UNIT – IV

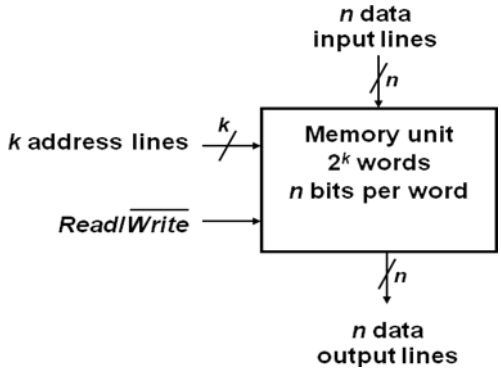
THE MEMORY SYSTEM: Some Basic Concepts, Semiconductor RAM Memories, Read-Only Memories, Speed, Size, and Cost, Cache Memories, Performance Considerations, Virtual memories, Memory Management requirements, Secondary Storage.

2 MARKS

1. What is memory system?

Every Computer contains several types of devices to store the instructions and data for its operation. These storage devices plus the algorithm implements by hardware and software needed to manage the stored information from the memory system of computer.

2. Draw the block diagram for memory unit.



3. Define addressing scheme?

Addressing scheme is a scheme used in any computer to determine and maximum size of the memory.

4. What are the two registers involved in data transfer between the memory and the processor?

The registers used to transfer data are,

MAR (Memory address register)

MDR (Memory data register).

5. Give the classification of memory (Apr 11)

a. CPU Register

b. Main memory

c. Secondary Memory

d. Cache.

6. Define Static Memories and Dynamic Memories.

Memories that consist of circuits capable of retaining their state as long as power is applied are known as static memories. In Dynamic Memories such cells do not retain their state indefinitely.

7. What is read access time?

A basic performance measure is the average time to read a fixed amount of information for instance, one word from the memory. This parameter is called the read access time.

8. Define RAM

In storage location can be accessed in any order and access time is independent of the location being accessed, the memory is termed as random access memory

9. What are ROM & PROMs?

Memories whose content cannot be altered online if they can be altered at all are read only memories.

Semi conductor ROMs whose contents can be changed offline-with some difficulties is called PROMs.

10. Difference between static RAM and Dynamic RAM.

S.no	STATIC RAM	DYNAMIC RAM
1.	They are fast	They are slow
2.	They are very expensive	They are less expensive
3.	They require several transistors	They require less no several transistors
4.	They retain their state indefinitely	They do not retain their state indefinitely

11. Differentiate asynchronous DRAM with synchronous DRAM?

S.no.	asynchronous DRAM	synchronous DRAM
1.	The timing of the memory device is controlled asynchronously	The timing of the memory device controlled synchronously.
2	There is specialized memory controlled circuit that provides the necessary control information	The memory operations are synchronized with a clock signals.
3	Separate refresh circuit is not used	It uses the separate refresh circuit.

12. List the differences between SRAM AND DRAM?

SRAM: Static random access memory. It tends to be faster and they require no refreshing

DRAM: Dynamic random access memory. Data is stored in the form of charges. So continuous refreshing is needed.

13. What is volatile memory?

A memory is volatile if the loss of power destroys the stored information. Information can be stored indefinitely in a volatile memory by providing battery backup or other means to maintain a continuous supply of power.

14. What are the types of memory? / What are the categories of memories?

- SRAM (Satic Random Access Memory)
- DRAM (Dynamic Random Access Memory)

15. What is flash memory?

A recent semiconductor technology called flash memory of a same non-volatility as a PROM, but it can be done a bit at a time.

16. What is cache memory? (Apr 11)

Memory word are stored in cache data memory and are grouped into small pages called cache blocks or line. The contents of the caches data memory are thus copies of a set of main memory blocks.

17. Mention two system organizations for caches.

Two system organizations for caches are

- look aside
- look through

18. What is RAMBUS memory?

The key feature of Rambus technology is a fast signaling method used to transfer information between chip using narrow bus.

19. What is write-through protocol?

For write operation, the cache location and the main memory location are updated simultaneously

20. Give the difference between EEPROM and Flash memory?

The primary difference between EEPROM and flash memory is that flash restricts writes to multiple kilobytes blocks, increasing the memory capacity per chip by reducing area of control.

21. Differences between cache memory and virtual memory

- In caches, replacement is primarily controlled by the hardware. In VM, replacement is primarily controlled by the os.
- The Number of bits in the address determines the size of VM, where as cache size is independent of the address size.
- But there is only one class of cache.

22. Uses of Virtual Memory.

Protection: VM is often used to protect one program from others in the system

Base and Bounds: this method allows relocation. User processes cannot be allowed to change these registers, but the OS must be able to do so on a process switch.

23. What is meant by Interleaved Memory Or What is memory interleaving?

Banks of memory are often one word wide , so bus width need not be changed to access memory. However several independent areas of memory can be accessed simultaneously by using interleaved memory.

24. What is write back protocol?

In this scheme, only the block in cache is modified. The main memory when the block must be replaces in the cache. This requires the use of a dirty bit to keep track of blocks, that have been modified.

25. What is virtual memory and what are the benefits of virtual memory?

Virtual memory is a computer system technique which gives an application program the impression that it has contiguous working memory (an address space), while in fact it may be physically fragmented and may even overflow on to disk storage.

Benefits

- Programs can be larger than physical memory
- Entire program need not be in memory

26. Give the features of a ROM cell

- ROM is Read only memory
- It is usually non-volatile memory meaning that when you turn power off to the electronic device the ROM memory retains its contents.
- This is typically found on computer mother boards were start-up instructions are installed.
- ROM is permanent memory, so it never loses what is stored in it.

27. Define Locality of Reference.

Many applications continually reference large amounts of data. Often, the link to this data is slow, as in the cases of primary or secondary memory references, database queries, or network requests. This leads to poor performance in the application. To improve application efficiency by exploiting the principle of Locality of Reference also Known as Caching

28. What is Translation Look aside Buffer? Or what is TLB?

A translation look aside buffer (TLB) is a cache that memory management hardware uses to improve virtual address translation speed. All current desktop, notebook, and server processors use a TLB to map virtual and physical address spaces, and it is nearly always present in any hardware which utilizes virtual memory.

29. Define data transfer rate.

The data transfer rate (DTR) is the amount of digital data that is moved from one place to another in a given time.

30. Define memory access time?

The time required to access one word is called the memory access time. Or It is the time that elapses between the initiation of an operation and the completion of that operation.

31. Define memory cycle time?

It is the minimum time delay required between the initiations of two successive memory operations.

Eg. The time between two successive read operations.

32. When is a memory unit called as RAM?

A memory unit is called as RAM if any location can be accessed for a read or writes operation in some fixed amount of time that is independent of the location's address.

33. What is MMU?

MMU is the Memory Management Unit. It is a special memory control circuit used for implementing the mapping of the virtual address space onto the physical memory.

- Programs can be larger than physical memory
- Entire program need not be in memory

34. Define memory cell?

A memory cell is capable of storing one bit of information. It is usually organized in the form of an array.

35. What is a word line?

In a memory cell, all the cells of a row are connected to a common line called as word line.

36. What are the Characteristics of semiconductor RAM memories?

- They are available in a wide range of speeds.
- Their cycle time range from 100ns to less than 10ns.
- They replaced the expensive magnetic core memories.
- They are used for implementing memories.

37. Why SRAMs are said to be volatile?

Because of their contents are lost when power is interrupted. So SRAMs are said to be volatile.

38. What are the Characteristics of SRAMs?

- SRAMs are fast.
- They are volatile.
- They are of high cost.
- Less density.

39. What are the Characteristics of DRAMs?

- Low cost.
- High density.
- Refresh circuitry is needed.

40. Define Refresh Circuit?

It is a circuit which ensures that the contents of a DRAM are maintained when each row of cells are accessed periodically.

41. Define Memory Latency?

It is used to refer to the amount of time it takes to transfer a word of data to or from the memory.

42. What are asynchronous DRAMs?

In asynchronous DRAMs, the timing of the memory device is controlled asynchronously. A specialized memory controller circuit provides the necessary control signals RAS and CAS that govern the timing. The processor must take into account the delay in the response of the memory. such memories are asynchronous DRAMs .

43. What are synchronous DRAMs?

Synchronous DRAMs are those whose operation is directly synchronized with a clock signal.

44. Define Bandwidth?

When transferring blocks of data, it is of interest to know how much time is needed to transfer an entire block. Since blocks can be variable in size it is useful to define a performance measure in terms of number of bits or bytes that can be transferred in one second. This measure is often referred to as the memory bandwidth.

45. What is double data rate SDRAMs? Or what is DDR SDRAM?

Double data rates SDRAMs are those which can transfer data on both edges of the clock and their bandwidth is essentially doubled for long burst transfers.

46. Differentiate static RAM and dynamic RAM?

Static RAM

- They are fast
- They are very expensive
- They retain their state indefinitely.
- They require several transistors
- Low density

Dynamic RAM

- They are slow
- They are less expensive
- They don't retain their state indefinitely
- They require less no transistors.
- High density

47. What is Ram Bus technology?

The key feature of Ram bus technology is a fast signaling method used to transfer information between chips. Instead of using signals that have voltage levels of either 0 or Vsupply to represent the logic

values, the signals consist of much smaller voltage swings around a reference voltage , vref. Small voltage swings make it possible to have short transition times, which allows for a high speed of transmission.

48. What are RDRAMs?

RDRAMs are Rambus DRAMs. Rambus requires specially designed memory chips. These chips use cell arrays based on the standard DRAM technology. Multiple banks of cell arrays are used to access more than one word at a time. Circuitry needed to interface to the Rambus channel is included on the chip. Such chips are known as RDRAMs.

49. What are the special features of Direct RDRAMs?

- It is a two channel Rambus.
- It has 18 data lines intended to transfer two bytes of data at a time.
- There are no separate address lines.

50. What are the disadvantages of EPROM?

The chip must be physically removed from the circuit for reprogramming and its entire contents are erased by the ultraviolet light.

51. Differentiate flash devices and EEPROM devices.

Flash devices

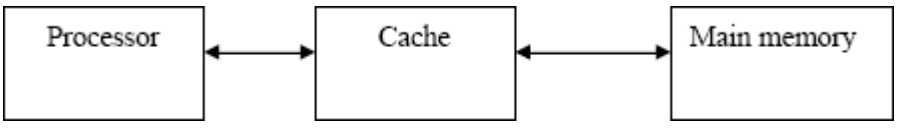
- It is possible to read the contents of a single cell, but it is only possible to write an entire block of cells.
- Greater density which leads to higher capacity.
- Lower cost per bit.
- Consumes less power in their operation and makes it more attractive for use in portable equipments that is battery driven.

EEPROM devices

- It is possible to read and write the contents of a single cell.
- Relatively more cost
- Consumes more power.

52. What is cache memory?

It is a small, fast memory that is inserted between the larger, slower main memory and the processor. It reduces the memory access time.



53. Define flash memory?

It is an approach similar to EEPROM technology. A flash cell is based on a single transistor controlled by trapped charge just like an EEPROM cell.

54. What are the two aspects of locality of reference? Define them.

Two aspects of locality of reference are temporal aspect and spatial aspect. Temporal aspect is that a recently executed instruction is likely to be executed again very soon.

The spatial aspect is that instructions in close proximity to a recently executed instruction are also to be executed soon.

55. Define cache line.

Cache block is used to refer to a set of contiguous address locations of some size. Cache block is also referred to as cache line.

56. What are the two ways in which the system using cache can proceed for a write operation?

- Write through protocol technique.
- Write-back or copy back protocol technique.

57. What is write-through protocol?

For a write operation using write through protocol during write hit: the cache location and the main memory location are updated simultaneously.

For a write miss, the information is written directly to the main memory.

58. What is write-back or copy back protocol?

For a write operation using this protocol during **write hit:** the technique is to update only the cache location and to mark it as updated with an associated flag bit, often called the dirty or modified bit. The main memory location of the word is updated later, when the block containing this marked word is to be removed from the cache to make room for a new block.

For a write miss: the block containing the addressed word is first brought into the cache, and then the desired word in the cache is overwritten with the new information.

59. What are the mapping technique? Discuss the different mapping techniques used in cache memory

- Direct mapping
- Associative mapping
- Set Associative mapping

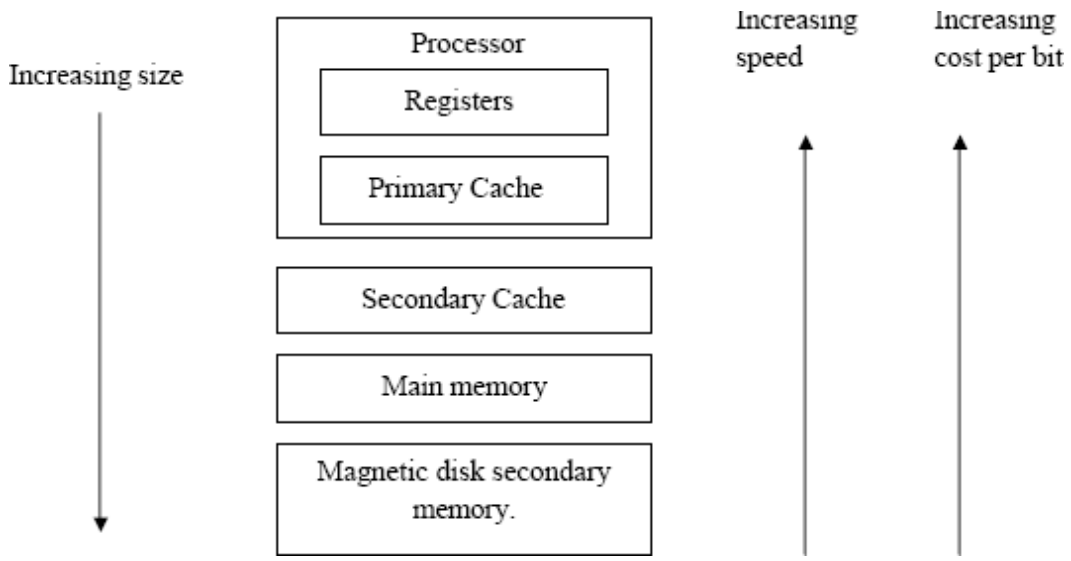
60. What is a hit?

