MODULE 5

SYLLABUS

The Memory System – basic concepts, semiconductor RAM memories - organization – static and dynamic RAM, Structure of larger memories, semiconductor ROM memories, Speed, Size and cost ,Cache memory – mapping functions – replacement algorithms , Virtual memory – paging and segmentation.

BASIC CONCEPTS

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

Address	Memory Locations
16 Bit	2^{16} = 64 K
32 Bit	2^{32} = 4G (Giga)
40 Bit	2^{40} = 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 2^K 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 initiation 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.

RAM (Random Access Memory):

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 program 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.

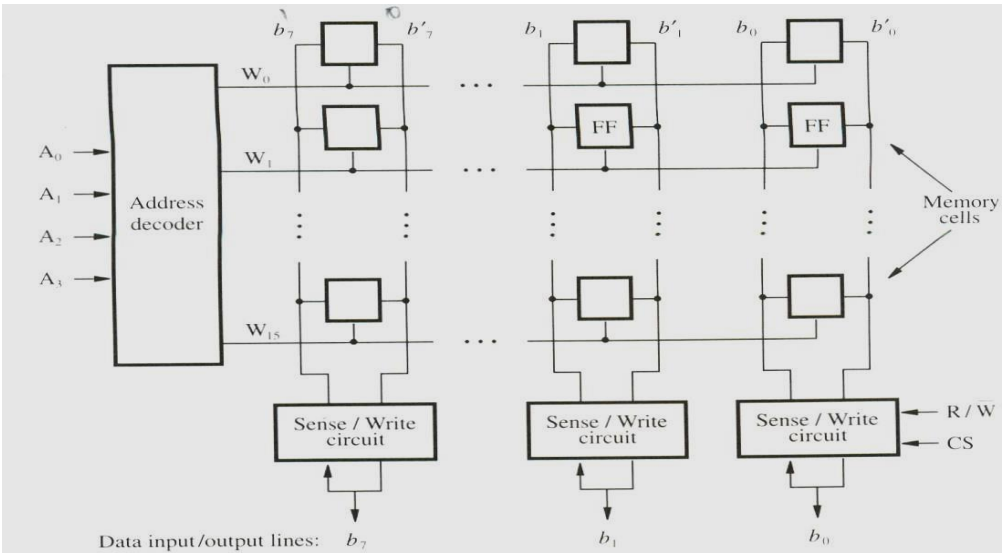
SEMI CONDUCTOR RAM MEMORIES:

- 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 constitutes a memory word and all cells of a row are connectedto 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.

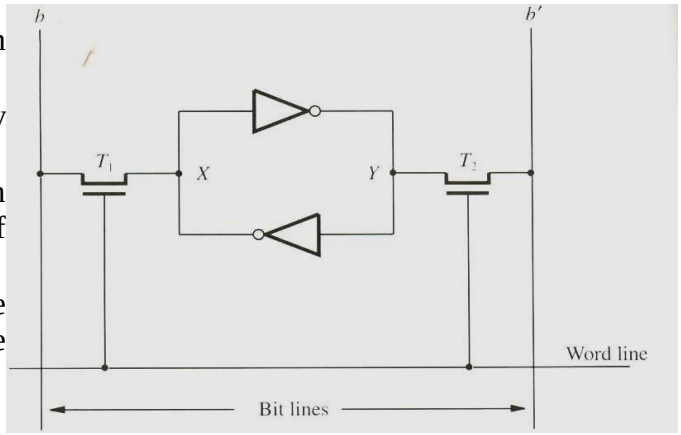
CS →**Chip Select** input selects a given chip in the multi-chip memory system

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

Static Memories:

Memories that consists 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 T_1 and T_2.
  These transistors act as switches that can be opened / closed under the control of the word line.
- When the word line is at ground level, the transistors are turned off and the latchretain its state.

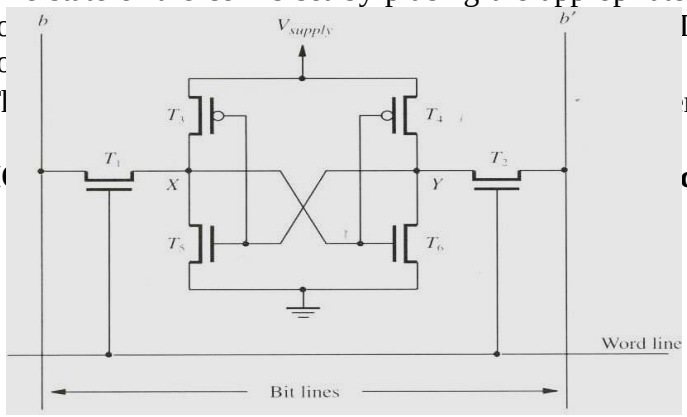
Read Operation:

- In order to read the state of the SRAM cell, the word line is activated to close switches T_1 and T_2.
- 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 co This forces the cell into the co
- Th nse / Write circuit.

