

Chapter – 6 Memory System

6.1 Microcomputer Memory

- Memory is an essential component of the microcomputer system. It stores binary instructions and datum for the microcomputer.
- The memory is the place where the computer holds current programs and data that are in use.
- None technology is optimal in satisfying the memory requirements for a computer system.
- Computer memory exhibits perhaps the widest range of type, technology, organization, performance and cost of any feature of a computer system.
- The memory unit that communicates directly with the CPU is called main memory.
- Devices that provide backup storage are called auxiliary memory or secondary memory.

6.2 Characteristics of memory systems

The memory system can be characterised with their Location, Capacity, Unit of transfer, Access method, Performance, Physical type, Physical characteristics, Organisation.

Location

- Processor memory: The memory like registers is included within the processor and termed as processor memory.
- Internal memory: It is often termed as main memory and resides within the CPU.
- External memory: It consists of peripheral storage devices such as disk and magnetic tape that are accessible to processor via i/o controllers.

Capacity

- Word size: Capacity is expressed in terms of words or bytes.
 - The natural unit of organisation
- Number of words: Common word lengths are 8, 16, 32 bits etc.
 - or Bytes

Unit of Transfer

- Internal: For internal memory, the unit of transfer is equal to the number of data lines into and out of the memory module.
- External: For external memory, they are transferred in block which is larger than a word.
- Addressable unit
 - Smallest location which can be uniquely addressed
 - Word internally
 - Cluster on Magnetic disks



Access Method

- Sequential access: In this access, it must start with beginning and read through a specific linear sequence. This means access time of data unit depends on position of records (unit of data) and previous location.
 - e.g. tape
- Direct Access: Individual blocks of records have unique address based on location. Access is accomplished by jumping (direct access) to general vicinity plus a sequential search to reach the final location.
 - e.g. disk
- Random access: The time to access a given location is independent of the sequence of prior accesses and is constant. Thus any location can be selected out randomly and directly addressed and accessed.
 - e.g. RAM
- Associative access: This is random access type of memory that enables one to make a comparison of desired bit locations within a word for a specified match, and to do this for all words simultaneously.
 - e.g. cache

Performance

- Access time: For random access memory, access time is the time it takes to perform a read or write operation i.e. time taken to address a memory plus to read / write from addressed memory location. Whereas for non-random access, it is the time needed to position read / write mechanism at desired location.
 - Time between presenting the address and getting the valid data
- Memory Cycle time: It is the total time that is required to store next memory access operation from the previous memory access operation.

Memory cycle time = access time plus transient time (any additional time required before a second access can commence).

 - Time may be required for the memory to “recover” before next access
 - Cycle time is access + recovery
- Transfer Rate: This is the rate at which data can be transferred in and out of a memory unit.
 - Rate at which data can be moved
 - For random access, $R = 1 / \text{cycle time}$
 - For non-random access, $T_n = T_a + N / R$; where T_n – average time to read or write N bits, T_a – average access time, N – number of bits, R – Transfer rate in bits per second (bps).

Physical Types

- Semiconductor
 - RAM
- Magnetic
 - Disk & Tape
- Optical
 - CD & DVD
- Others



- Bubble
- Hologram

Physical Characteristics

- Decay: Information decays mean data loss.
- Volatility: Information decays when electrical power is switched off.
- Erasable: Erasable means permission to erase.
- Power consumption: how much power consumes?



Organization

- Physical arrangement of bits into words
- Not always obvious
 - e.g. interleaved

6.3 The Memory Hierarchy

- ↵ Capacity, cost and speed of different types of memory play a vital role while designing a memory system for computers.
- ↵ If the memory has larger capacity, more application will get space to run smoothly.
- ↵ It's better to have fastest memory as far as possible to achieve a greater performance. Moreover for the practical system, the cost should be reasonable.
- ↵ There is a tradeoff between these three characteristics cost, capacity and access time. One cannot achieve all these quantities in same memory module because
 - ↵ If capacity increases, access time increases (slower) and due to which cost per bit decreases.
 - ↵ If access time decreases (faster), capacity decreases and due to which cost per bit increases.
- ↵ The designer tries to increase capacity because cost per bit decreases and the more application program can be accommodated. But at the same time, access time increases and hence decreases the performance.

So the best idea will be to use memory hierarchy.