A successful access to data in cache memory is called hit.

61. Define hit rate?

The number of hits stated as a fraction of all attempted access.

62. Describe the memory hierarchy?



63. Define miss rate?

It is the number of misses stated as a fraction of attempted accesses.

64. Define access time for magnetic disks?

The sum of seek time and rotational delay is called as access time for disks. Seek time is the time required to move the read/write head to the proper track. Rotational delay or latency is the amount of time that elapses after the head is positioned over the correct track until the starting position of the addressed sector passes under the read/write head.

65. What is the formula for calculating the average access time experienced by the processor?

t_{ave}=hc +(1-h)M

Where,

h =Hit rate

M=miss penalty

C=Time to access information in the cache.

66. What is the formula for calculating the average access time experienced by the processor in a system with two levels of caches?

tave =h1c1(1-h1)h2c2+(1-h1)(1-h2)M

where,

h1=hit rate in L1 cache

h2=hit rate in L2 cache

C1=Time to access information in the L1 cache.

C2=Time to access information in the L2 cache.

67. What are pages?

All programs and data are composed of fixed length units called pages. each consists of blocks of words that occupies contiguous locations in main memory.

68. What is replacement algorithm?

When the cache is full and a memory word that is not in the cache is referenced, the cache control hardware must decide which block should be removed to create space for the new block that contains the reference word .The collection of rules for making this decision constitutes the replacement algorithm.

69. What is meant by internal and external fragmentation?

Fragmentation occurs in a dynamic memory allocation system when many of the free blocks are too small to satisfy any request.

External Fragmentation: External Fragmentation happens when a dynamic memory allocation algorithm allocates some memory and a small piece is left over that cannot be effectively used. If too much external fragmentation occurs, the amount of usable memory is drastically reduced. Total memory space exists to satisfy a request, but it is not contiguous.

Internal Fragmentation: Internal fragmentation is the space wasted inside of allocated memory blocks because of restriction on the allowed sizes of allocated blocks. Allocated memory may be slightly larger than requested memory; this size difference is memory internal to a partition, but not being used

70. What is content addressable memory and what are the advantages of this memory?

Content-addressable memory, also referred to as associative memory or abbreviated CAM, is a mechanism for storing information that can be retrieved based on its content, not its storage location.

Advantage

It is typically used for high-speed storage and retrieval of fixed content, such as documents stored for compliance with government regulations.

71. Define interleaving

Cell array can be organized in two banks. Each bank can be accessed separately. So the consecutive words of given block are stored in different banks. It is known as interleaving of words. It increases the transfer rate.

72. Define SIMM and DIMM

SIMM –> Single In – line memory modules.

DIMM-> Dual In – line memory modules.

SIMM means single in line memory module and DIMM means dual in line memory module. These two large memories are created by using the DRAM. This module is an assembly of several memory chips on a separate small board that plugs vertically into a single socket on the mother board.

73.What is Memory controller?

A memory controller is a circuit which is interposed between the processor and the dynamic memory. It is used for performing multiplexing of address bits.

74. Draw backs present in the DRAM.

All dynamic memories have to be refreshed and it does not have a refreshing capability. So the memory controller has to provide all the information needed to control the refreshing operation. This increases the over head of the controller circuit.

75. How many memory chips are needed to construct 2M*16 memory system using 512K * 8 static memory chips?

4096/512 =16 chips 16 memory chips are needed to construct 2M*16 memory system using 512K * 8 static memory chips

76. How is disk access time calculated?

Disk access time can be calculated by adding up the 'seek time' and the average 'latency time'

- **Seek time**: the time needed for the read/write arm to look for the desired track
- **Latency time** (also called rotational delay): the time it takes for the desired sector on the track to spin around to the read/write arm (since the piece of data required might be on the other side of the disk relative to the read/write at the moment).
- **Transfer time**: the time a hard disk drive needs to read and transmit one block of data
- Disk access time in hard disk drives is measured in milliseconds
- Even though this might seem fast, CPUs are still able to calculate much faster than this, causing hard disk drives to be slow in comparison

Disk access time = seek time + latency time + transfer time

77. What is the use of EEPROM?

- EEPROM stands for Electrically Erasable Programmable Read-Only Memory
- It is a type of non-volatile memory used in computers and other electronic devices to store small amounts of data that must be saved when power is removed
- e.g., calibration tables or device configuration.

78. State the hardware needed to implement the LRU in replacement algorithm.

- The hardware is equipped with a counter (typically 64 bits). After each instruction the counter is incremented.
- In addition, each page table entry has a field large enough to accommodate the counter. Every time the page is referenced the value from the counter is copied to the page table field. When a page fault occurs the operating system inspects all the page table entries and selects the page with the lowest counter. This is the page that is evicted as it has not been referenced for the longest time.

79. What is DDR SDRAM?

- Double data rate synchronous dynamic random-access memory (DDR SDRAM) is a class of memory integrated circuits used in computers.
- The SDRAM transfer data on the both edges of the clock, their bandwidth is essentially doubled for long burst transfers. Such devices are known as double-data-rate SDRAMs.

80. An address space is specified by 24 bits and the corresponding memory space by 16 bits: How many words are there in the virtual memory and in the main memory?OR

An address space is specified by 24 bits & the corresponding memory space is 16 bits.

a) How many words are there in address space?

b) How many words are there in memory space?

c) If a page has 2k words, how many pages & blocks are in the system?

Solution:-

a) Address space = 24 bits

224 = 24.220 = 16M words

b) Memory space: 16 bits

216 = 64k words

c) page consists of 2k words

Number of pages in add space = 16M/2K = 8000

Number of blocks = 64k/2k = 32 blocks

81. What is data stripping?

In a write system a single large file is stored in several separate disk units by breaking the file up in to a number of small pieces and stored these pieces on different disk.

82. Define hit ratio (Nov 13)

Hit rate=No of hits/no of bus cycles *100%

83. Define the terms hit, miss and ratio with respect to cache

Cache is a small high-speed memory. Stores data from some frequently used addresses (of main memory).

Cache hit: Data found in cache. Results in data transfer at maximum speed.

Cache miss: Data not found in cache. Processor loads data from M and copies into cache. This results in extra delay, called miss penalty.

Hit ratio = percentage of memory accesses satisfied by the cache.

Miss ratio = 1-hit ratio

84. Which type of memory provides backup storage? (Nov 12)

External storage provides data back-up. It includes floppy disks, tapes.

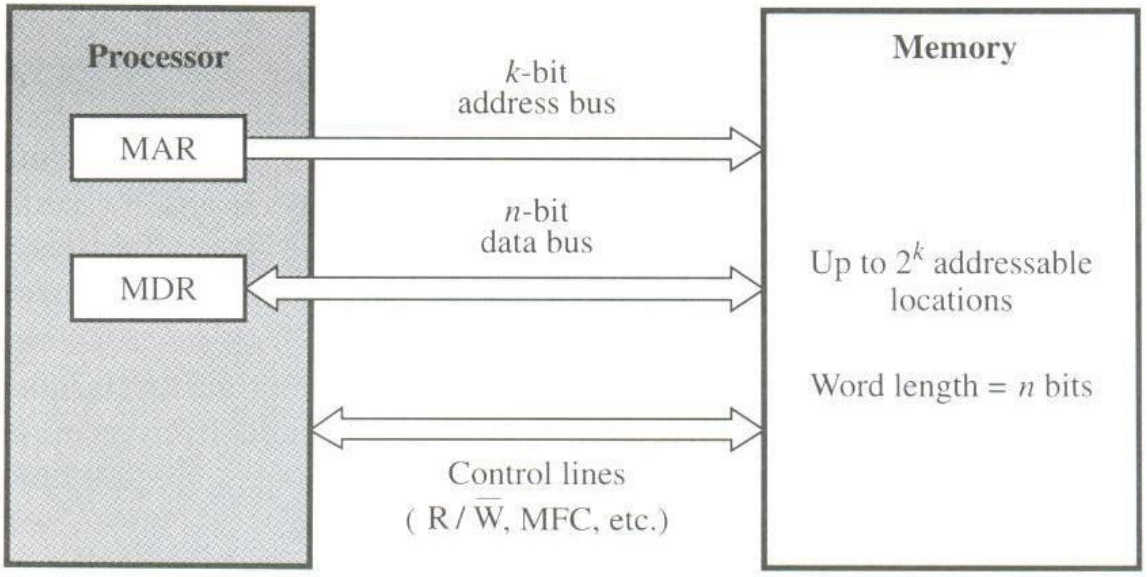
11 MARKS

1. Describe in detail the concepts of memory.

The maximum size of the memory that can be used in any computer is determined by the addressing scheme.

Address	**Memory Locations**
16 Bit	216 = 64 K
32 Bit	232 = 4G (Giga)
40 Bit	240 = IT (Tera)

Fig: Connection of Memory to Processor:



If MAR is k bits long and MDR is n bits long, then the memory may contain upto 2K addressable locations and the n-bits of data are transferred between the memory and processor. This transfer takes place over the processor bus.

The processor bus has,

- Address Line
- Data Line
- Control Line (R/W, MFC – Memory Function Completed)

➢ The control line is used for co-ordinating data transfer.

- The processor reads the data from the memory by loading the address of the required memory location into MAR and setting the R/W line to 1.
- The memory responds by placing the data from the addressed location onto the data lines and confirms this action by asserting MFC signal.
- Upon receipt of MFC signal, the processor loads the data onto the data lines into MDR register.
- The processor writes the data into the memory location by loading the address of this location into MAR and loading the data into MDR sets the R/W line to 0.

Memory Access Time → It is the time that elapses between the intiation of an Operation and the completion of that operation.

Memory Cycle Time → It is the minimum time delay that required between the initiation of the two successive memory operations.

In RAM, if any location that can be accessed for a Read/Write operation in fixed amount of time, it is independent of the location" s address.

Cache Memory:

- It is a small, fast memory that is inserted between the larger slower main memory and the processor.
- It holds the currently active segments of a pgm and their data.

Virtual memory:

- The address generated by the processor does not directly specify the physical locations in the memory.
- The address generated by the processor is referred to as a virtual / logical address.
- The virtual address space is mapped onto the physical memory where data are actually stored.
- The mapping function is implemented by a special memory control circuit is often called the memory management unit.
- Only the active portion of the address space is mapped into locations in the physical memory.
- The remaining virtual addresses are mapped onto the bulk storage devices used, which are usually magnetic disk.
- As the active portion of the virtual address space changes during program execution, the memory management unit changes the mapping function and transfers the data between disk and memory.
- Thus, during every memory cycle, an address processing mechanism determines whether the addressed in function is in the physical memory unit.
- If it is, then the proper word is accessed and execution proceeds.
- If it is not, a page of words containing the desired word is transferred from disk to memory.
- This page displaces some page in the memory that is currently inactive.

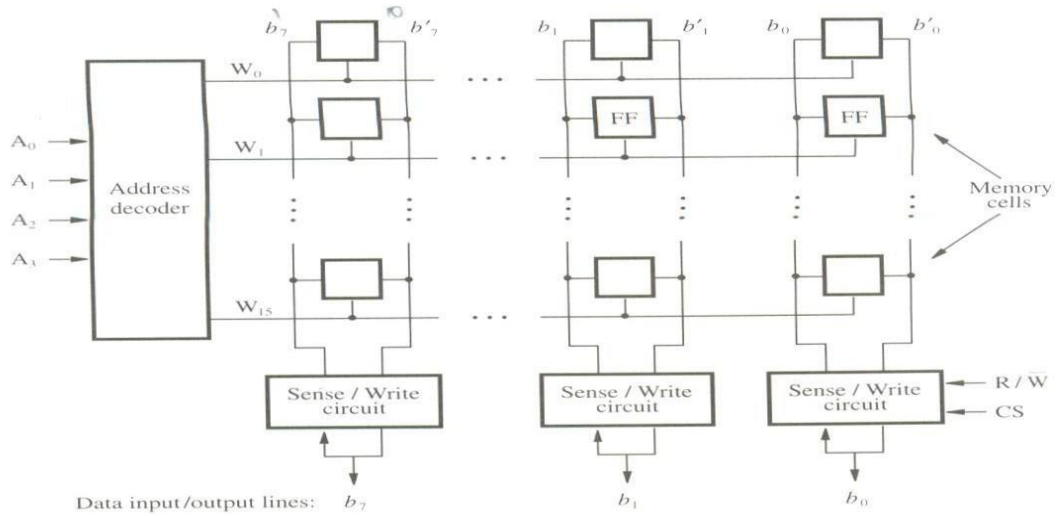
2. Describe in detail about Semiconductor RAM memories (Apr 11)

Semi-Conductor memories are available is a wide range of speeds. Their cycle time ranges from 100ns to 10ns

Internal Organization Of Memory Chips

- Memory cells are usually organized in the form of array, in which each cell is capable of storing one bit of information.
- Each row of cells constitute a memory word and all cells of a row are connected to a common line called as word line.
- The cells in each column are connected to Sense / Write circuit by two bit lines.
- The Sense / Write circuits are connected to data input or output lines of the chip.
- During a write operation, the sense / write circuit receive input information and store it in the cells of the selected word.

Fig: Organization of bit cells in a memory chip



- The data input and data output of each senses / write ckt are connected to a single bidirectional data line that can be connected to a data bus of the cptr.

R / W - Specifies the required operation.