Fig: CM(ctor):



- Transistor pairs (T_3, T_5) and (T_4, T_6) form the inverters in the latch.
- In state 1, the voltage at point X is high by having T_5, T_6 on and T_4, T_5 are OFF.
- Thus T_1, and T_2 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.

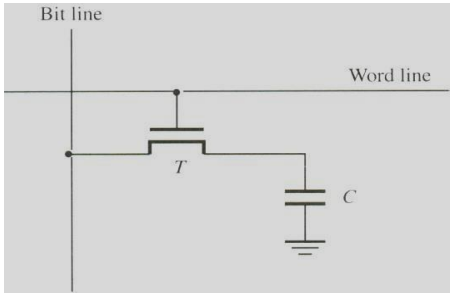
Demerit:

- 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



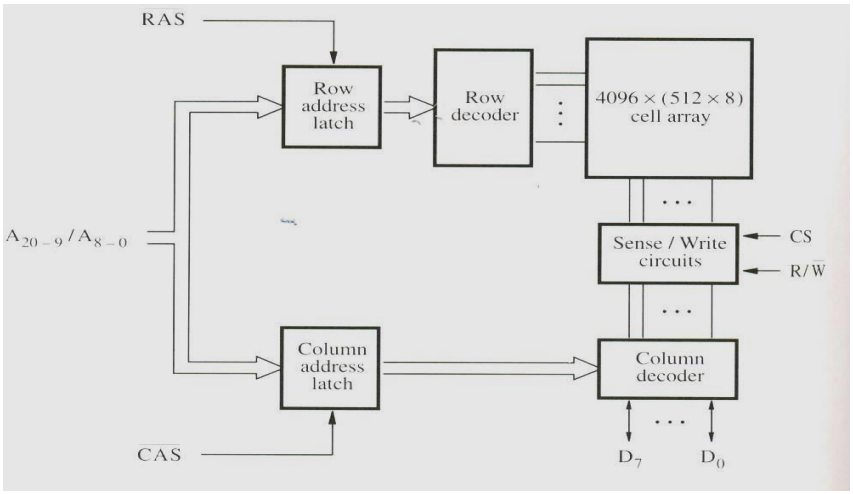
- In order to store information in the cell, the transistor T is turned „on‟ & the appropriate voltage is applied to the bit line, which charges the capacitor.
- After the transistor is turned off, the capacitor begins to discharge which is caused by the capacitors own leakage resistance.
- Hence the 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).

  A_{8-0} →Row address of a byte.
  A_{20-9} →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 theaddress 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.**

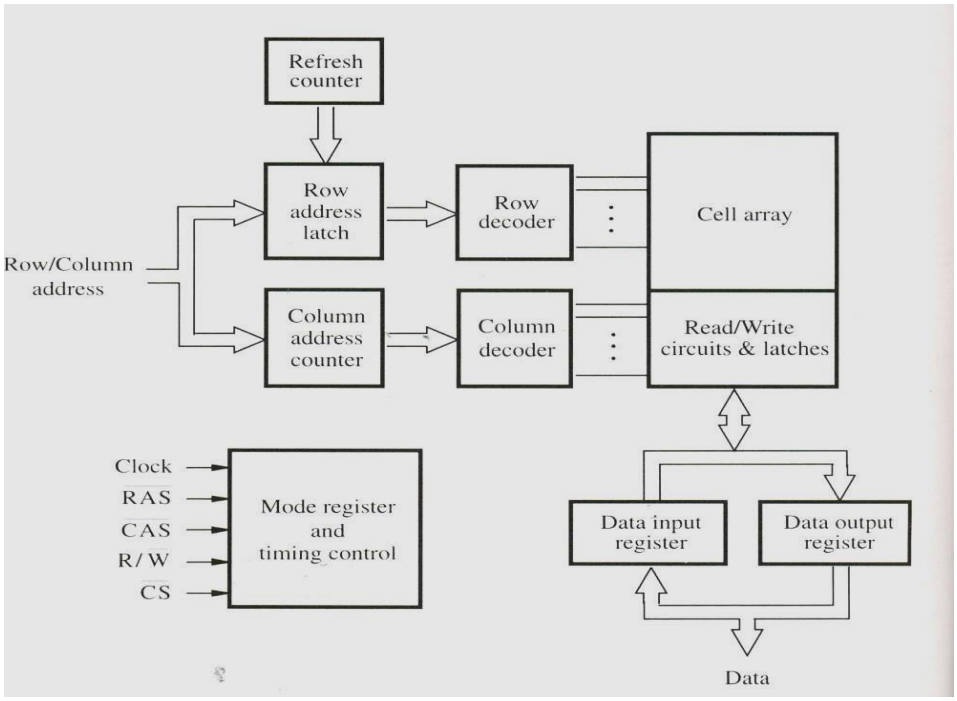
Fast Page Mode:

- 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:

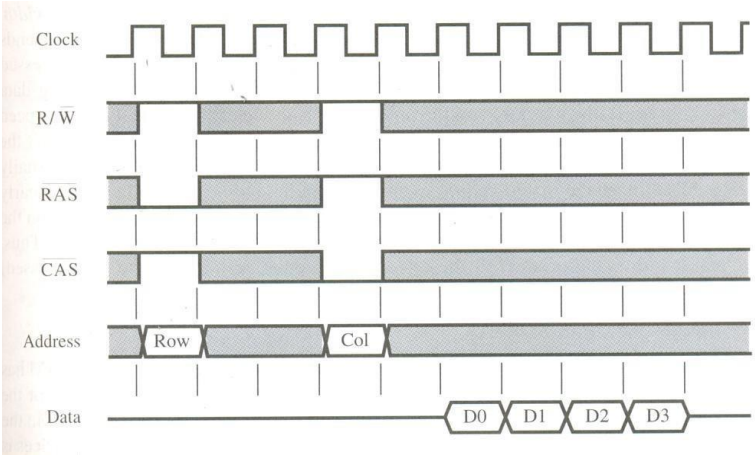
- Here the operations e directly synchronized with clock signal.
- The address and data connections are buffered by means of registers.
- The output of each sense amplifier is connected to a latch.
- A Read operation causes the contents of all cells in the selected row to be loaded in these latches.

Fig: Synchronous DRAM


Refresh counter
Row address latch
Row decoder
Cell array
Row/Column address
Column address counter
Column decoder
Read/Write circuits & latches
Clock
RAS
CAS
R/W
CS
Mode register and timing control
Data input register
Data output register
Data

- 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: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 ofmemory performance.
- For a block transfer, the latency denotes the time it takes to transfer the first wordof 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.

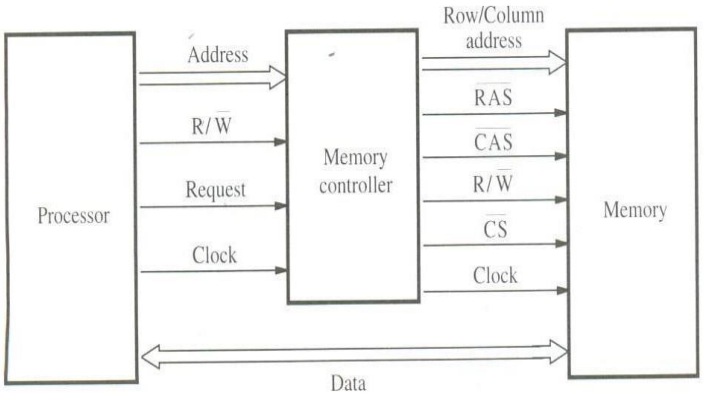
MEMORY SYSTEM CONSIDERATION:

- 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=$32768/133*10^6$
$=246*10^{-6}$ sec
=0.246sec
Refresh Overhead =0.246/64
Refresh Overhead =0.0038

Rambus Memory:

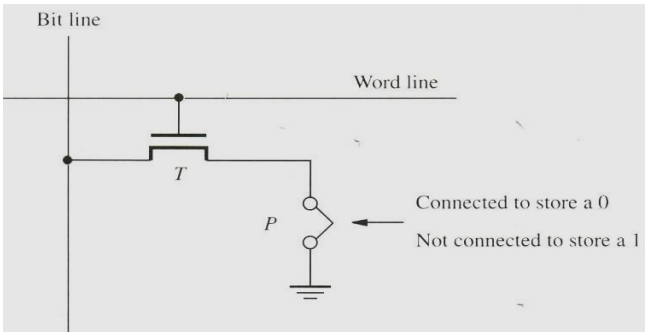
- The usage of wide bus is expensive.
- Rambus developed the implementation of narrow bus.
- Rambus technology is a fast signaling method used to transfer information between chips.