- ↵ Memory Hierarchy is to obtain the highest possible access speed while minimizing the total cost of the memory system.
- ↵ Not all accumulated information is needed by the CPU at the same time.
- ↵ Therefore, it is more economical to use low-cost storage devices to serve as a backup for storing the information that is not currently used by CPU
- ↵ The memory unit that directly communicate with CPU is called the *main memory*
- ↵ Devices that provide backup storage are called *auxiliary memory*
- ↵ The memory hierarchy system consists of all storage devices employed in a computer system from the slow by high-capacity auxiliary memory to a relatively faster main memory, to an even smaller and faster cache memory
- ↵ The main memory occupies a central position by being able to communicate directly with the CPU and with auxiliary memory devices through an I/O processor
- ↵ A special very-high-speed memory called **cache** is used to increase the speed of processing by making current programs and data available to the CPU at a rapid rate

- ⌘ CPU logic is usually faster than main memory access time, with the result that processing speed is limited primarily by the speed of main memory
- ⌘ The cache is used for storing segments of programs currently being executed in the CPU and temporary data frequently needed in the present calculations
- ⌘ The memory hierarchy system consists of all storage devices employed in a computer system from slow but high capacity auxiliary memory to a relatively faster cache memory accessible to high speed processing logic. The figure below illustrates memory hierarchy.

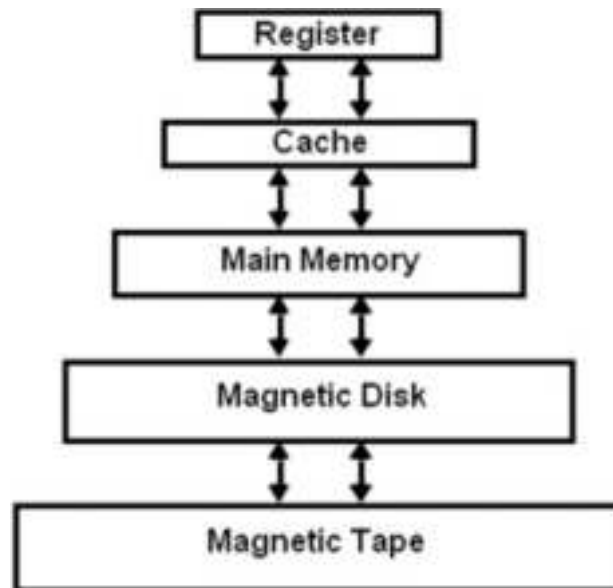
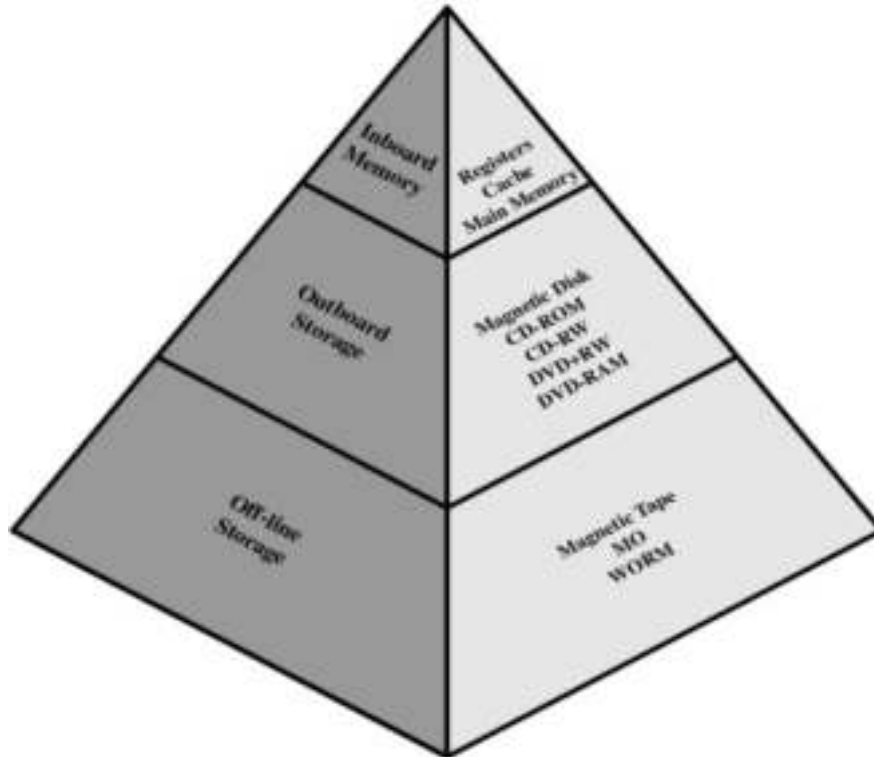


