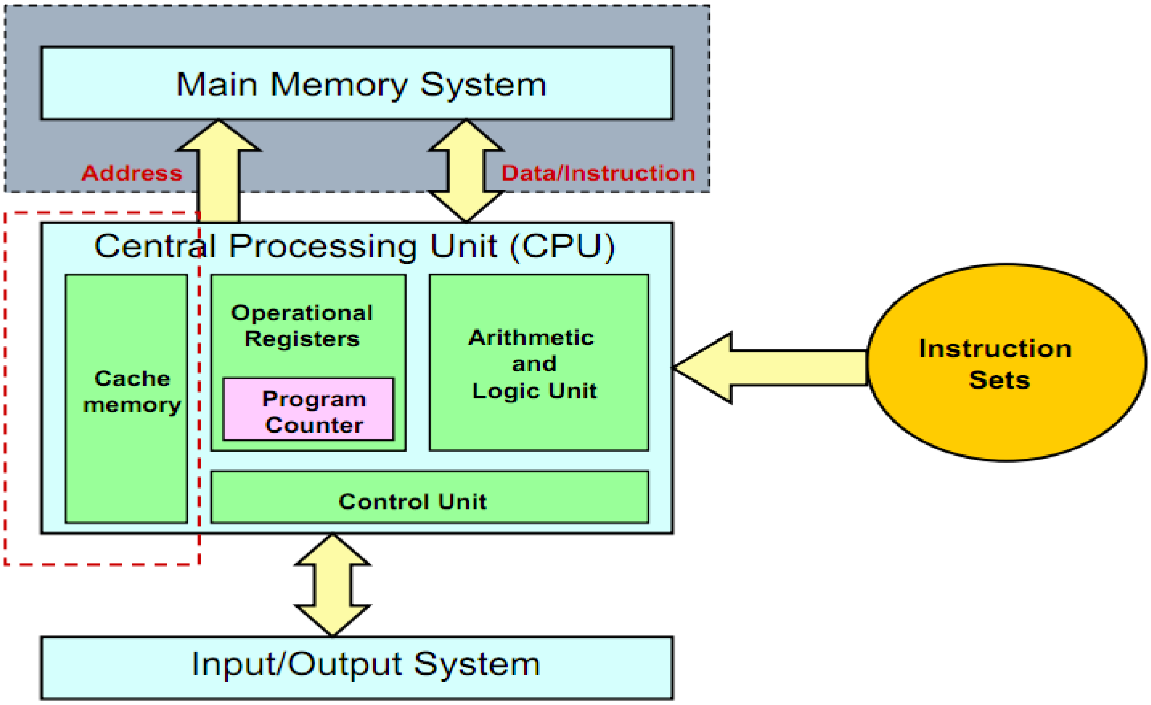
COMPUTER - MEMORY

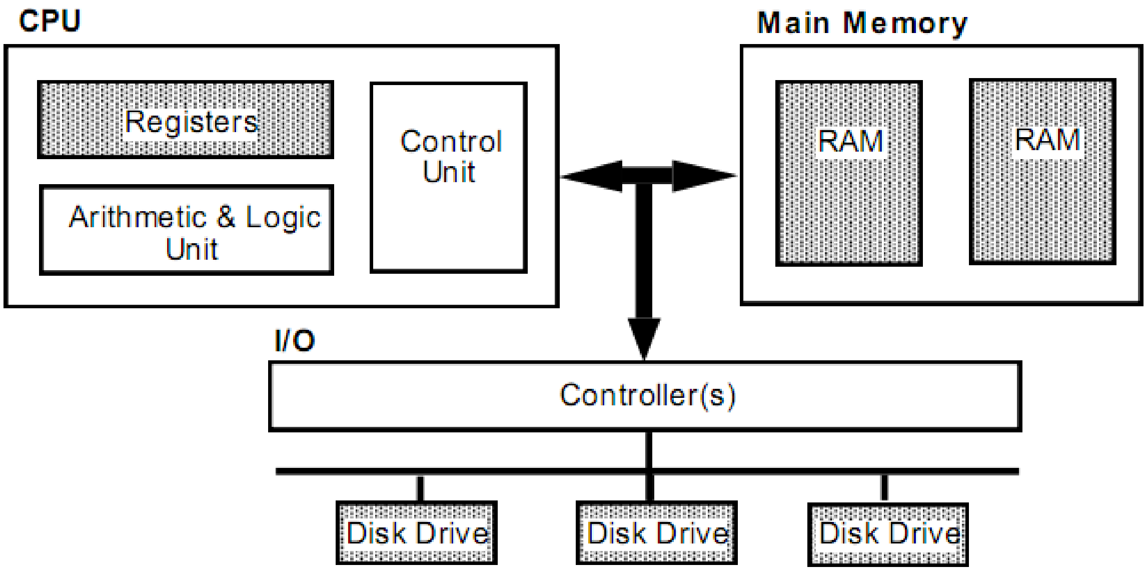
A memory is just like a human brain. It is used to store data and instructions. Computer memory is the storage space in computer where data is to be processed and instructions required for processing are stored. The memory is divided into large number of small parts called cells. Each location or cell has a unique address which varies from zero to memory size minus one. For example if computer has 64k words, then this memory unit has 64 * 1024=65536 memory locations. The address of these locations varies from 0 to 65535.

Memory circuits can largely be separated into two major groups: dyanamic memories that store data for use in a computer system (such as the RAM in a PC); and static memories that store information that defines the operating state of a digital system.



Main Memory (RAM) Organisation

Computers employ many different types of memory (semi-conductor, magnetic disks, USB sticks, DVDs etc.) to hold data and programs. Each type has its own characteristics and uses. We will look at the way that Main Memory (RAM) is organised and briefly at the characteristics of Register Memory and Disk Memory. Let us locate these 3 types of memory in a simplified model of a computer:



Basic Concepts

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

- For example, a 16-bit computer that generates 16-bit addresses is capable of addressing up to 2^{16}=64K memory locations.
- Similarly, machines whose instructions generate 32-bit addresses can utilize a memory that contains up to 2^{32}=4G memory locations.
- Data transfer between the memory and processor takes place through the use of two processor registers, MAR and MDR.
- One way to reduce the memory access time is to use a cache memory.
- Cache memory is a small, fast memory that is inserted between the larger, smaller main memory and the processor.

Main Memory (RAM)

If we were to sum all the bits of all registers within CPU, the total amount of memory probably would not exceed 5,000 bits. Most computational tasks undertaken by a computer require a lot more memory. Main memory is the next fastest memory within a computer and is much larger in size.

Typical main memory capacities for different kinds of computers are: PC 512MB5 , fileserver 4GB , database server 8GB.

Computer architectures also impose an architectural constraint on the maximum allowable RAM. This constraint is normally equal to 2WordSize memory locations.

