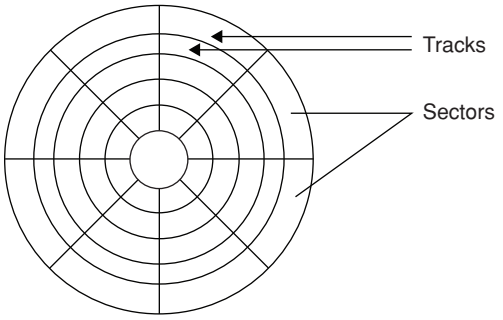
Number of caches

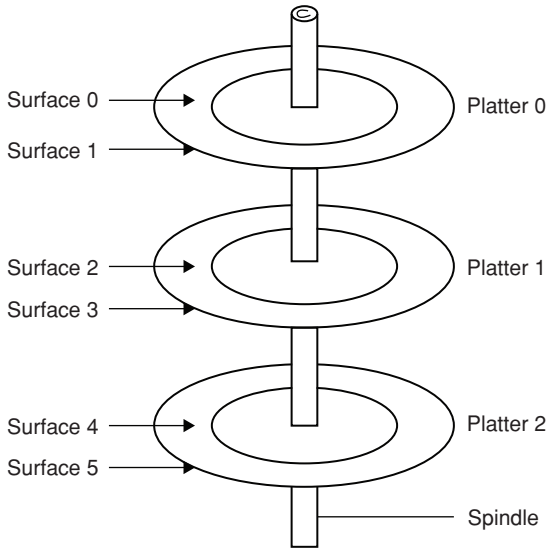
Multilevel caches We may have on-chip cache as well as external cache. This is a two level cache organization, with the internal cache designated as level 1, and external cache designated as level 2.

Secondary Storage

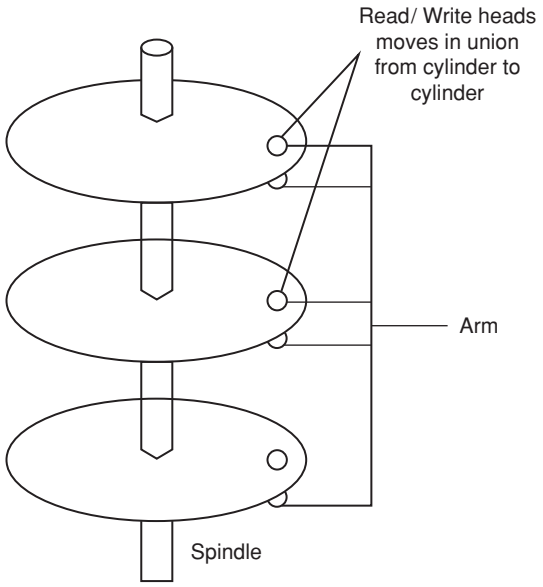
Disk

Disk consists of platters, each with two surfaces. Each surface consists of concentric rings called tracks. Each track consists of sectors separated by gaps.





Disk operation: The disk surface spins at a fixed rotational rate. There is a read/write head attached to the end of the arm and flies over the disk surface on a thin cushion of air. By moving radially the arm can position the read/write head over any track.



Disk access time: Average time to access some target sector:

$$T_{ae} = T_{\text{avg seek}} + T_{\text{avg rotation}} + T_{\text{avg transfer}}$$

Where $T_{\text{avg seek}}$ is typical 9 ms.

$$T_{\text{avg rotation}} = \frac{1}{2} \times \frac{1}{\text{RPMS}} \times 60 \text{ Sec/1min}$$

$$T_{\text{avg rotation}} = \frac{1}{\text{RPM}} \times \frac{1}{(\text{avg sector/track})} \times 60 \text{ Sec/1min}$$

Notes:

1. Seek time is the Time to position heads over cylinder containing target sectors ($T_{\text{avg seek}}$).
2. Rotational Latency is the time waiting for first bit of target sector to pass under read/write head. ($T_{\text{avg rotation}}$).
3. Transfer Time is the time to read the bits in the target sector ($T_{\text{avg transfer}}$).
   - Data are recorded on the surface of a hard disk made of metal coated with magnetic material.
   - The disks and the drive are usually built together and encased in an air tight container to protect the disk from pollutants such as smoke particle and dust. Several disks are usually started on a common drive shaft with each disk having its own read/write head.