Fig: Memory Hierarchy

- ↯ As we go down in the hierarchy
 - ↯ Cost per bit decreases
 - ↯ Capacity of memory increases
 - ↯ Access time increases
 - ↯ Frequency of access of memory by processor also

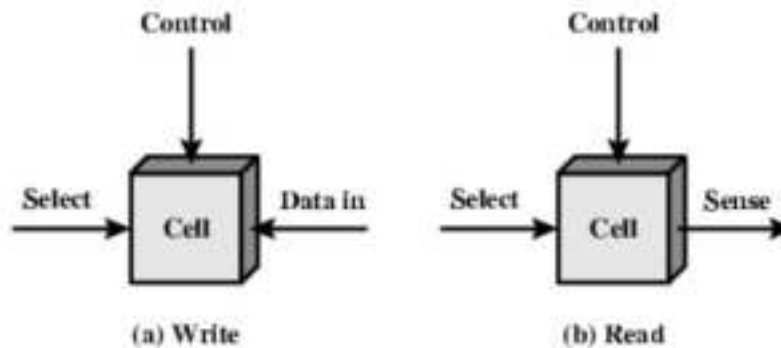
decreases. **Hierarchy List**

- ↯ Registers
- ↯ L1 Cache
- ↯ L2 Cache
- ↯ Main memory
- ↯ Disk cache
- ↯ Disk
- ↯ Optical
- ↯ Tape



6.4 Internal and External memory Internal or Main Memory

- ↯ The main memory is the central unit of the computer system. It is relatively large and fast memory to store programs and data during the computer operation. These memories employ semiconductor integrated circuits. The basic element of the semiconductor memory is the memory cell.
- ↯ The memory cell has three functional terminals which carries the electrical signal.
 - The select terminal: It selects the cell.
 - The data in terminal: It is used to input data as 0 or 1 and data out or sense terminal is used for the output of the cell's state.
 - The control terminal: It controls the function i.e. it indicates read and write.



- ↯ Most of the main memory in a general purpose computer is made up of RAM integrated circuits chips, but a portion of the memory may be constructed with ROM chips

RAM– Random Access memory

- ↯ Memory cells can be accessed for information transfer from any desired random location.
- ↯ The process of locating a word in memory is the same and requires of locating a word in memory is the same and requires an equal amount of time no matter where the cells are located physically in memory thus named 'Random Access'.
- ↯ Integrated RAM are available in two possible operating modes, *Static and Dynamic*

Static RAM (SRAM)

- ↯ The static RAM consists of flip flop that stores binary information and this stored information remains valid as long as power is applied to the unit.