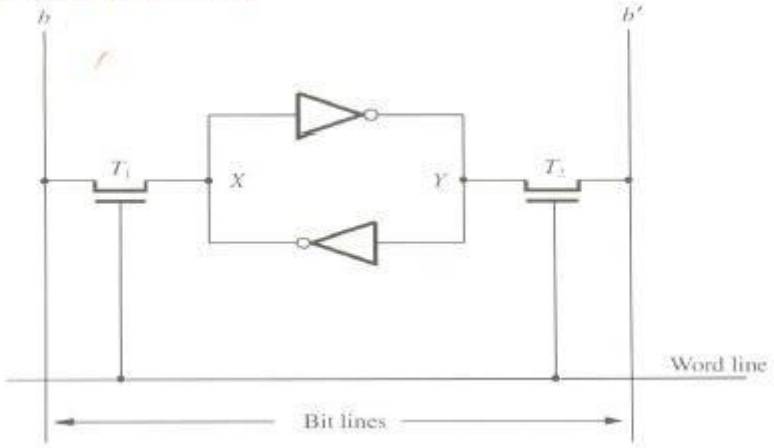
CS - **Chip Select** input selects a given chip in the multi-chip memory system

Bit Organization	Requirement of external connection for address, data and control lines
128 (16x8)	14
(1024) 128x8(1k)	19

Static Memories:

Memories that consist of circuits capable of retaining their state as long as power is applied are known as static memory.

Fig:Static RAM cell



- Two inverters are cross connected to form a batch
- The batch is connected to two bit lines by transistors T1 and T2.
- These transistors act as switches that can be opened / closed under the control of the word line.
- When the wordline is at ground level, the transistors are turned off and the latch retain its state.

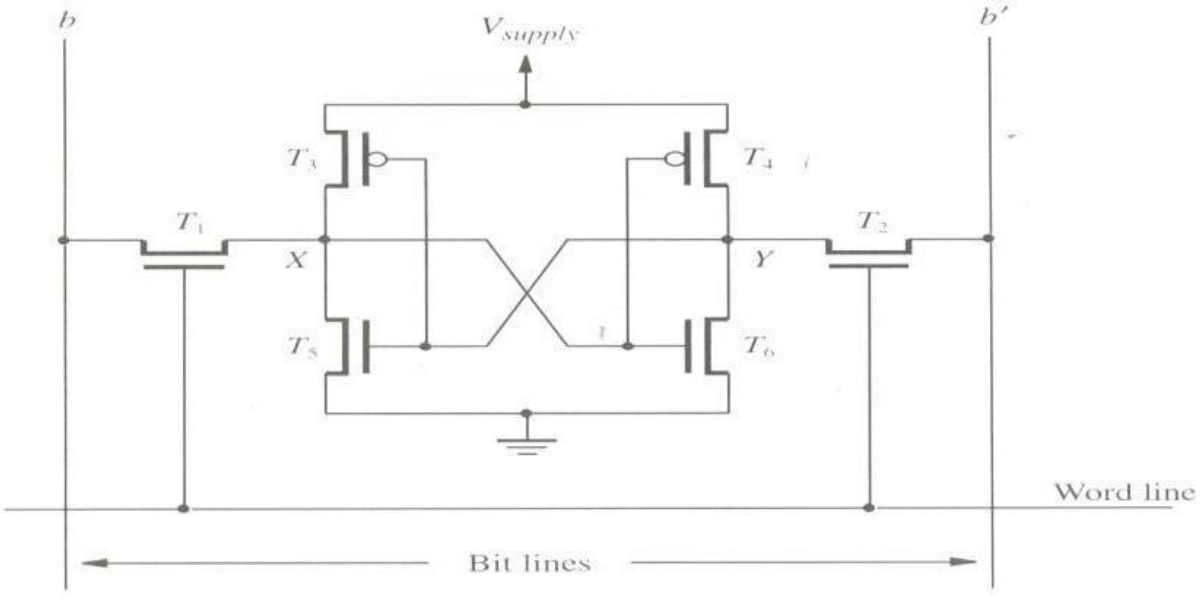
Read Operation:

- In order to read the state of the SRAM cell, the word line is activated to close switches T1 and T2.
- If the cell is in state 1, the signal on bit line b is high and the signal on the bit line b is low.Thus b and b are complement of each other.
- Sense / write circuit at the end of the bit line monitors the state of b and b‟ and set the output accordingly.

Write Operation:

- The state of the cell is set by placing the appropriate value on bit line b and its complement on b and then activating the word line. This forces the cell into the corresponding state.
- The required signal on the bit lines are generated by Sense / Write circuit.

Fig:CMOS cell (Complementary Metal oxide Semi Conductor):



- Transistor pairs (T3, T5) and (T4, T6) form the inverters in the latch.
- In state 1, the voltage at point X is high by having T5, T6 on and T4, T5 are OFF.
- Thus T1, and T2 returned ON (Closed), bit line b and b will have high and low signals respectively.
- The CMOS requires 5V (in older version) or 3.3.V (in new version) of power supply voltage.
- The continuous power is needed for the cell to retain its state

Merit :

It has low power consumption because the current flows in the cell only when the cell is being activated accessed.

Static RAM" s can be accessed quickly. It access time is few nanoseconds.

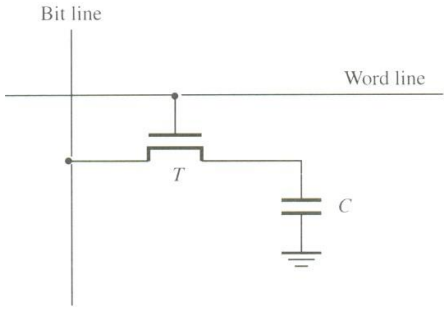
Demerit:

SRAM" s are said to be volatile memories because their contents are lost when the power is interrupted.

Asynchronous DRAMS:-

- Less expensive RAM" s can be implemented if simplex calls are used such cells cannot retain their state indefinitely. Hence they are called Dynamic RAM's (DRAM).
- The information stored in a dynamic memory cell in the form of a charge on a capacitor and this charge can be maintained only for tens of Milliseconds.
- The contents must be periodically refreshed by restoring by restoring this capacitor charge to its full value.

Fig:A single transistor dynamic Memory cell

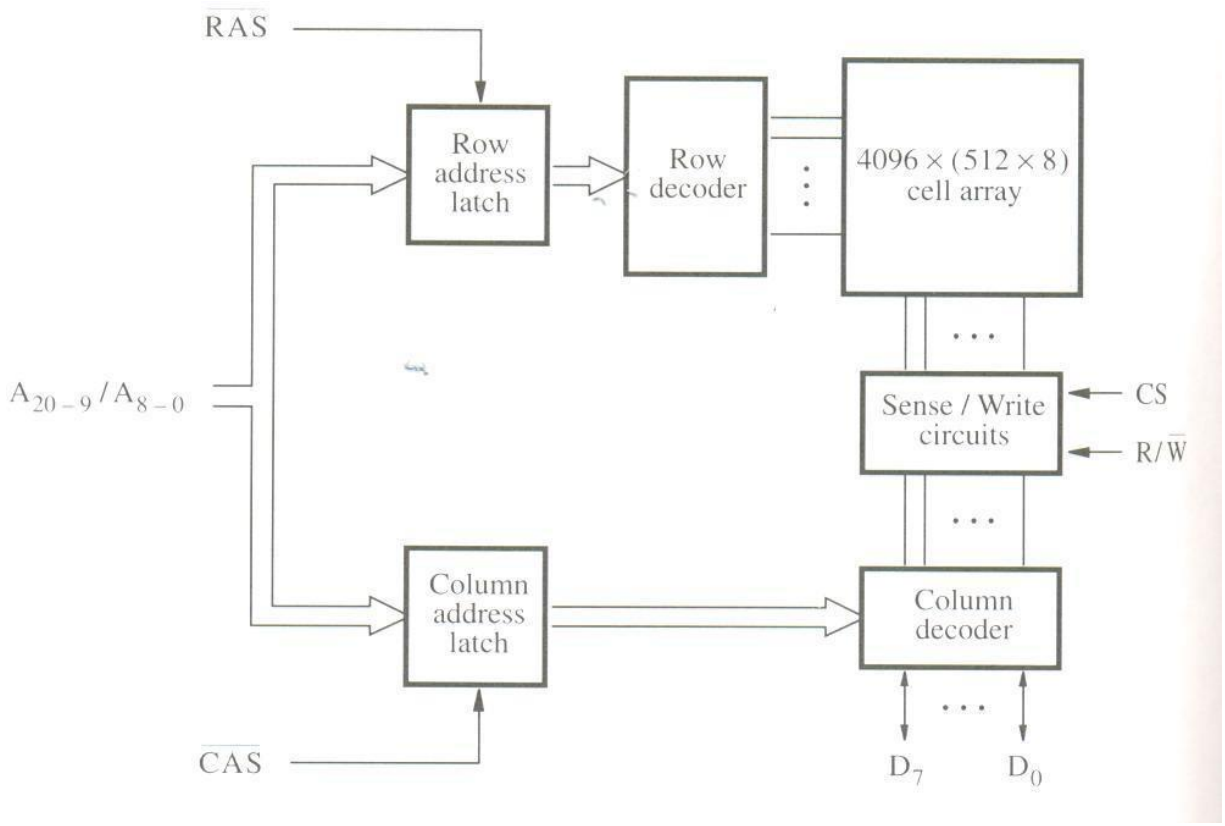


- In order to store information in the cell, the transistor T is turned „on" & the appropriate voltage is applied to the bit line, which charges the capacitor.
- After the transistor is turned off, the capacitor begins to discharge which is caused by the capacitor" s own leakage resistance.
- Hence the information stored in the cell can be retrieved correctly before the threshold value of the capacitor drops down.

- During a read operation, the transistor is turned „on" & a sense amplifier connected to the bit line detects whether the charge on the capacitor is above the threshold value.
  - If charge on capacitor > threshold value -> Bit line will have logic value „1" .

- If charge on capacitor < threshold value -> Bit line will set to logic value „0" .

Fig:Internal organization of a 2M X 8 dynamic Memory chip.



DESCRIPTION:

- The 4 bit cells in each row are divided into 512 groups of 8.
- 21 bit address is needed to access a byte in the memory(12 bit To select a row,9 bit Specify the group of 8 bits in the selected row).
- A8-0 Row address of a byte.
- A20-9 Column address of a byte.
- During Read/ Write operation ,the row address is applied first. It is loaded into the row address latch in response to a signal pulse on Row Address Strobe(RAS) input of the chip.
- When a Read operation is initiated, all cells on the selected row are read and refreshed.
- Shortly after the row address is loaded,the column address is applied to the address pins & loaded into Column Address Strobe(CAS).
- The information in this latch is decoded and the appropriate group of 8 Sense/Write circuits are selected.
- R/W =1(read operation) The output values of the selected circuits are transferred to the data lines D0 - D7.
- R/W =0(write operation) The information on D0- D7 are transferred to the selected circuits.
- RAS and CAS are active low so that they cause the latching of address when they change from high to low. This is because they are indicated by RAS & CAS.

- To ensure that the contents of a DRAM „s are maintained, each row of cells must be accessed periodically.
- Refresh operation usually perform this function automatically.
- A specialized memory controller circuit provides the necessary control signals RAS & CAS, that govern the timing.
- The processor must take into account the delay in the response of the memory. Such memories are referred to as Asynchronous DRAM's**.**

Fast Page Mode:

- Transferring the bytes in sequential order is achieved by applying the consecutive sequence of column address under the control of successive CAS signals.
- This scheme allows transferring a block of data at a faster rate. The block of transfer capability is called as Fast Page Mode.

Synchronous DRAM:

- Here the operations e directly synchronized with clock signal.
- The address and data connections are buffered by means of registers.
- The output of each sense amplifier is connected to a latch.
- A Read operation causes the contents of all cells in the selected row to be loaded in these latches
- Data held in the latches that correspond to the selected columns are transferred into the data output register, thus becoming available on the data output pins.

.