Diskette

Data are recorded on the surface of a floppy disk made of polyester coated with magnetic material.

A special diskette drive must be used to access data stored in the floppy disk. It works much like a record turntable of Gramophone.

Main features

- Direct access
- Cheap
- Portable, convenient to use

Main Standards

- $5^1/_4$ inch capacity ≅ 360 KB/ disk
- $3^1/_2$ inch capacity ≅ 1.44 MB/disk (about 700 pages of A_4 text)

Magnetic Tape

Magnetic tape is made up from a layer of plastic which is coated with iron oxide. The oxide can be magnetized in different directions to represent data. The operation uses a similar principle as in the case of a tape recorder.

Main features

- Sequential access (access time about 1.55)
- High value of storage (50 MB/tape)
- Inexpensive

It is often used for Batch up or archive purpose.

Optimal Memory

CD-ROM (Compact disk ROM): The disk surface is imprinted with microscopic holes which record digital information. When a low-powered power beam shines on the surface, the intensity of the reflected light changes as it encounters a hole. The change is detected by a photo sensor and converted into a digital signal.

- Huge capacity: 775 MB/disk (≈550 diskette)
- Inexpensive replication, cheap production.
- Removable, read only.
- Long access time (could be half a second)

WORM (Write Once Read Memory) CD: A lower beam of modest intensity equipped in the disk drive is used to imprint the hole pattern.

- Good for archival storage by providing a permanent record of large volumes of data.

Erasable Optical Disk: Combination of Laser technology and magnetic surface technique.

- Can be repeatedly written and overwritten.
- High reliability and longer life than magnetic disks.

Exercise

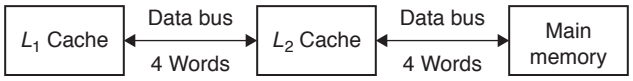
Practice Problems I

Directions for questions 1 to 20: Select the correct alternative from the given choices.

1. Find the number of bits in the cache index and tag for a direct mapped cache of size 32 KB with block size of 32 bytes. The CPU generates 48-bit addresses.
(A) 33,15 (B) 15,10
(C) 10,33 (D) 15,33

2. Given the cache access time is 200 ns and the memory access time is 400 ns. If the effective access time is 20% greater than the cache access time, what is the hit ratio?
(A) 80% (B) 20%
(C) 40% (D) 100%

3. A computer system has an L_1 cache, an L_2 cache and a main memory unit connected as shown below. The block size in L_1 cache is 2 words. The block size in L_2 cache is 8 words. The memory access times are 2 nanoseconds, 20 nanoseconds and 200 nanoseconds for L_1 cache, L_2 cache and main memory unit, respectively.



When there is a miss in L_1 cache and a hit in L_2 cache, a block is transferred from L_2 cache to L_1 cache. What is the time taken for this transfer?
(A) 22 ns (B) 44 ns
(C) 66 ns (D) 88 ns

4. In direct memory management, CPU references address of 15-bits. Main memory size is 512 * 8 and cache memory size is 128 * 8. Tag and line are respectively
(A) 9, 7 (B) 7, 9
(C) 15, 7 (D) 7, 15

5. Consider a cache with 64 blocks and a block size of 16 bytes. The byte address of 1200 maps to ____ block number.
(A) 10 (B) 11
(C) 64 (D) 16

6. In a cache memory, cache line is 64 bytes. The main memory has latency of 32 ns and bandwidth of 1 GB/sec. Then the time required to fetch the entire cache line from main memory is
(A) 32 ns (B) 64 ns
(C) 96 ns (D) 128 ns

7. A set associative cache consists of 64 lines or slots divided into four-line sets. Main memory contains 4K blocks of 128 word each. Then the number of bits present in tag, set and word fields are respectively.
(A) 7, 6, 7 (B) 6, 7, 7
(C) 4, 8, 7 (D) 8, 4, 7