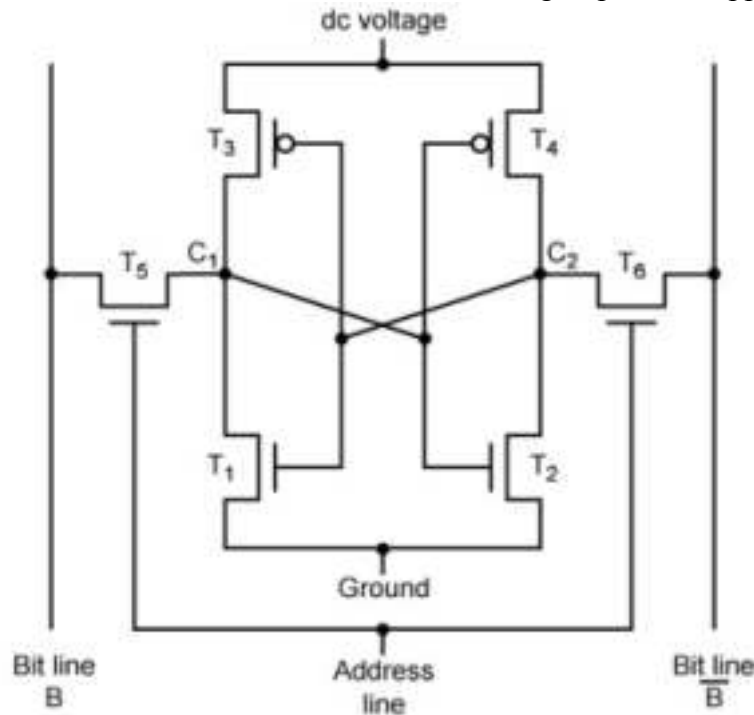


Fig: SRAM structure

- ↯ Four transistors T1, T2, T3 and t4 are cross connected in an arrangement that produces a stable logical state.
- ↯ In logic state 1, point C1 is high and point C2 is low. In this state, T1 & T4 are off and T2 & T3 are on.
- ↯ In logic state 0, point C1 is low and C2 is high. In this state, T1 & T4 are on and T2 & T3 are off.
- ↯ The address line controls the two transistors T5 & T6. When a signal is applied to this line, the two transistors are switched on allowing for read and write operation.
- ↯ For a write operation, the desired bit value is applied to line B while it's complement is applied to line B complement. This forces the four transistors T1, T2, T3 & T4 into a proper state.
- ↯ For the read operation, the bit value is read from line B.

Dynamic RAM (DRAM)

- ↯ The dynamic RAM stores the binary information in the form of electrical charges and capacitor is used for this purpose.
- ↯ Since charge stored in capacitor discharges with time, capacitor must be periodically recharged and which is also called refreshing memory.

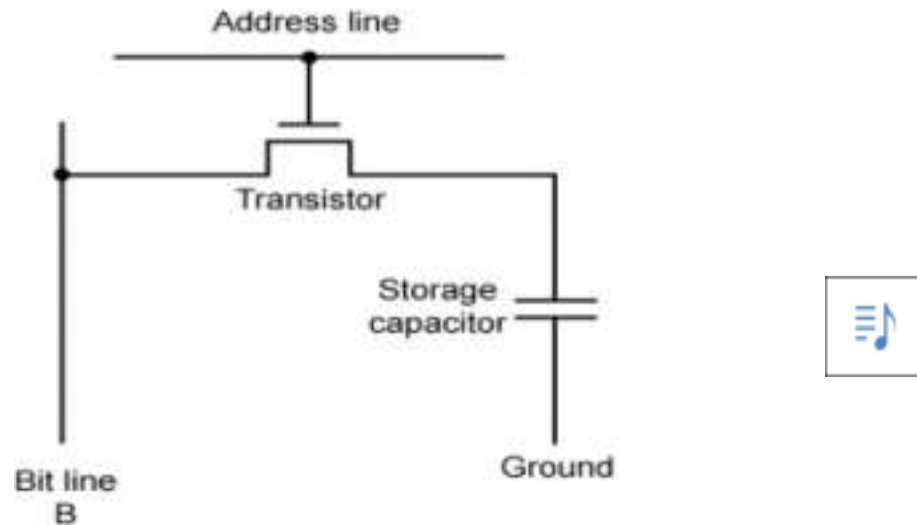


Fig: DRAM structure

- ↯ The address line is activated when the bit value from this cell is to be read or written.
- ↯ The transistor acts as switch that is closed i.e. allowed current to flow, if voltage is applied to the address line; and opened i.e. no current to flow, if no voltage is present in the address line.

For DRAM writing

- ↯ The address line is activated which causes the transistor to conduct.
- ↯ The sense amplifier senses the content of the data bus line for this cell.
- ↯ If the bus line is low, then amplifier will ground the bit line of cell and any charge in capacitor is addressed out.
- ↯ If data bus is high, then a +5V is applied on bit line and voltage will flow through transistor and charge the capacitor.

For DRAM reading

- ↯ Address line is activated which causes the transistor to conduct.
- ↯ If there is charge stored in capacitor, then current will flow through transistor and raise the voltage in bit line. The amplifier will store the voltage and place a 1 on data out line.
- ↯ If there is no charge stored in capacitor, then no current will flow through transistor and voltage bit line will not be raised. The amplifier senses that there is no charge and places a 0 on data out line.