Fig:Synchronous DRAM

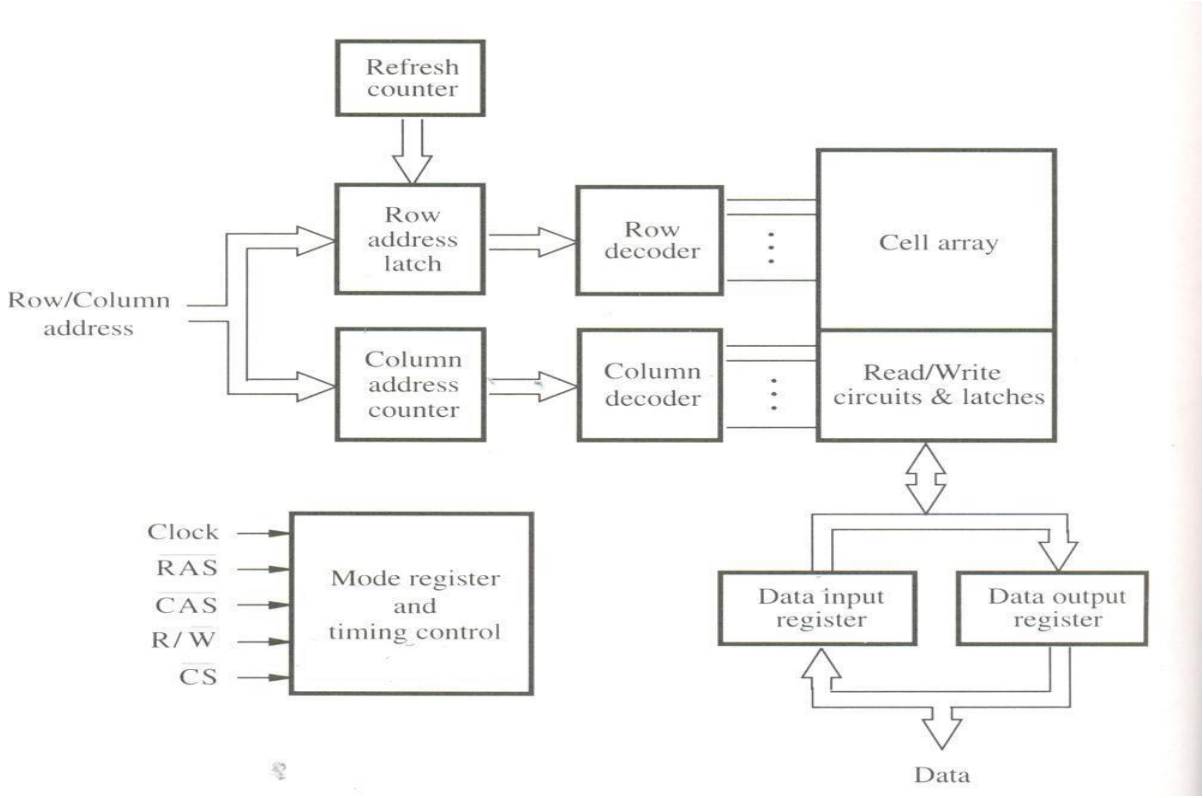
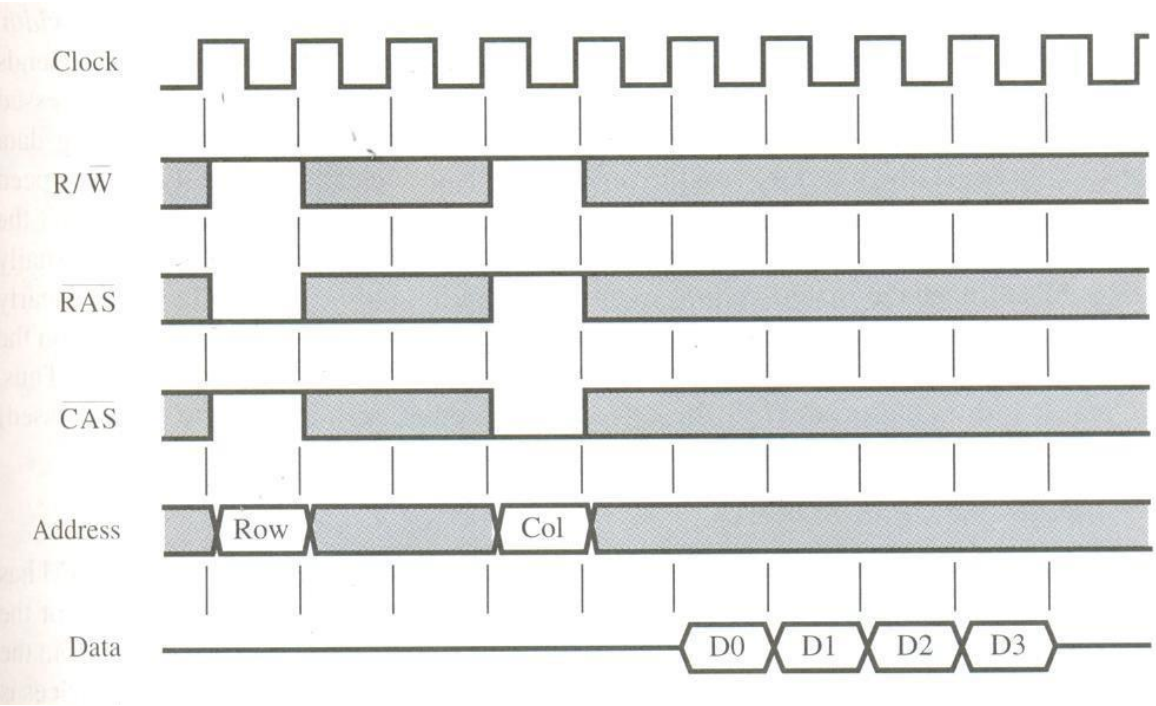


Fig:Timing Diagram Burst Read of Length 4 in an SDRAM



- First ,the row address is latched under control of RAS signal.
- The memory typically takes 2 or 3 clock cycles to activate the selected row.
- Then the column address is latched under the control of CAS signal.
- After a delay of one clock cycle,the first set of data bits is placed on the data lines.
- The SDRAM automatically increments the column address to access the next 3 sets of bits in the selected row, which are placed on the data lines in the next 3 clock cycles.

Latency & Bandwidth:

A good indication of performance is given by two parameters.They are,

- Latency
- Bandwidth

Latency:

It refers to the amount of time it takes to transfer a word of data to or from the memory.

For a transfer of single word,the latency provides the complete indication of memory performance.

For a block transfer,the latency denote the time it takes to transfer the first word of data.

Bandwidth:

It is defined as the number of bits or bytes that can be transferred in one second

Bandwidth mainly depends upon the speed of access to the stored data & on the number of bits that can be accessed in parallel.

Double Data Rate SDRAM(DDR-SDRAM):

- The standard SDRAM performs all actions on the rising edge of the clock signal.
- The double data rate SDRAM transfer data on both the edges(loading edge, trailing edge).

- The Bandwidth of DDR-SDRAM is doubled for long burst transfer.
- To make it possible to access the data at high rate , the cell array is organized into two banks.
- Each bank can be accessed separately.
- Consecutive words of a given block are stored in different banks.
- Such interleaving of words allows simultaneous access to two words that are transferred on successive edge of the clock.

Larger Memories:

Dynamic Memory System:

- The physical implementation is done in the form of Memory Modules.
- If a large memory is built by placing DRAM chips directly on the main system printed circuit board that contains the processor ,often referred to as Motherboard;it will occupy large amount of space on the board.
- These packaging consideration have led to the development of larger memory units known as SIMM‟s & DIMM‟s .

- SIMM-Single Inline memory Module
- DIMM-Dual Inline memory Module
- SIMM & DIMM consists of several memory chips on a separate small board that plugs vertically into single socket on the motherboard.

MEMORY SYSTEM CONSIDERATION:

To reduce the number of pins, the dynamic memory chips use multiplexed address inputs.

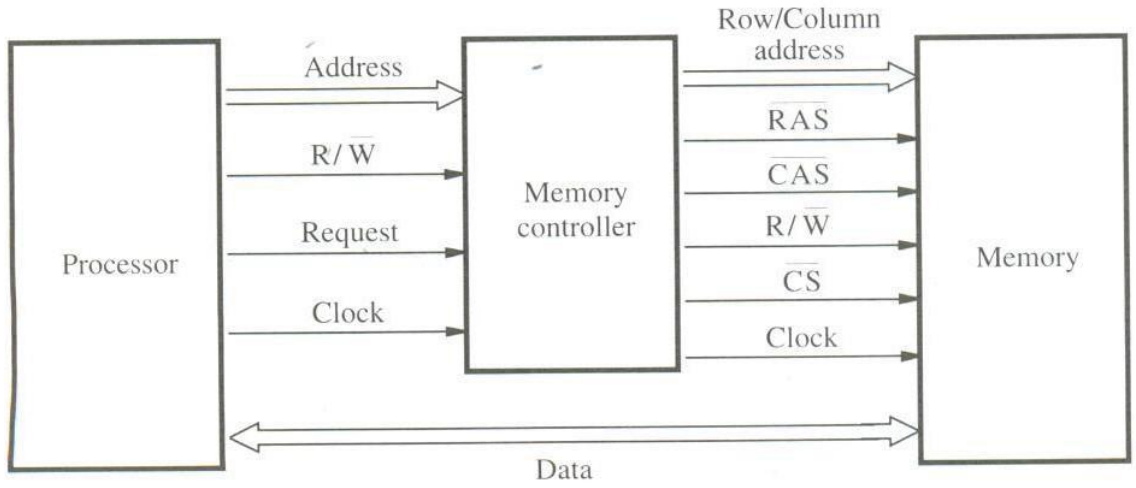
The address is divided into two parts.They are,

High Order Address Bit(Select a row in cell array & it is provided first and latched into memory chips under the control of RAS signal).

Low Order Address Bit(Selects a column and they are provided on same address pins and latched using CAS signals).

The Multiplexing of address bit is usually done by Memory Controller Circuit.

Fig:Use of Memory Controller



- The Controller accepts a complete address & R/W signal from the processor, under the control of a Request signal which indicates that a memory access operation is needed.
- The Controller then forwards the row & column portions of the address to the memory and generates RAS &CAS signals.
- It also sends R/W &CS signals to the memory.
- The CS signal is usually active low, hence it is shown as CS.

Refresh Overhead:

All dynamic memories have to be refreshed.

In DRAM ,the period for refreshing all rows is 16ms whereas 64ms in SDRAM.

Eg:Given a cell array of 8K(8192).

Clock cycle=4

Clock Rate=133MHZ

No of cycles to refresh all rows =8192*4

=32,768

Time needed to refresh all rows=32768/133*106

=246*10-6 sec

=0.246sec

Refresh Overhead =0.246/64

Refresh Overhead =0.0038

Rambus Memory:

- The usage of wide bus is expensive.
- Rambus developed the implementation of narrow bus.
- Rambus technology is a fast signaling method used to transfer information between chips.
- Instead of using signals that have voltage levels of either 0 or Vsupply to represent the logical values, the signals consists of much smaller voltage swings around a reference voltage Vref.

- .The reference Voltage is about 2V and the two logical values are represented by 0.3V swings above and below Vref..
- This type of signaling is generally is known as Differential Signalling.
- Rambus provides a complete specification for the design of communication links(Special Interface circuits) called as Rambus Channel.
- Rambus memory has a clock frequency of 400MHZ.
- The data are transmitted on both the edges of the clock so that the effective data transfer rate is 800MHZ.
- The circuitry needed to interface to the Rambus channel is included on the chip.Such chips are known as Rambus DRAM‟ s(RDRAM).
- Rambus channel has,
    - 9 Data lines(1-8 Transfer the data,9th line Parity checking).
    - Control line
    - Power line

A two channel rambus has 18 data lines which has no separate address lines.It is also called as

Direct RDRAM's.

Communication between processor or some other device that can serves as a master and RDRAM modules are serves as slaves ,is carried out by means of packets transmitted on the data lines.

There are 3 types of packets.They are,

□ Request
□ Acknowledge
□ Data

3. Describe in detail about Read-Only Memories

READ ONLY MEMORY:

Both SRAM and DRAM chips are volatile,which means that they lose the stored information if power is turned off.

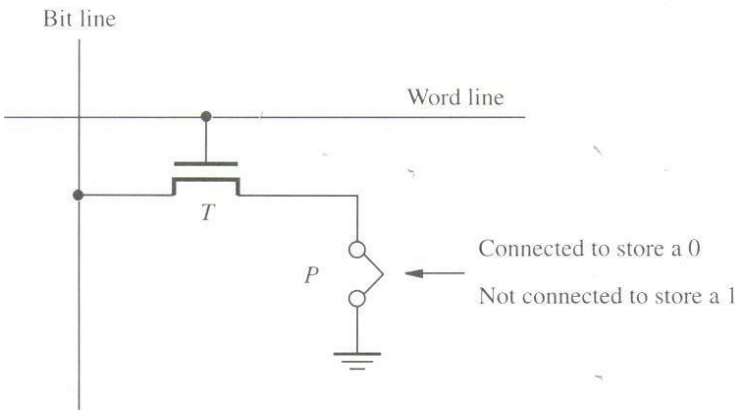
Many application requires Non-volatile memory (which retain the stored information if power is turned off).

Eg:Operating System software has to be loaded from disk to memory which requires the program that boots the Operating System ie. It requires non-volatile memory.

Non-volatile memory is used in embedded system.

Since the normal operation involves only reading of stored data ,a memory of this type is called ROM

Fig:ROM cell



At Logic value '0' – Transistor(T) is connected to the ground point(P).
Transistor switch is closed & voltage on bitline nearly drops to zero.
At Logic value '1' – Transistor switch is open.
The bitline remains at high voltage.
To read the state of the cell,the word line is activated.
A Sense circuit at the end of the bitline generates the proper output value.

Types of ROM:

Different types of non-volatile memory are,

- PROM
- EPROM
- EEPROM
- Flash Memory

PROM:-Programmable ROM:

- PROM allows the data to be loaded by the user.
- Programmability is achieved by inserting a „fuse" at point P in a ROM cell.
- Before it is programmed, the memory contains all 0" s
- The user can insert 1" s at the required location by burning out the fuse at these locations using high-current pulse.
- This process is irreversible.

Merit:

- It provides flexibility.
- It is faster.
- It is less expensive because they can be programmed directly by the user.

EPROM:-Erasable reprogrammable ROM:

- EPROM allows the stored data to be erased and new data to be loaded.

- In an EPROM cell, a connection to ground is always made at „P" and a special transistor is used, which has the ability to function either as a normal transistor or as a disabled transistor that is always turned „off" .

- This transistor can be programmed to behave as a permanently open switch, by injecting charge into it that becomes trapped inside.
- Erasure requires dissipating the charges trapped in the transistor of memory cells. This can be done by exposing the chip to ultra-violet light, so that EPROM chips are mounted in packages that have transparent windows.

Merits:

- It provides flexibility during the development phase of digital system.
- It is capable of retaining the stored information for a long time.

Demerits:

- The chip must be physically removed from the circuit for reprogramming and its entire contents are erased by UV light.

EEPROM:-Electrically Erasable ROM:

Merits:

- It can be both programmed and erased electrically.
- It allows the erasing of all cell contents selectively.

Demerits:

- It requires different voltage for erasing ,writing and reading the stored data.

Flash Memory:

- In EEPROM, it is possible to read & write the contents of a single cell.
- In Flash device, it is possible to read the contents of a single cell but it is only possible to write the entire contents of a block.
- Prior to writing,the previous contents of the block are erased.
- Eg.In MP3 player,the flash memory stores the data that represents sound.
- Single flash chips cannot provide sufficient storage capacity for embedded system application.
- There are 2 methods for implementing larger memory modules consisting of number of chips.They are,
    - Flash Cards
    - Flash Drives.