8. A 2-way set-associative cache has lines of 32 bytes and a total size of 16 KB. The 32 MB main memory is byte addressable. Then which of the following two memory addresses mapped to same set?
(A) 10D6A32, 035C3A2
(B) 2A36D01, 2A3C530
(C) 10D63A2, 035C3A0
(D) 2A36D08, 0A3C538

9. Let the cache memory capacity is 64 KB and main memory capacity is 16 MB. Let block size is 4 bytes. Then the tag, line, word fields in hexadecimal notation for the main memory address cccccc using direct mapped cache will be
(A) cc, ccc, c (B) cc, 3333, 0
(C) cc, ccc, 0 (D) cc, 333, 30

10. Consider a 32-bit microprocessor that has an on-chip 16 KB four-way set associative cache. Assume that the cache has a line size of four 32-bit words. Then the word in the memory location ABCDE8F8 will be mapped to
(A) 143rd set (B) 815th set
(C) 255th set (D) 0th set

11. Given the following specifications for an external cache memory:
Four-way set associative, Line size of two 16-bit words; Able to accommodate a total of 4K 32-bit words from

main memory. Used with a 16-bit processor that issues 24-bit address. Then the number of bits used to represent set field is
(A) 2-bits (B) 10-bits
(C) 12-bits (D) 14-bits

12. 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. Then in what line would bytes with the address 1100 0011 0011 0100 is stored?
(A) slot 3 (B) slot 4
(C) slot 6 (D) slot 12

13. A computer system contains a main memory of 32 K 16-bit words. It also has a 4 K-word cache divided into four line sets with 64 words per line. The processor fetches words from locations 0, 1, 2…, 4351 in that order. It then repeats this fetch sequence 10 more times. The cache is 10 times faster than main memory. Then the improvement resulting from the use of the cache is (assume an LRU policy is used for block replacement)
(A) 0.63 (B) 0.45
(C) 1.21 (D) 2.18

14. Consider an L_1 cache with an access time of 1ns and a hit ratio of $H = 0.95$. Suppose that we can change the cache design such that we increase H to 0.98, but increase access time to 1.5ns. Which of the following condition is met for this change to result in improved performance?
(A) Next level memory access time must be less than 16.67
(B) Next level memory access time must be greater than 16.67
(C) Next level memory access time must be less than 50
(D) Next level memory access time must be greater than 50

15. Consider a single-level cache with an access time of 2.5 ns and a line size of 64 bytes and a hit ratio of $H = 0.95$. Main memory uses a block transfer capability that has a first-word (4 bytes) access time of 50 ns and an access time of 5ns for each word thereafter. What is the access time when there is a cache miss?
(A) 130 ns (B) 149.4 ns
(C) 2.375 ns (D) 8.875 ns

16. The tag, block and word fields of main memory address using direct mapping technique for 2048 main memory blocks, 128 blocks of cache memory and block size of 16:
(A) 4, 7, 4 (B) 7, 4, 4
(C) 11, 7, 4 (D) Data insufficient

17. Let H_1 is level 1 cache hit ratio, H_2 is level 2 cache hit ratio, C_1 is the time required to access Level 1 cache, C_2 is the time required to access Level 2 cache and M is the time required to access Main memory. Then the average access time required by the processor is
(A) $H_1C_1 + (1 - H_1)\, H_2(C_2) + (1 - H_1)\, (1 - H_2)\, (M)$
(B) $H_1C_1 + (1 - H_1)\, H_2(C_1 + C_2) + (1 - H_1)\, (1 - H_2)\, (C_1 + C_2 + M)$
(C) $H_1C_1 + H_1H_2(C_1 + C_2) + H_1H_2(C_1 + C_2 + M)$
(D) $H_1C_1 + (1 - H_1)\, H_2(C_1.C_2) + (1 - H_1)\, (1 - H_2)\, (C_1.C_2.M)$

18. If $p = 2^m$ be the number of lines in cache and $b = 2^n$ be the size of each block, then total words that can be stored in cache memory is given by
(A) 2^{m+n} (B) 2^{m-n}
(C) $m + n$ (D) $p + b$

19. Cache memory enhances
(A) memory capacity
(B) memory access time
(C) secondary storage capacity
(D) secondary storage access time

20. Which of the following property allows the processor to execute a number of clustered locations?
(A) Spatial (B) Temporal
(C) Inclusion (D) Coherence

Practice Problems 2

Directions for questions 1 to 20: Select the correct alternative from the given choices.

