Computer System Architecture
UNIT – 4
Memory System

Syllabus

- Memory Characteristics
- Memory Hierarchy
- RAM and ROM Organization
- Interleaved Memory
- Cache Memory
- Virtual Memory

Memory Characteristics

Classification of Memory System Based on Key Characteristics

- Location
  - Processor
  - Internal
  - External
- Capacity
  - Word Length
  - Number of words
- Unit of Transfer
  - Word
  - Block
- Access Method
  - Sequential
  - Direct
  - Random
  - Associative
- Performance Measure
  - Access Time
  - Cycle Time
  - Transfer Rate
- Physical Type
  - Semiconductor
  - Magnetic
  - Optical
  - Magneto-Optical
- Retention Characteristics
  - Volatile / Non-volatile
  - Erasable / Non-erasable
- Organization

Memory Characteristics

Classification of Memory System Based on **Location**

- Processor
  - This is often in the form of CPU registers and small amount of cache.
- Internal or Main Memory
  - This is the main memory like RAM or ROM. The CPU can directly access the main memory.
- External or Secondary Memory
  - It comprises of secondary storage devices like hard disks, magnetic tapes.
  - The CPU doesn’t access these devices directly.
  - It uses device controllers to access secondary storage devices.

Memory Characteristics

Classification of Memory System Based on **Capacity**

- The capacity of any memory device is expressed in terms of
  - Word size:
    - Words are expressed in bytes (8 bits).
    - A word can however mean any number of bytes.
    - Commonly used word sizes are 1 byte (8 bits), 2bytes (16 bits) and 4 bytes (32 bits).
  - Number of words:
  - This specifies the number of words available in the particular memory device.
  - For example, if a memory device is given as 4K x 16.
    - This means the device has a word size of 16 bits and a total of 4096(4K) words in memory.

Memory Characteristics

Classification of Memory System Based on **Unit of Transfer**

- Unit of Transfer is the maximum number of bits that can be read or written into the memory at a time.
- In case of main memory, it is mostly equal to word size.
- In case of external memory, unit of transfer is not limited to the word size; it is often larger and is referred to as blocks.

Memory Characteristics

Classification of Memory System Based on **Access Method**

- Access Method is a fundamental characteristic of memory devices.
- It is the sequence or order in which memory can be accessed.
- There are three types of access methods:
  - Sequential
    - Access is made in a specific linear sequence.
    - Access time is highly variable.
  - Direct Access
  - Random Access
  - Associative

Memory Characteristics

Classification of Memory System Based on **Performance Measure**

- Performance of the memory system is determined using the following three parameters
- Access Time
  - For random access memories, it is the time taken by memory to complete the read/write operation from the instant that an address is sent to the memory.
  - For non-random access memories, it is the time taken to position the read write head at the desired location. Access time is widely used to measure performance of memory devices.
- Memory cycle time:
  - It is defined only for Random Access Memories and is the sum of the access time and the additional time required before the second access can commence.
- Transfer rate:
  - It is defined as the rate at which data can be transferred into or out of a memory unit.

Memory Characteristics

Classification of Memory System Based on **Performance Measure**

- Transfer rate:
  - It is defined as the rate at which data can be transferred into or out of a memory unit.

For a random access memory it is equals to 1.

For a non random access memory It can be calculated as-

$$TN = TA + N/R$$

where,	TN	-	Avg time to read or write
	TA	-	Average access time
	N	-	Number of bits
	R	-	Transfer rate in bits per second

Memory Characteristics

Classification of Memory System Based on **Physical Type**

- Semiconductor
- Magnetic
- Optical
- Magneto-Optical

Memory Characteristics

Classification of Memory System Based on **Retention Characteristics**

- Volatile / Non-volatile
  - Volatile : Continues to hold data even after system restart
  - Non-volatile : Data is lost on system restart