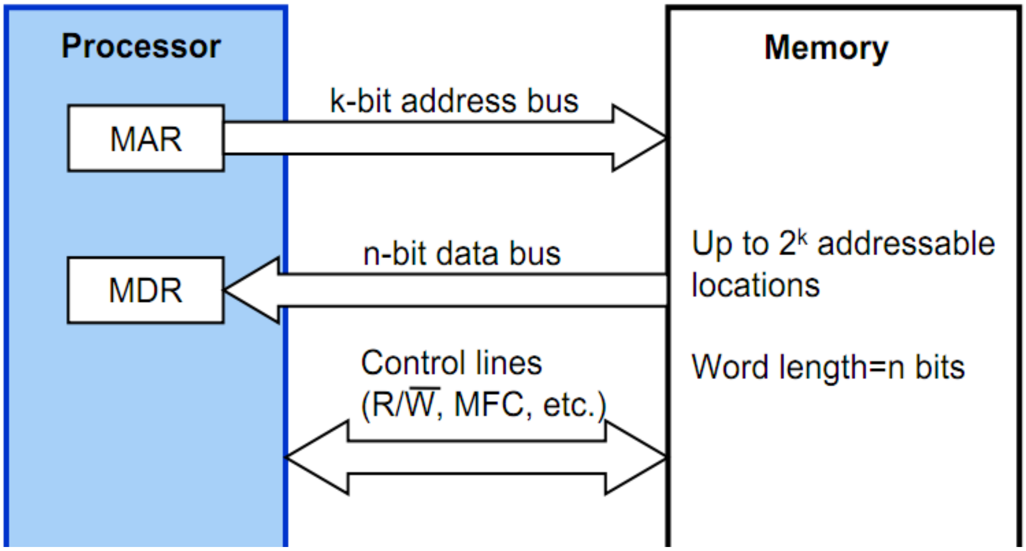
RAM (Random Access Memory) is the most common form of Main Memory. RAM is normally located on the motherboard and so is typically less than 12 inches from the CPU.

ROM (Read Only Memory) is like RAM except that its contents cannot be overwritten and its contents are not lost if power is turned off (ROM is non-volatile).Although slower than register memory, the contents of any location in RAM can still be "read" or "written" very quickly . The time to read or write is referred to as the access time and is constant for all RAM locations.

In contrast to register memory, RAM is used to hold both program code (instructions) and data (numbers, strings etc). Programs are "loaded" into RAM from a disk prior to execution by the CPU.

Locations in RAM are identified by an addressing scheme e.g. numbering the bytes in RAM from 0 onwards [10]. Like registers, the contents of RAM are lost if the power is turned off.

Connection of the Memory to the Processor



Memory is primarily of three types

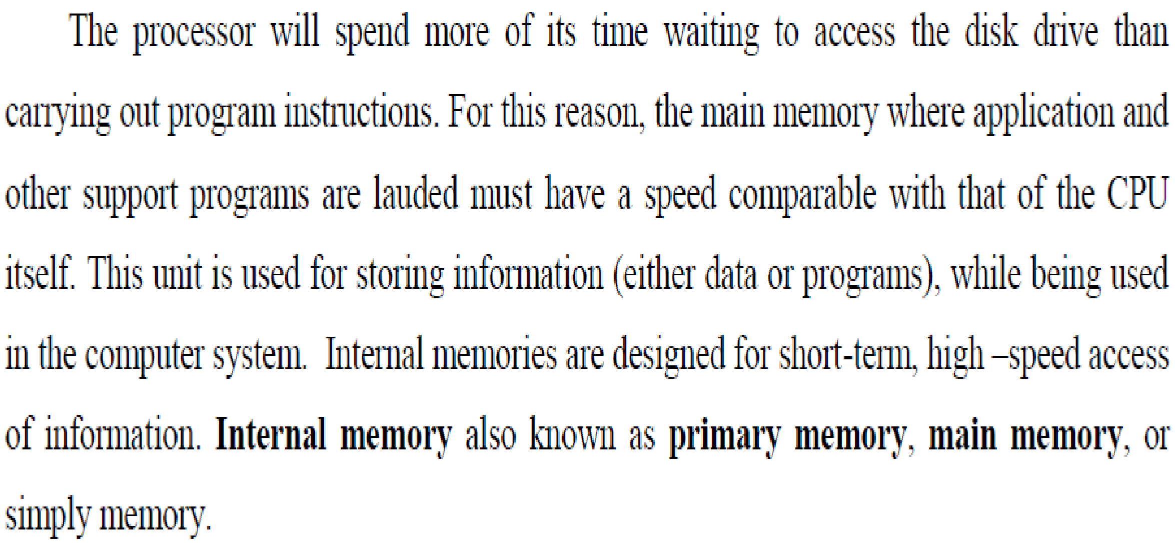
1) Primary Memory/Main Memory

2) Cache Memory

3) Secondary Memory

Primary Memory (Main Memory)

Primary memory holds only those data and instructions on which computer is currently working. It has limited capacity and data is lost when power is switched off. It is generally made up of semiconductor device. These memories are not as fast as registers. The data and instruction required to be processed reside in main memory.



The processor will spend more of its time waiting to access the disk drive than carrying out program instructions. For this reason, the main memory where application and other support programs are lauded must have a speed comparable with that of the CPU itself. This unit is used for storing information (either data or programs), while being used in the computer system. Internal memories are designed for short-term, high –speed access of information. **Internal memory** also known as **primary memory**, **main memory**, or simply memory.

Internal memory(Primary memory) is divided into two subcategories RAM and ROM : **RAM** (Random Accesses Memory) and **ROM** (Read Only Memory).

1) Read-Write or Random-Access-Memory(RAM)

It has a variable content , also it is generally used to store the variable data-RAM can also be used to store frequently changed programs and other information. RAM allows the computer to store information quickly for later reference, so that(in most personal computer). RAM holds:-

- The active part of the operating system, the fundamental program that control the operation of the computer.
- The application programs being executed(for example a word processing program).
- Data used by the application program(for example a letter being written with the word processing program.
- A representation of the data being presented on the video display.

A computer's memory is in constant communication with the CPU, as the program being executed. This type of memory contains a large number of semiconductor storage cells, each capable of storing a one-bit of instruction. A bit which is short of binary digit which is either 1 or 0(either full or empty).

A group of eight bits is called a **byte**. Since a bite represents only a very small a mount of information,. The usual approach is to deal them in a groups of fixed size. For this purpose, the memory is organized so that a group of n-bits referred to as a **word** of information and n is called the **word size**(word length).

