Name of Faculty: Dr Jaikaran Singh

Designation : Professor

Department : ECE

Subject: LNCTS CS-4th Sem. CSO

Unit : 4

Topic : Notes for Unit -IV

Contents : Main memory-RAM, ROM, Secondary Memory –Magnetic Tape, Disk, Optical Storage, Cache Memory: Cache Structure and Design, Mapping Scheme, Replacement Algorithm, Improving Cache Performance, Virtual Memory, memory management hardware

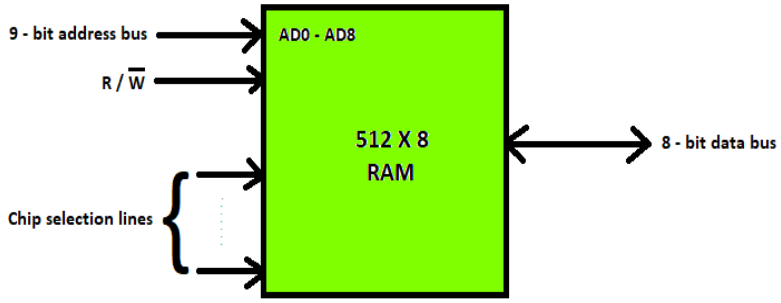
Random Access Memory (RAM)

RAM(Random Access Memory) is a part of computer's Main Memory which is directly accessible by CPU. RAM is used to Read and Write data into it which is accessed by CPU randomly. RAM is volatile in nature, it means if the power goes off, the stored information is lost. RAM is used to store the data that is currently processed by the CPU. Most of the programs and data that are modifiable are stored in RAM.

Integrated RAM chips are available in two form:

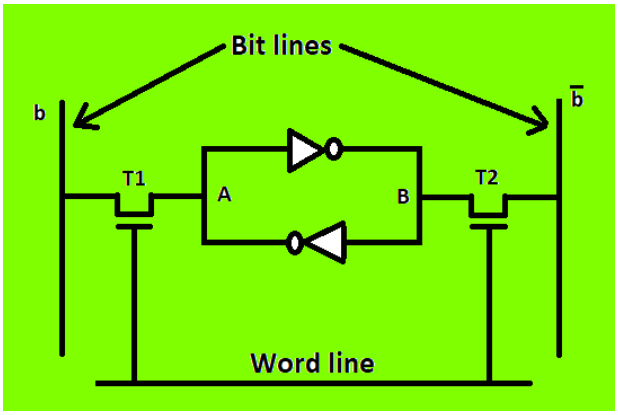
1. SRAM(Static RAM)
2. DRAM(Dynamic RAM)

The block diagram of RAM chip is given below.



The SRAM memories consist of circuits capable of retaining the stored information as long as the power is applied. That means this type of memory requires constant power. SRAM memories are used to build Cache Memory.

SRAM Memory Cell: Static memories (SRAM) are memories that consist of circuits capable of retaining their state as long as power is on. Thus this type of memories is called volatile memories. The below figure shows a cell diagram of SRAM. A latch is formed by two inverters connected as shown in the figure. Two transistors T1 and T2 are used for connecting the latch with two bit lines. The purpose of these transistors is to act as switches that can be opened or closed under the control of the word line, which is controlled by the address decoder. When the word line is at 0-level, the transistors are turned off and the latch remains its information. For example, the cell is at state 1 if the logic value at point A is 1 and at point B is 0. This state is retained as long as the word line is not activated.



For **Read operation**, the word line is activated by the address input to the address decoder. The activated word line closes both the transistors (switches) T1 and T2. Then the bit values at points A and B can transmit to their respective bit lines. The sense/write circuit at the end of the bit lines sends the output to the processor.

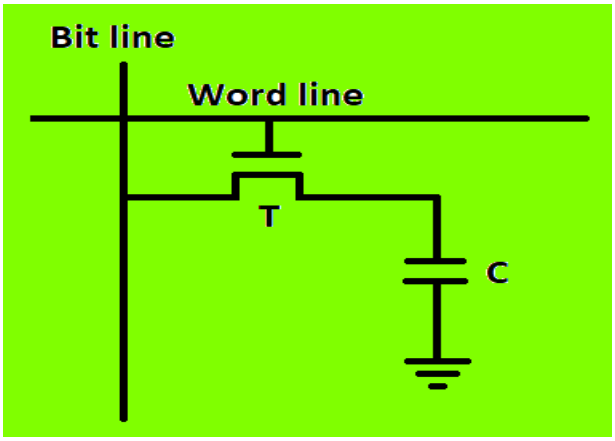
For **Write operation**, the address provided to the decoder activates the word line to close both the switches. Then the bit value that to be written into the cell is provided through the sense/write circuit and the signals in bit lines are then stored in the cell.

DRAM

DRAM stores the binary information in the form of electric charges that applied to capacitors. The stored information on the capacitors tend to lose over a period of time and thus the capacitors must be periodically recharged to retain their usage. The main memory is generally made up of DRAM chips.

DRAM Memory Cell: Though SRAM is very fast, but it is expensive because of its every cell requires several transistors. Relatively less expensive RAM is DRAM, due to the use of one transistor and one capacitor in each cell, as shown in the below figure., where C is the capacitor and T is the transistor. Information is stored in a DRAM cell in the form of a charge on a capacitor and this charge needs to be periodically recharged.

For storing information in this cell, transistor T is turned on and an appropriate voltage is applied to the bit line. This causes a known amount of charge to be stored in the capacitor. After the transistor is turned off, due to the property of the capacitor, it starts to discharge. Hence, the information stored in the cell can be read correctly only if it is read before the charge on the capacitors drops below some threshold value.



Types of DRAM

There are mainly 5 types of DRAM:

1. **Asynchronous DRAM (ADRAM):** The DRAM described above is the asynchronous type DRAM. The timing of the memory device is controlled asynchronously. A specialized memory controller circuit generates the necessary control signals to control the timing. The CPU must take into account the delay in the response of the memory.

2. **Synchronous DRAM (SDRAM):** These RAM chips' access speed is directly synchronized with the CPU's clock. For this, the memory chips remain ready for operation when the CPU expects them to be ready. These memories operate at the CPU-memory bus without imposing wait states. SDRAM is commercially available as modules incorporating multiple SDRAM chips and forming the required capacity for the modules.

3. **Double-Data-Rate SDRAM (DDR SDRAM):** This faster version of SDRAM performs its operations on both edges of the clock signal; whereas a standard SDRAM performs its operations on the rising edge of the clock signal. Since they transfer data on both edges of the clock, the data transfer rate is doubled. To access the data at high rate, the memory cells are organized into two groups. Each group is accessed separately.

4. **Rambus DRAM (RDRAM):** The RDRAM provides a very high data transfer rate over a narrow CPU-memory bus. It uses various speedup mechanisms, like synchronous memory interface, caching inside the DRAM chips and very fast signal timing. The Rambus data bus width is 8 or 9 bits.

5. **Cache DRAM (CDRAM):** This memory is a special type DRAM memory with an on-chip cache memory (SRAM) that acts as a high-speed buffer for the main DRAM.

Difference between SRAM and DRAM

Below table lists some of the differences between SRAM and DRAM:

SRAM	DRAM
1. SRAM has lower access time, so it is faster compared to DRAM.	1. DRAM has higher access time, so it is slower than SRAM.
2. SRAM is costlier than DRAM.	2. DRAM costs less compared to SRAM.
3. SRAM requires constant power supply, which means this type of memory consumes more power.	3. DRAM offers reduced power consumption, due to the fact that the information is stored in the capacitor.
4. Due to complex internal circuitry, less storage capacity is available compared to the same physical size of DRAM memory chip.	4. Due to the small internal circuitry in the one-bit memory cell of DRAM, the large storage capacity is available.
5. SRAM has low packaging density.	5. DRAM has high packaging density.

Source : https://www.geeksforgeeks.org/different-types-ram-random-access-memory/

Read Only Memory

Read Only Memory (ROM) is an integrated circuit which is pre-programmed with specific functional data at manufacturing time. It is also called Firmware. These ROMs are not limited to computers only. Most of the electronic gadgets utilize its flexible functionality. In this post we will try to understand What is ROM, How it works, its types, applications, advantages and disadvantages.