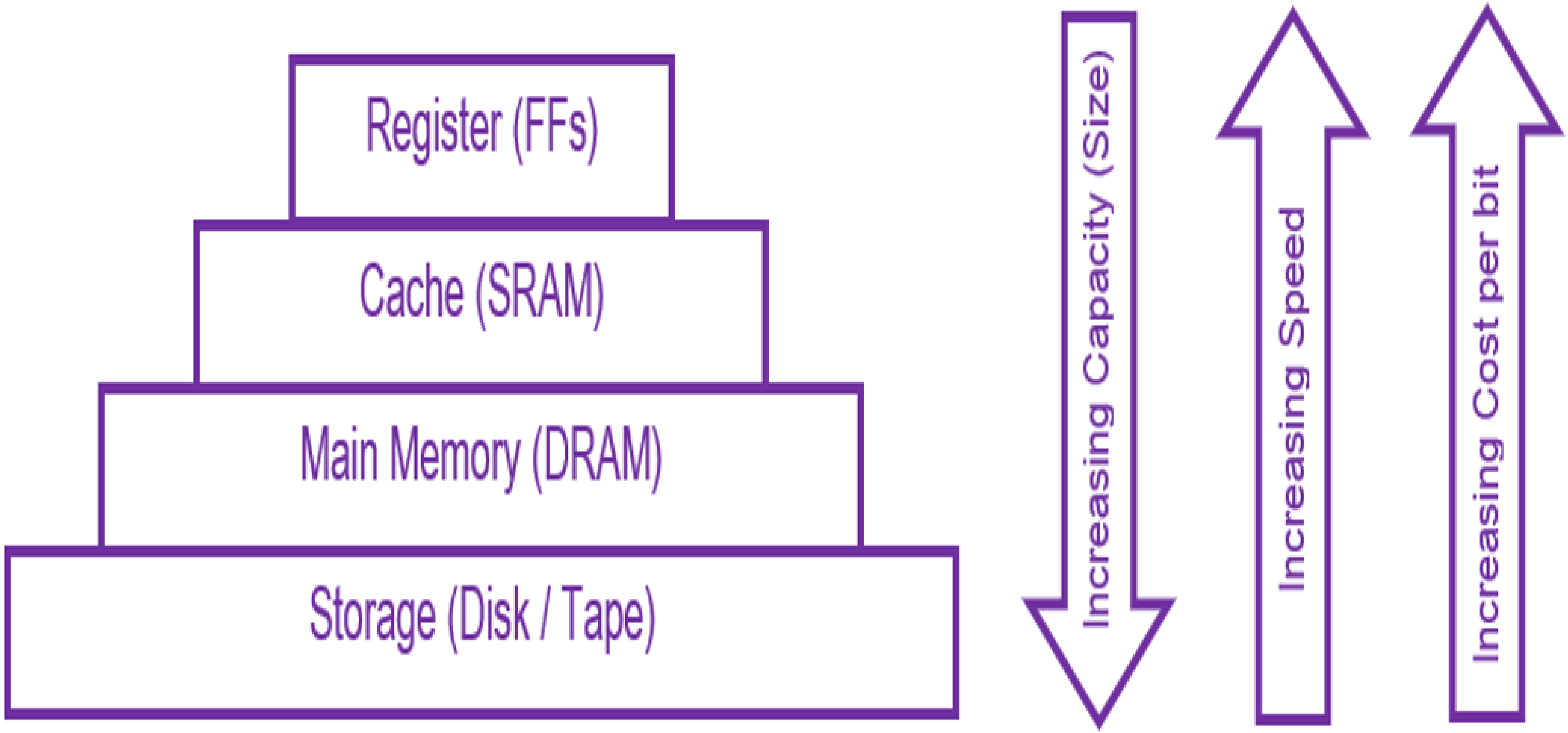
- Erasable / Non-erasable
  - Erasable : Data can be erased upon requirement.
  - Non-erasable : Data can not be erased once it is stored into these memory.

Memory Characteristics

Organization

- It is the physical arrangement of bits to form word.
- The memory is organized in the form of a cell.
- Each cell is able to be identified with a unique number called address.
- Organization is the key concept in RAM design.

Memory Hierarchy



Memory Hierarchy

- Principle of locality (temporal locality & spatial locality) states that programs do not access code and data uniformly.
- Smaller hardware is faster.
- Faster hardware is expensive.
- Performance enhancement is realizable by
- keeping frequently used code and data in fast memory such as register, cache, and main memory
- and the rest of the code and data in slower memory such as disk and tape drives

Memory Hierarchy

Speed, Size and Cost of Memory

- Faster Access Time → Higher Cost per bit
- Greater Capacity → Lower Cost per bit
- Greater Capacity → Slower Access Time
- To obtain optimum performance, at a reasonable cost,
  a smaller more expensive and faster memory
  is often supplemented by
  larger, cheaper and slower memory