Memory chips:

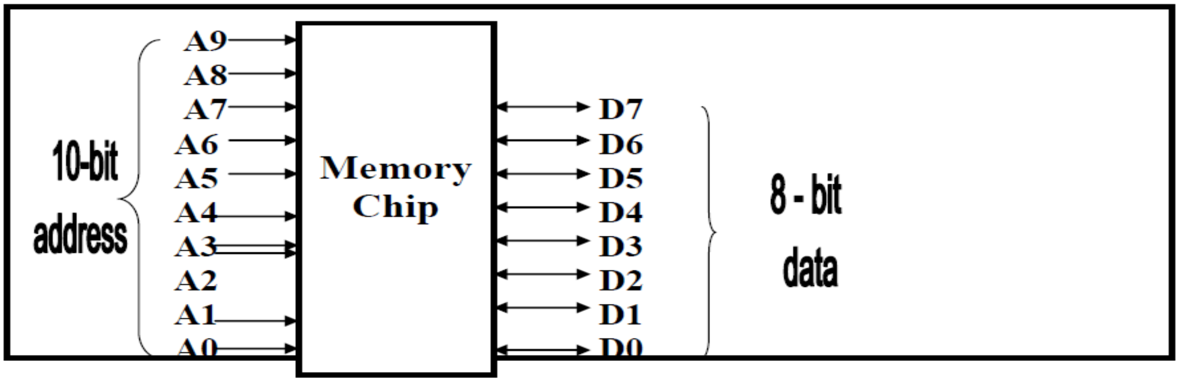
Memory chips have two main properties that determine their application, storage capacity or size and access time or speed.

A memory chip contains a number of locations, each of which stores one or more bits of data known as its **bit width**. The storage capacity of a memory chip is the product of the number of locations and the bit width. For example, a chip with 512 locations and a 2-bit data width, has a memory size of 512×2=1024 bits.

Since the standard unit of data is a byte(8 bits), the above storage capacity is normally given as 1024/8 =128 bytes.

Since the standard unit of data is a byte(8 bits), the above storage capacity is normally given as 1024/8 =128 bytes.

The number of locations may be obtained from the address width of the chip. For example, a chip with 10 address lines has 210 = 1024 or 1 k locations. Given an 8-bit data width, a 10- bit address chip has a memory size of 210 ×8 = 1024×8 = 1k ×1 byte = 1 k byte or 1KB.



The computer's word size can be expressed in bytes as well as in bits.

For example, a word size of 8-bit is also a word size of one byte, a word size of 16- bit is a word size of two byte. Computers are often described in terms of their word size, such as an 8-bit computer, a 16-bit computer and so on.

For example, a 16-bit computer is one in which the instruction data are stored in memory as 16-bit units, and processed by the CPU in 16-bit units. The word size also indicates the size of the data. Bus which carries data between the CPU and memory and between the CPU and I/O devices.

To access the memory, to store or retrieve a single word of information, it is necessary to have a unique address. The word address is the number that identifies the location of a word in a memory.

Each word stored in a memory device has a unique address. Address are always expressed as binary number, although hexadecimal and decimal numbers are often used for convenience.

NOTS:

- The large computer(mainframes) have word-sizes that are usually in the 32-to-64 –bits range.
- Mini computers have a word sizes from 8-to-32-bits range.
- Microcomputers have a word sizes from 4-to-32-bits range.

In general a computer with a larger word size, can execute programs of instruction at a fast rate because more data and more instruction are stuffed into one word. The larger word sizes, however, mean more lines making up the data bus, and therefore more interconnections between the CPU and memory and I/O devices.

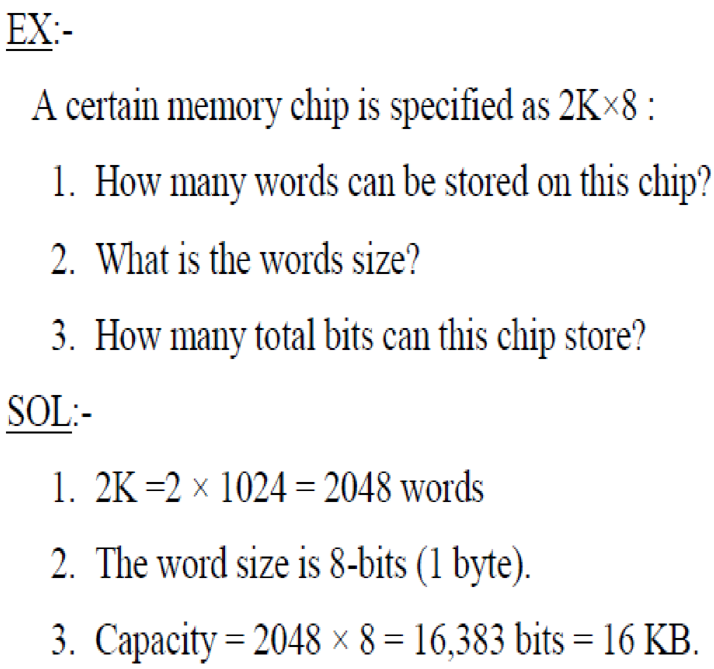
A more important than a computer's word size is the a mount of memory the computer has, (i.e.) the memory capacity.

A memory capacity is a way of specifying how many bits can be stored in a particular memory device or complete memory system. The capacity of memory depends on **two** parameters, **the number of words**(m) and **the number of bits per word** (n).

Memory capacity = (number of word) × (number of bits per word)

= m (word) * n (bits)

= m*n bits



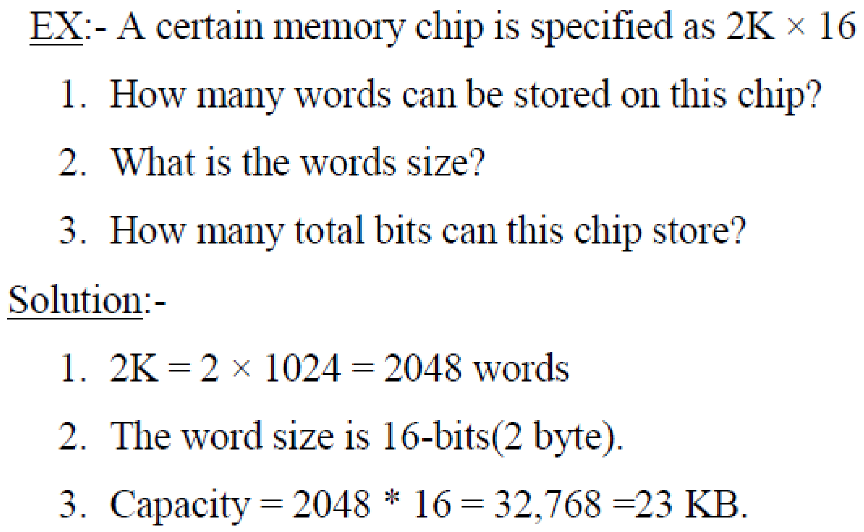
EX:-

A certain memory chip is specified as 2K×8 :

1. How many words can be stored on this chip?
2. What is the words size?
3. How many total bits can this chip store?

SOL:-

1. 2K =2 × 1024 = 2048 words
2. The word size is 8-bits (1 byte).
3. Capacity = 2048 × 8 = 16,383 bits = 16 KB.



EX:- A certain memory chip is specified as 2K × 16

1. How many words can be stored on this chip?
2. What is the words size?
3. How many total bits can this chip store?

Solution:-

1. 2K = 2 × 1024 = 2048 words
2. The word size is 16-bits(2 byte).
3. Capacity = 2048 * 16 = 32,768 =23 KB.

The Maximum size of the memory in any computer is determined by the number address lines, provided by processor used in the computer. For ex: if processor has 20 address lines, it is capable of addressing 2^{20} = 1M (mega) memory locations.