Merits:

- Flash drives have greater density which leads to higher capacity & low cost per bit.
- It requires single power supply voltage & consumes less power in their operation.

Flash Cards:

- One way of constructing larger module is to mount flash chips on a small card.
- Such flash card have standard interface.

- The card is simply plugged into a conveniently accessible slot.
- Its memory size are of 8,32,64MB.
- Eg:A minute of music can be stored in 1MB of memory. Hence 64MB flash cards can store an hour of music.

Flash Drives:

- Larger flash memory module can be developed by replacing the hard disk drive.
- The flash drives are designed to fully emulate the hard disk.
- The flash drives are solid state electronic devices that have no movable parts.

Merits:

- They have shorter seek and access time which results in faster response.
- They have low power consumption which makes them attractive for battery driven application.
- They are insensitive to vibration.

Demerit:

- The capacity of flash drive (<1GB) is less than hard disk(>1GB).
- It leads to higher cost perbit.
- Flash memory will deteriorate after it has been written a number of times(typically atleast 1 million times.)

4. What is cache memory? Describe in detail. (Apr 13)

CACHE MEMORIES

The effectiveness of cache mechanism is based on the property of „**Locality of reference'.**

Locality of Reference:

Many instructions in the localized areas of the program are executed repeatedly during some time period and remainder of the program is accessed relatively infrequently.

It manifests itself in 2 ways.They are,

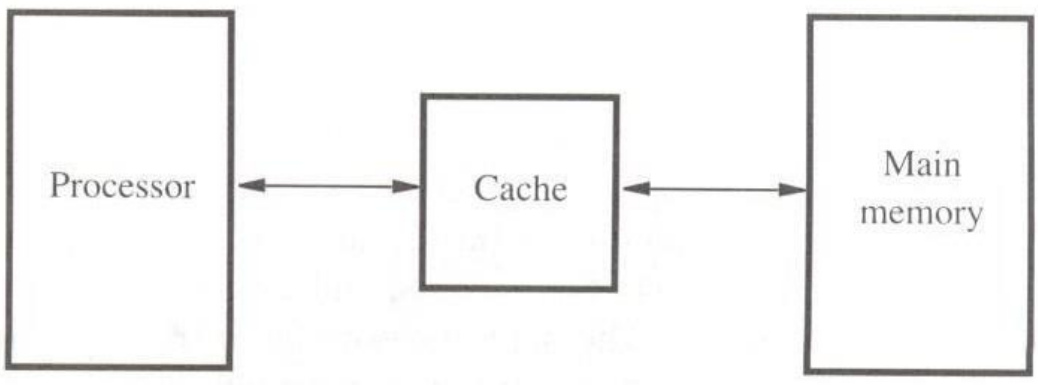
Temporal(The recently executed instruction are likely to be executed again very soon.)

Spatial(The instructions in close proximity to recently executed instruction are also likely to be executed soon.)

If the active segment of the program is placed in cache memory, then the total execution time can be reduced significantly. The term Block refers to the set of contiguous address locations of some size.

The cache line is used to refer to the cache block.

Fig:Use of Cache Memory



- The Cache memory stores a reasonable number of blocks at a given time but this number is small compared to the total number of blocks available in Main Memory.
- The correspondence between main memory block and the block in cache memory is specified by a mapping function.
- The Cache control hardware decide that which block should be removed to create space for the new block that contains the referenced word.
- The collection of rule for making this decision is called the replacement algorithm.
- The cache control circuit determines whether the requested word currently exists in the cache.
- If it exists, then Read/Write operation will take place on appropriate cache location. In this case Read/Write hit will occur.
- In a Read operation, the memory will not involve.

The write operation is proceed in 2 ways.They are,

- Write-through protocol
- Write-back protocol

Write-through protocol:
Here the cache location and the main memory locations are updated simultaneously.

Write-back protocol:
- This technique is to update only the cache location and to mark it as with associated flag bit called dirty/modified bit.
- The word in the main memory will be updated later,when the block containing this marked word is to be removed from the cache to make room for a new block.
- If the requested word currently not exists in the cache during read operation,then read miss will occur.
- To overcome the read miss Load –through / Early restart protocol is used.

Read Miss:
The block of words that contains the requested word is copied from the main memory into cache.

Load –through:
- After the entire block is loaded into cache,the particular word requested is forwarded to the processor.
- If the requested word not exists in the cache during write operation,then Write Miss will occur.
- If Write through protocol is used,the information is written directly into main memory.
- If Write back protocol is used then block containing the addressed word is first brought intothe cache and then the desired word in the cache is over-written with the new information.

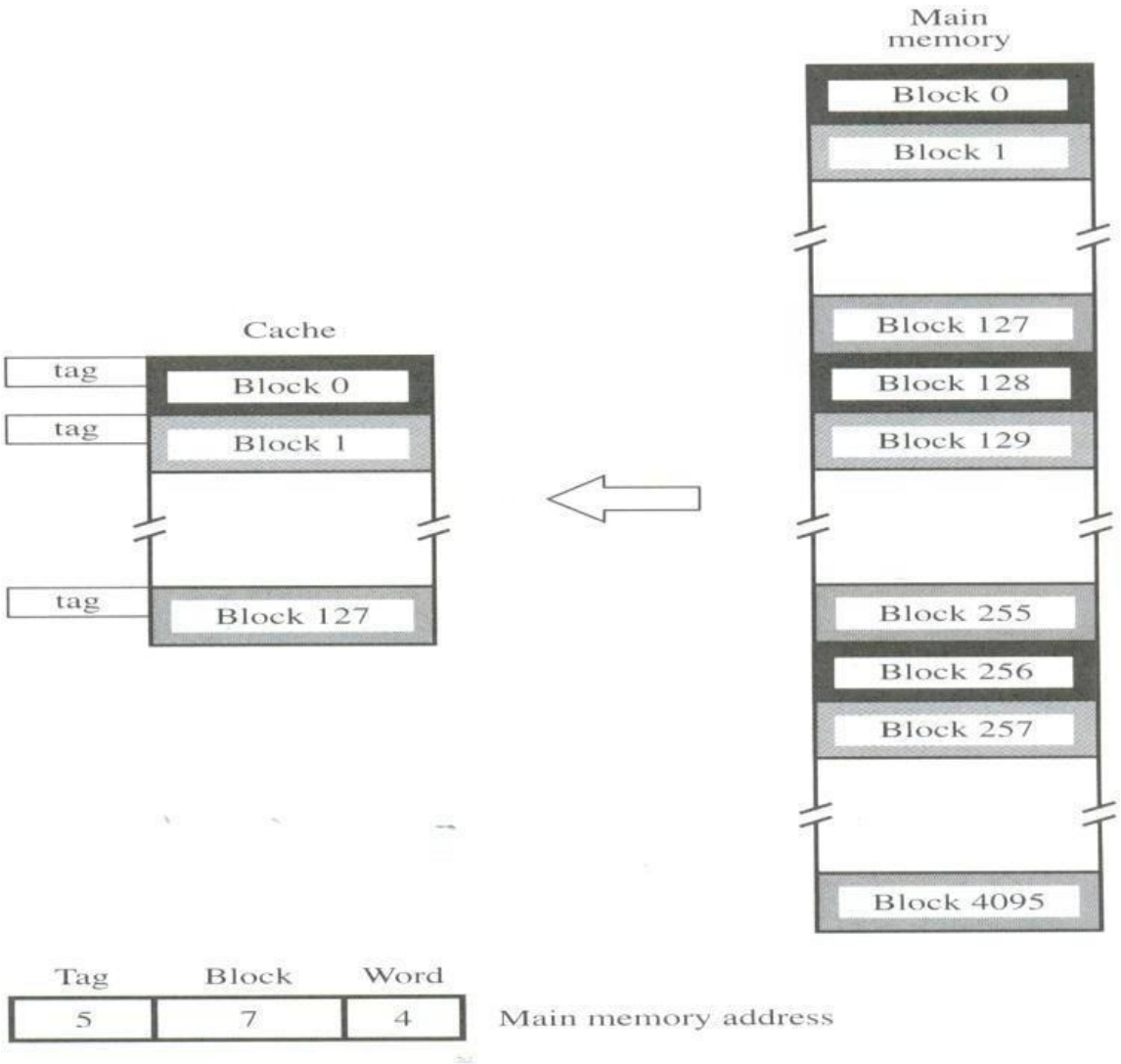
5. Explain different types of mapping functions in cache memory (Apr 11)
Mapping Function:
Direct Mapping:
- It is the simplest technique in which block j of the main memory maps onto block „j" modulo 128 of the cache.

- Thus whenever one of the main memory blocks 0,128,256 is loaded in the cache,it is stored in block 0.
- Block 1,129,257 are stored in cache block 1 and so on.

The contention may arise when,

- When the cache is full
- When more than one memory block is mapped onto a given cache block position.
- The contention is resolved by allowing the new blocks to overwrite the currently resident block.
- Placement of block in the cache is determined from memory address.

Fig: Direct Mapped Cache



The memory address is divided into 3 fields.They are,

Low Order 4 bit field(word) Selects one of 16 words in a block.

7 bit cache block field When new block enters cache,7 bit determines the cache position in which this block must be stored.

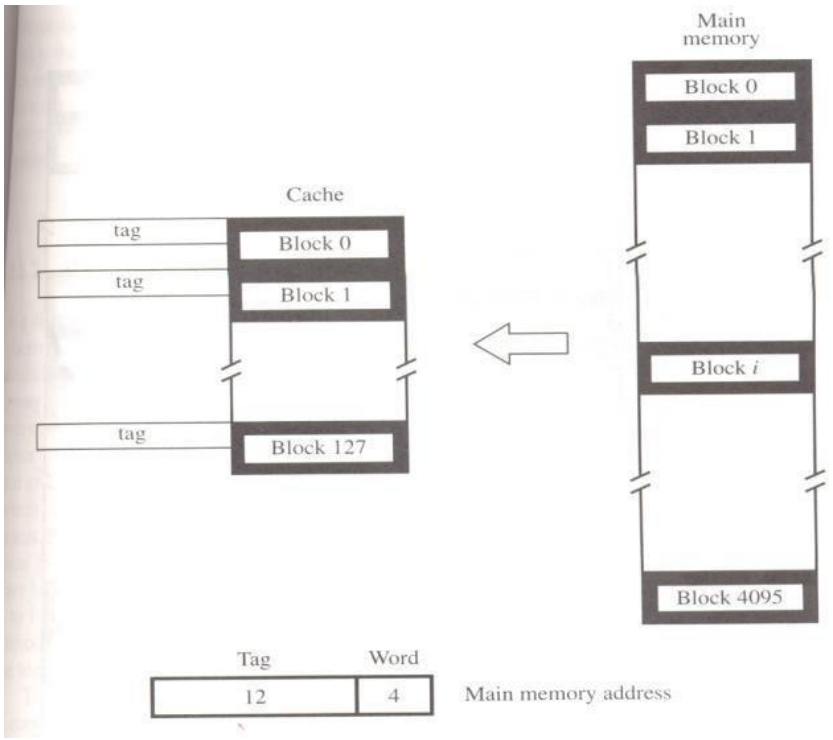
5 bit Tag field The high order 5 bits of the memory address of the block is stored in 5 tag bits associated with its location in the cache.

As execution proceeds, the high order 5 bits of the address is compared with tag bits associated with that cache location. If they match,then the desired word is in that block of the cache. If there is no match,then the block containing the required word must be first read from the main memory and loaded into the cache.

Merit:
It is easy to implement.

Demerit:
It is not very flexible.

Associative Mapping:
In this method, the main memory block can be placed into any cache block position.
Fig:Associative Mapped Cache.



- 12 tag bits will identify a memory block when it is resolved in the cache.
- The tag bits of an address received from the processor are compared to the tag bits of each block of the cache to see if the desired block is persent.This is called associative mapping.
- **I**t gives complete freedom in choosing the cache location.
- A new block that has to be brought into the cache has to replace(eject)an existing block if the cache is full.
- In this method,the memory has to determine whether a given block is in the cache.
- A search of this kind is called an associative Search.

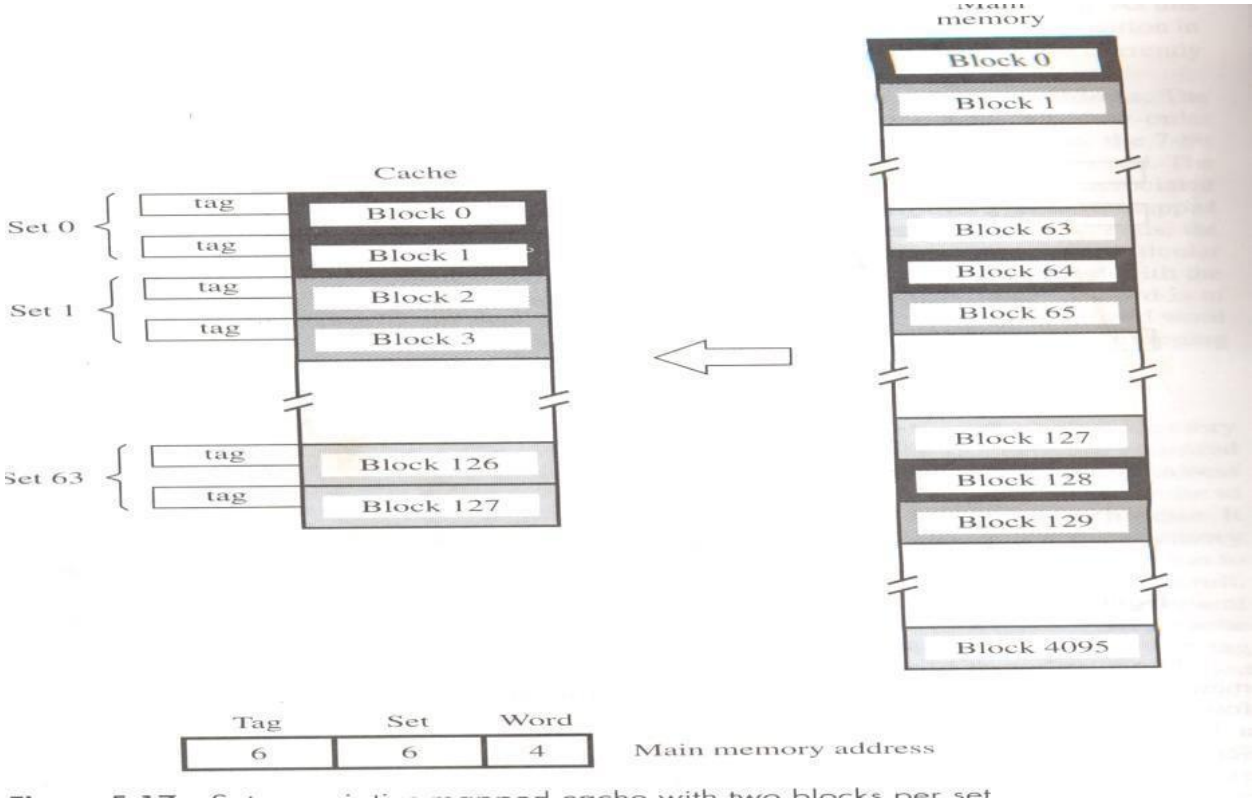
Merit:
It is more flexible than direct mapping technique.