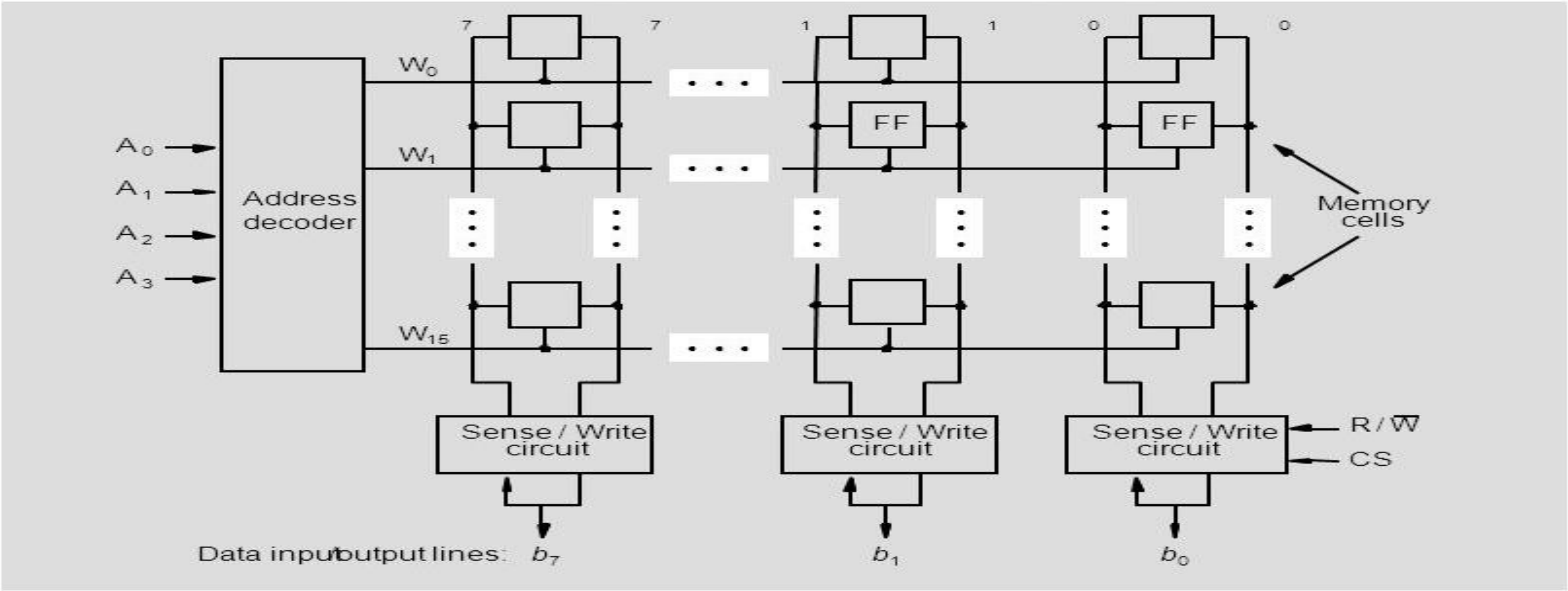
RAM and ROM Organization

RAM

- Memory cells are organized in the form of an array of cells.
- Each cell capable of storing one bit of information.