The maximum bits that can be transferred from memory or to the memory depend on the data lines supported by the processor. From the system standpoint, the memory unit is viewed as a black box. Data transfer between the memory and the processor takes place through the two processor registers AR(Address Register) and DR(Data Register).

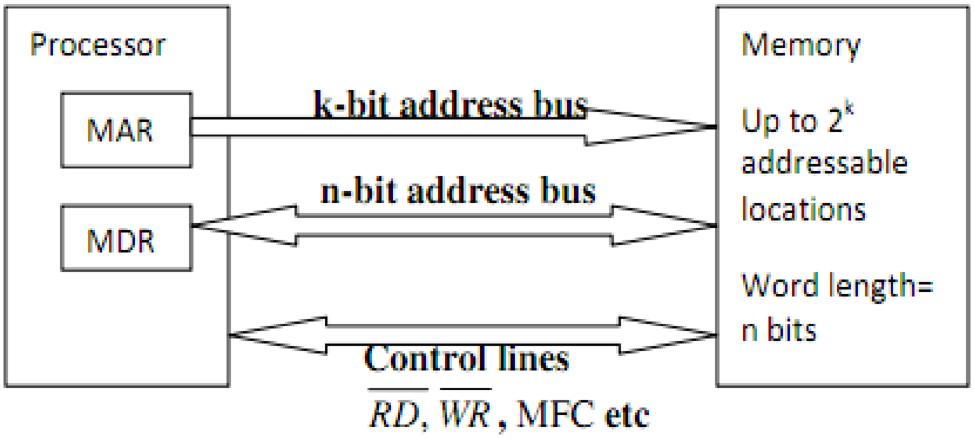
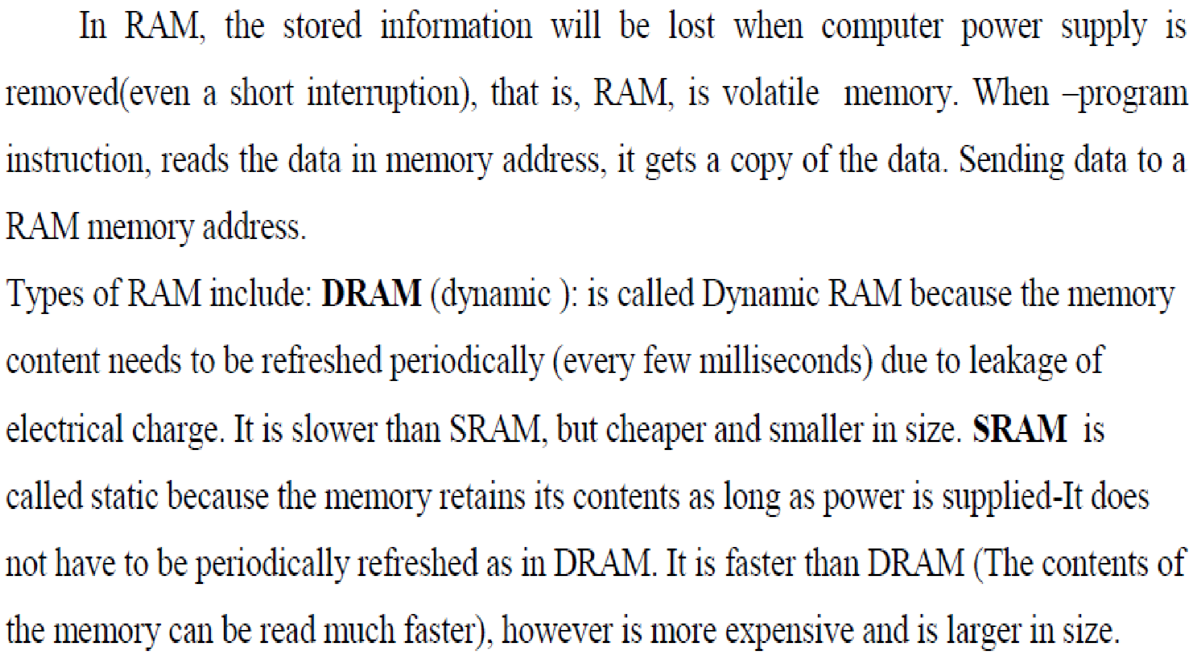


Figure 1: Connection of the memory to the processor

The processor **writes** the data into a memory location by loading the address of this location into AR and loading the data into DR. Random access memory (RAM) is the best known form of computer memory. RAM is considered "random access" because you can access any memory cell directly if you know the row and column that intersect at that cell. RAM data, on the other hand, can be accessed in any order.

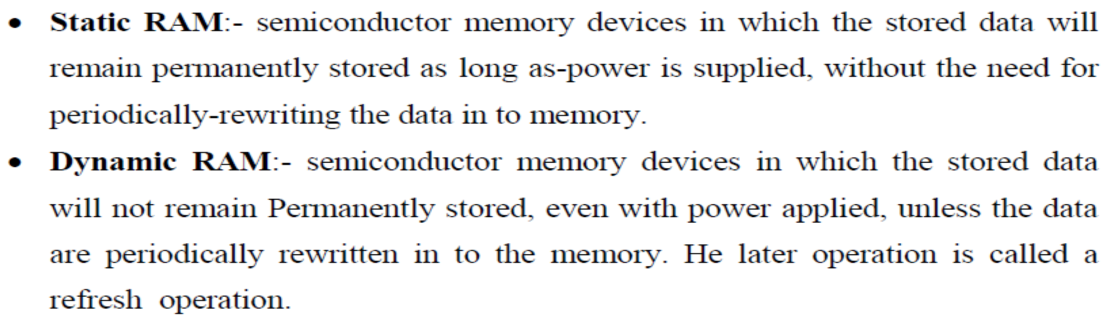
RAM memory consists of memory cells. Each memory cell represents a single bit of data (logic 1 or logic 0). Memory cells are etched onto a silicon wafer in an array of columns (bit lines) and rows (word lines). The intersection of a bit line and word line constitutes the address of the memory cell.

There are many kinds of RAM and new ones are invented all the time. One aim is to make RAM access as fast as possible in order to keep up with the increasing speed of CPUs.

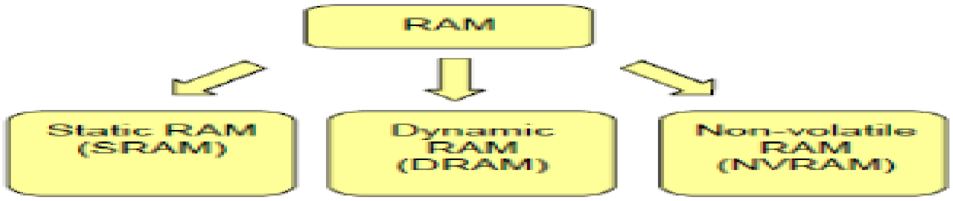


In RAM, the stored information will be lost when computer power supply is removed(even a short interruption), that is, RAM, is volatile memory. When –program instruction, reads the data in memory address, it gets a copy of the data. Sending data to a RAM memory address.