Demerit:
Its cost is high.

Set-Associative Mapping:

- It is the combination of direct and associative mapping.
- The blocks of the cache are grouped into sets and the mapping allows a block of the main memory to reside in any block of the specified set.
- In this case,the cache has two blocks per set,so the memory blocks 0,64,128……..4032 maps into cache set „0" and they can occupy either of the two block position within the set.
- 6 bit set field Determines which set of cache contains the desired block .
- 6 bit tag field The tag field of the address is compared to thetags of the two blocks of
- the set to clock if the desired block is present.

Fig: Set-Associative Mapping:



No of blocks per set	no of set field
2	6
3	5
8	4
128	no set field

- The cache which contains 1 block per set is called direct Mapping.
- A cache that has „k" blocks per set is called as „k-way set associative cache" .
- Each block contains a control bit called a valid bit.
- The Valid bit indicates that whether the block contains valid data.
- The dirty bit indicates that whether the block has been modified during its cache residency.

- Valid bit=0 When power is initially applied to system
- Valid bit =1 When the block is loaded from main memory at first time.
- If the main memory block is updated by a source & if the block in the source is already exists in the cache,then the valid bit will be cleared to „0".
- If Processor & DMA uses the same copies of data then it is called as the Cache Coherence Problem.

Merit:

The Contention problem of direct mapping is solved by having few choices for block placement.

The hardware cost is decreased by reducing the size of associative search.

Replacement Algorithm:

- In direct mapping, the position of each block is pre-determined and there is no need of replacement strategy.
- In associative & set associative method,the block position is not pre-determined;ie..when the cache is full and if new blocks are brought into the cache, then the cache controller must decide which of the old blocks has to be replaced.
- Therefore,when a block is to be over-written,it is sensible to over-write the one that has gone the longest time without being referenced.This block is called Least recently Used(LRU) block & the technique is called LRU algorithm.
- The cache controller track the references to all blocks with the help of block counter.

Eg:

Consider 4 blocks/set in set associative cache,

2 bit counter can be used for each block.

When a **'hit'** occurs,then block counter=0;The counter with values originally lower than the referenced one are incremented by 1 & all others remain unchanged.

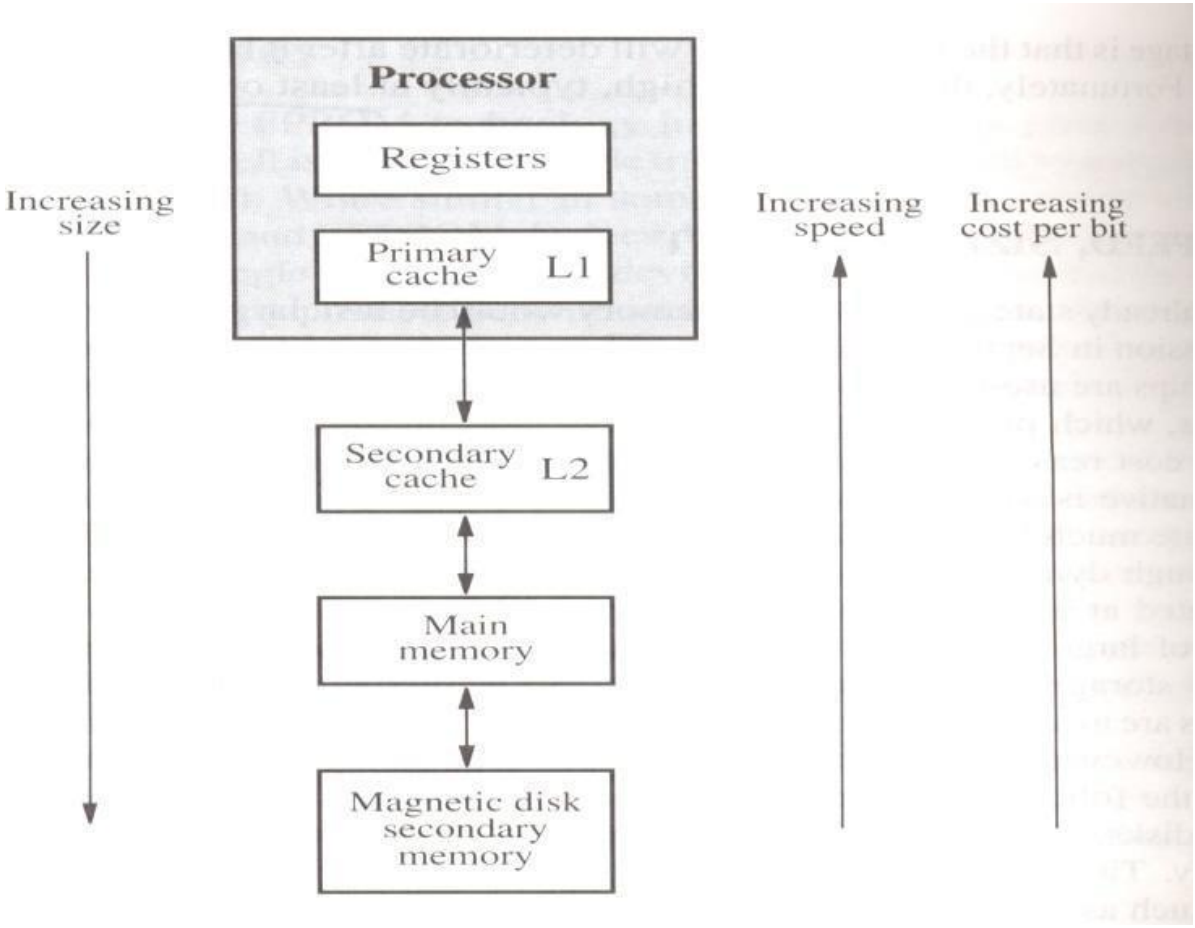
When a **'miss'** occurs & if the set is full,the blocks with the counter value 3 is removed,the new block is put in its place & its counter is set to „0" and other block counters are incremented by 1.

Merit:

The performance of LRU algorithm is improved by randomness in deciding which block is to be over-written.

6. Write about memory hierarchy.

Fig:Memory Hierarchy



Characteristics	SRAM	DRAM	Magnetis Disk
Speed	Very Fast	Slower	Much slower than DRAM
Size	Large	Small	Small
Cost	Expensive	Less Expensive	Low price

Magnetic Disk:

A huge amount of cost effective storage can be provided by magnetic disk;The main memory can be built with DRAM which leaves SRAM‟ s to be used in smaller units where speed is of essence.

Memory	Speed	Size	Cost
Registers	Very high	Lower	Very Lower
Primary cache	High	Lower	Low
Secondary cache	Low	Low	Low
Main memory	Lower than Seconadry cache	High	High
Secondary Memory	Very low	Very High	Very High

Types of Cache Memory:

The Cache memory is of 2 types.They are,

Primary /Processor Cache(Level1 or L1 cache)

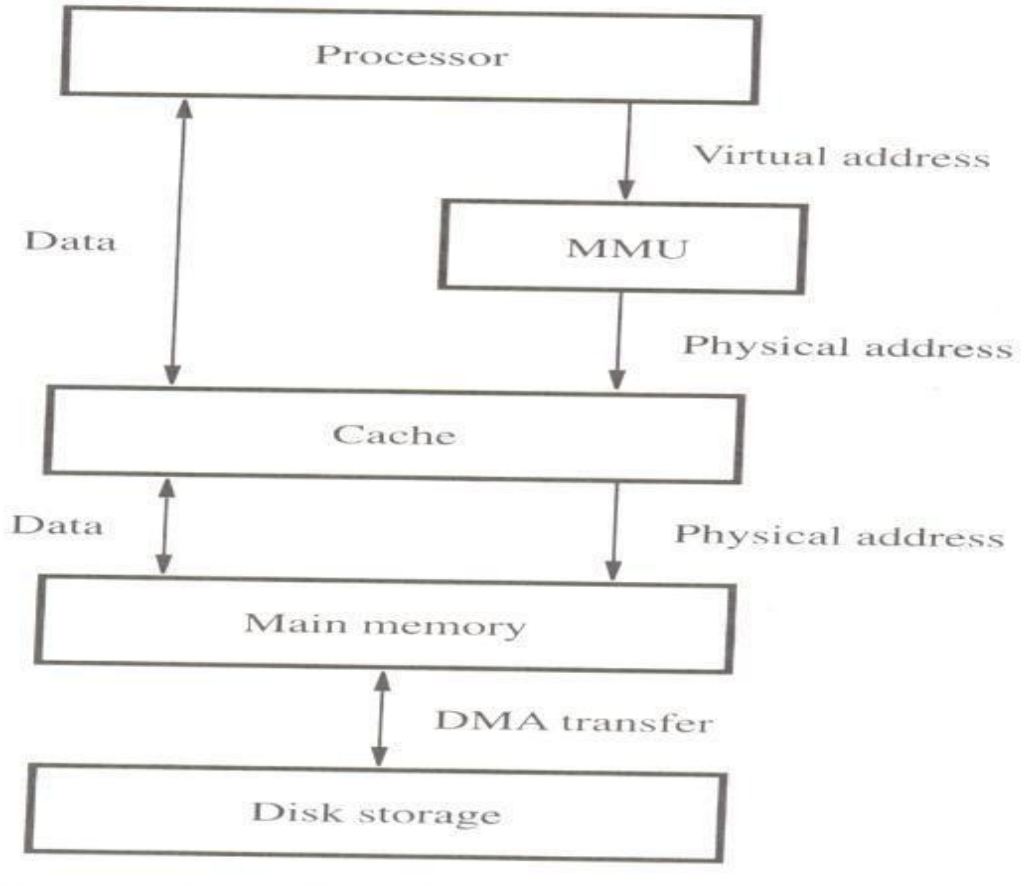
Secondary Cache(Level2 or L2 cache)

Primary Cache --- It is always located on the processor chip.
Secondary Cache ---It is placed between the primary cache and the rest of the memory.
The main memory is implemented using the dynamic components(SIMM,RIMM,DIMM).
The access time for main memory is about 10 times longer than the access time for L1 cache.

7. Explain in detail the concept of Virtual Memory (Apr 13)
VIRTUAL MEMORY:

- Techniques that automatically move program and data blocks into the physical main memory when they are required for execution is called the Virtual Memory**.**
- The binary address that the processor issues either for instruction or data are called the virtual / Logical address.
- The virtual address is translated into physical address by a combination of hardware and software components.This kind of address translation is done by MMU(Memory Management Unit).
- When the desired data are in the main memory ,these data are fetched /accessed immediately.
- If the data are not in the main memory,the MMU causes the Operating system to bring the data into memory from the disk.
- Transfer of data between disk and main memory is performed using DMA scheme.

Fig:Virtual Memory Organisation



Address Translation:

- In address translation,all programs and data are composed of fixed length units called Pages.
- The Page consists of a block of words that occupy contiguous locations in the main memory.
- The pages are commonly range from 2K to 16K bytes in length.
- The cache bridge speed up the gap between main memory and secondary storage and it is implemented in software techniques.
- Each virtual address generated by the processor contains virtual Page number(Low order bit) and offset(High order bit)
- Virtual Page number+ Offset Specifies the location of a particular byte (or word) within a page.

Page Table:
It contains the information about the main memory address where the page is stored & the current status of the page.

Page Frame:
An area in the main memory that holds one page is called the page frame.

Page Table Base Register:
It contains the starting address of the page table.

Virtual Page Number+Page Table Base register -- Gives the address of the corresponding entry in the page table.ie)it gives the starting address of the page if that page currently resides in memory.

Control Bits in Page Table:
The Control bits specifies the status of the page while it is in main memory.

Function:

- The control bit indicates the validity of the page ie)it checks whether the page is actually loaded in the main memory.
- It also indicates that whether the page has been modified during its residency in the memory;this information is needed to determine whether the page should be written back to the disk before it is removed from the main memory to make room for another page.

- The Page table information is used by MMU for every read & write access.
- The Page table is placed in the main memory but a copy of the small portion of the page table is located within MMU.
- This small portion or small cache is called Translation LookAside Buffer(TLB).
- This portion consists of the page table enteries that corresponds to the most recently accessed pages and also contains the virtual address of the entry.