1. If average access time of CPU is 20 ns, access time of main memory is 110 ns and the cache access time is 10 ns. What is the hit ratio?
(A) 100% (B) 90%
(C) 80% (D) 70%

2. A hard disk spins at 180 revolutions per minute. What is the average rotational latency?
(A) 0.16 sec (B) 0.32 sec
(C) 0.2 sec (D) 0.4 sec

3. A disk pack have 16 surfaces, with 128 tracks per surface and 256 sectors per track. 512 bytes of data are stored in a bit serial manner in a sector. The number of bits required to specify a particular sector in the disk is
(A) 4 (B) 7
(C) 11 (D) 19

4. A disk has 19456 cylinders, 16 heads and 63 sectors per track. The disk spins at 5400 rpm. Seek time between adjacent tracks is 2 ms. Assuming the read/write head is already positioned at track 0, how long does it take to read the entire disk?
(A) 48 min (B) 58 min
(C) 64 min (D) 72 min

5. A certain moving arm disk storage with one head has following specifications:
Number of tracks/recording surface = 200
Disk Rotation Speed = 2400 rpm
Track storage capacity = 62500-bits
The average latency time (assuming that head can move from one track to another only by traversing the entire track) is
(A) 0.125 sec (B) 1.25 sec
(C) 0.0125 sec (D) 12.5 sec

6. In Memory management system, cache memory access time is 100 ns and main memory access time is 200 ns. Number of CPU references is 100 and number of hits is 10. Average access time is
(A) 150 ns (B) 100 ns
(C) 190 ns (D) 280 ns

7. The seek time of disk is 40 m sec. It rotates at the rate of 40 rps. The capacity of each track is 400 words. The access time is
(A) 50 m sec (B) 53 m sec
(C) 60 m sec (D) 63 m sec

8. An Associated cache and one million word main memory are divided into 256 word blocks. How many blocks are there?
(A) 2^8 (B) 2^{12}
(C) 2^{20} (D) 2^{28}

9. The average access time of a disk is
(A) Seek time + Rotational latency time
(B) Seek time
(C) Rotational latency + transfer time + seek time
(D) Rotation latency + transfer time.

10. What will be the size of the memory whose last memory location is FFFF?
(A) 64 k (B) 32 k
(C) 10 k (D) 24 k

11. Data from a cassette tape is obtained by ____ accessing method.
(A) Parallel (B) Serial
(C) Sequential (D) Random

12. For a memory system, the desirable characteristics is/are
(A) Speed and reliability
(B) Durability and compactness
(C) Low power consumption
(D) All of these

13. The memory that has the shortest access time is
(A) Magnetic bubble (B) Magnetic core memory
(C) Cache memory (D) RAM

14. Cache memory
(A) has greater capacity than RAM.
(B) enhances secondary storage access time.
(C) is faster to access than registers.
(D) is faster to access than main memory

15. Consider a disk pack with 16 surfaces 128 tracks per surface and 256 sectors per track. 512 bytes of data are stored in bit and serial manner, Then the capacity of the disk is
(A) 256 MB (B) 256 KB
(C) 512 MB (D) 64 MB

16. Principle of locality justifies the use of
(A) Cache (B) DMA
(C) Disk (D) RAM

17. The main memory of a computer has $2ab$ blocks while cache has $2a$ blocks. If the cache uses the set associative mapping scheme with two blocks per set, then block k of main memory maps to the set:
(A) (k mod b) of the cache (B) (k mod a) of cache
(C) (k mod $2a$) of cache (D) (k mod $2ab$) of cache