Types of RAM include: **DRAM** (dynamic): is called Dynamic RAM because the memory content needs to be refreshed periodically (every few milliseconds) due to leakage of electrical charge. It is slower than SRAM, but cheaper and smaller in size. **SRAM** is called static because the memory retains its contents as long as power is supplied-It does not have to be periodically refreshed as in DRAM. It is faster than DRAM (The contents of the memory can be read much faster), however is more expensive and is larger in size.

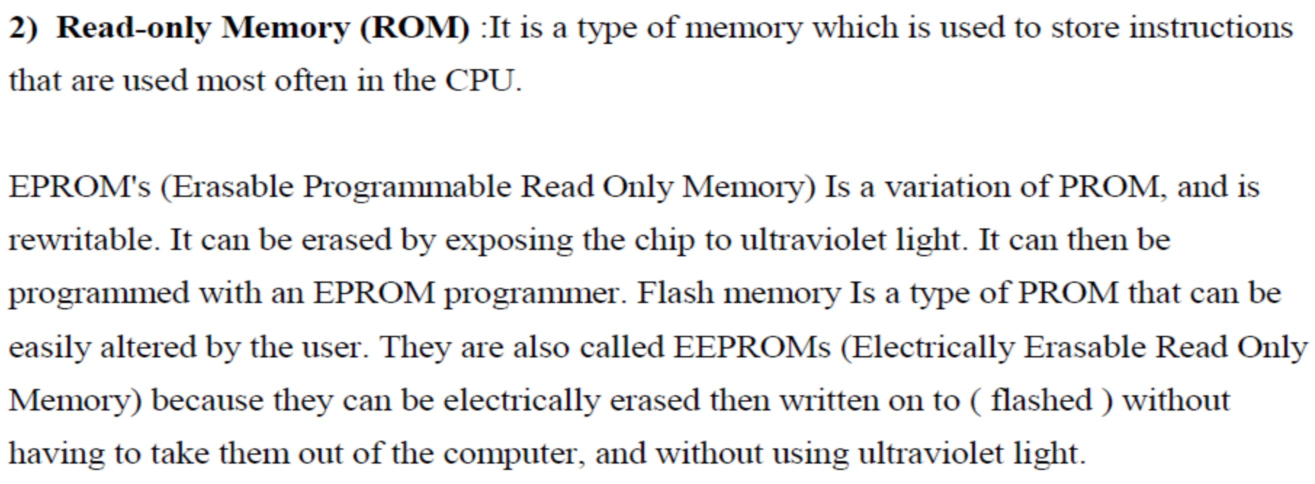


- **Static RAM**:- semiconductor memory devices in which the stored data will remain permanently stored as long as-power is supplied, without the need for periodically-rewriting the data in to memory.
- **Dynamic RAM**:- semiconductor memory devices in which the stored data will not remain Permanently stored, even with power applied, unless the data are periodically rewritten in to the memory. He later operation is called a refresh operation.



Characteristics of Main Memory

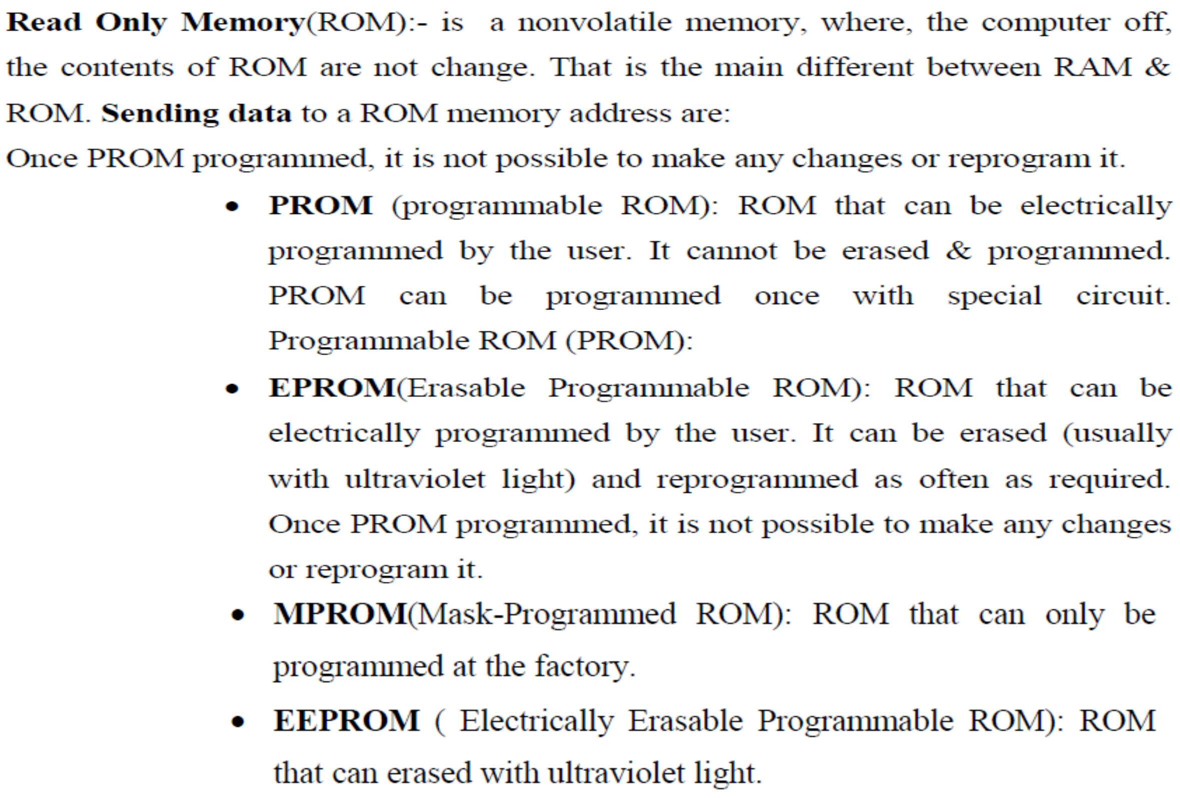
- It is known as main memory.
- Usually volatile memory.
- Data is lost in case power is switched off .
- It is working memory of the computer.
- Faster than secondary memories.
- A computer cannot run without primary memory.



2) Read-only Memory (ROM) :It is a type of memory which is used to store instructions that are used most often in the CPU.

EPROM's (Erasable Programmable Read Only Memory) Is a variation of PROM, and is rewritable. It can be erased by exposing the chip to ultraviolet light. It can then be programmed with an EPROM programmer. Flash memory Is a type of PROM that can be easily altered by the user. They are also called EEPROMs (Electrically Erasable Read Only Memory) because they can be electrically erased then written on to (flashed) without having to take them out of the computer, and without using ultraviolet light.

ROM was used to store the “boot” or start-up program (so called firmware) that a computer executes when powered on, although it has now fallen out-of-favour to more flexible memories that support occasional writes. ROM is still used in systems with fixed functionalities.



Read Only Memory(ROM):- is a nonvolatile memory, where, the computer off, the contents of ROM are not change. That is the main different between RAM & ROM. **Sending data** to a ROM memory address are:
Once PROM programmed, it is not possible to make any changes or reprogram it.