What is Read Only Memory (ROM)

ROM is the acronym for Read Only Memory. ROM is a type of Primary Memory. As the name suggests its contents can be read only but cannot write on it. It is a non-volatile memory and so the data is retained even when the power is switched off.

The data that is required to be stored inside ROM is written during manufacturing phase. It stores such programs that are essential for the booting process of the computer. It generally cannot be altered. However, technologies are available to program these types of ROM.

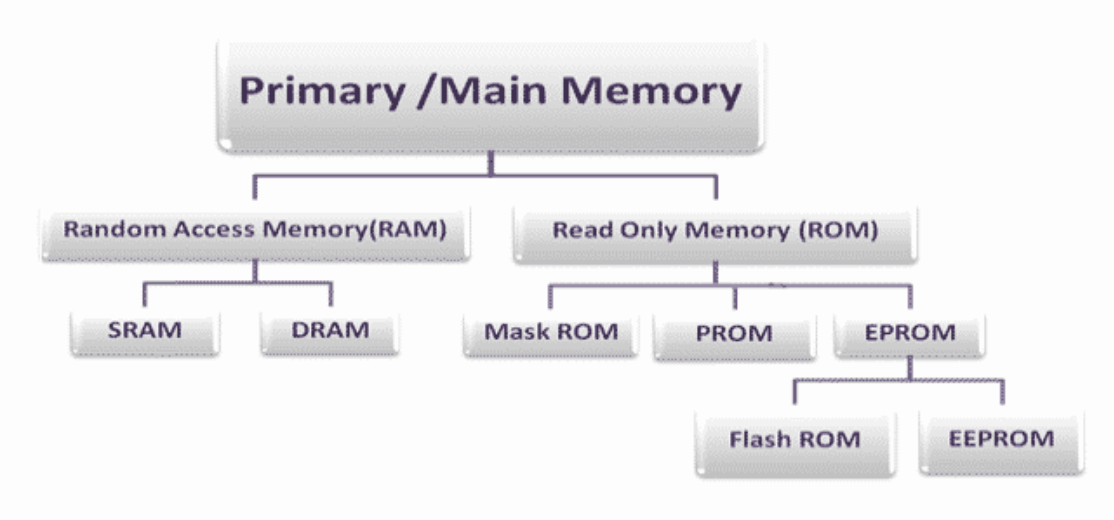


Fig.1 – Classification of Primary Memory

How Read Only Memory (ROM) Works

A ROM operates like an array. ROM chips contain a grid of rows and columns to turn ON or OFF. It uses a diode to connect the lines if the value is 1. If the value is 0, then these lines are not connected at all. Each element of the array corresponds to one storage element in the memory chip.

The address input to the chip is employed to pick out a specific memory location (corresponding to the array index). The value read from the memory chip corresponds to the contents of the selected element of the array.

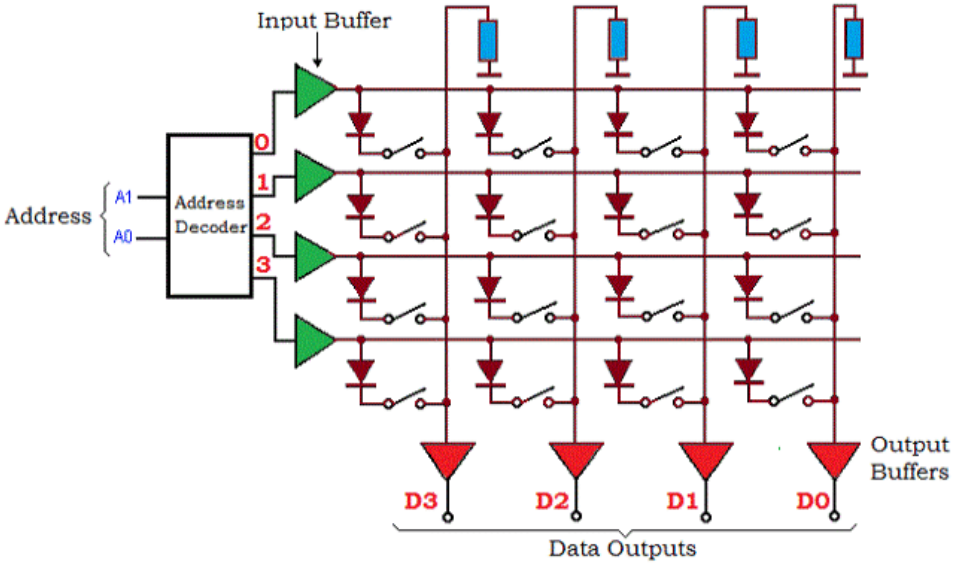


Fig.2 – Diode Grid in Read Only Memory (ROM)

ROM consists of two basic components- Decoder and OR gates. In ROM, the input to decoder will be in binary form and output will be its decimal equivalent. All the OR Gates present in the ROM will take decoders output as their input.

Types of Read Only Memory (ROM)

ROM is differentiated on the basis of methods used to write data on ROM chips and the number of times they can be written. It can be classified into following types : –

- Mask Read-Only Memory (MROM)
- Programmable Read-Only Memory (PROM)
- Erasable Programmable Read-Only Memory (EPROM)
- Electrically Erasable Programmable Read-Only Memory (EEPROM)
- Flash Read-Only Memory (Flash ROM)

Mask Read Only Memory (MROM)

MROM stands for Mask Read Only Memory. It is a memory chip that is manufactured with its contents. These are inexpensive and are the very first ROMs which were hard wired devices that contain a pre-programmed set of data or instructions.

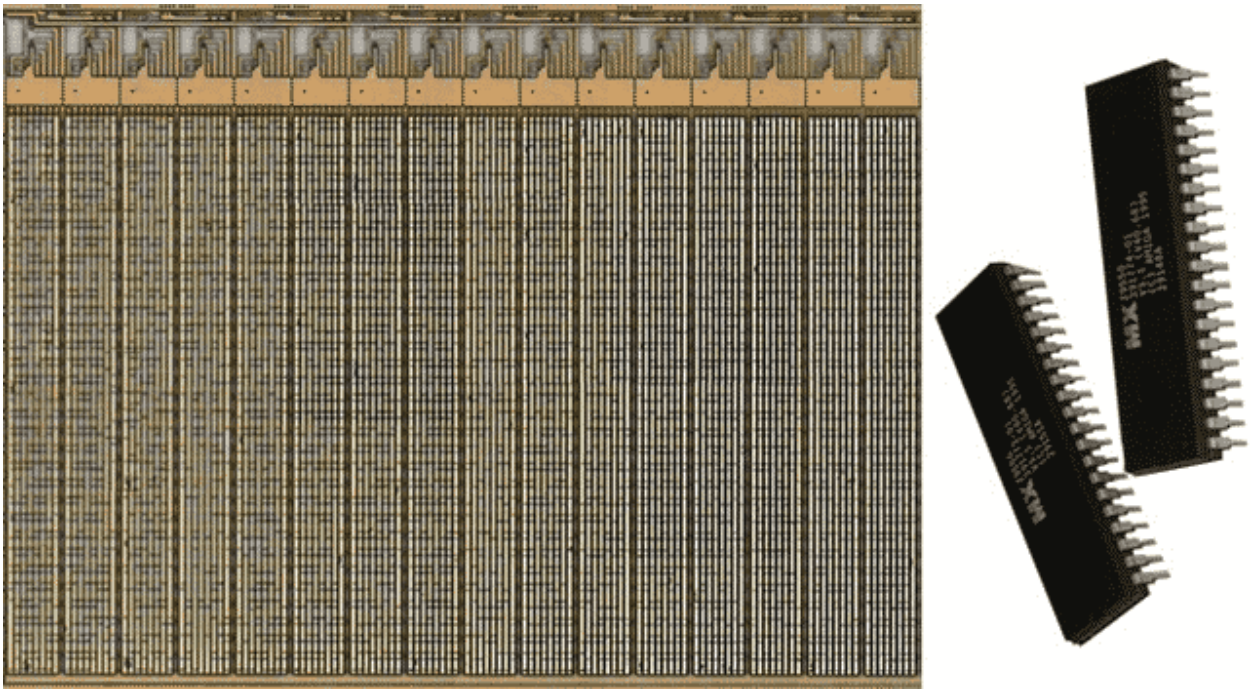


Fig. 4 – Programmable Read Only Memory (PROM)

The process of programming a PROM is called burning the PROM. There are tiny fuses in a PROM chip which are burnt open during programming. The data can be programmed only once and cannot be altered. So it is called one- time programming device.