18. Which of the following factors do not affect the hit ration of cache?
(A) Block replacement algorithms.
(B) Block frame size
(C) Cycle counts
(D) Main memory size

19. In which of the following mapping function, there is no need of replacement algorithm?
(A) Direct Mapping
(B) Set-associative mapping
(C) Full associative mapping
(D) Both (A) and (B)

20. In a direct mapping, the index field equals to
(A) Sum of tag and word fields
(B) Sum of block and word fields
(C) Sum of tag an block fields
(D) Same as block field

Previous Years' Questions

1. Consider a small two-way set-associative cache memory, consisting of four blocks. For choosing the block to be replaced, use the least recently used (LRU) scheme. The number of cache misses for the following sequence of block addresses is 8, 12, 0, 12, 8 **[2004]**

(A) 2 (B) 3
(C) 4 (D) 5

2. Consider a direct mapped cache of size 32 KB with block size 32 bytes. The CPU generates 32 bit addresses. The number of bits needed for cache indexing and the number of tag bits are respectively. **[2005]**

(A) 10, 17 (B) 10, 22
(C) 15, 17 (D) 5, 17

Common data for questions 3 and 4: Consider two cache organizations: The first one is 32 KB 2-way set associative with 32-byte block size. The second one is of the same size but direct mapped. The size of an address is 32 bits in both cases. A 2-to-1 multiplexer has a latency of 0.6 ns while a k-bit comparator has a latency of k/10 ns. The hit latency of the set associative organization is h_1 while that of the direct mapped one is h_2.

3. The value of h_1 is: **[2006]**

(A) 2.4 ns (B) 2.3 ns
(C) 1.8 ns (D) 1.7 ns

4. The value of h_2 is: **[2006]**

Data for question 5: Consider a machine with a byte addressable main memory of 2^{16} bytes. Assume that a direct mapped data cache consisting of 32 lines of 64 bytes each is used in the system. A 50 × 50 two-dimensional array of bytes is stored in the main memory starting from memory location 1100 H. Assume that the data cache is initially empty. The complete array is accessed twice. Assume that the contents of the data cache do not change in between the two accesses.

5. Which of the following lines of the data cache will be replaced by new blocks in accessing the array for the second time? **[2007]**

(A) line 4 to line 11 (B) line 4 to line 12
(C) line 0 to line 7 (D) line 0 to line 8

6. For inclusion to hold between two cache levels L_1 and L_2 in a multi-level cache hierarchy, which of the following are necessary? **[2008]**

(i) L_1 must be a write-through cache
(ii) L_2 must be a write-through cache
(iii) The associativity of L_2 must be greater than that of L_1
(iv) The L_2 cache must be atleast as large as the L_1 cache

(A) (iv) only (B) (i) and (iv) only
(C) (i), (ii) and (iv) only (D) (i), (ii), (iii) and (iv)

Common data for questions 7, 8 and 9: Consider a machine with a 2-way set associative data cache of size 64K-bytes and block size 16-bytes. The cache is managed using 32-bit virtual addresses and the page size is 4Kbytes. A program to be run on this machine begins as follows:

```
double ARR [1024] [1024] ;
       int i, j;
/* Initialize array ARR to 0.0 */
 for (i = 0; i < 1024; i++)
 for (j = 0; j < 1024; j++)
 ARR [i] [j] = 0.0;
```

The size of double is 8 Bytes. Array ARR is located in memory starting at the beginning of virtual page 0XFF000 and stored in row major order. The cache is initially empty and no pre-fetching is done. The only data memory references made by the program are those to array ARR.

7. The total size of the tags in the cache directory is **[2008]**

(A) 32K-bits (B) 34K-bits
(C) 64K-bits (D) 68K-bits

8. Which of the following array elements has the same cache index as ARR [0] [0]? **[2008]**

(A) ARR [0] [4] (B) ARR [4] [0]
(C) ARR [0] [5] (D) ARR [5] [0]

9. The cache hit ratio for this initialization loop is **[2008]**

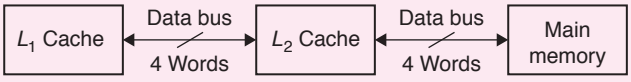
(A) 0% (B) 25%
(C) 50% (D) 75%

10. Consider a 4-way set associative cache (initially empty) with total 16 cache blocks. The main memory consists of 256 blocks and the request for memory blocks is in the following order: **[2009]**
0, 255, 1, 4, 3, 8, 133, 159, 216, 129, 63, 8, 48, 32, 73, 92, 155.