SRAM versus DRAM

- ↗ Both volatile
 - Power needed to preserve data
- ↗ Static RAM
 - Uses flip flop to store information
 - Needs more space
 - Faster, digital device
 - Expensive, big in size
 - Don't require refreshing circuit
 - Used in cache memory
- ↗ Dynamic RAM
 - Uses capacitor to store information
 - More dense i.e. more cells can be accommodated per unit area
 - Slower, analog device
 - Less expensive, small in size
 - Needs refreshing circuit
 - Used in main memory, larger memory units

**ROM– Read Only memory**

- ↗ Read only memory (ROM) contains a permanent pattern of data that cannot be changed.
- ↗ A ROM is non-volatile that is no power source is required to maintain the bit values in memory.
- ↗ While it is possible to read a ROM, it is not possible to write new data into it.
- ↗ The data or program is permanently presented in main memory and never be loaded from a secondary storage device with the advantage of ROM.
- ↗ A ROM is created like any other integrated circuit chip, with the data actually wired into the chip as part of the fabrication process.
- ↗ It presents two problems
 - The data insertion step includes a relatively large fixed cost, whether one or thousands of copies of a particular ROM are fabricated.
 - There is no room for error. If one bit is wrong, the whole batch of ROM must be thrown out.

Types of ROM

- ↗ Programmable ROM (PROM) It is non-volatile and may be written into only once. The writing process is performed electrically and may be performed by a supplier or customer at a time later than the original chip fabrication.
- ↗ Erasable Programmable ROM (EPROM)
 - It is read and written electrically. However, before a write operation, all the storage cells must be erased to the same initial state by exposure of the packaged chip to ultraviolet radiation (UV ray). Erasure is performed by shining an intense ultraviolet light through a window that is designed into the memory chip. EPROM is optically managed and more expensive than PROM, but it has the advantage of the multiple update capability.

- ↯ Electrically Erasable programmable ROM (EEPROM)
 - This is a read mostly memory that can be written into at any time without erasing prior contents, only the byte or byte addresses are updated. The write operation takes considerably longer than the read operation, on the order of several hundred microseconds per byte. The EEPROM combines the advantage of non-volatility with the flexibility of being updatable in place, using ordinary bus control, addresses and data lines. EEPROM is more expensive than EPROM and also is less dense, supporting fewer bits per chip.
- ↯ Flash Memory
 - Flash memory is also the semiconductor memory and because of the speed with which it can be reprogrammed, it is termed as flash. It is interpreted between EPROM and EEPROM in both cost and functionality. Like EEPROM, flash memory uses an electrical erasing technology. An entire flash memory can be erased in one or a few seconds, which is much faster than EPROM. In addition, it is possible to erase just blocks of memory rather than an entire chip. However, flash memory doesn't provide byte level erasure, a section of memory cells are erased in an action or 'flash'.

External Memory

- ↯ The devices that provide backup storage are called external memory or auxiliary memory. It includes serial access type such as magnetic tapes and random access type such as magnetic disks.

Magnetic Tape

- ↯ A magnetic tape is the strip of plastic coated with a magnetic recording medium. Data can be recorded and read as a sequence of character through read / write head. It can be stopped, started to move forward or in reverse or can be rewound. Data on tapes are structured as number of parallel tracks running length wise. Earlier tape system typically used nine tracks. This made it possible to store data one byte at a time with additional parity bit as 9th track. The recording of data in this form is referred to as parallel recording.

Magnetic Disk

- ↯ A magnetic disk is a circular plate constructed with metal or plastic coated with magnetic material often both side of disk are used and several disk stacked on one spindle which Read/write head available on each surface. All disks rotate together at high speed. Bits are stored in magnetize surface in spots along concentric circles called tracks. The tracks are commonly divided into sections called sectors. After the read/write head are positioned in specified track the system has to wait until the rotating disk reaches the specified sector under read/write head. Information transfer is very fast once the beginning of sector has been reached. Disk that are permanently attached to the unit assembly and cannot be used by occasional user are called hard disk drive with removal disk is called floppy disk.



Optical Disk

- ↯ The huge commercial success of CD enabled the development of low cost optical disk storage technology that has revolutionized computer data storage. The disk is form from resin such as polycarbonate. Digitally recorded information is imprinted as series of microscopic pits on the surface of poly carbonate. This is done with the finely focused high intensity leaser. The pitted surface is then coated with reflecting surface usually aluminum or gold. The shiny surface is protected against dust and scratches by the top coat of acrylic.
- ↯ Information is retrieved from CD by low power laser. The intensity of reflected light of laser changes as it encounters a pit. Specifically if the laser beam falls on pit which has somewhat rough surface the light scatters and low intensity is reflected back to the surface. The areas between pits are called lands. A land is a smooth surface which reflects back at higher intensity. The change between pits and land is detected by photo sensor and converted into digital signal. The sensor tests the surface at regular interval.