- Instead of using signals that have voltage levels of either 0 or V_{supply} to represent the logical values, the signals consists of much smaller voltage swings around a reference voltage $V_{ref.}$
- The reference Voltage is about 2V and the two logical values are represented by 0.3V swings above and below $V_{ref..}$
- 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,9^{th} 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

READ ONLY MEMORY:

- Both SRAM and DRAM chips are volatile, which means that they lose the storedinformation 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 bit line 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 has standard interface.

- The card is simply plugged into a conveniently accessible slot. Its
- memory size is of 8,32,64MB.
- Eg: A minute of music can be stored in 1MB of memory. Hence 64MB flash cardscan 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 per bit.
- Flash memory will deteriorate after it has been written a number of times (typically at least 1 million times.)

SPEED,SIZE COST:

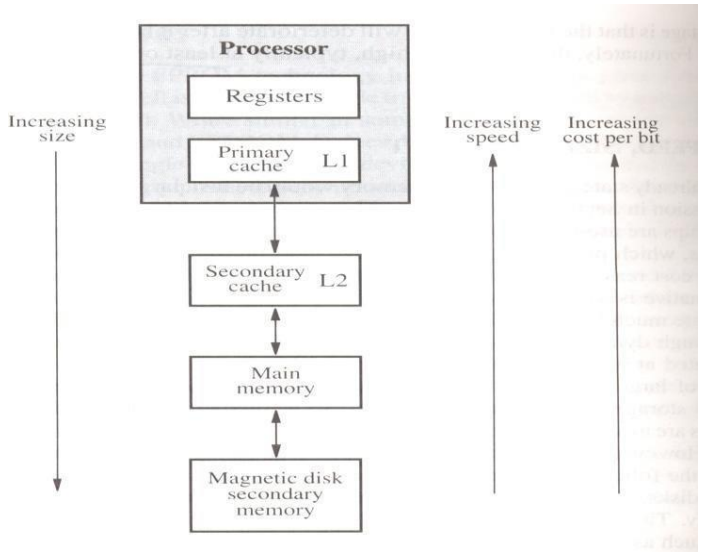
Characteristics	SRAM	DRAM	Magnetis Disk
Speed	Very Fast	Slower	Much slower than DRAM
Size	Large	Small	Small
Cost	Expensive	Less Expensive	Low price

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 higher
Primary cache	High	Lower	Higher
Secondary cache	Low	Low	Low
Main memory	Lower than Seconadry cache	High	Low
Secondary Memory	Very low	Very High	Very lower

Fig:Memory Hierarchy



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.

CACHE MEMORIES

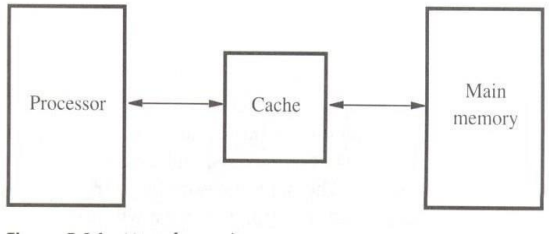
- The effectiveness of cache mechanism is based on the property of „**Locality of reference'.**

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 againvery soon.)
  - **Spatial (**The instructions in close proximity to recently executed instructionare 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 proceeding 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:

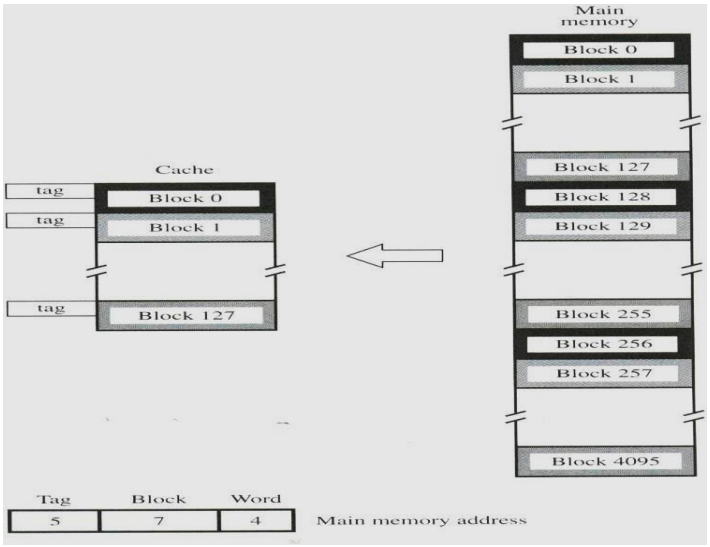
- After the entire block is loaded into cache, the particular word requested isforwarded to the processor.
- If the requested word not exists in the cache during write operation, then **WriteMiss** will occur.
- If Write through protocol is used, the information is written directly into mainmemory.
- If Write back protocol is used, then block containing the addressed word is first brought into the cache and then the desired word in the cache is over-written withthe new information.