Which one of the following memory block will NOT be in cache if LRU replacement policy is used?

(A) 3 (B) 8
(C) 129 (D) 216

Common data questions 11 and 12: A computer system has an L_1 cache, an L_2 cache, and a main memory unit connected as shown below. The block size in L_1 cache is 4 words. The block size in L_2 cache is 16 words. The memory access times are 2 nanoseconds. 20 nanoseconds and 200 nanoseconds, for L_1 cache, L_2 cache and main memory unit respectively.



11. When there is a miss in L_1 cache and a hit in L_2 cache, a block is transferred from L_2 cache to L_1 cache. What is the time taken for this transfer? **[2010]**
(A) 2 ns (B) 20 ns
(C) 22 ns (D) 88 ns

12. When there is a miss in both L_1 cache and L_2 cache, first a block is transferred from main memory to L_2 cache, and then a block is transferred from L_2 cache to L_1 cache. What is the total time taken for these transfers? **[2010]**
(A) 222 ns (B) 888 ns
(C) 902 ns (D) 968 ns

13. An 8 KB direct-mapped write-back cache is organized as multiple blocks, each of size 32 bytes. The processor generates 32-bit addresses. The cache controller maintains the tag information for each cache block comprising of the following.
1 Valid bit
1 Modified bit
As many bits as the minimum needed to identify the memory block mapped in the cache.
What is the total size of memory needed at the cache controller to store meta-data (tags) for the cache? **[2011]**
(A) 4864 bits (B) 6144 bits
(C) 6656 bits (D) 5376 bits

Common data for questions 14 and 15: A computer has a 256 KB, 4-way set associative, write back data cache with block size of 32 bytes. The processor sends 32 bit addresses to the cache controller. Each cache tag directory entry contains, in addition to address tag, 2 valid bits, 1 modified bit and 1 replacement bit.

14. The number of bits in the tag field of an address is **[2012]**
(A) 11 (B) 14
(C) 16 (D) 27

15. The size of the cache tag directory is **[2012]**
(A) 160 K-bits (B) 136 K-bits
(C) 40 K-bits (D) 32 K-bits
(A) 2.4 ns (B) 2.3 ns
(C) 1.8 ns (D) 1.7 ns

16. In a k-way set associative cache, the cache is divided into v sets, each of which consists of k lines. The lines of a set are placed in sequence one after another. The lines in set s are sequenced before the lines in set $(s + 1)$. The main memory blocks are numbered 0 onwards. The main memory block numbered j must be mapped to any one of the cache lines from **[2013]**
(A) $(j \bmod v) * k$ to $(j \bmod v) * k + (k - 1)$
(B) $(j \bmod v)$ to $(j \bmod v) + (k - 1)$
(C) $(j \bmod k)$ to $(j \bmod k) + (v - 1)$
(D) $(j \bmod k) * v$ to $(j \bmod k) * v + (v - 1)$

17. An access sequence of cache block addresses is of length N and contains n unique block addresses. The number of unique block addresses between two consecutive accesses to the same block address is bounded above by K. what is the miss ratio if the access sequence is passed through a cache of associativity $A \geq K$ exercising least-recently used replacement policy? **[2014]**
(A) n/N (B) $1/N$
(C) $1/A$ (D) K/n

18. A 4-way set -associative cache memory unit with a capacity of 16 KB is built using a block size of 8 words. The word length is 32-bits. The size of the physical address space is 4 GB. The number of bits for the TAG field is _____. **[2014]**

19. In designing a computer's cache system, the cache block (or cache line) size is an important parameter. Which one of the following statements is correct in this context? **[2014]**
(A) A smaller block size implies better spatial locality.
(B) A smaller block size implies a smaller cache tag and hence lower cache tag overhead.
(C) A smaller block size implies a larger cache tag and hence lower cache hit time.
(D) A smaller block size incurs a lower cache miss penalty.