DVD-Technology

- ↯ Multi-layer
- ↯ Very high capacity (4.7G per layer)
- ↯ Full length movie on single disk
- ↯ Using MPEG compression
- ↯ Finally standardized (honest!)
- ↯ Movies carry regional coding
- ↯ Players only play correct region

films

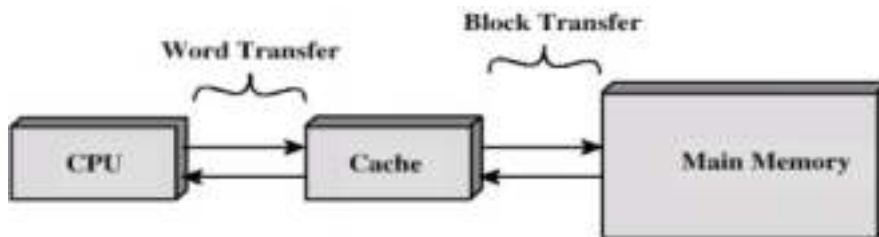
DVD-Writable

- ↯ Loads of trouble with standards
- ↯ First generation DVD drives may not read first generation DVD-W disks
- ↯ First generation DVD drives may not read CD-RW disks

6.5 Cache memory principles**Principles**

- *Intended to give memory speed approaching that of fastest memories available but with large size, at close to price of slower memories*
- *Cache is checked first for all memory references.*
- *If not found, the entire block in which that reference resides in main memory is stored in a cache slot, called a line*
- *Each line includes a tag (usually a portion of the main memory address) which identifies which particular block is being stored*
- *Locality of reference implies that future references will likely come from this block of memory, so that cache line will probably be utilized repeatedly.*
- *The proportion of memory references, which are found already stored in cache, is called the hit ratio.*

- ↯ Cache memory is intended to give memory speed approaching that of the fastest memories available, and at the same time provide a large memory size at the price of less expensive types of semiconductor memories. There is a relatively large and slow main memory together with a smaller, faster cache memory contains a copy of portions of main memory.
- ↯ When the processor attempts to read a word of memory, a check is made to determine if the word is in the cache. If so, the word is delivered to the processor. If not, a block of main memory, consisting of fixed number of words is read into the cache and then the word is delivered to the processor.
- ↯ The locality of reference property states that over a short interval of time, address generated by a typical program refers to a few localized area of memory repeatedly. So if programs and data which are accessed frequently are placed in a fast memory, the average access time can be reduced. This type of small, fast memory is called cache memory which is placed in between the CPU and the main memory.



- ↯ When the CPU needs to access memory, cache is examined. If the word is found in cache, it is read from the cache and if the word is not found in cache, main memory is accessed to read word. A block of word containing the one just accessed is then transferred from main memory to cache memory.

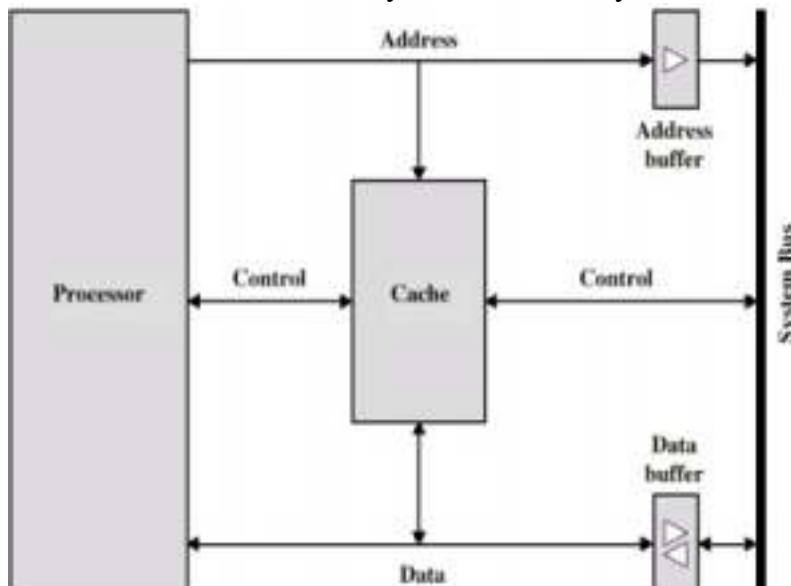


Fig: Typical Cache organization

- ↯ Cache connects to the processor via data control and address line. The data and address lines also attached to data and address buffer which attached to a system bus from which main memory is reached.