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, itis 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 readfrom the main memory and loaded into the cache.

Merit:

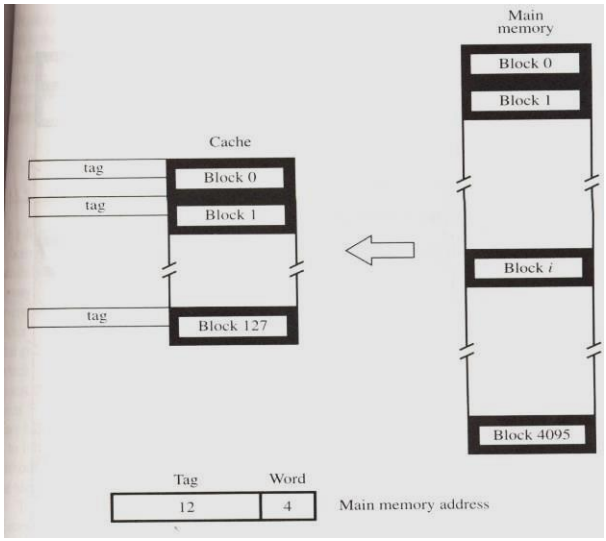
- It is easy to implement.

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 present. 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:

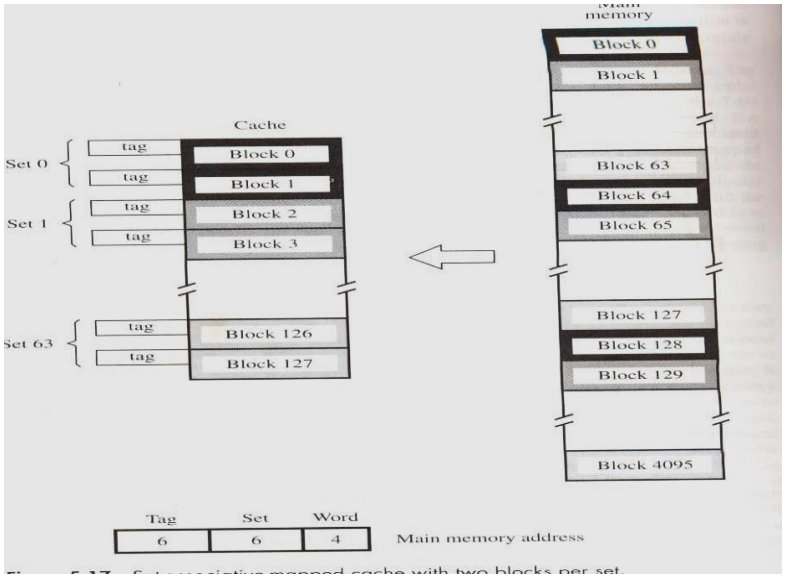
- Its cost is high.

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 the tags of the two blocks of the set to clock if the desired block is present.

Fig: Set-Associative Mapping:



<u>No of blocks per set no of set field</u>

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 **CacheCoherence 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- determinedly. 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 **LeastRecently Used(LRU) block** & the technique is called **LRU algorithm**.
- The cache controller tracks the references to all blocks with the help of blockcounter.

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 originallylower 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 isremoved, the new block is put in its place & its counter is set to „0" and other block counters are incremented by 1.

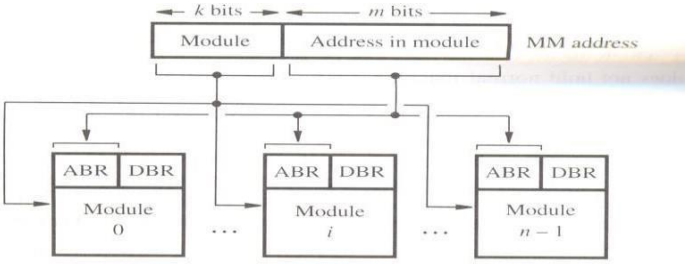
Merit:

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

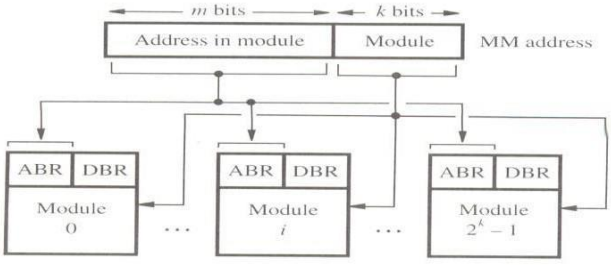
PERFORMANCE CONSIDERATION:

- Two Key factors in the commercial success are the performance & cost ie the best possible performance at low cost.
- A common measure of success is called the **Price Performance ratio**.
- Performance depends on how fast the machine instruction are brought to the processor and how fast they are executed.
- To achieve parallelism (i.e. Both the slow and fast units are accessed in the samemanner), **interleaving** is used.