- **PROM** (programmable ROM): ROM that can be electrically programmed by the user. It cannot be erased & programmed. PROM can be programmed once with special circuit. Programmable ROM (PROM):
- **EPROM**(Erasable Programmable ROM): ROM that can be electrically programmed by the user. It can be erased (usually with ultraviolet light) and reprogrammed as often as required. Once PROM programmed, it is not possible to make any changes or reprogram it.
- **MPROM**(Mask-Programmed ROM): ROM that can only be programmed at the factory.
- **EEPROM** (Electrically Erasable Programmable ROM): ROM that can erased with ultraviolet light.

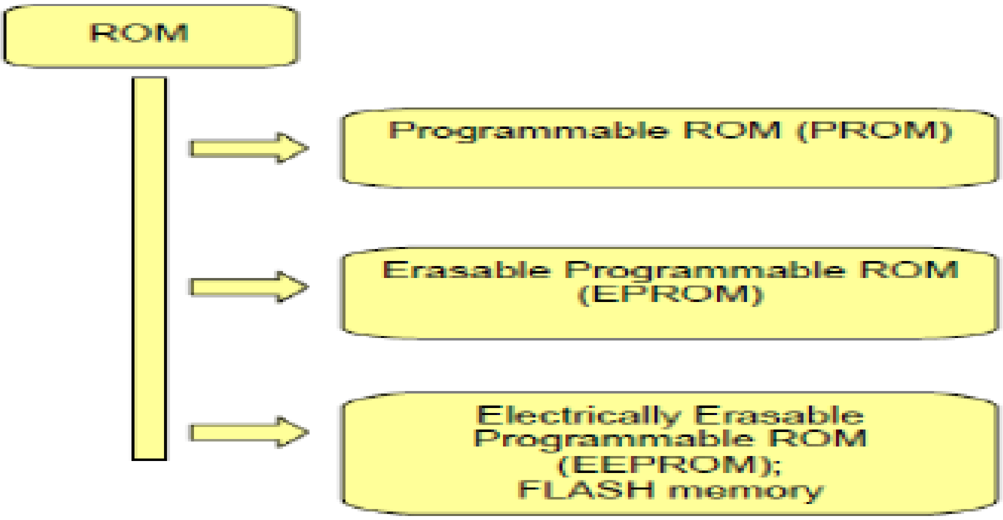
Read Only Memories

Programmable read only memories (PROM) - are programmed during manufacturing process. The contents of each memory cell is locked by a fuse or antifuse (diodes). PROMs are used for permanent data storage.

Erasable read only memories (EPROM) - there is a possibility to erase EPROM with ultraviolet light (about 20 minutes) what sets all bits in memory cells to 1. Programming requires higher voltage. Memory cells are built with floating gate transistors. Data can be stored in EPROMs for about 10 years.

Example of EPROM chip with glass window admitting UV light

Electrically erasable read only memories (EEPROM) - erasing does not require ultraviolet light but higher voltage and can be applied not to the whole circuit but to each memory cell separately.



ROM

Programmable ROM (PROM)

Erasable Programmable ROM (EPROM)

Electrically Erasable Programmable ROM (EEPROM); FLASH memory

Cache Memory

Cache memory is located between CPU and RAM. Cache memory is faster than RAM because the instructions travel shorter distance to the CPU. The advantage of cache memory is increasing the computer processing speed. The disadvantages of cache are that it is expensive and has small size. processor is usually faster than main memory access time with the result that result that processing speed is mostly limited by the speed of main memory. If the average memory access time can be reduced this would reducing , the total exaction time of the program. Normally, this would required that all of the internal memory have an operating speed comparable to that of the CPU in order to achieve maximum system operation.

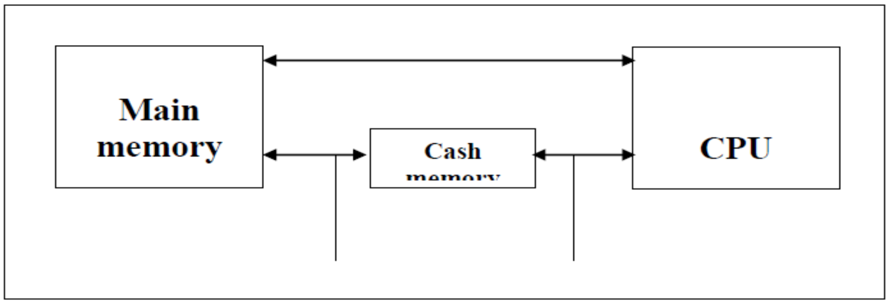
Kinds of Cache: There are three types of cache:-

Level-1 (L1) Cache : part of microprocessor chip It is also called **Internal cache** and is built in processor chip ranging from 8 to 256 Kilobytes its capacity is less than L2 cache.

Level-2 (L2) Cache: not a part of microprocessor chip It is also called **External cache** and is built in SRAM chips its capacity ranges from 64 Kilobytes to 2 megabytes.

Level-3 (L3) Cache: This kind of Cache is separated from processor chip on the motherboard. It is found only on very high –end computers.

In many system it is not economical to use high-speed memory devices for all of the internal memory. Instead, system designers use a small block of high speed cache memory which might hold, say, 512 words. The cache memory access time is less than access time of main memory by a factor 5 to 10.

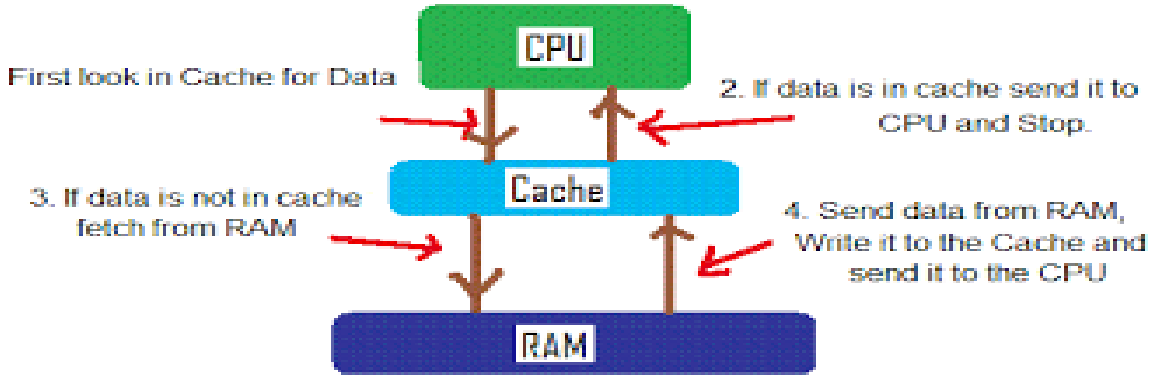


Cache memory is a very high speed semiconductor memory which can speed up CPU. It acts as a buffer between the CPU and main memory. It is used to hold those parts of data and program which are most frequently used by CPU. The parts of data and programs are transferred from disk to cache memory by operating system, from where CPU can access them.

