25/02/2020
Tuesday

UNIT - 3

Memory Organization

operating system

Memory Hierarchy

Hardware and software interface

The memory unit is an essential ~~unit~~ component in any digital computer and is needed for storing programs and data

Q5

The memory unit that communicates directly with the CPU is called primary memory or main memory.

i- RAM, Rom

direct access to processor

Devices that provide backup storage are called auxilary memory or secondary memory.

2°

Harddisk
Floppy
USB

The most common auxilary devices used is Harddisk drive (HDD), Magnetic disk, magnetic tapes, CD, floppy disk. etc

size ↑ 2°
i

speed ↑ i°
2°

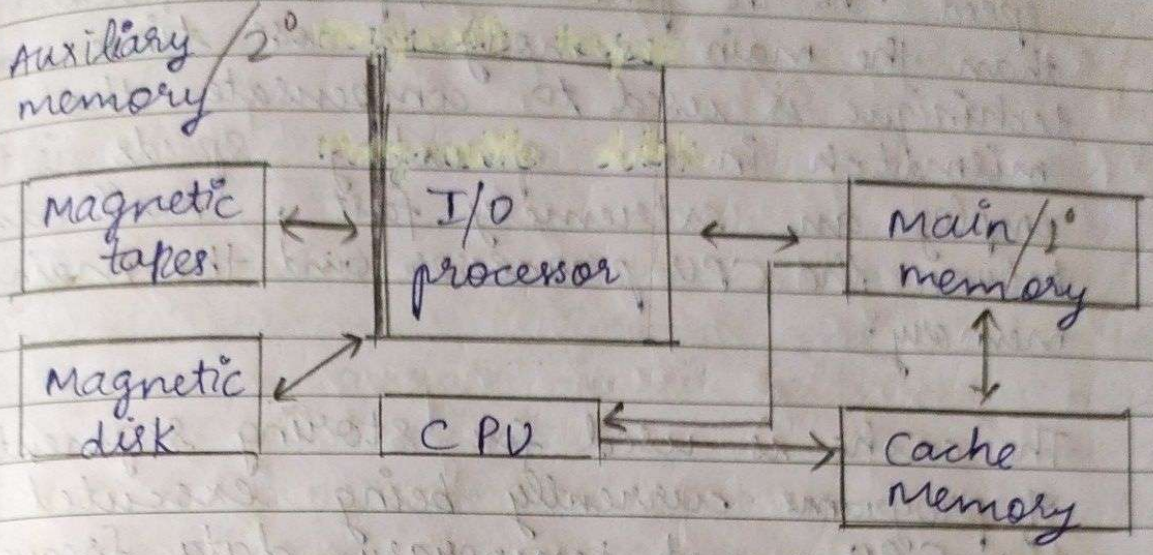
These types of memories are used for storing system files, large data files and other backup information

Not all accumulated information is needed by the processor at the same time

Only the programs and data currently needed by the processor reside in the main memory.

All the information stored in the auxilary memory are transferred in

the main memory when needed



28/02/2020 Friday

Main memory occupies a central position by being able to communicate directly with the CPU and with the auxilary memory through an input-output processor (I/O processor).

Only programs and data currently needed by the processor reside in the main memory

When programs are not residing in the main memory, and are needed by the CPU, they are brought in from the auxilary memory.

Programs not currently needed in main memory are transferred into auxilary memory to provide space for currently used programs and data.

A special high speed memory called a cache memory is sometimes used to increase the speed of processing by making current programs and data available to the CPU at a rapid rate. The cache memory is employed in computer systems to compensate the speed difference between the main

memory access time and processor speed. As the processor speed is faster than the main memory access time, a technique is used to compensate the mismatch in the operating speeds is to employ an extremely fast, small cache between the CPU processor and the 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.

→ While the input-output processor manages data transfer between auxiliary memory and the main memory, the cache organisation is concerned with the transfer of information between main memory and CPU.

Thus as the memory access speed increases, so does its relative cost

The cache memory is very small, relatively expensive and has very high access speed. It holds those part of the program and data that are mostly used, while the auxiliary memory holds those path parts that are not presently used by the CPU.