Fig:Virtual Memory Address Translation

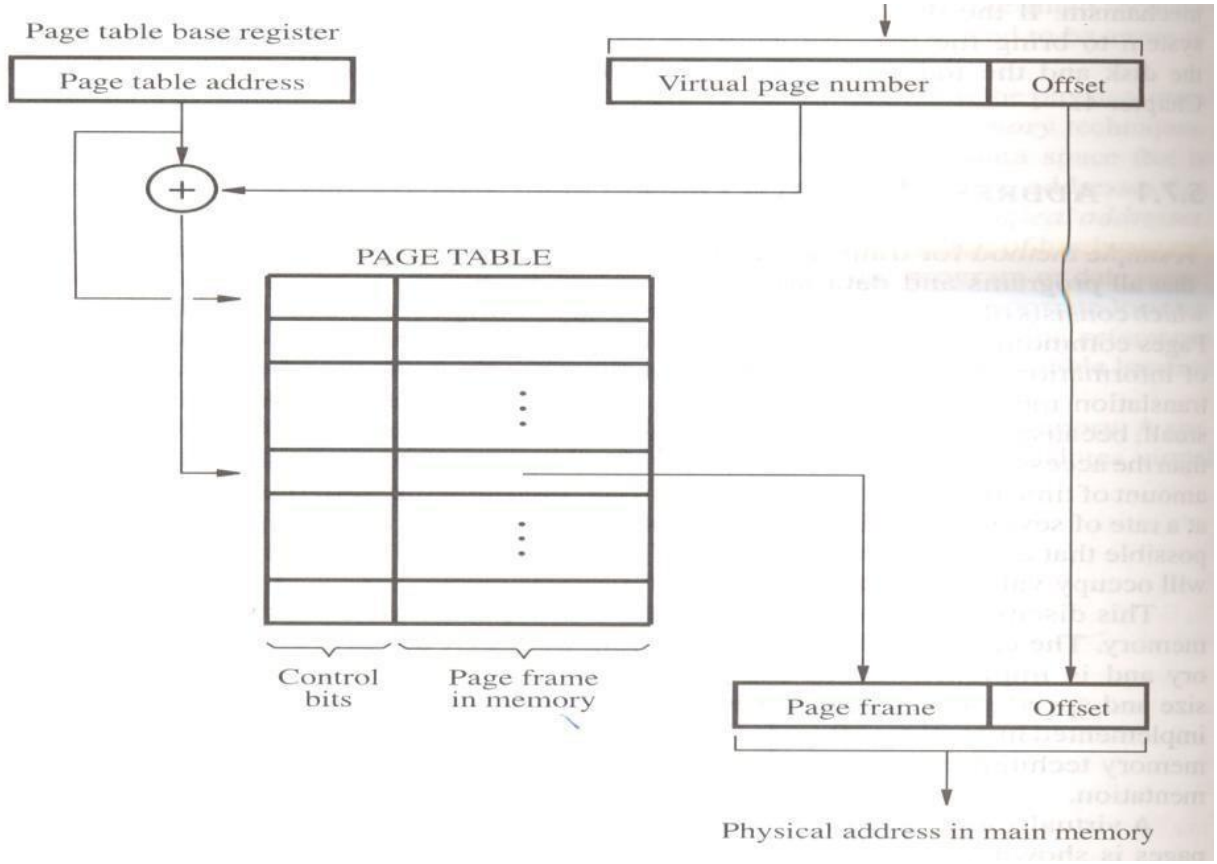
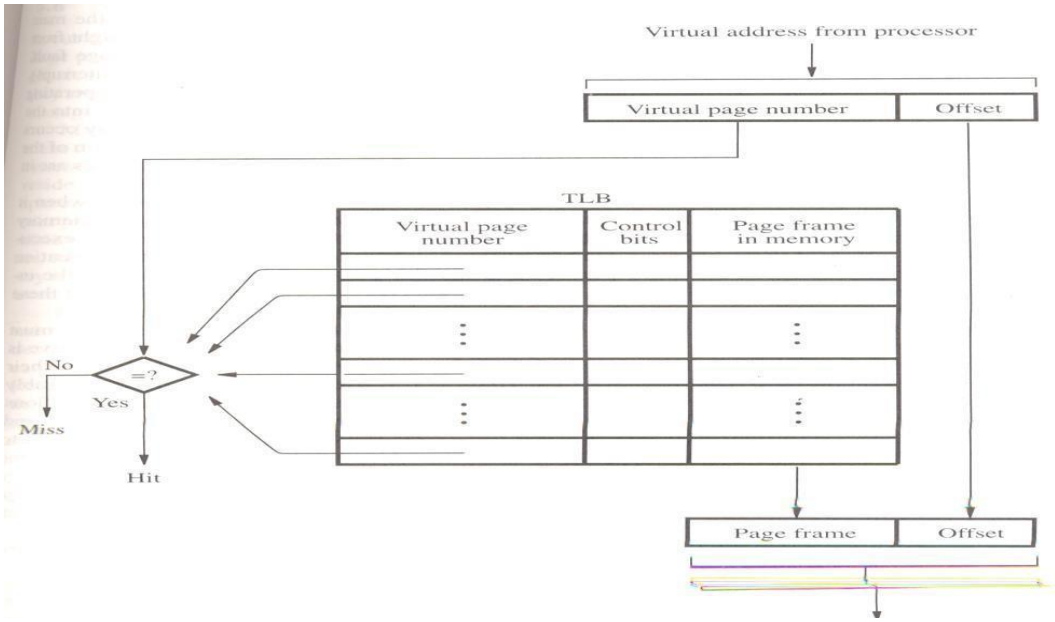


Fig:Use of Associative Mapped TLB



- When the operating system changes the contents of page table ,the control bit in TLB will invalidate the corresponding entry in the TLB.
- Given a virtual address,the MMU looks in TLB for the referenced page.
- If the page table entry for this page is found in TLB,the physical address is obtained immediately.
- If there is a miss in TLB,then the required entry is obtained from the page table in the main memory & TLB is updated.
- When a program generates an access request to a page that is not in the main memory ,then Page Fault will occur.
- The whole page must be broght from disk into memry before an access can proceed.
- When it detects a page fault,the MMU asks the operating system to generate an interrupt.
- The operating System suspend the execution of the task that caused the page fault and begin execution of another task whose pages are in main memory because the long delay occurs while page transfer takes place.
- When the task resumes,either the interrupted instruction must continue from the point of interruption or the instruction must be restarted.
- If a new page is brought from the disk when the main memory is full,it must replace one of the resident pages.In that case,it uses LRU algorithm which removes the least referenced Page.
- A modified page has to be written back to the disk before it is removed from the main memory. In that case,write –through protocol is used.

8. Discuss in detail about the secondary storage devices

SECONDARY STORAGE:

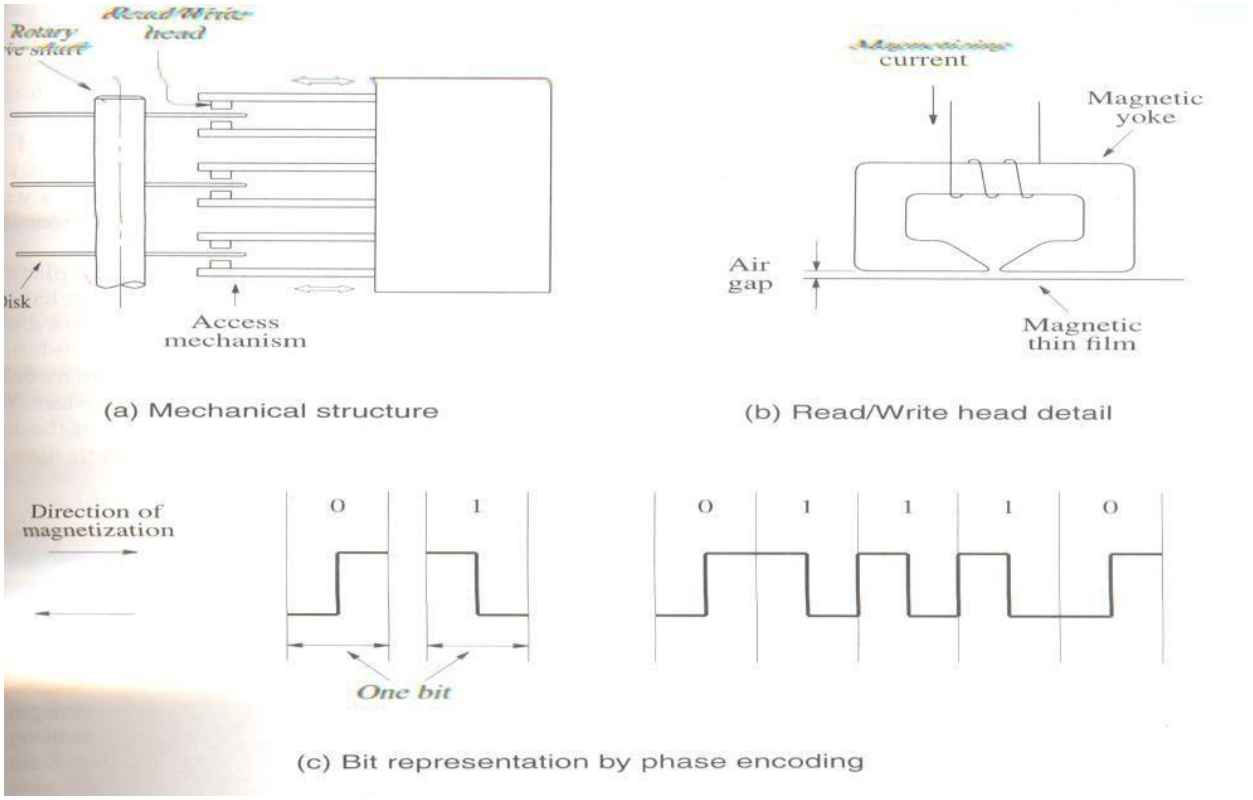
- The Semi-conductor memories donot provide all the storage capability.
- The Secondary storage devices provide larger storage requirements.
- Some of the Secondary Storage devices are,
    - Magnetic Disk
    - Optical Disk
    - Magnetic Tapes.

Magnetic Disk:

- Magnetic Disk system consists o one or more disk mounted on a common spindle.
- A thin magnetic film is deposited on each disk, usually on both sides.
- The disk are placed in a rotary drive so that the magnetized surfaces move in close proximity to read /write heads.
- Each head consists of magnetic yoke & magnetizing coil.
- Digital information can be stored on the magnetic film by applying the current pulse of suitable polarity to the magnetizing coil.
- Only changes in the magnetic field under the head can be sensed during the Read operation.

- Therefore if the binary states 0 & 1 are represented by two opposite states of magnetization, a voltage is induced in the head only at 0-1 and at 1-0 transition in the bit stream.
- A consecutive (long string) of 0‟ s & 1‟ s are determined by using the clock which is mainly used for synchronization.
- Phase Encoding or Manchester Encoding is the technique to combine the clocking information with data.
- The Manchester Encoding describes that how the self-clocking scheme is implemented.

Fig:Mechanical Structure



- The Read/Write heads must be maintained at a very small distance from the moving disk surfaces in order to achieve high bit densities.
- When the disk are moving at their steady state, the air pressure develops between the disk surfaces & the head & it forces the head away from the surface.
- The flexible spring connection between head and its arm mounting permits the head to fly at the desired distance away from the surface.

Wanchester Technology:

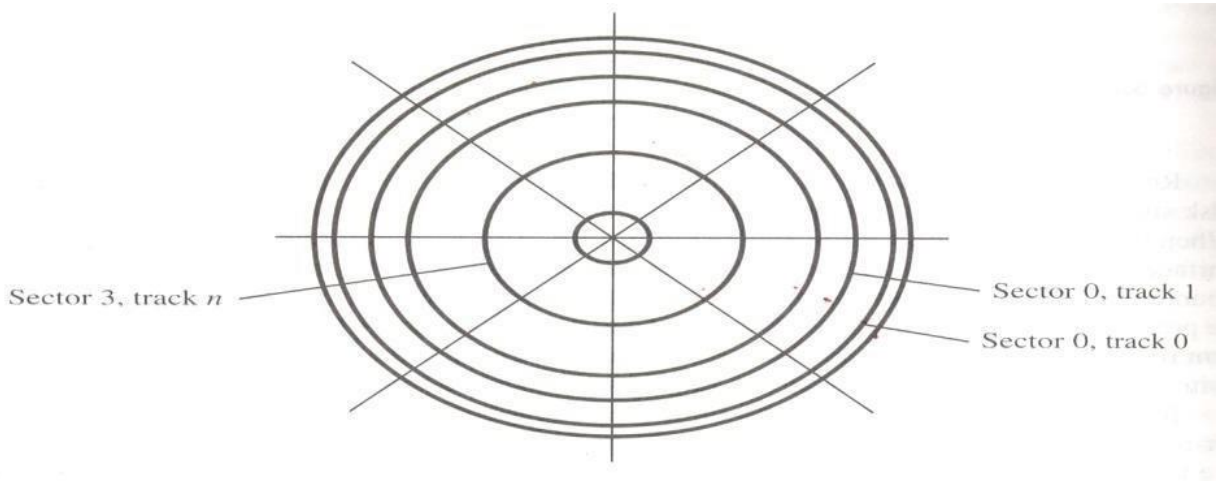
- Read/Write heads are placed in a sealed, air –filtered enclosure called the Wanchester Technology.
- In such units, the read/write heads can operate closure to magnetic track surfaces because the dust particles which are a problem in unsealed assemblies are absent.

Merits:

- It have a larger capacity for a given physical size.
- The data intensity is high because the storage medium is not exposed to contaminating elements.
- The read/write heads of a disk system are movable.
- The disk system has 3 parts.They are,

Disk Platter(Usually called Disk)
Disk Drive(spins the disk & moves Read/write heads)
Disk Controller(controls the operation of the system.)