Semiconductor RAM memories

Internal organization of memory chips

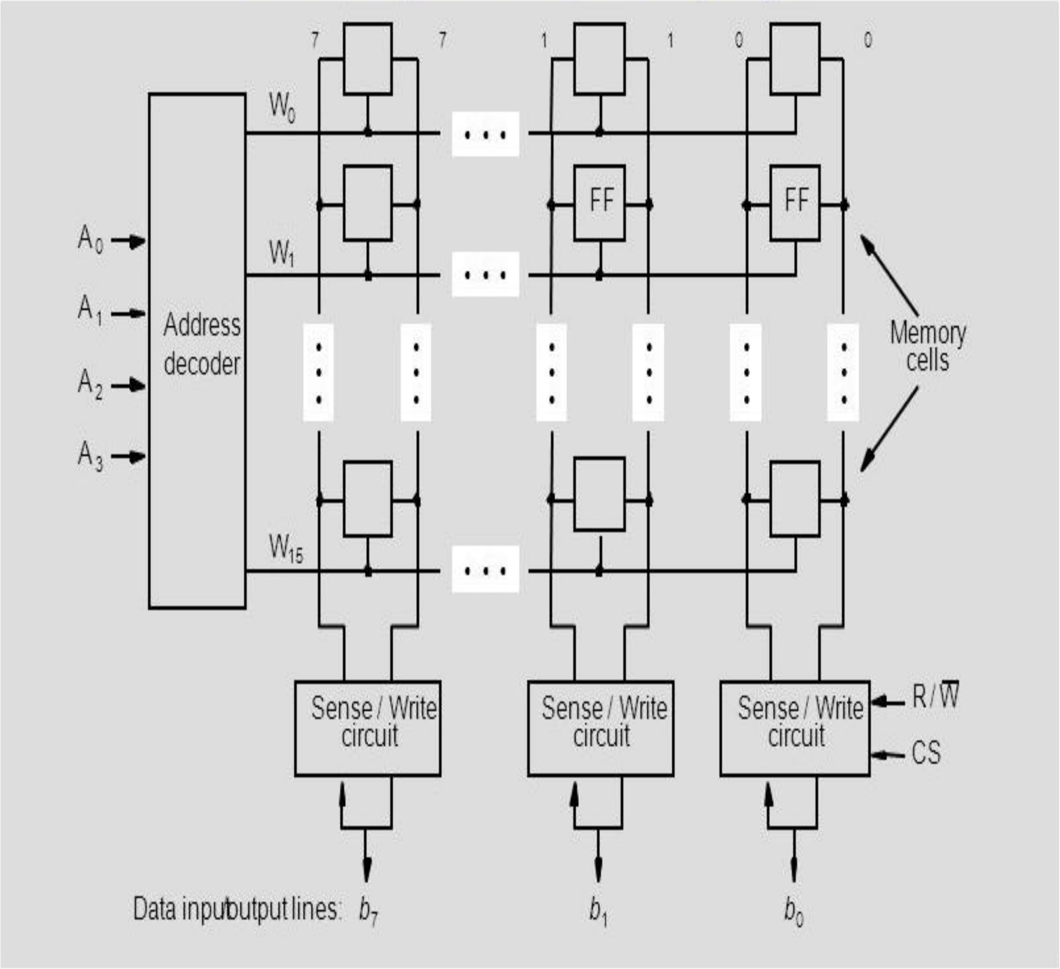


RAM and ROM Organization

- Each row of cells constitute a memory word of 8 bits b0-b7 and cells of a row are connected to a common signal line called “word line”, which is driven by the address decoder.
- Two ‘bit lines' connect the cells in each column to a sense /write circuit.
- The sense/write circuit are connected to the data I/O lines.
- During a 'Read' operation, these circuits read the information stored and transmits this information to the output data line
- During a 'Write' operation, the sense/write circuit receive input information and store it in the selected cell.
- This following organization of a very small memory chip consisting of 8bits of 16 word i.e. 16×8 organization.