Applications of Programmable Read Only Memory (PROM)

The Programmable ROM (PROM) are used in:

- Mobile Phones for providing User Specific Selections.
- Video game consoles
- Implantable Medical devices.
- Radio-Frequency Identification (RFID)tags.
- High definition Multimedia Interfaces(HDMI)

Advantages of Programmable Read Only Memory (PROM)

The advantages of Programmable ROM (PROM) are: –

- The programming can be done using many types of software and does not rely on hard wiring of the program to the chip.
- Since it is not possible to un-blow the fuse, so the authenticity of the data remains intact and it is impossible to remove or alter the contents.

Disadvantage of Programmable Read Only Memory (PROM)

The biggest disadvantage of PROM is that the data once burnt cannot be erased or changed when detected with errors.

Erasable Programmable Read Only Memory (EPROM)

EPROM stands for Erasable Programmable Read-Only Memory. It is a non volatile memory i.e. it can retain data even if the power supply is cut off. The basic limitation being encountered in PROM is that once it is programmed, it cannot be changed or altered. This limitation has been overcame by EPROM.

EPROM can be erased by exposing it to ultra violet light for a particular length of time using an EPROM eraser. After exposing, the chip returns to its initial state and can be reprogrammed.

This procedure can be carried out many times but repeated erasing and rewriting can eventually render the chip useless. Once written, data can be retained for about 10 years.

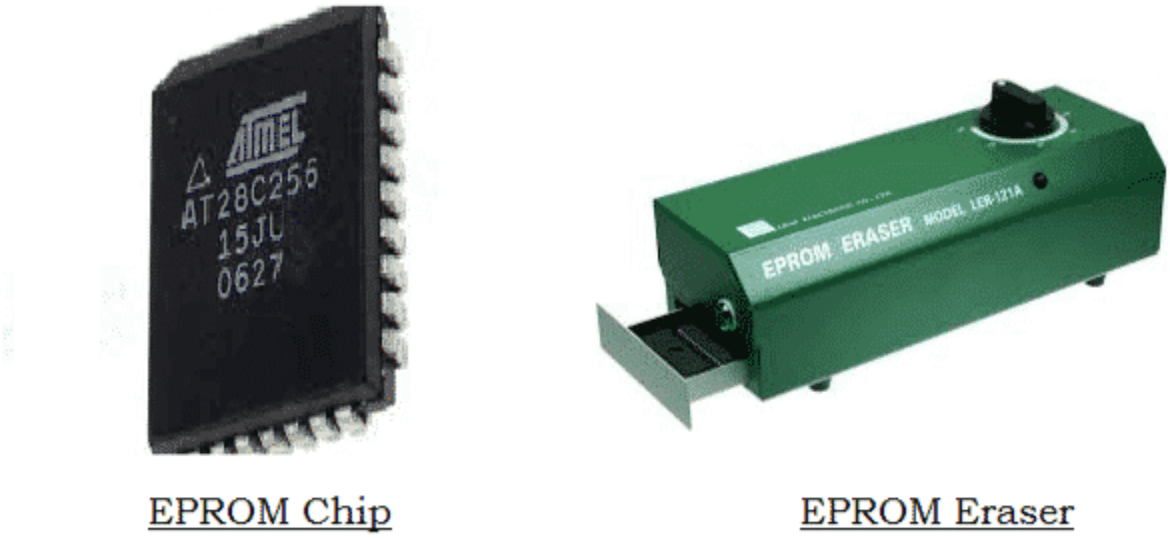


Fig. 5 – Erasable Programmable Read Only Memory (EPROM)

Applications of Erasable Programmable Read Only Memory (EPROM)
The applications of Erasable Programmable ROM (EPROM) includes:

- As program storage chip in Micro controllers.
- For debugging.
- For program development.
- As BIOS chip in computers.
- As program storage chip in modem, video card and many electronic gadgets.

Advantages of Erasable Programmable Read Only Memory (EPROM)
The advantages of Erasable Programmable ROM (EPROM) are:

- It is non-volatile.
- It can be erased and re -programmed.
- It is cost effective as compared to PROM.

Disadvantages of Erasable Programmable Read Only Memory (EPROM)
The disadvantages of Erasable Programmable ROM (EPROM) are:

- The static power consumption is high as the transistors used have higher resistance.
- It is not possible for a particular byte to be erased, instead the entire content is erased.
- UV based EPROM takes time to erase the content.

Electrically Erasable Programmable Read Only Memory (EEPROM)

EEPROM is the short form for Electrically Erasable Programmable Read Only Memory. It is similar to EPROM and thus developed to overcome the drawbacks of EPROMs. It is erased and programmed electrically i.e. it uses electrical signals instead of ultra violet rays.

The erasing and programming of data takes 4 to 10 milliseconds. Any byte can be erased at a time instead of the entire chip. The chip can be erased and re programmed for around ten thousand times, though the process is flexible but slow.

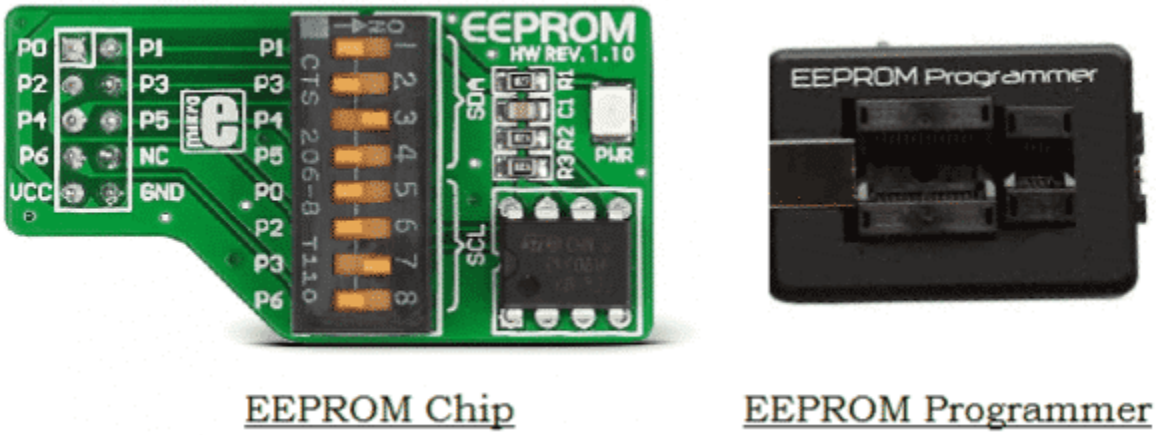


Fig. 6 –Electrically Erasable Programmable Read Only Memory (EEPROM)

Applications of Electrically Erasable Programmable Read Only Memory (EEPROM)

The applications of Electrically Erasable Programmable ROM (EPROM) includes:

- As BIOS chip in computers
- As storage for re-programmable calibration information in test-equipment.
- As storage for in-built self learning functionality in remote operated transmitters.

Advantages of Electrically Erasable Programmable Read Only Memory (EEPROM)

The advantages of Electrically Erasable Programmable ROM (EEPROM) are:

- The method of erasing is electrical and instant.
- Chip can be reprogrammed infinite number of times.
- Byte wise data can be erased instead of entire content on the board.
- To change the data, additional devices are not required.

Disadvantages of Electrically Erasable Programmable Read Only Memory (EEPROM)

The disadvantages of Electrically Erasable Programmable ROM (EEPROM) are:

- Different voltages are required for erasing, reading and writing the data.
- The data retention period of EEPROM is limited i.e 10 years approx.
- EEPROM devices are expensive compared to others.

Flash Read Only Memory (Flash ROM)

It is a universal flash programming non volatile utility, used in computer as a storage medium. It can be electrically erased and reprogrammed. In this, memory blocks of data (512 bytes) can be deleted and written at a particular time.