Interleaving:
Fig: Consecutive words in a Module



(a) Consecutive words in a module



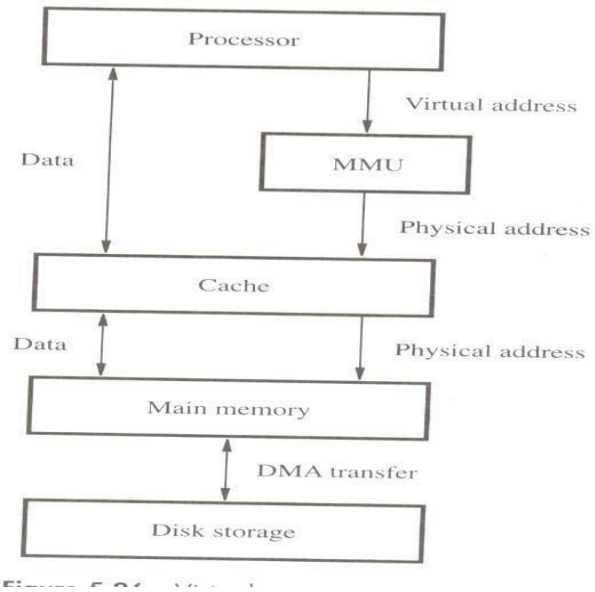
(b) Consecutive words in consecutive modules

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 tobring the data into memory from the disk.

- Transfer of data between disk and main memory is performed using DMA scheme.

Fig: Virtual Memory Organization



Address Translation:

- In address translation, all programs and data are composed of fixed length unitscalled **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** speeds up the gap between main memory and secondary storageand 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 thatpage currently resides in memory.

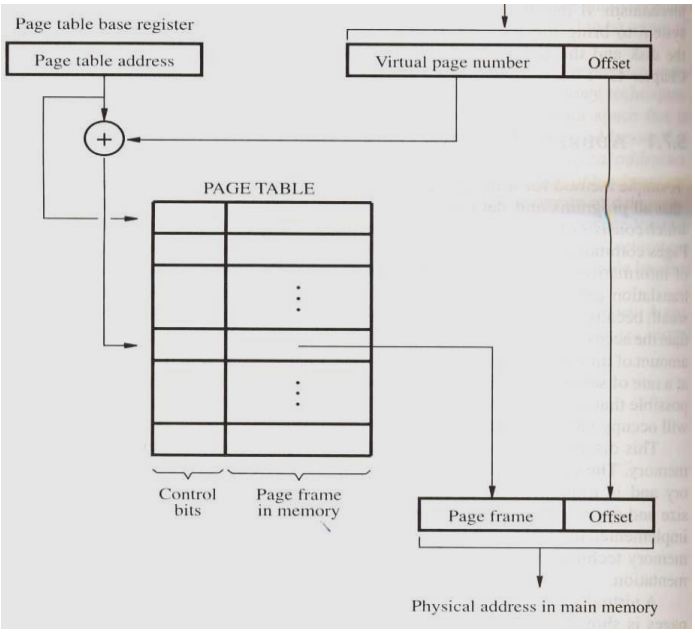
Control Bits in Page Table:

- The Control bits specifies the status of the page while it is in main memory.

Function:

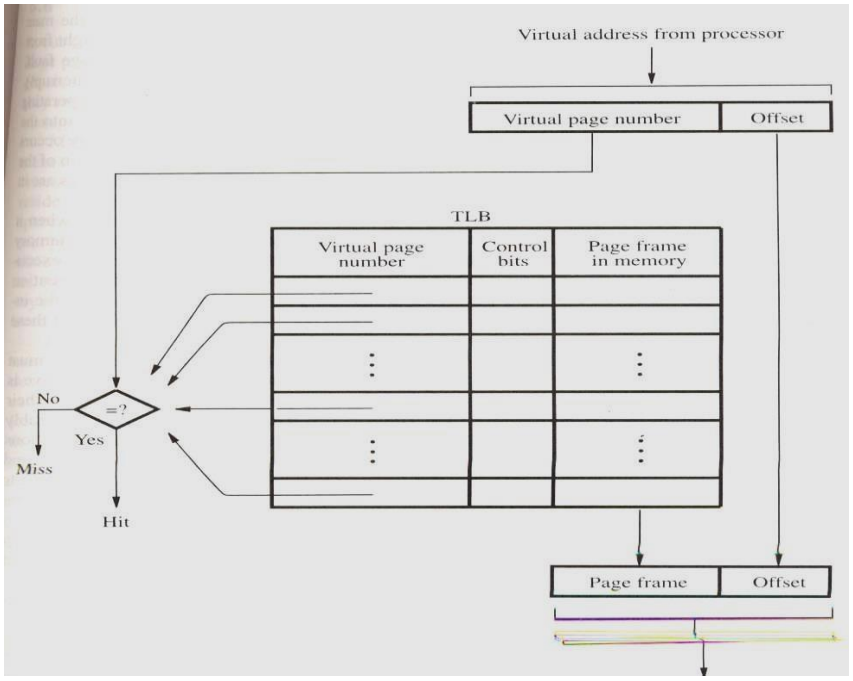
- The control bit indicates the validity of the page ie) it checks whether the page isactually loaded in the main memory.
- It also indicates that whether the page has been modified during its residency in the memory; this information is needed to determine whether the page should be written back to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

Fig: Virtual Memory Address Translation



- 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 entries that corresponds to the most recently accessed pages and also contains the virtual address of the entry.

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 brought from disk into memory 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.

MEMORY MANAGEMENT REQUIREMENTS:

Management routines are part of the Operating system.

Assembling the OS routine into virtual address space is called „**System Space**".

The virtual space in which the user application program reside is called the „**User Space'**.

Each user space has a separate page table.

The MMU uses the page table to determine the address of the table to be used in the translation process.

Hence by changing the contents of this register, the OS can switch from one space to another.

The process has two stages. They are,

- User State
- Supervisor state.

User State:

In this state, the processor executes the user program.

Supervisor State:

When the processor executes the operating system routines, the processor will bein supervisor state.

Privileged Instruction:

In user state, the machine instructions cannot be executed. Hence a user program isprevented from accessing the page table of other user spaces or system spaces. The control bits in each entry can be set to control the access privileges granted to each program.

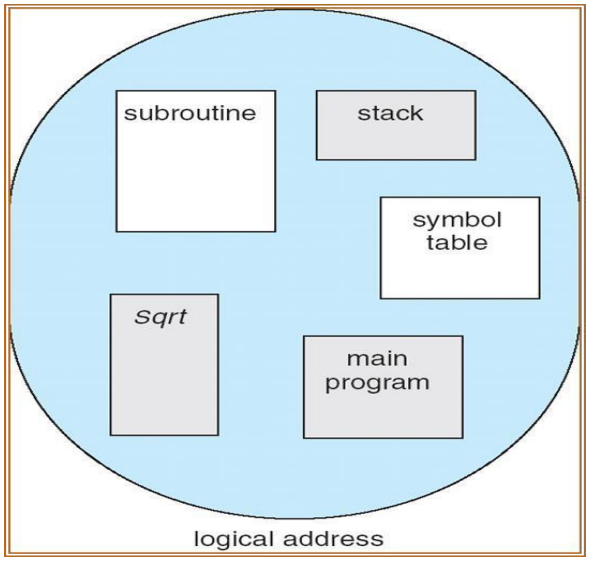
I.e.) One program may be allowed to read/write a given page, while the otherprograms may be given only red access.

SEGMENTATION

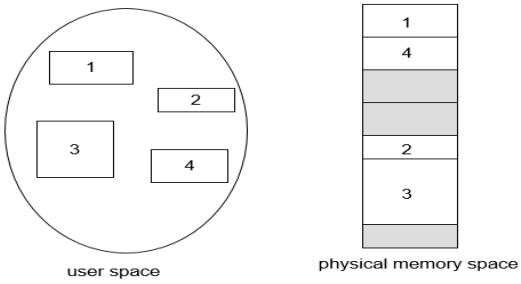
- Memory-management scheme that supports user view of memory.
- A program is a collection of segments.
- A Segment is a logical unit such as:
    - Main Program
    - Procedure

- Function
- Method
- Object
- Local Variables
- Global Variables
- Common Block
- Stack
- Symbol Table
- Arrays