Semiconductor RAM memories

Internal organization of memory chips

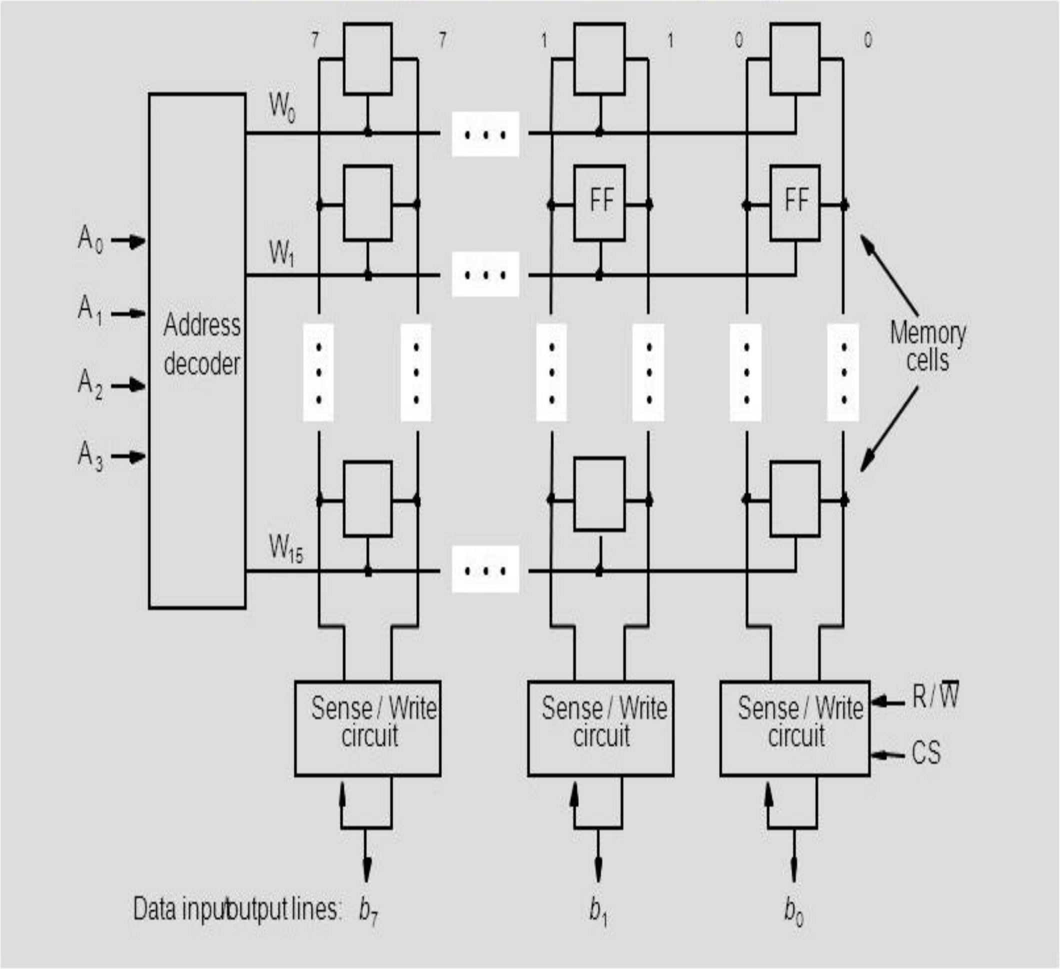


RAM and ROM Organization

- The data I/O of each sense/write circuit are connected to a single bidirectional data lines that are connected to the data bus of the computer.
- In addition there are also two control lines R/W and CS(chip select)
- The R/W signal line specifies the required operations and CS selects a given chip in multi -chip memory system.
- This memory circuit stores 128 bit and it requires 14 external connection for address, data and control line. It also needs two lines for power supply and ground connection.
- For a larger memory circuit let 1k(1024) memory cell can be organized as 128×8 memory require 19 external connection.

Semiconductor RAM memories

Internal organization of memory chips



RAM and ROM Organization

RAM

- Static RAM
  - Memory capable of current state as long as power is applied
- Dynamic RAM
  - Memory that is not as expensive as static memory and not capable of storing information indefinitely. Contents are periodically refreshed to keep the information for longer time period.
- Asynchronous DRAM
  - Does not use system clock to coordinate the memory access
- Synchronous DRAM (SDRAM)
  - Uses the system clock to coordinate the memory access
- Double-Data-Rate SDRAM (DDR SDRAM)
  - Transfers data on both edges of the clock
- Latency
  - Amount of time it takes to transfer a word of data to or from memory
- Bandwidth
  - Amount of data (in bits or Bytes) that can be transferred in a unit time (per second)

RAM and ROM Organization

ROM

- Both SRAM and DRAM chips are volatile in nature as they lose the stored information if the power is removed.
- Many applications need memory devices to retain the stored information even if power is removed.
- When a computer is turned on the operating system software has to be loaded from the disk into the memory.
- The boot program is quite large and most of it is stored on the disk, but the processor should execute some instructions that loads the boot program;
- So a small amount of non-volatile memory is required to hold the instruction which helps in loading the boot program from the disk.
- Since, it’s normal operation involves only reading of stored data, a memory of this type is called Read only memory.

RAM and ROM Organization

ROM

- PROM (Programmable ROM)
- EPROM (Erasable PROM)
- EEPROM (Electronically EPROM)
- Flash Memory
- Flash Card
- Flash Drive

Interleaved Memory

- The two key factors in the success of computer are performance and cost.
- This can be achieved through parallelism.
- In parallel processing or pipeline environment, the main memory is the prime system resource, which is normally shared by all processor or stages of the pipeline.
- In such cases there may be memory interference, which as a result degrades the performance.
- So to avoid this problem new method is adopted which is known as “Memory interleaving”.
- The Memory interleaving means the main memory of the computer is partitioned into a no of modules and distributing the address among those modules.
- Each memory module has its own Address Buffer Register(ABR) or Memory Address Register(MAR) and Data Buffer Register(DBR) or Memory Buffer Register(MBR) or Memory Data Register(MDR).
- There are two memory interleaving layouts :
  - High order interleaving(Consecutive words in a module)
  - Low order interleaving (consecutive words in consecutive module)