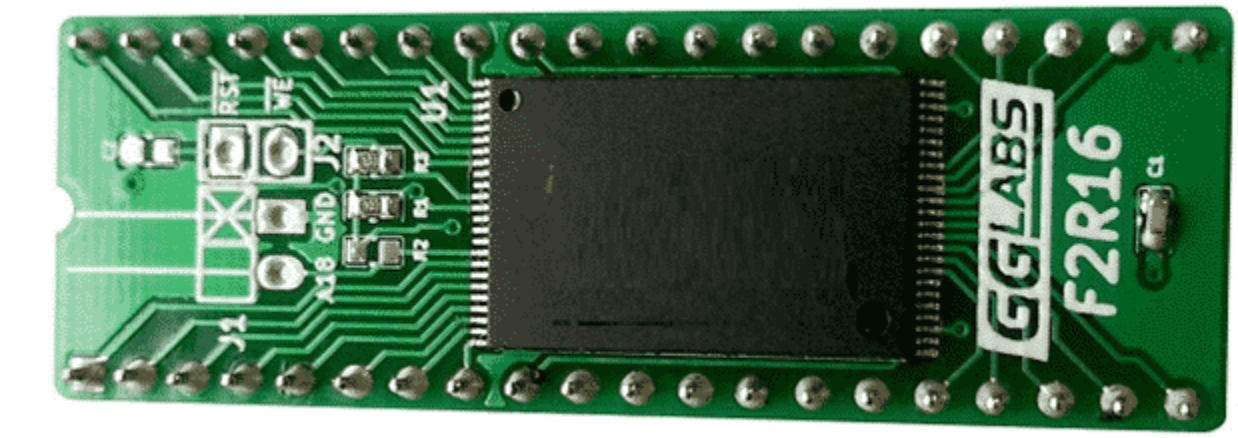


Fig. 7 – Flash Read Only Memory (Flash-ROM)

Applications of Flash Read Only Memory (Flash ROM)

The applications of Flash Read-Only Memory (Flash ROM) are:

- The latest technology computers use BIOS stored on a flash memory chip, called as flash BIOS.
- Modems, pen drives, small cards use flash ROM.

Advantages of Flash Read Only Memory (Flash ROM)

The Advantages of Flash Read-Only Memory (Flash ROM) are:

- High transferring speed.
- It saves data when turns OFF, preserve its state without power.
- Less prone to damage.

- ComparatiAely economical to other drives in small storage capacities.

Disadvantages of Flash Read Only Memory (Flash ROM)

The disadvantages of Flash Read-Only Memory (Flash ROM) are:

- Comparatively costly than hard disk.
- Number of read/writes are limited.

Introduction of Secondary Memory

Primary memory has limited storage capacity and is volatile. Secondary memory overcome this limitation by providing permanent storage of data and in bulk quantity. Secondary memory is also termed as external memory and refers to the various storage media on which a computer can store data and programs. The Secondary storage media can be fixed or removable. Fixed Storage media is an internal storage medium like hard disk that is fixed inside the computer. Storage medium that are portable and can be taken outside the computer are termed as removable storage media.

Uses of Secondary Media:

- **Permanent Storage:** Primary Memory (RAM) is volatile, i.e. it loses all information when the electricity is turned off, so in order to secure the data permanently in the device, Secondary storage devices are needed.

- **Portability:** Storage medium, like the CDs, flash drives can be used to transfer the data from one devive to another.

Fixed and Removable Storage

Fixed Storage-

A Fixed storage is an internal media device that is used by a computer system to store data, and usually these are referred to as the Fixed Disks drives or the Hard Drives. Fixed storage devices are literally not fixed, obviously these can be removed from the system for repairing work, maintenance purpose, and also for upgrade etc. But in general, this can't be done without a proper toolkit to open up the computer system to provide physical access, and that needs to be done by an engineer.

Technically, almost all of the data i.e. being processed on a computer system is stored on some type of a built-in fixed storage device.

Types of fixed storage:

- Internal flash memory (rare)
- SSD (solid-state disk) units
- Hard disk drives (HDD)

Removable Storage-

A Removable storage is an external media device that is used by a computer system to store data, and usually these are referred to as the Removable Disks drives or the External Drives. Removable storage is any type of storage device that can be removed/ejected from a computer system while the system is running. Examples of external devices include CDs, DVDs and Blu-Ray disk drives, as well as diskettes and USB drives. Removable storage makes it easier for a user to transfer data from one computer system to another.
In a storage factors, the main benefit of removable disks is that they can provide the fast data transfer rates associated with storage area networks (SANs)

Types of Removable Storage:

- Optical discs (CDs, DVDs, Blu-ray discs)
- Memory cards
- Floppy disks
- Magnetic tapes
- Disk packs
- Paper storage (punched tapes , punched cards)

Secondary Storage Media

There are the following main types of storage media:

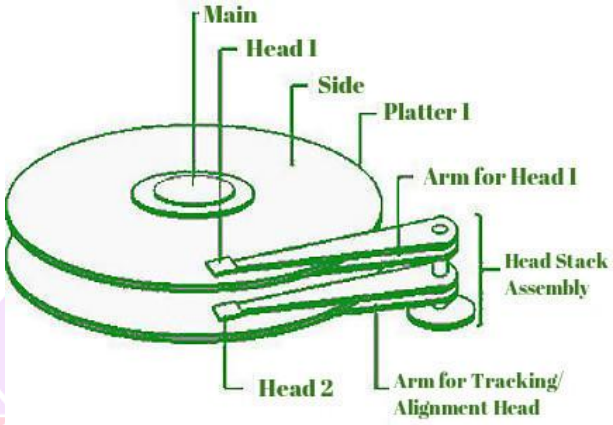
1. Magnetic storage media: Magnetic media is coated with a magnetic layer which is magnetized in clockwise or anticlockwise directions. When the disk moves, the head interprets the data stored at a specific location in binary 1s and 0s at reading. **Examples:** hard disks, floppy disks and magnetic tapes.

- **Floppy Disk:** A floppy disk is a flexible disk with a magnetic coating on it. It is packaged inside a protective plastic envelope. These are one of the oldest type of portable storage devices that could store up to 1.44 MB of data but now they are not used due to very less memory storage.

- **Hard disk:** A hard disk consists of one or more circular disks called platters which are mounted on a common spindle. Each surface of a platter is coated with a magnetic material. Both surfaces of each disk are capable of storing data except the top and bottom disk where only the inner surface is used. The information is recorded on the surface of the rotating disk by magnetic read/write heads. These heads are joined to a common arm known as access arm.

Hard disk drive components:

Most of the basic types of hard drives contains a number of disk platters that are placed around a spindle which is placed inside a sealed chamber. The chamber also includes read/write head and motors. Data is stored on each of these disks in the arrangement of concentric circles called tracks which are divided further into sectors. Though internal Hard drives are not very portable and used internally in a computer system, external hard disks can be used as a substitute for portable storage. Hard disks can store data upto several terabytes.



2. Optical storage media: In optical storage media information is stored and read using a laser beam. The data is stored as a spiral pattern of pits and ridges denoting binary 0 and binary 1. Examples: CDs and DVDs

- **Compact Disk:** A Compact Disc drive(CDD) is a device that a computer uses to read data that is encoded digitally on a compact disc(CD). A CD drive can be installed inside a computer's compartment, provided with an opening for easier disc tray access or it can be used by a peripheral device connected to one of the ports provided in the computer system.A compact disk or CD can store approximately 650 to 700 megabytes of data. A computer should possess a CD Drive to read the CDs. There are three types of CDs:

CD- ROM	CD-R	CD-RW
It stands for Compact Disk – Read Only Memory	It stands for Compact Disk- Recordable.	It stands for Compact Disk-Rewritable.
Data is written on these disks at the	Data can be recorded on	It can be read or

CD- ROM	CD-R	CD-RW
time of manufacture. This data cannot be changed, once is it written by the manufacturer, but can only be read. CD- ROMs are used for text, audio and video distribution like games, encyclopedias and application software.	these disks but only once. Once the data is written in a CD-R, it cannot be erased/modified.	written multiple times but a CD-RW drive needs to be installed on your computer before editing a CD-RW.

- **DVD:**It stands for Digital Versatile Disk or Digital Video Disk. It looks just like a CD and use a similar technology as that of the CDs but allows tracks to be spaced closely enough to store data that is more than six times the CD's capacity. It is a significant advancement in portable storage technology. A DVD holds 4.7 GB to 17 GB of data.

