Chapter 6: Memory Units

Introduction

Memory units are the internal storage areas in computer. They are the locations which hold actual data and information either for short time temporarily or for long time permanently. The temporary memory is referred to as main memory and the permanent memory is called auxiliary memory. RAM and ROM are the main memory while secondary storage devices like hard disk, floppy disk, CDs, USB are the auxiliary memory.

Main Memory

- is the workspace for the computer's processor
- CPU needs to have millions of bytes of randomly accessed space where it can quickly read or write programs and data while they are being used.
- its storage is considered temporary because the data and programs will remain there only as long as the computer has electrical power or is not reset.

RAM (Random Access Memory)

- It is a volatile memory and holds data on a temporary basis
- When power is turned off its contents are erased.
- RAM holds programs and data which are currently being used.
- We can randomly (and quickly), directly access any location in memory.
- RAM is used to store:
    - Instruction awaiting to be obeyed
    - Instruction currently being obeyed
    - Data currently being processed
    - Data awaiting output.
- RAM can be assumed as the set of boxes, the boxes are numbered from zero upwards so that each box can be identified and located.
- Each bit of semi-conductor memory is represented by a single cell which may be regarded as Microscopic electronic circuits with two distinguishable stages used to represent 0 and 1.

Types of RAM

DRAM and SRAM

DRAM (Dynamic RAM)

- It is dense, meaning that we can pack a lot of bits into a very small chip and it is inexpensive which makes it affordable for large amount of memory.
- DRAM gets its name from the fact that it must be refreshed frequently. (The term refreshing means recharging the RAM chips with electricity.)
- DRAM chips must be recharged many times each second or they will lose their contents.
- The memory cells in a DRAM chip are tiny capacitors that retain a charge to indicate a bit.
- DRAM must be constantly refreshed or the electrical charges in the individual memory capacitors will drain and the data will be lost.
- The charge slowly leaks from the cells and has to be topped up constantly called "Refreshing".

SRAM (Static RAM)

- It does not need the periodic refresh rates like DRAM.
- Due to design of SRAM, not only are refresh rates unnecessary but SRAM is much faster than DRAM.
- Transistors are used instead of capacitors in SRAM.
- Transistors do not loose their charge.
- SRAM is much faster but lower in density and more expensive.
- The lower in density means that SRAM chips are both physically larger and store many less bit overall.
- Much more expensive that DRAM.

ROM (Read Only Memory)

- is a non-volatile type of memory that can permanently or semi permanently hold data
- is called ROM because it is either impossible or needs a special device to write to.
- contents or data in ROM will remain even if power is turned off.
- is an ideal place to put the computer's startup instructions that is , the software that are required to boots the system (are called firm ware)
- otherwise the processor would have no program in memory to execute when it is powered on.

Types of ROM

- PROM
- EPROM
- EEPROM

PROM (Programmable Read Only Memory)

- It is blank when new and must be programmed with what ever data is necessary
- They are technically preloaded with binary 1s.
- 1MB ROM chip would come with about 1 million bit location each containing 1.
- A blank PROM can be programmed, using a special machine called ROM programmer or ROM burner.
- Each binary 1 bit can be thought of as a fuse that is in fact (un burned).
- Most chip run on 5 volts but when we program a PROM we place a higher voltage normally 12 volt at various address with the chip
- The higher voltage actually blows or burns the fuses at the location we desire thus turning any given 1 into 0.
- PROM chips are often called OTP i.e. One Time Programmable chips because we cannot convert a 0 back into a 1.
- That is they can be programmed once and never erased.

EPROM (Erasable PROM)

- PROM that is erasable.
- EPROM is erased by the exposure to intense UV (Ultra Voilet) light.

- UV light erases the chip by causing a chemical reaction that essentially melts the fuse back together, thus any binary 0s in the chip become 1s and the chip is restored to new condition with binary 1s in all location.

EEPROM (Electrically Erasable PROM)

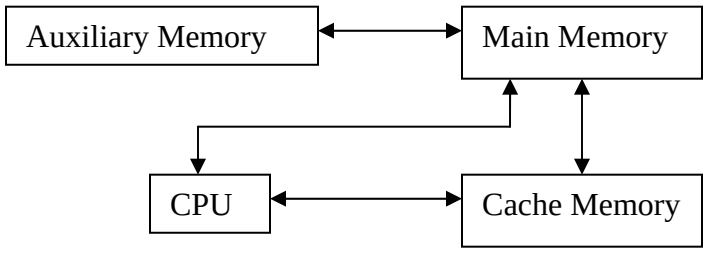
- Can be erased with electrical signals instead of UV light.
- Also known as Flash ROM.
- Are characterized by their capability to be erased and reprogrammed directly in the circuit board in which they are installed with no special equipment required.
- By using EEPROM it is possible to erase and reprogram the ROM in a computer without removing the chip from the System.

FLASH MEMORY

- Special types of EEPROM that can be erased and reprogrammed in blocks instead of 1 byte at ta time.
- Many modern computers have their BIOS stored on a flash memory chip so that it can be easily updated of necessary.
- Also popular in modems because it enables the modem manufacturers to support new protocols as they become standardized.

CACHE MEMORY

- Moving data between RAM and the CPU's register is one of the most time consuming operations a CPU must perform, simply RAM is much slower than the CPU.
- The cache memory is placed between CPU and main memory.
- It is a semiconductor memory similar to RAM except that it is extremely fast compared to normal memory.
- It is a special and high speed memory used to increase the speed of processing by making current program and data available to the CPU at a rapid rate.
- Cache memory is used in computer systems to compensate for the speed difference between main memory access time and processor logic.
- It stores instruction code and data, which are to be currently executed by the CPU.
- It is used to reduce the average access time for instructions and data which are normally stored in the main memory.



- The modern 32- bit and 64 – bit microprocessors operate at a very high speed.

- The memory matching with high – speed microprocessor must be very fast. But very fast memory is very expensive.
- If a fast microprocessor operates with conventional main memory it has to operate with several wait state, this will reduce the speed of the computer.
- A high speed cache memory is used to supply currently needed instructions and date to CPU.
- The main memory stores programs and data which is to be processed by the CPU.
- The currently needed instructions and data of the program are loaded into the cache from the main memory.
- The technique of accessing a cache memory differs from that of the main memory.
- To access main memory the CPU sends an address to it, in response of this the memory sends data contained at the specified memory address.
- On the other hand cache memory used parallel searching of data.
- It first compares the incoming address to the address present in the cache, it the address matches it is said that a "hit" has occurred then the corresponding data is read by the CPU.
- If the address does not match, it is said the a "miss" has occurred when a miss occurs, data is read from the main memory but it also loads a copy of the data to the cache memory so that when the CPU needs the same data or address, it finds it in the cache memory i.e. hit occurs and saves the time needed to load the data from the memory.
- Usually hit ratio is above 90%.
- The hit ratio is defined as the ratio of the number of hits to the total number of read requests sent to the cache by the CPU.
- Cache memory is placed at two or three levels they are called first level cache (L1), second level cache (L2) and third level cache (L3).
- Some microprocessor contains L1, L2, and L3 within the microprocessor.
- Cache within the microprocessor is called internal cache
- Cache outside the processor is called external cache.
- L1 → up to 256 KB built-in
- L2→ 2 MB
- Use of cache memory let the processor work in its actual speed