User's View of a Program

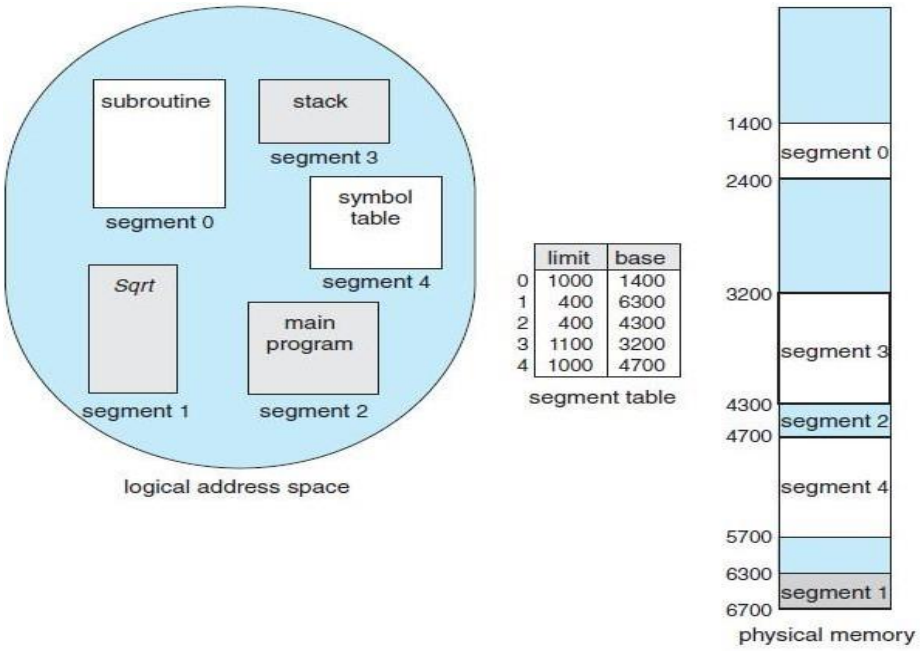


Logical View of Segmentation



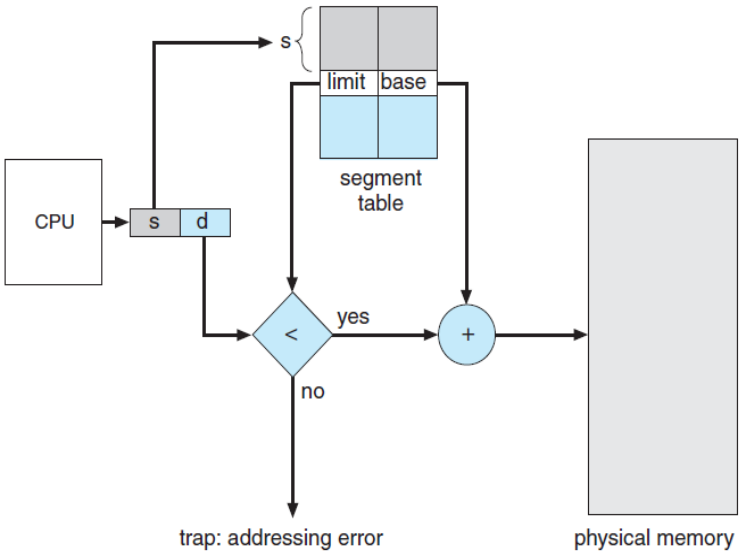
Segmentation Architecture

- A logical address space is a collection of segments.
- Logical address consists of a two tuple:
  - <*segment-number, offset*>
- *Segment table* maps two-dimensional physical addresses into one- dimensional physical addresses.
- Each table entry has:
  - *Base*: contains the starting physical address where the segments reside in memory.
  - *Limit*: specifies the length of the segment.



Segmentation Hardware

- Segment-table base register (STBR) points to the segment table's location in memory.
- Segment-table length register (STLR) indicates number of segments used by a program;
    - segment number s is legal if, s < STLR
- Since segments vary in length, memory allocation has tobe dynamic storage-allocation.



Advantages of segmentation:

1) Association of protection with the segments.

- With each entry in segment table associate:
    - validation bit = 0, it's illegal segment
    - read/write/execute privileges
- Protection bits are associated with segments, to prevent illegal access to memory.
- By placing an array in its own segment, the memory management hardware will automatically check that array indexes are legal & do not stay outside the array boundaries.

2) sharing of code or data:

- Code sharing occurs at segment level.
- Segments are shared when entries in the segment tables of two different processes point to the same physical location.

Fragmentation:

- Since segments vary in length, memory allocation is a dynamic storage-allocation problem, usually solved with best fit or first fit algorithm.
- Since segments are variable size, we have the problem of external fragmentation.