- **Blue Ray Disk:** This is the latest optical storage media to store high definition audio and video. It is similar to a CD or DVD but can store up to 27 GB of data on a single layer disk and up to 54 GB of data on a dual layer disk. While CDs or DVDs use red laser beam, the blue ray disk uses a blue laser to read/write data on a disk.

3. Solid State Memories: Solid-state storage devices are based on electronic circuits with no moving parts like the reels of tape, spinning discs etc. Solid-state storage devices use special memories called flash memory to store data. Solid state drive (or flash memory) is used mainly in digital cameras, pen drives or USB flash drives.

Pen Drives: Pen Drives or Thumb drives or Flash drives are the recently emerged portable storage media. It is an EEPROM based flash memory which can be repeatedly erased and written using electric signals. This memory is accompanied with a USB connector which enables the pendrive to connect to the computer. They have a capacity smaller than a hard disk but greater than a CD. Pendrive has following advantages:

- **Transfer Files:** A pen drive being plugged into a USB port of the system can be used as a device to transfer files, documents and photos to a PC and also vice versa. Similarly, selected files can be transferred between a pen drive and any type of workstation.

- **Portability:** The lightweight nature and smaller size of a pen drive make it possible to carry it from place to place which makes data transportation an easier task.

- **Backup Storage:** Most of the pen drives now come with a feature of having password encryption, important information related to family, medical records and photos can be stored on them as a backup.

- **Transport Data:** Professionals/Students can now easily transport large data files and video/audio lectures on a pen drive and gain access to them from anywhere. Independent PC technicians can store work-related utility tools, various programs and files on a high-speed 64 GB pen drive and move from one site to another.

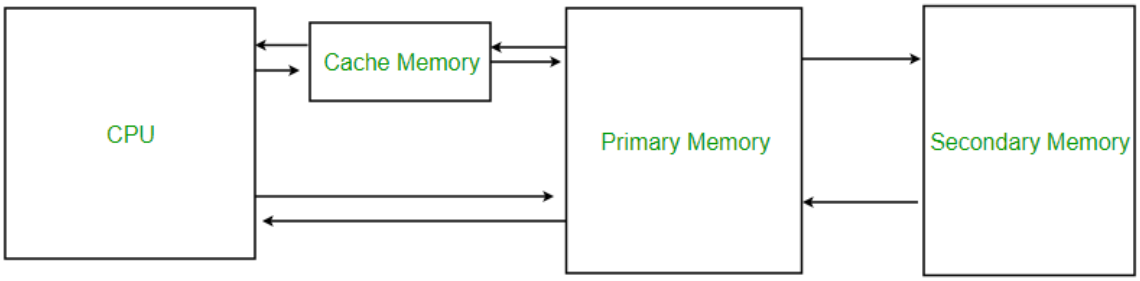
Difference between Primary Memory and Secondary Memory:	
PRIMARY MEMORY	SECONDARY MEMORY
Primary memory is directly accessed by the Central Processing Unit(CPU).	Secondary memory is not accessed directly by the Central Processing Unit(CPU). Instead, data accessed from a secondary memory is first loaded into Random Access Memory(RAM) and is then sent to the Processing Unit.
RAM provides much faster accessing speed to data than secondary memory. By loading software programs and required files into primary memory(RAM), computer can process data much more quickly.	Secondary Memory is slower in data accessing. Typically primary memory is six times faster than the secondary memory.
Primary memory, i.e. Random Access Memory(RAM) is volatile and gets completely erased when a computer is shut down.	Secondary memory provides a feature of being non-volatile, which means it can hold on to its data with or without electrical power supply.

Cache Memory

Cache Memory is a special very high-speed memory. It is used to speed up and synchronizing with high-speed CPU. Cache memory is costlier than main memory or disk memory but

economical than CPU registers. Cache memory is an extremely fast memory type that acts as a buffer between RAM and the CPU. It holds frequently requested data and instructions so that they are immediately available to the CPU when needed.

Cache memory is used to reduce the average time to access data from the Main memory. The cache is a smaller and faster memory which stores copies of the data from frequently used main memory locations. There are various different independent caches in a CPU, which store instructions and data.



Levels of memory:

- **Level 1 or Register :** It is a type of memory in which data is stored and accepted that are immediately stored in CPU. Most commonly used register is accumulator, Program counter, address register etc.
- **Level 2 or Cache memory :** It is the fastest memory which has faster access time where data is temporarily stored for faster access.
- **Level 3 or Main Memory :** It is memory on which computer works currently. It is small in size and once power is off data no longer stays in this memory.
- **Level 4 or Secondary Memory :**It is external memory which is not as fast as main memory but data stays permanently in this memory.

Cache Performance: When the processor needs to read or write a location in main memory, it first checks for a corresponding entry in the cache.

- If the processor finds that the memory location is in the cache, a **cache hit** has occurred and data is read from cache
- If the processor **does not** find the memory location in the cache, a **cache miss** has occurred. For a cache miss, the cache allocates a new entry and copies in data from main memory, then the request is fulfilled from the contents of the cache.

The performance of cache memory is frequently measured in terms of a quantity called **Hit ratio.**

Hit ratio = hit / (hit + miss) = no. of hits/total accesses

We can improve Cache performance using higher cache block size, higher associativity, reduce miss rate, reduce miss penalty, and reduce Reduce the time to hit in the cache.

Cache Mapping: There are three different types of mapping used for the purpose of cache memory which are as follows: Direct mapping, Associative mapping, and Set-Associative mapping. These are explained below.

1. **Direct Mapping :** The simplest technique, known as direct mapping, maps each block of main memory into only one possible cache line. or
In Direct mapping, assigne each memory block to a specific line in the cache. If a line is previously taken up by a memory block when a new block needs to be loaded, the old block is trashed. An address space is split into two parts index field and a tag field. The cache is used to store the tag field whereas the rest is stored in the main memory. Direct mapping`s performance is directly proportional to the Hit ratio.

2. i = j modulo m

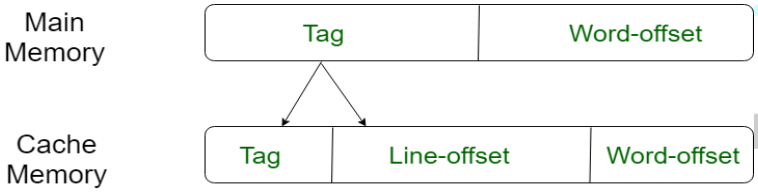
3. where

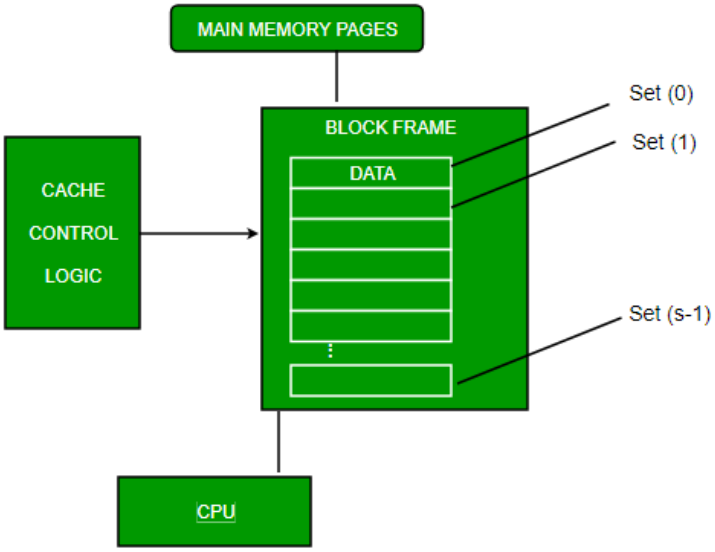
4. i=cache line number

5. j= main memory block number

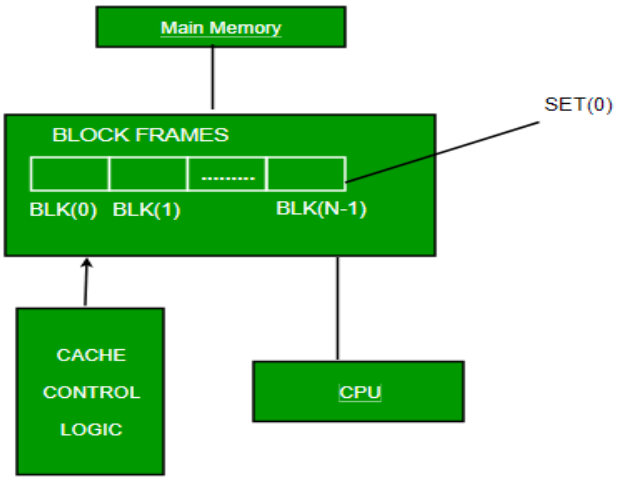
m=number of lines in the cache

For purposes of cache access, each main memory address can be viewed as consisting of three fields. The least significant w bits identify a unique word or byte within a block of main memory. In most contemporary machines, the address is at the byte level. The remaining s bits specify one of the 2^s blocks of main memory. The cache logic interprets these s bits as a tag of s-r bits (most significant portion) and a line field of r bits. This latter field identifies one of the $m=2^r$ lines of the cache.