The CPU has direct access to both cache and main memory but not to auxiliary memory.

The transfer from auxiliary memory to main memory is usually done by means of direct memory access of large data blocks.

Main Memory

Main memory is the central storage unit in a computer system. Programs and data that the processor operate on is held in the main memory. (It is relatively large and fast memory used to store data and programs during computer operation)
Physically the main memory of a computer system is a collection of chips that are plugged into the mother board.
The principal technology used for the main memory is based on semi-conductor integrated circuits.

The main memory is broadly classified into2:- RAM and ROM.

RAM (Random Access Memory)

The RAM chips are available for two types of ~~operin~~ Operating modes.
(i) Static RAM - (SRAM)
(ii) Dynamic RAM - (DRAM)

The S-RAM consists essentially of internal flip-flops that store binary information. The stored information remains valid as long as the power is applied to the unit.

The static RAM is easier to use and has shorter read and write cycles.

The capacitors are provided inside the chip by MOS transistors

The D-RAM stores the binary information in form of electric charges on the that are applied to capacitors. The stored charge on the capacitor tend to discharge with time and the capacitors must be periodically recharged by refreshing the dynamic memory (DRAM)

Refreshing is done in every few milliseconds to restore the decaying charge. DRAM offers reduced power consumption and the larger storage capacity in a single memory chip.

ROM : (Read only memory)

The ROM is a portion of the main memory. which ~~It~~ is needed for ~~an~~ storing an initial program called boot strap loader.

The boot-strap loader is a program whose function is to start the computer software and load the O.S when the power is turned on.

RAM → Since the RAM is volatile, its contents are destroyed when power is switched off. The contents of ROM ~~do~~ remains unchanged even if the power is off and on again.

The startup of a computer consist of

turning the power on and starting the execution of boot-strap loader. initial program

Thus when the power is turned on, the hardware of the computer sets PC (program counter) to the first address of the boot strap loader. program

The boot strap loader loads the operating system from the secondary memory to the main memory and the control is then transferred to the O.S, which prepares the computer for general use.

RAM and ROM chips

RAM and ROM chips are available in a variety of sizes.

If the memory needed for the computer is larger than the capacity of one chip, it is necessary to compile a number of chips to form the required memory size.

Organisation of RAM

A RAM chip is suited for communication with the CPU. It has one or more control inputs. Another common feature is a bi-directional data bus that allows the transfer of data either from the memory to CPU during the read operation or from CPU to memory during a write operation.

A bidirectional bus can be constructed with

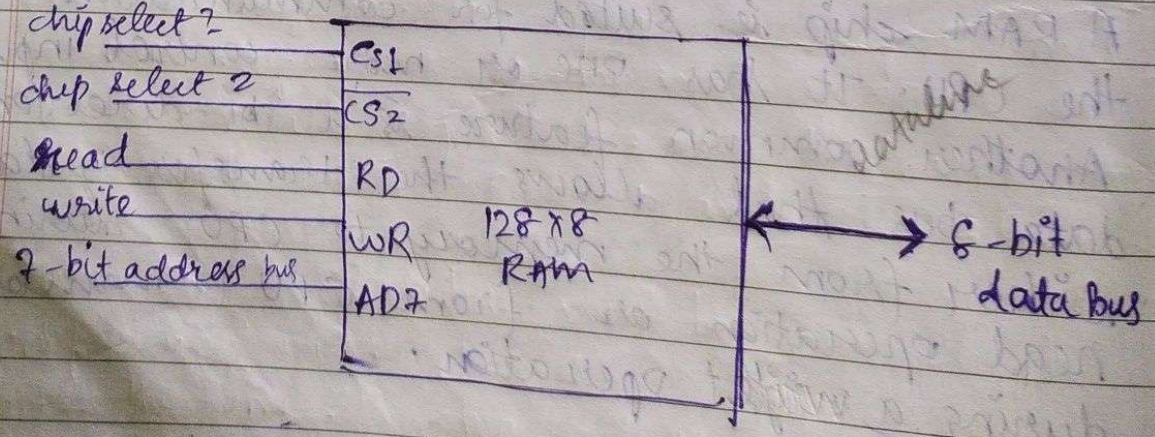
three state stage buffers - A three state stage buffer output can be placed in one of three possible states: A signal equivalent to logic 1, a signal equivalent to logic 0 or a high impedance state.