Interleaved Memory

High Order Interleaving

- Memory is divided into M modules where the consecutive address lies in a single module.
- The higher order bits are used for indicating the module no. and the lower order bits are used for the words in the module .
- Each memory address is of n bit out of which the higher order m bits are used for interleaving and n-m bits are used for the words in particular module.
- The m bits are being decided by the decoder which will specify the particular module no. and n-m bit specify the words in the module .
- Every memory module has it’s own MAR and MBR.
- Advantage
  - Permits easy expansion of memory by addition of one or more memory module as needed to a maximum of m-1.
  - Better system reliability in case of a failed module as it affects only a localized area of address space.
- Disadvantage:-
  - When consecutive location are to be accessed then only one module is involved

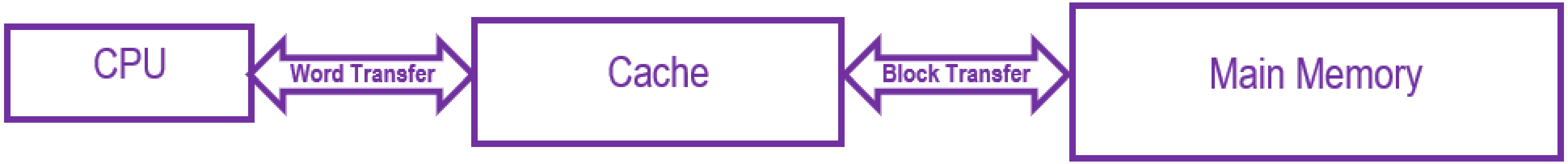
Interleaved Memory

Low Order Interleaving

- Consecutive words are distributed in consecutive modules.
- Here the higher order n-m bits are used for address of words in a module while m lower bits are used for module no.
- This method is efficient way to address the module.
- Here any request for accessing consecutive words can keep several modules busy at the same time.
- This is faster than the previous one and so used frequently.

Cache Memory

- The speed of main memory is very slow in compare to the speed of processor. So for better performance, a high speed memory is used in between main memory and CPU. That is called as cache memory.
- The cache memory comes from the word cache meaning to hide.
- The basic idea behind a cache is simple i.e. the most heavily used memory word are kept in the cache, when the CPU needs a word it will first look in the cache, only if the word is not there, it goes to the main memory.
- Analysis of any standard program shows that the maximum program execution time spent in those portion in which many instruction were executed repeatedly as in loops.
- Hence the execution of the programs forms a localize area, where the programs or instruction executed repeatedly and the remainder of the programs are executed relatively less frequently that is called locality of reference.



Cache Memory

Read operation

- When the CPU needs to access memory, the cache is first searched.
- if the word is found, then it is known as “cache hit”
- If the word is not found, then it is known as “cache miss”
- When a “cache miss” occurs it initiate to access main memory to transfer the required byte or word from main memory to cache.
- The performance of the cache memory is known as hit ratio.

Problem: Cache Access Time = 100hs; Main memory access time is 1000ns; Hit Ratio=0.9

- Average memory access time = h* cache access time + (1-h)*main memory access time

= 0.9 * 100 + (1-0.9)*100 = 90 + 100 =190ns

Cache Memory

Write operation

- During read operation, when the CPU finds a word in cache memory, then the main memory is not involved in the transfer.
- But in case of write operation there are two ways of writing:
  - **Write through policy**
    - The simplest and most commonly used procedure is to update main memory with every memory write operation with cache memory being update in parallel.
    - **Advantage:** This method is the most important characteristics of direct memory access transfer.
  - **Write back policy**
    - If the cache follows this policy then the cache is updated during write operation and the location is marked by a flag.
    - When the block of the cache containing the flagged word is required to transfer the main memory at that time it is updated in main memory.
    - **Advantage:** Whenever the word is updated several times, it is better to use write back policy.

Virtual Memory

- Virtual memory is the most fundamental memory management concept to be implemented for memory management functions such as space allocation, program relocation, code sharing and protection.
- The key idea is to allow a user program more memory locations than those available in physical memory.
- A virtual address is generated by a user program and the set of virtual addresses constitute the virtual address space.
- The main memory of a computer contains a fixed number of memory locations and a set of these locations are the physical address space.