2. Associative Mapping : In this type of mapping, the associative memory is used to store content and addresses of the memory word. Any block can go into any line of the cache. This means that the word id bits are used to identify which word in the block is needed, but the tag becomes all of the remaining bits. This enables the placement of any word at any place in the cache memory. It is considered to be the fastest and the most flexible mapping form.



3. Set-associative Mapping : This form of mapping is an enhanced form of direct mapping where the drawbacks of direct mapping are removed. Set associative addresses the problem of possible thrashing in the direct mapping method. It does this by saying that instead of having exactly one line that a block can map to in the cache, we will group a few lines together creating a ***set***. Then a block in memory can map to any one of the lines of a specific set..Set-associative mapping allows that each word that is present in the cache can have two or more words in the main memory for the same index address. Set associative cache mapping combines the best of direct and associative cache mapping techniques.

In this case, the cache consists of a number of sets, each of which consists of a number of lines. The relationships are

m = v * k

i= j mod v

where

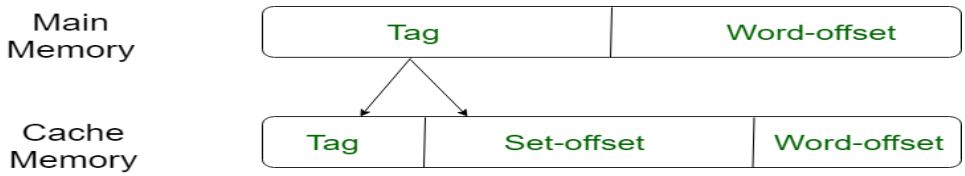
i=cache set number

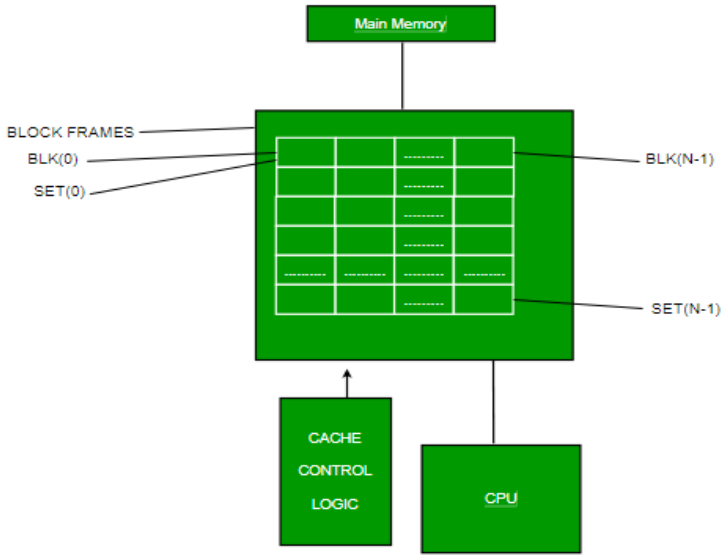
j=main memory block number

v=number of sets

m=number of lines in the cache number of sets

k=number of lines in each set





Application of Cache Memory –

1. 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.
2. The correspondence between the main memory blocks and those in the cache is specified by a mapping function.

Types of Cache –

- **Primary Cache :** A primary cache is always located on the processor chip. This cache is small and its access time is comparable to that of processor registers.
- **Secondary Cache :** Secondary cache is placed between the primary cache and the rest of the memory. It is referred to as the level 2 (L2) cache. Often, the Level 2 cache is also housed on the processor chip.

Source : https://www.geeksforgeeks.org/cache-memory-in-computer-organization/

Cache Structure and Design

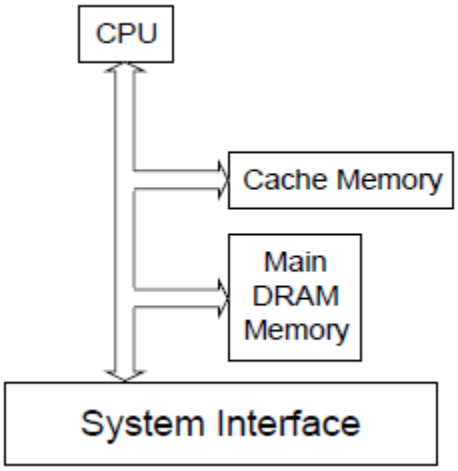


Figure 2-1 Basic Cache Model

Figure 2-1 shows a simplified diagram of a system with cache. In this system, every time the CPU performs a read or write, the cache may intercept the bus transaction, allowing the cache to decrease the response time of the system. Before discussing this cache model, lets define some of the common terms used when talking about cache.

2.1.1 Cache Hits

When the cache contains the information requested, the transaction is said to be a cache hit.

2.1.2 Cache Miss

When the cache does not contain the information requested, the transaction is said to be a cache miss.

2.1.3 Cache Consistency

Since cache is a photo or copy of a small piece main memory, it is important that the cache always reflects what is in main memory. Some common terms used to describe the process of maintaining cache consistency are:

2.1.3.1 Snoop

When a cache is watching the address lines for transaction, this is called a snoop. This function allows the cache to see if any transactions are accessing memory it contains within itself.

2.1.3.2 Snarf

When a cache takes the information from the data lines, the cache is said to have snarfed the data. This function allows the cache to be updated and maintain consistency.

Snoop and snarf are the mechanisms the cache uses to maintain consistency. Two other terms are commonly used to describe the inconsistencies in the cache data, these terms are:

2.1.3.3 Dirty Data

When data is modified within cache but not modified in main memory, the data in the cache is called “dirty data.”

2.1.3.4 Stale Data

When data is modified within main memory but not modified in cache, the data in the cache is called stale data.

Now that we have some names for cache functions lets see how caches are designed and how this effects their function.

2.2 Cache Architecture

Caches have two characteristics , a read architecture and a write policy. The read architecture may be either “Look Aside” or “Look Through.” The write policy may be either “Write-Back” or “Write-Through.” Both types of read architectures may have either type of write policy, depending on the design. Write policies will be described in more detail in the next section. Lets examine the read architecture now.

2.2.1 Look Aside

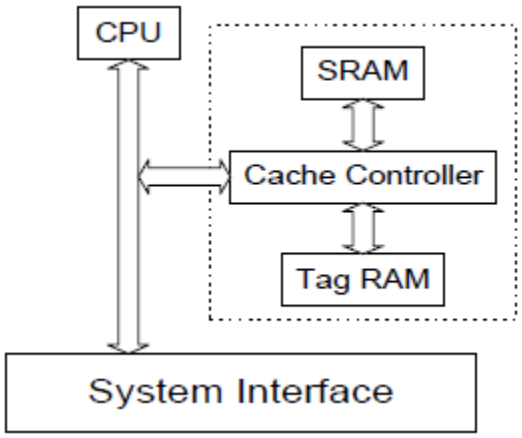


Figure 2-2 Look Aside Cache

Figure 2-2 shows a simple diagram of the “look aside “cache architecture. In this diagram, main memory is located opposite the system interface. The discerning feature of this cache unit is that it sits in parallel with main memory. It is important to notice that both the main memory and the cache see a bus cycle at the same time. Hence the name “look aside.”

2.2.1.1 Look Aside Cache Example

When the processor starts a read cycle, the cache checks to see if that address is a

cache hit.

HIT: If the cache contains the memory location, then the cache will respond to the read cycle and terminate the bus cycle.