First generation processors, those designed with vacuum tubes in 1950 or those designed with integrated circuits in 1965 or those designed as microprocessors in 1980 were generally about the same speed as main memory. On such processors, this naive model was perfectly reasonable. By 1970, however, transistorized supercomputers were being built where the central processor was significantly faster than the main memory, and by 1980, the difference had increased, although it took several decades for the performance difference to reach today's extreme.

Solution to this problem is to use what is called a cache memory between the central processor and the main memory. Cache memory takes advantage of the fact that, with any of the memory technologies available for the past half century, we have had a choice between building large but slow memories or small but fast memories.

A cache memory sits between the central processor and the main memory. During any particular memory cycle, the cache checks the memory address being issued by the processor. If this address matches the address of one of the few memory locations held in the cache, the cache handles the memory cycle very quickly; this is called a cache hit. If the address does not, then the memory cycle must be satisfied far more slowly by the main memory; this is called a cache miss.


CPU
First look in Cache for Data
2. If data is in cache send it to CPU and Stop.
Cache
3. If data is not in cache fetch from RAM
4. Send data from RAM, Write it to the Cache and send it to the CPU
RAM

The basic characteristic of cache memory is its fast access time, therefore very little or no time must be wasted when searching forwards in the cache memory.

The speed of the main memory is very low in comparison with the speed of modern processors Hence,

- it is important to devise a scheme that reduces the time needed to access the necessary information
- Since the speed of main memory unit is limited by electronic and packaging constraints, the solution must be sought in a different architectural arrangement.
- An efficient solution is to use a fast cache memory which essentially makes the main memory appear to the processor to be faster than it really is Usually, the cache memory can store a reasonable number of blocks at any given time, but this number is small compared to the total number of blocks in the main memory. The correspondence between the main memory blocks and those in the cache is specified by a mapping function.

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 referenced word. The collection of rules for making this decision constitutes the replacement algorithm.

The transformation of data from the MM to cache memory is referred to as mapping process, three types of mapping procedures are of practical interest when considering the organization of cache memory are:

1. Associative Mapping
2. Direct Mapping
3. Set-Associative Mapping.

Advantages

The advantages of cache memory are as follows:

- Cache memory is faster than main memory.
- It consumes less access time as compared to main memory.
- It stores the program that can be executed within a short period of time.
- It stores data for temporary use.

Disadvantages

The disadvantages of cache memory are as follows:

- Cache memory has limited capacity.
- It is very expensive.

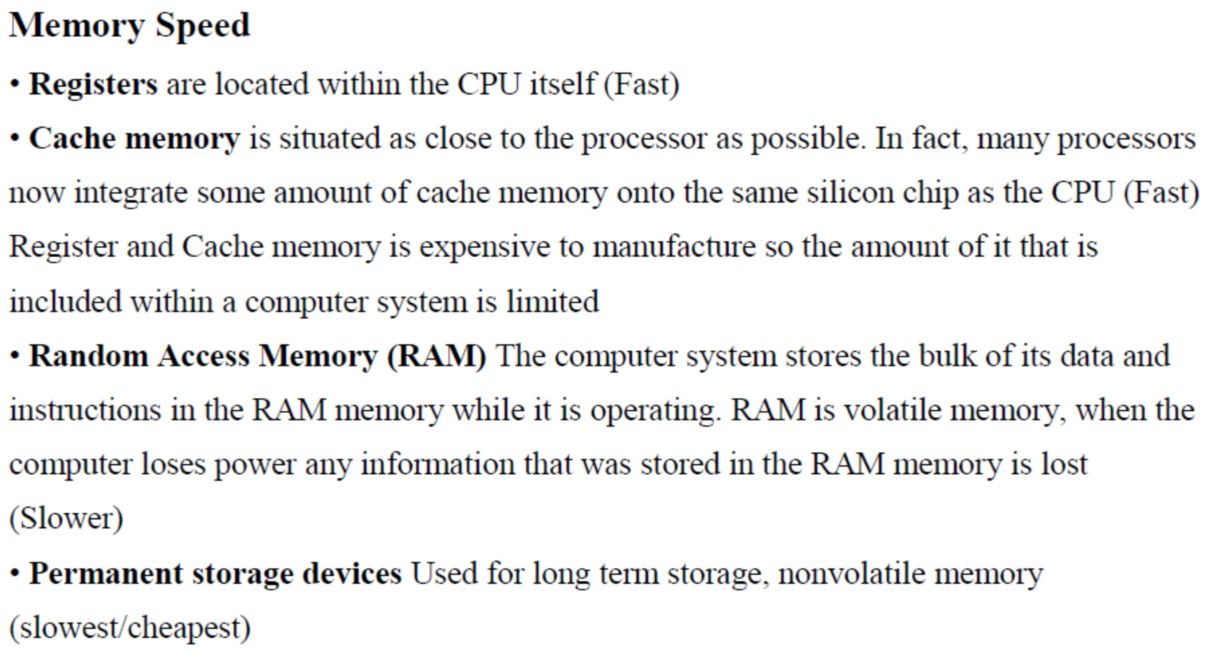
Computer memories exhibit the following features:

1. Memory is organized in equal-size units called words (or cells). If there are M words in memory, each consisting of N bits, then memory is said to be of size M×N.

2. Only binary information (bits, zeros and ones) can be stored in a computer memory.

3. There are only two memory operations, read and write. Reading a word from memory brings in a copy of the data. Writing into a word destroys its original data.

4. Memory is accessed on a word basis, so each word should be uniquely identified. This is done by assigning an address to each word.



Memory Speed

- **Registers** are located within the CPU itself (Fast)
- **Cache memory** is situated as close to the processor as possible. In fact, many processors now integrate some amount of cache memory onto the same silicon chip as the CPU (Fast) Register and Cache memory is expensive to manufacture so the amount of it that is included within a computer system is limited
- **Random Access Memory (RAM)** The computer system stores the bulk of its data and instructions in the RAM memory while it is operating. RAM is volatile memory, when the computer loses power any information that was stored in the RAM memory is lost (Slower)
- **Permanent storage devices** Used for long term storage, nonvolatile memory (slowest/cheapest)