Virtual Memory

- In the early days of computer development, programmers wrote large volumes of program code that could not fit into the main memory.
- A programmer had to design overlays to make them independent to each other.
- With this kind of provision, one can successively bring each overlay into the main memory and execute them in a sequence.
- This considerably increases the program development time.

Virtual Memory

- In a system using virtual memory, the size of the virtual address space is much larger than the physical address space.
- In this system, a programmer does not have to worry about overlay design and can write assuming unlimited address space.
- In a virtual memory system, the programming effort can be greatly simplified, as the actual number of physical addresses available is considerably less than the number of virtual addresses provided by the system.
- There should be some mechanism for dividing a large program into small overlays automatically.
- A virtual memory system performs a series of mapping operations that mechanizes the process of overlay generation.
- A virtual memory system can be configured as follows:

  Memory Paging **Memory Segmentation.**

Virtual Memory

Memory Paging

- Virtual memory is divided into blocks of equal size, called pages.
- Physical memory is also divided into frames in the same way.
- Page size = Frame size = 512 / 1024 / 2048 words.
- Each virtual address can be considered to be an ordered pair <p, n>

  where, p = page number, n = offset
- A user program consists of sequence of pages and a complete copy of the program is always held in a magnetic drum or disk.
- The user program can be placed in any available page frame of the main memory.
- Pages are brought from secondary memory and are stored in main memory in a dynamic manner.
- All addresses given by a user program must be translated into physical addresses in the memory.

Virtual Memory

Memory Paging

- When a running program tries to access a virtual memory location v, the mapping algorithm returns the virtual page p as mapped to the physical frame p.
- Then the physical address is determined by appending p to n.
- This dynamic address translator implemented using page table is maintained in the main memory.
- This table will contain one entry for each virtual page of the virtual address space.

Virtual Memory

Memory Segmentation

- Paging concept is viewed as a one-dimensional technique as virtual addresses generated by a program increase linearly from 0 to some maximum value.
- It is desirable to have a multidimensional virtual address space and this is the key idea behind segmentation system.
- In a segmentation system, each logical entity like stack, array, subroutine etc. have a separate virtual address space.
- The virtual address is called a segment and each segment can grow from zero to a maximum value independently without affecting other segments.
- Segment details (which are called as segment descriptors) are stored in segment table.

Virtual Memory

Memory Segmentation

- Typically, a segment descriptor consists of the following information:
  - Segment base address b (starting address of the segment in main memory)
  - Segment length l (size of a segment)
  - Segment presence bit (virtual address present in memory)
  - Protection bits (protection protocols)
- Each segment can be of different size.
- If all segments are of same size in the segmentation system, then it is called paging system.

Virtual Memory

Paging Vs Segmentation

Parameters	Paging	Segmentation
Individual Memory	In Paging, we break a process address space into blocks known as pages.	In the case of Segmentation, we break a process address space into blocks known as sections.
Memory Size	The pages are blocks of fixed size.	The sections are blocks of varying sizes.
Accountability	The OS divides the available memory into individual pages.	The compiler mainly calculates the size of individual segments, their actual address as well as virtual address.
Speed	This technique is comparatively much faster in accessing memory.	This technique is comparatively much slower in accessing memory than Paging.
Size	The available memory determines the individual page sizes.	The user determines the individual segment sizes.
Fragmentation	The Paging technique may underutilize some of the pages- thus leading to internal fragmentation.	The Segmentation technique may not use some of the memory blocks at all. Thus, it may lead to external fragmentation.
Logical Address	A logical address divides into page offset and page number in the case of Paging.	A logical address divides into section offset and section number in the case of Segmentation.
Data Storage	Page table leads to the storage of the page data.	Segmentation table leads to the storage of the segmentation data.

Important Questions

- Describe characteristics of memory system.
- Describe memory hierarchy.
- Define the following:
  - Sequential access
  - Direct access
  - Random access
  - Associative access
  - Access time
  - Cycle time
  - Transfer rate
  - Latency
  - Bandwidth
- Explain various RAM types.
- Explain various ROM types.
- Write short notes on the following:
  - MAR & MDR
  - Big-endian & little-endian
  - Virtual memory
  - Cache memory
  - Memory interleaving
  - Paging
  - Segmentation
  - Page table
  - Segmentation table
- Differentiate between the following:
  - RAM vs ROM
  - Segmentation vs Paging