MISS: If the cache does not contain the memory location, then main memory will respond to the processor and terminate the bus cycle. The cache will snarf the data, so next time the processor requests this data it will be a cache hit.

Look aside caches are less complex, which makes them less expensive. This architecture also provides better response to a cache miss since both the DRAM and the cache see the bus cycle at the same time. The draw back is the processor cannot access cache while another bus master is accessing main memory.

2.2.2 Read Architecture: Look Through

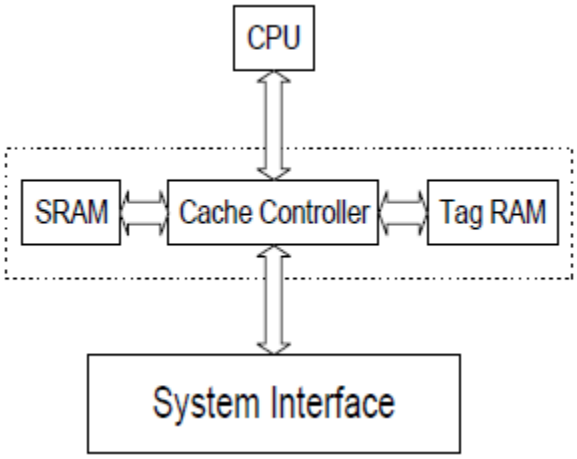


Figure 2-3 Look Through Cache

Figure 2-3 shows a simple diagram of cache architecture. Again, main memory is located opposite the system interface. The discerning feature of this cache unit is that it sits between the processor and main memory. It is important to notice that cache sees the processors bus cycle before allowing it to pass on to the system bus.

2.2.2.1 Look Through Read Cycle Example

When the processor starts a memory access, the cache checks to see if that address is a cache hit.

HIT: The cache responds to the processor's request without starting an access to main memory.

MISS: The cache passes the bus cycle onto the system bus. Main memory then responds to the processors request. Cache snarfs the data so that next time the processor requests this data, it will be a cache hit.

This architecture allows the processor to run out of cache while another bus master is accessing main memory, since the processor is isolated from the rest of the system. However, this cache architecture is more complex because it must be able to control accesses to the rest of the system. The increase in complexity increases the cost. Another down side is that memory accesses on cache misses are slower because main memory is not accessed until after the cache is checked. This is not an issue if the cache has a high hit rate and their are other bus masters.

2.2.3 Write Policy:

A write policy determines how the cache deals with a write cycle. The two common write policies are Write-Back and Write-Through.

In Write-Back policy, the cache acts like a buffer. That is, when the processor starts a write cycle the cache receives the data and terminates the cycle. The cache then writes the data back to main memory when the system bus is available. This method provides the greatest performance by allowing the processor to continue its tasks while main memory is updated at a later time. However, controlling writes to main memory increase the cache's complexity and cost.

The second method is the Write-Through policy. As the name implies, the processor writes through the cache to main memory. The cache may update its contents, however the write cycle does not end until the data is stored into main memory. This method is less complex and
An Overview of Cache therefore less expensive to implement. The performance with a Write-Through policy is lower since the processor must wait for main memory to accept the data.

2.3 Cache Components

The cache sub-system can be divided into three functional blocks: SRAM, Tag RAM, and the Cache Controller. In actual designs, these blocks may be implemented by multiple chips or all may be combined into a single chip.

2.3.1 SRAM

Static Random Access Memory (SRAM) is the memory block which holds the data. The size of the SRAM determines the size of the cache.

2.3.2 Tag RAM

Tag RAM (TRAM) is a small piece of SRAM that stores the addresses of the data that is stored in the SRAM.

2.3.3 Cache Controller

The cache controller is the brains behind the cache. Its responsibilities include: performing the
snoops and snarfs, updating the SRAM and TRAM and implementing the write policy. The cache controller is also responsible for determining if memory request is cacheable2 and if a request is a cache hit or miss.

Mapping of Cache Memory

Cache Mapping Techniques

Today in this cache mapping techniques based tutorial for Gate CSE Exam we will learn about different type of cache memory mapping techniques. These techniques are used to fetch the information from main memory to cache memory.

There are three type of mapping techniques used in cache memory. Let us see them one by one. Three types of mapping procedures used for cache memory are as follows;

What is cache memory mapping?

Cache memory mapping is a method of loading the data of main memory into cache memory. In more technical sense content of main memory is brought into cache memory which is referenced by the cpu. This can be done in three ways ;

(i) Associative mapping Technique

The fastest and most flexible computer used associative memory in computer organization. The associative memory has both the address and memory word. In associative mapping technique any memory word from main memory can be store at any location in cache memory.

The address value of 15 bits is shown as a five-digit octal number and itscorresponding 12 bit word is shown as a four-digit octal number and its corresponding 12-bit word is shown as a four-digit octal number. CPU generated 15 bits address is placed in the argument register and the associative memory is searched for a matching address.

If the address is present then corresponding 12-bit data is read from it and sent to the CPU. But If no match occurs for that address, in that case required word is accessed from the main memory , after that this address-data pair is sent to the associative cache memory.

Fig. Associative mapping cache (all numbers in octal)

CPU address (15 bits)

↓

Argument register

← Address →	← Data →
01000	3450
02777	6710
22345	1234

Suppose that cache is full then question arises that where to store this address-data pair. In this condition this concept of replacement algorithms comes into existence.

Replacement algorithm determines that which existing data in cache is remove from cache and make a space free so that required data can be placed in cache.

A simple procedure is to replace cells of the cache is round-robin order whenever a new word is requested from main memory. This constitutes a first-in first-out (FIFO) replacement policy.

(ii) Direct mapping technique

Associative memories are more costly as compared to random-access memories because of logic is added in with each cell. The 15 bits address generated by the cpu is divided into two fields. The nine lower bits represents the *index* field and the remaining six bits form the *tag* field. The figure shows that main memory required an address that includes tag and the index bits.

The index field bits represent the number of address bits required to fetch the cache memory.

Consider a case where there are 2k words in cache memory and 2*n* words in main memory. The n bit memory address is divided into two fields: k bits for the index field and the *n*-k bits for the tag field. The direct mapping cache organization uses the *n*-k bits for the tag field.

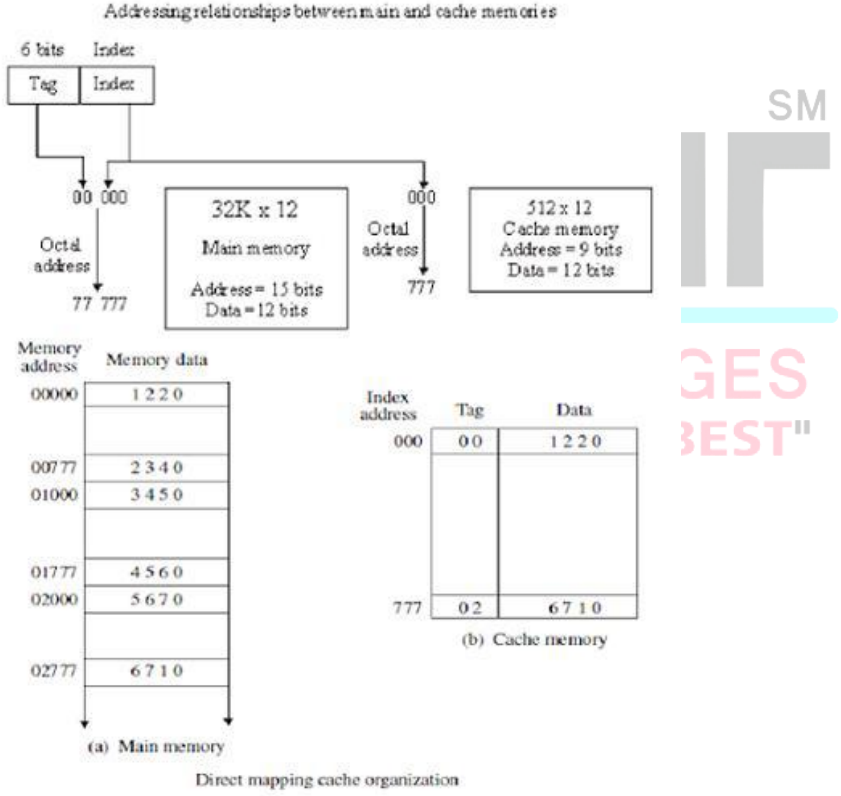
In this direct mapped cache tutorial it is also explained the direct mapping technique in cache organization uses the *n* bit address to access the main memory and the k-bit index to access the cache. The internal arrangement of the words in the cache memory is as shown in figure.