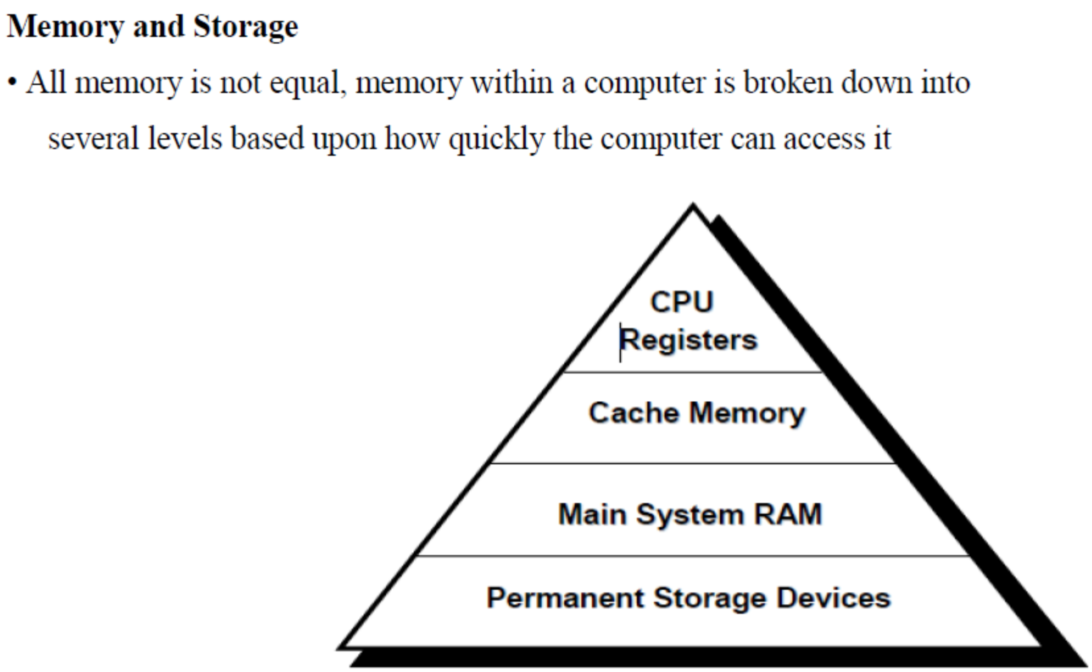
Speed, Size and Cost

Ideally, computer memory should be fast, large and inexpensive. Unfortunately, it is impossible to meet all the three requirements simultaneously. Increased speed and size are achieved at increased cost. Very fast memory systems can be achieved if SRAM chips are used. These chips are expensive and for the cost reason it is impracticable to build a large main memory using SRAM chips. The alternative used to use DRAM chips for large main memories.

The processor fetches the code and data from the main memory to execute the program. The DRAMs which form the main memory are slower devices. So it is necessary to insert wait states in memory read/write cycles. This reduces the speed of execution. The solution for this problem is in the memory system small section of SRAM is added along with the main memory, referred to as cache memory. The program which is to be executed is loaded in the main memory, but the part of the program and data accessed from the cache memory. The cache controller looks after this swapping between main memory and cache memory with the help of

DMA controller, Such cache memory is called secondary cache. Recent processor have the built in cache memory called primary cache. The size of the memory is still small compared to the demands of the large programs with the voluminous data. A solution is provided by using secondary storage, mainly magnetic disks and magnetic tapes to implement large memory spaces, which is available at reasonable prices.

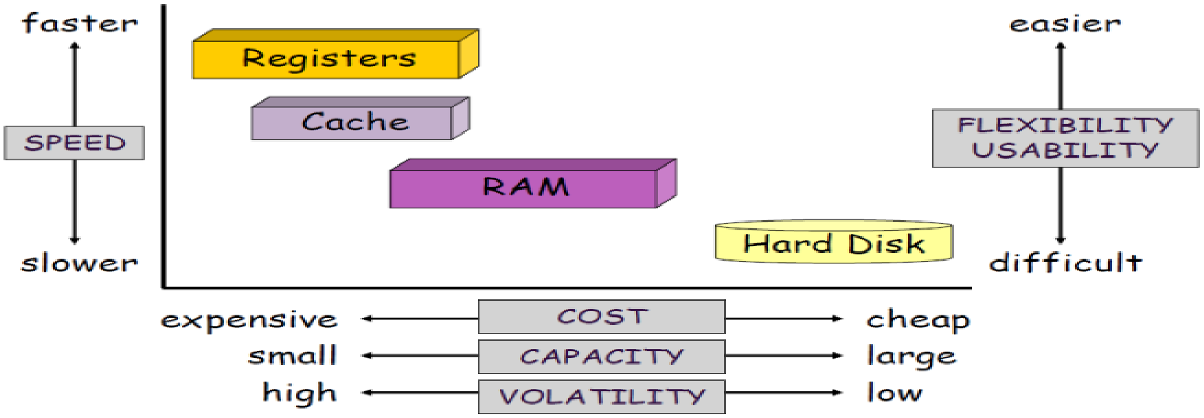
To make efficient computer system it is not possible to rely on a single memory component, but to employ a memory hierarchy which uses all different types of memory units that gives efficient computer system. A typical memory hierarchy is illustrated below in the figures :

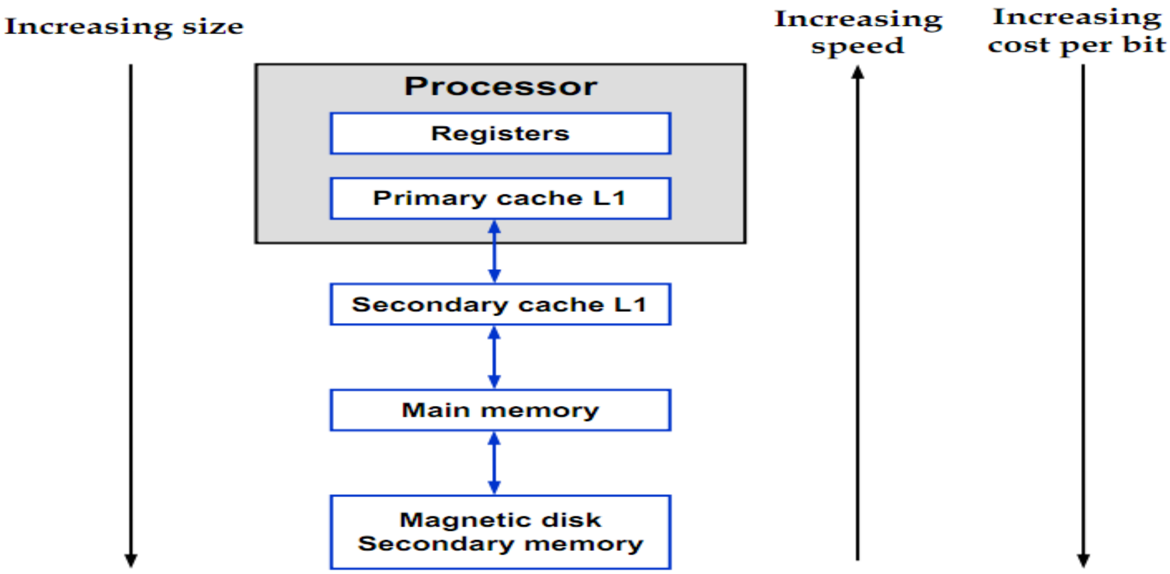


Memory and Storage

- All memory is not equal, memory within a computer is broken down into several levels based upon how quickly the computer can access it

Summary of Characteristics


faster
SPEED
slower
Registers
Cache
RAM
Hard Disk
easier
FLEXIBILITY
USABILITY
difficult
expensive
COST
cheap
small
CAPACITY
large
high
VOLATILITY
low


Increasing size
Processor
Registers
Primary cache L1
Secondary cache L1
Main memory
Magnetic disk
Secondary memory
Increasing speed
Increasing cost per bit