Fig:Organizing & Accessing the data on disk



- Each surface is divided into concentric tracks.
- Each track is divided into sectors.
- The set of corresponding tracks on all surfaces of a stack of disk form a logical cylinder.
- The data are accessed by specifying the surface number,track number and the sector number.
- The Read/Write operation start at sector boundaries.
- Data bits are stored serially on each track.
- Each sector usually contains 512 bytes.
- Sector header -> contains identification information.
- It helps to find the desired sector on the selected track.
- ECC (Error checking code)- used to detect and correct errors.
- An unformatted disk has no information on its tracks.
- The formatting process divides the disk physically into tracks and sectors and this process may discover some defective sectors on all tracks.
- The disk controller keeps a record of such defects.
- The disk is divided into logical partitions. They are,
  - ✓ Primary partition
  - ✓ Secondary partition
- In the diag, Each track has same number of sectors.
- So all tracks have same storage capacity.
- Thus the stored information is packed more densely on inner track than on outer track.

Access time

There are 2 components involved in the time delay between receiving an address and the beginning of the actual data transfer. They are,

- ✓ Seek time
- ✓ Rotational delay / Latency

Seek time – Time required to move the read/write head to the proper track.

Latency – The amount of time that elapses after the head is positioned over the correct track until the starting position of the addressed sector passes under the read/write head.

Seek time + Latency = Disk access time

Typical disk

One inch disk- weight=1 ounce, size -> comparable to match book

Capacity -> 1GB

Inch disk has the following parameter

Recording surface=20

Tracks=15000 tracks/surface

Sectors=400.

Each sector stores 512 bytes of data

Capacity of formatted disk=20x15000x400x512=60x109 =60GB

Seek time=3ms

Platter rotation=10000 rev/min

Latency=3ms

Internet transfer rate=34MB/s

Data Buffer / cache

- A disk drive that incorporates the required SCSI circuit is referred as SCSI drive.
- The SCSI can transfer data at higher rate than the disk tracks.
- An efficient method to deal with the possible difference in transfer rate between disk and SCSI bus is accomplished by including a data buffer.
- This buffer is a semiconductor memory.
- The data buffer can also provide cache mechanism for the disk (ie) when a read request arrives at the disk, then controller first check if the data is available in the cache(buffer).
- If the data is available in the cache, it can be accessed and placed on SCSI bus . If it is not available then the data will be retrieved from the disk.

Disk Controller

The disk controller acts as interface between disk drive and system bus.

The disk controller uses DMA scheme to transfer data between disk and main memory.

When the OS initiates the transfer by issuing Read/Write request, the controllers register will load the following information. They are,

Main memory address(address of first main memory location of the block of words involved in the transfer)

Disk address(The location of the sector containing the beginning of the desired block of words)

9. Discuss the following

(i) Interleaving

(ii) Hit rate and Miss penalty

(iii) Pre-fetching

IMPROVING CACHE PERFORMANCE:

- Two key factors in the commercial success of a computer are performance and cost. The objective is the best possible Performance at the lowest cost.
- The challenge in design alternative is to improve the performance without increasing the cost. A common measure of success is the price/performance ratio.
- The memory hierarchy shows the best price/performance ratio. Each level of hierarchy plays an important role. The speed and efficiency of data transfer between various level of hierarchy are having great significance. Both is not possible if, both the slow and the fast units are accessed in the same manner, but can be achieved by the parallelism in the organization of the slower unit. An effective way to introduce parallelism is to use an "interleaved organization"

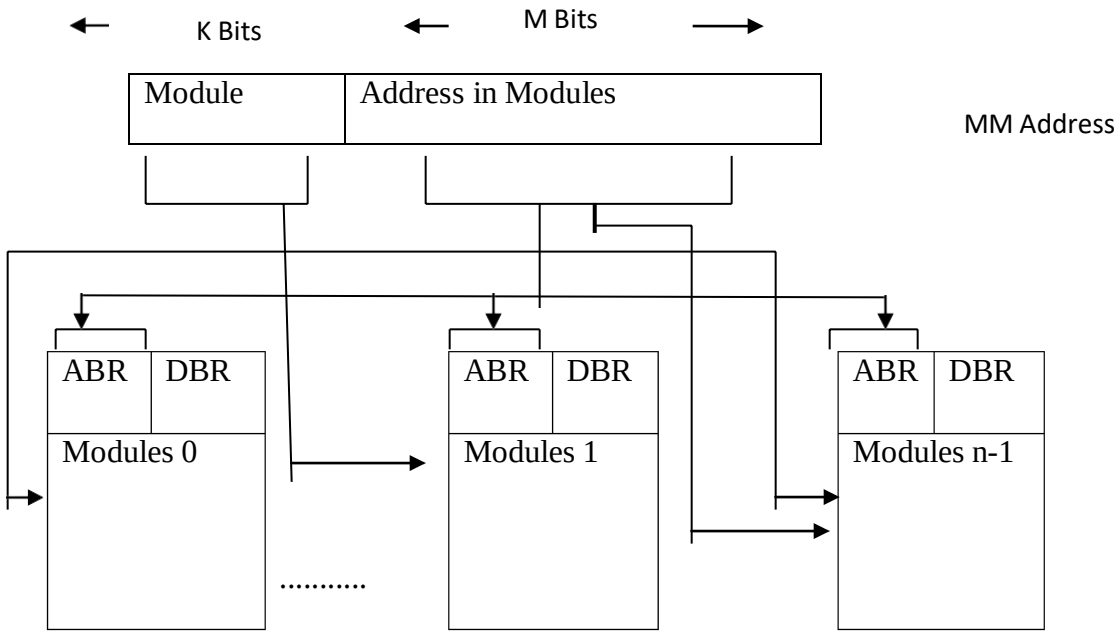
INTERLEAVING

- If the main memory of computer is structured as a collection of physical separate modules , each with its own address buffer register(ABR) and data buffer register(DBR), memory address operations may proceed in more than one module at the same time.
- How individual address are distributed over the modules is critical. Two methods of address layout are there.
- In the first case, the memory address generated by the processor is decoded as shown in fig. the higher-order k bits name one of n modules, and the lower-order m bits name a particular word in that module.
- When consecutive locations are accessed , when a block of data is transferred to a cache, only one module is involved. At the same time, devices with direct memory access(DMA) ability may be accessing information in other memory modules.

The second way to address the modules is shown in fig. It is called memory interleaving. The low-order k bits of the memory address select a module, and the higher-order m bits name a location within that module. In this way , consecutive addresses are located in successive modules. This results in faster access to a block of data and higher average utilization of the memory system as a whole.

Interleaving is used to within SDRAM chips to improve the speed of accessing successive word of data.

Fig(a). Consecutive words in a modules



Fig(b). Consecutive words in Consecutive a modules

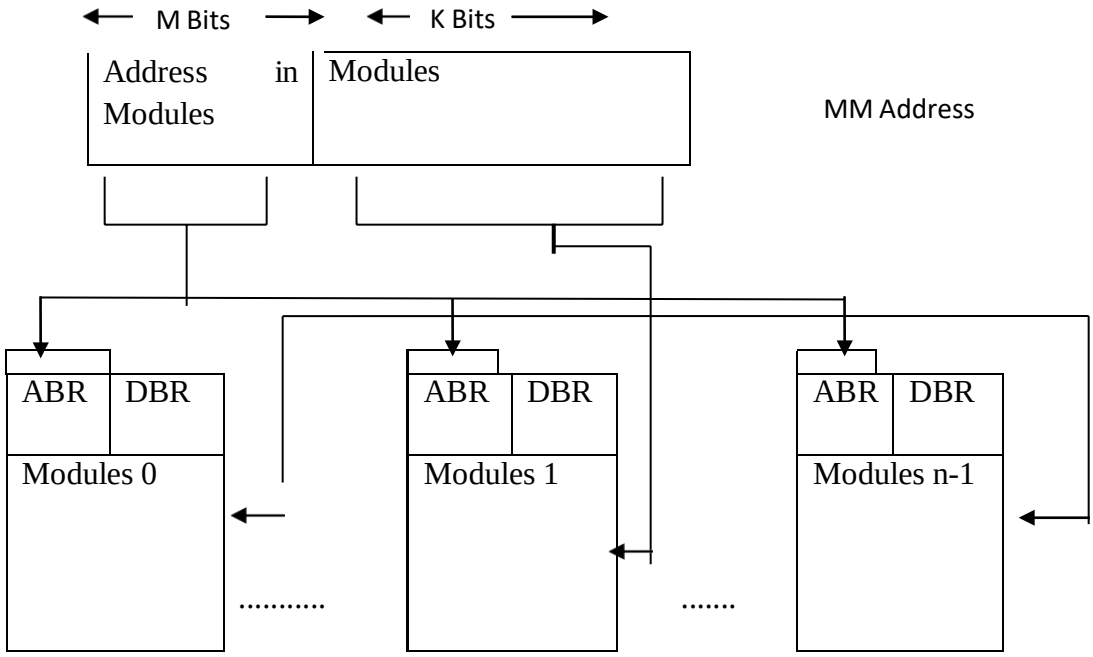


Fig : Addressing multiple-module memory sytems

HIT RATE AND MISS PENALTY

- The effective implementation of the memory hierarchy is the success rate in accessing information at various levels of hierarchy.

- The successful access to data in cache is called a hit. The number of hits are stated as a fraction of all attempted accesses is call the hit rate, and the misses rate is the number of misses stated as a fraction of attempted accesses.
- High hit rates are essential for high-performance computers , well over 0.9.
- Performance is affected by the miss. The extra time needed to bring the desired information into the catch is called the miss penalty. This penalty make the processor to stall.
- In general, the miss penalty is the time needed to bring a block of data from a slower unit to a faster unit. The miss penalty is reduced, if the efficient mechanisms are implemented.
- The performance of a computer is affected positively by increased hit rate and negatively by increased miss penalty, the block sizes that are neither very small nor large give the best results.
- In practice, block sizes in the range of 16 to 128 bytes have been the most popular choices.

PREFETCHING

- The new data are brought into the cache when they are first needed. A read miss occurs , and the desired data are loaded from the main memory. The processor has to pause until the new data arrive.
- A special prefetch instruction may be provided in the instruction set of the processor. Executing this instruction causes the addressed data to the loaded into the cache, in the case of read miss.
- Prefetch instructions can be inserted into a program either programmer or by the compiler.
- However the overall effect of the software prefetching on performance is positive and it is supported by machine instructions of many processors.
- Prefetching can also done through hardware . this involves adding circuitry thet attempts to discover a pattern in memory references and then prefetches data according to that.
- Example:
- Intel's Pentium 4 processor has facilitates for prefetching information into caches using both software and hardware approaches. There are special prefetch instructions that can be included in programs to bring a block of data into the desired level of cache.

CACHES ON THE PROCESSOR CHIP

- The space on the processor chip is needed for many other functions, this limits the size of the cache that can be accommodated.
- All high-performance processor chips include some form of cache. Some manufacturers have chosen to implement two separate caches, one for instructions and another for data.
- Example: 68404, Pentium 3 and Pentium 4 processors.
- Example: ARM710T Processor.

- A combined cache provides better bit rate and it offers greater flexibility in mapping new information into the cache.

DISADVANTAGES:

- Increase parallelism comes at the expense of more complex circuitry.
- In high-performance processors two levels oh caches are normally used. The L1 caches(s) is on the processor chip. The L2 cache, which is much larger, may be implemented externally.
- If both L1 and L2 caches are used, the L1 cache should be designed to do fast access. A practical way to speed up access to the cache is to access more than one word simultaneously and let the processor use them one at a time.
- The average access time experienced by the processor in a system with two levels of caches is
- $T_{ave}=h_1C_1+(1-h_1)h_2C_2+(1-h_1)(1-h_2)M$

Where h_1 is the hit rate in the L_1 cache

h_2 is the hit rate in the L_2 cache

C_1 is the time to access information in the L_1 cache

C_2 is the time to access information in the L_2 cache

M is the time to access information in the main memory

Other Enhancements:

Several other possibilities exist for enhancing performance.

Write buffer

- When a write-through protocol is user, each write operation results in writing new value into the memory.
- If the processor must wait for the memory function to be completed , the it slowed down by all write requests.
- It is not necessary for the processor to wait for the write request to be completed.
- To improve performance , a write buffer can be included for temporary storage of write requests.
- The processor [places each write request into this buffer and continues execution if the next instruction.
- The write requests stored in the write buffer are sent to the main memory when it is not having any reading operation.
- The write buffer may hold a number of write requests.

LOOKUP-FREE CACHE

- A cache that can support multiple outstanding misses is called lookup-free. Since it can service only one miss at a time, it must include circuitry that keeps track of all outstanding misses. They may be done with special registers that hold the pertinent information about these misses.

10. Explain about memory management requirements

1. Virtual memory gives us an illusion that deals with only one large program. But that is not the case. If all of the programs do not fit into the available memory space, parts of it are swapped to and from the disk into the main memory. These are done with the help of management routine that resides as a part of the operating system.

2. There are two spaces available in the memory. They are namely

    - System space
    - User space

3. The operating system routines reside in system space.

4. The user application programs reside in user space.

5. There may be a number of user spaces, one for each user.These are arrnaged by providing a separate page table for each user program. The MMU uses a page table base register to determine the address of the table to be used in the translation process

6. By changing the contents of the page table base register, the oprating system can switch from one space to another. Thus the physical main memory is shared by the active pages of the system space and several user spaces.

7. But these two spaces must not be overlapped as its leads to serious threats. Hence a notion of promotion must be addressed. This kind of protection can be provided in several ways. The basic form of protection utilizes two kinds of states namely

    - Supervisor state
    - User state

8. The processor is always placed in the supervisor state while executing the oprating system routines and placed in the user state while executing the user programs

9. There are some machine instructions known as privileged instructions cannot be executed in the user state. These can be executed only in the supervisor state thereby a user program is prevented from accessing the page tables of other user spaces of the system space.

10. It is sometimes needed for one application program to have access to certain pages belonging to another program. The operating system can arrange this by having these pages to appear in both spaces. The shared pages will therefore have entries in two different page tables.
11. The control bits in each table entry can be set to control the access privileges granted to each program. For Example, one program may be allowed to read and write a give page, while the other program may be given only read access.