20. Consider a main memory system that consists of 8 memory modules attached to the system bus, which is one word wide. When a write request is made, the bus is occupied for 100 nanoseconds (ns) by the data, address, and control signals. During the same 100 ns, and for 500 ns thereafter, the addressed memory module executes one cycle accepting and storing the data. The (internal) operation of different memory modules may overlap in time, but only one request can be on the bus at any time. The maximum number of stores (of one word each) that can be initiated in 1 millisecond is _____. **[2014]**

21. If the associativity of processor cache is doubled while keeping the capacity and block size unchanged, which one of the following is guaranteed to be NOT affected? **[2014]**
(A) Width of tag comparator
(B) Width of set index decoder
(C) Width of way selection multiplexer
(D) Width of processor to main memory data bus

22. The memory access time is 1 nanosecond for a read operation with a hit in cache, 5 nanoseconds for a read operation with a miss in cache, 2 nanoseconds for a write operation with a hit in cache and 10 nanoseconds for a write operation with a miss in cache. Execution of a sequence of instructions involves 100 instruction fetch operations. 60 memory operand read operations

and 40 memory operand write operations. The cache - hit ratio is 0.9. The average memory access time (in nanoseconds) in executing the sequence of instructions is _____. **[2014]**

23. Assume that for a certain processor, a read request takes 50 nanoseconds on a cache miss and 5 nanoseconds on a cache hit. Suppose while running a program, it was observed that 80% of the processor's read requests result in a cache hit. The average read access time in nanoseconds is _____ **[2015]**

24. A computer system implements a 40-bit virtual address, page size of 8 kilobytes, and a 128-entry translation look-aside buffer (TLB) organized into 32 sets each having four ways. Assume that the TLB tag does not store any process id. The minimum length of the TLB tag in bits is _____ **[2015]**

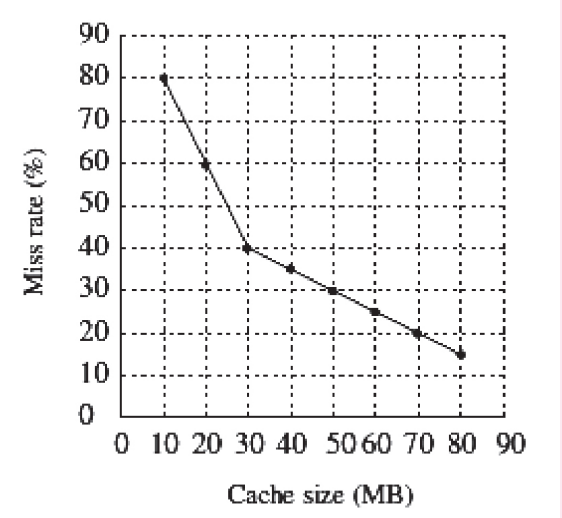
25. Consider a machine with a byte addressable main memory of 2^{20} bytes, block size of 16 bytes and a direct mapped cache having 2^{12} cache lines. Let the addresses of two consecutive bytes in main memory be $(E201F)_{16}$ and $(E2020)_{16}$. What are the tag and cache line address (in hex) for main memory address $(E201F)_{16}$? **[2015]**

(A) E, 201 (B) F, 201
(C) E, E20 (D) 2, 01F

26. The width of the physical address on a machine is 40 bits. The width of the tag field in a 512KB 8-way set associative cache is ____ bits. **[2016]**

27. A file system uses an in - memory cache to cache disk blocks. The miss rate of the cache is shown in the figure. The latency to read a block from the cache is 1 ms and to read a block from the disk is 10ms. Assume that the cost of checking whether a block exists in the cache is negligible. Available cache sizes are in multiples of 10MB.