It has been shown in cache that each word in cache consists of the data and tag associated with it.When a new word is loaded into the cache, then its tag bits are also stored alongside with the data bits. When the CPU generates a memory request, the index field is used for the address to access that cache.

The tag field of the address referred by the cpu are compared with the tag in the word read from the cache. If these two tags match it means that there is a hit and the desired data word is available in the cache.

If these two tags does not match then there is a miss and the required word is not present in cache and it is read from main memory. It is then stored in the cache memory along with the new tag.

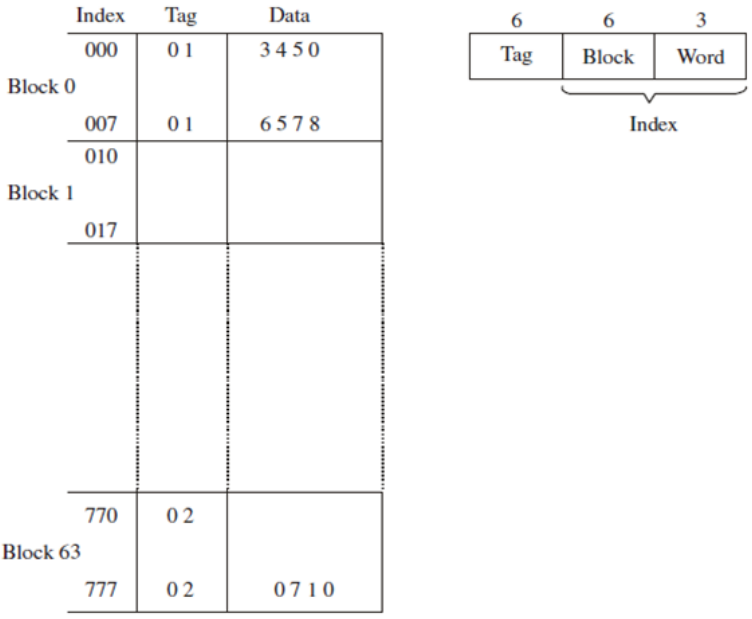
The disadvantage of direct mapping technique is that the hit ratio can drop considerably if two or more words whose addresses have the same index but different tags are accessed repeatedly.



Direct mapping cache organization

To see how the direct-mapping technique operates, consider the numerical example as shown. The word at address zero is presently stored in the cache with its index = 000, tag = 00, data = 1220. Assume that the CPU now wants to read the word at address 02000. Since the index address is 000, so it is used to read the cache and two tags are then compared.

Here we found that the cache tag is 00 but the address tag is 02 these two tags do not match, iss occurs. So the main memory is accessed and the data word 5670 is sent to the CPU. The cache word at index address 000 is then replaced with a tag of 02 and data of 5670.



Direct mapping cache with block size of 8 words

First part of the index filed is the block field and second is the word field. In a 512-word cache there are 64 blocks and size of each block is 8 words. Since there are 64 blocks in cache so 6 bits are used to identify a block within 64 blocks. So 6 bits are used to represent the block field and size of each block is 8 words so 3 bits are used to identify a word among these 8 words. (iii)Set associative mapping

Disadvantage of direct mapping techniques is that it required a lot of comparisons.

A third type of cache organization, called set associative mapping, is an improvement over the direct-mapping organization in that in set associative mapping technique. In this technique each

data word is stored together with its tag and the number of tag data items in one word of cache is said to form a set.

Here I have explained the concept of set associative memory with the help of an example. An example of a set associative cache is shown in figure. Each index address has two parts data words and their associated tags.

6 bits are required to represent the tag field and 12 bits are required to represent the word field. Word length is36 bits. Here an index 9 bits index address can have 512 words. So cache memory size is 512 x 36.

It can accommodate 1024 words of main memory since each word of cache contains two data words. In general, a set-associative cache of set size k will accommodate *k* words of main memory in each word of cache.

Index	Tag	Data	Tag	Data
000	01	3450	02	5670
777	02	6710	00	2340

Direct mapping cache with block size of 8 words

When the CPU generates a logical address to fetch a word from main memory then the index value of the address is used to access the cache. The tag field of the CPU generated address is then compared with both tags in the cache to determine weather they match or not.

The comparison logic is performed by an associative search of the tags in the set similar to an associative memory search so it is named as “set-associative.”

We can improve the cache performance of cache memory if we can improve the hit ratio and the hit ratio can be improve by improving the set size increases because more words with the same index but different tags can reside in cache.

When a miss occurs in a set-associative cache and the set if full, it is necessary to replace one of the tag-data items with a new value using cache replacement algorithms.

Cache Line Replacement Algorithms :

When a new line is loaded into the cache, one of the existing lines must be replaced. In a direct mapped cache, the requested block can go in exactly one position, and the block occupying that position must be replaced. In an associative cache we have a choice of where to place the requested block, and hence a choice of which block to replace. In a fully associative cache, all blocks are candidates for replacement. In a set associative cache, we must choose among the blocks in the selected set. Therefore a line replacement algorithm is needed which sets up well defined criteria upon which the replacement is made. A large number of algorithms are possible and many have been implemented. Four of the most common cache line replacement algorithms are:

- *Least Recently Used* (LRU) - the cache line that was last referenced in the most distant past is replaced.

- *FIFO* (First In- First Out) - the cache line from the set that was loaded in the most distant past is replaced.

- *LFU* (Least Frequently Used - the cache line that has been referenced the fewest

number of times is replaced.

Random - a randomly selected line form cache is replaced

The most commonly used algorithm is LRU. LRU replacement is implemented by keeping track of when each element in a set was used relative to the other elements in the set. For a two-way set associative cache, tracking when the two lines were used can be easily implemented in hardware by adding a single bit (*use* bit) to each cache line. Whenever a cache line is referenced its use bit is set to 1 and the use bit of the other cache line in the same set is set to 0. The line selected for replacement at any specific time is the line whose use bit is currently 0. The principle of the locality of reference means that a recently used cache line is more likely to be referenced again, LRU tends to give the best performance. In practice, as associativity increases,

LRU is too costly to implement, since tracking the information is costly. Even for four-way set associativity, LRU is often approximated – for example, by keeping track of which of a pair of blocks is LRU (which requires one bit), and then tracking which line in each pair is LRU (which requires one bit per pair). For large associativity, LRU is either approximated or random replacement is used.

The FIFO replacement policy is again easily implemented in hardware by the cache lines as queues. The LFU replacement algorithm is implemented by associating with each cache line a counter which increments on every reference to the line. Whenever a line needs to be replaced, the line with the smallest counter value is selected, as it will be the cache line that has experienced the fewest references.

Random replacement is simple to build in hardware While it may seem that this algorithm would be a poor replacement line selection method, in reality it performs only slightly worse than any of the other three algorithms that we mentioned. For a two-way set associative cache, random replacement has a miss rate only about 1.1 times higher than LRU replacement. The reason for this is easy to see. Since there are only two cache lines per set, any replacement algorithm must select one of the two, therefore the random selection method has a 50-50 chance of selecting the same one that the LRU algorithm would select yet the random algorithm has no overhead (i.e., there wouldn't be any use bit). As the caches become larger, the miss rate for both replacement strategies falls, and the absolute difference becomes small. In fact, random replacement is sometimes better than simple LRU approximations that can be easily implemented in hardware. [13]

Question Bank for Unit IV

1: (a) Explain paging. Explain how paging can be implemented in CPU to access virtual Memory

(b) Explain any three page replacement methods.

2 (a) What is cache memory? Explain hit ration and average access time

(b) Name three techniques of cache mapping and explain any one in detail.

3 (a) Draw block diagram of memory hierarchi of computer system. Explain why 3 level hierarchi is necessary.

(b) If cache access time is 100 ns, main memory access time is 1000 ns and the hit ration is 0.9. Find the average access time and also define hit ratio.

4 (a) Write comparison between serial and parallel data transfer.

(b) Draw the block diagram of DMA and explain

5 (a) What are different mapping scheme? Explain

(b) Explain the concept of virtual memory.