The smallest cache size required to ensure an average read latency of less than 6 ms is ____ MB. **[2016]**



28. Consider a two-level cache hierarchy with L1 and L2 caches. An application incurs **1.4** memory accesses per instruction on average. For this application, the miss rate of L1 cache is **0.1**; the L2 cache experiences on average 7 misses per 1000 instructions. The miss rate of L2 expressed correct to two decimal places is _____. **[2017]**

29. Consider a 2-way set associative cache with 256 blocks and uses LRU replacement. Initially the cache is empty. Conflict misses are those misses which occur due to contention of multiple blocks for the same cache set. Compulsory misses occur due to first time access to the block. The following sequence of accesses to memory blocks

(0, 128, 256, 128, 0, 128, 256, 128, 1, 129, 257, 129, 1, 129, 257, 129)

is repeated 10 times. The number of *conflict misses* experienced by the cache is _____. **[2017]**

30. A cache memory unit with capacity of N words and block size of B words is to be designed. If it is designed as a direct mapped cache, the length of the TAG field is 10 bits. If the cache unit is now designed as a 16-way set-associative cache, the length of the TAG filed is _____bits. **[2017]**

31. In a two-level cache system, the access times of L_1 and L_2 caches are 1 and 8 clock cycles, respectively. The miss penalty from the L_2 cache to main memory is 18 clock cycles. The miss rate of L_1 cache is twice that of L_2. The average memory access time (AMAT) of this cache system is 2 cycles. The miss rates of L_1 and L_2 respectively are: **[2017]**

(A) 0.111 and 0.056 (B) 0.056 and 0.111
(C) 0.0892 and 0.1784 (D) 0.1784 and 0.0892

32. The read access times and the hit ratios for different caches in a memory hierarchy are as given below.

Cache	Read access time (in nanoseconds)	Hit ratio
I-cache	2	0.8
D-cache	2	0.9
L2-cache	8	0.9

The read access time of main memory is 90 nanoseconds. Assume that the caches use the referred- word-first read policy and the write back policy. Assume that all the caches are direct mapped caches. Assume that the dirty bit is always 0 for all the blocks in the caches. In execution of a program. 60% of memory reads are for instruction fetch and 40% are for memory operand fetch. The average read access time in nanoseconds (up to 2 decimal places) is _____. **[2017]**

33. Consider a machine with a byte addressable main memory of 2^{32} bytes divided into blocks of size 32

bytes. Assume that a direct mapped cache having 512 cache lines is used with this machine. The size of the tag field in bits is_____. **[2017]**

34. The size of the physical address space of a processor is 2^P bytes. The word length is 2^W bytes. The capacity of cache memory is 2^N bytes. The size of each cache block is 2^M words. For a K-way set-associative cache memory, the length (in number of bits) of the tag field is: **[2018]**

(A) $P - N - \log_2 K$
(B) $P - N + \log_2 K$
(C) $P - N - M - W - \log_2 K$
(D) $P - N - M - W + \log_2 K$

Answer Keys

Exercises

Practice Problems 1

1. D	**2.** A	**3.** D	**4.** A	**5.** B	**6.** C	**7.** D	**8.** C	**9.** B	**10.** A
11. B	**12.** C	**13.** D	**14.** B	**15.** A	**16.** A	**17.** B	**18.** A	**19.** B	**20.** A

Practice Problems 2

1. B	**2.** A	**3.** D	**4.** B	**5.** C	**6.** D	**7.** B	**8.** B	**9.** C	**10.** A
11. C	**12.** D	**13.** C	**14.** D	**15.** A	**16.** A	**17.** B	**18.** D	**19.** A	**20.** B

Previous Years' Questions

1. C	**2.** A	**3.** A	**4.** D	**5.** C	**6.** B	**7.** D	**8.** B	**9.** C	**10.** D
11. D	**12.** D	**13.** D	**14.** C	**15.** A	**16.** A	**17.** A	**18.** 20	**19.** D	
20. 10000		**21.** D	**22.** 1.68	**23.** 14	**24.** 22	**25.** A	**26.** 24	**27.** 30	**28.** 0.05
29. 76	**30.** 14	**31.** A	**32.** 4.72	**33.** 18	**34.** B				