The logic 1 and 0 are normal digital signals. The high impedence state behaves like an open circuit, which means that the output does not carry a signal and has no logic significance.

The block diagram of a RAM chip is shown below:-

The capacity of the memory is 128 words of 8 bits (1 byte) per word. This require a 7 bit address and an 8 bit bi-directional data bus. The read and write inputs specifies the memory operation and the two chips select (CS) control inputs are for enabling the chip only when it is selected by the microprocessor.

The RD and WR signals control the memory operation as well as the bus buffers associated with the bidirectional data bus.



(one word has size 8 bits)

The availability of more than one control input to select the chip facilities the decoding of the address lines when multiple chips are used in the microcomputer.

The function table listed below specifies the operation of the RAM chip. The unit is in operation only when $CS1 = 1$ and $\overline{CS2} = 0$ the bar of the top of the second select variable indicates that this input is enabled when it is equal to zero.

If the chip select inputs are not enabled, or if they are enabled but the read or write inputs are not enabled, the memory is inhibited and its data bus is in a high impedence state. When $CS1 = 1$ and $\overline{CS2} = 0$, the memory can be placed in a write or read mode.

When the WR input is enabled, the memory stores a byte from the data bus into a location specified by the address input lines.

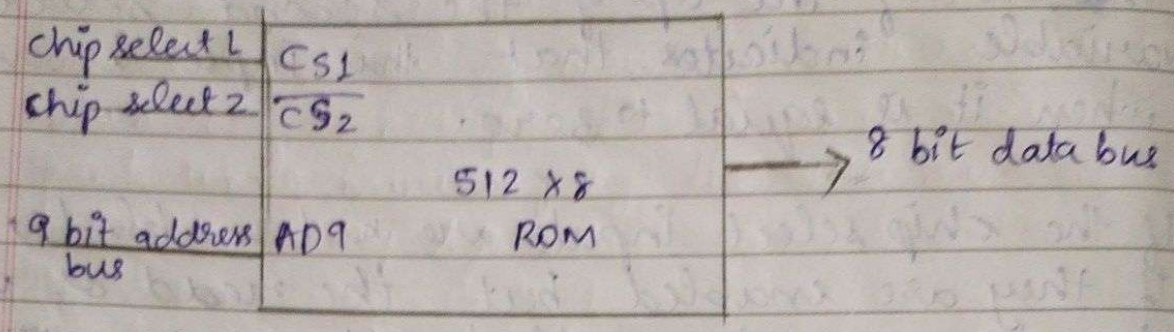
When the RD input is enabled, the content of the selected byte is placed into the data bus.

It is used in buses, when you want to transfer more than one signal through same wire without the loss of data

CS_1	$\overline{CS_2}$	RD	WR	Memory function	State of data bus
0	0	X	X	Inhibit	High - impedance
0	1	X	X	Inhibit	High - impedance
1	0	0	0	Inhibit	High - impedance
1	0	0	1	Write	Input data to RAM
1	0	1	X	Read	Output data from RAM
1	1	X	X	Inhibit	High - impedance

04/03/2020
Wednesday

ROM Organization



The block diagram of a ROM chip is shown above.
ROM is read only memory, ~~and~~ since the data bus can only be in output mode.
The figure shows a 512 bytes of ROM memory that means 512 words of 8 bit size can be stored in the above chip. There is two chip select signals CS1 and $\overline{CS2}$, these input lines help the unit to operate.

$2^9 = 512$

The ~~data~~ address bus is of 9 bit size which helps to locate the 512 different memory words. The data bus is of 8 bit size and is uni-directional because only reading operation is performed on ROM chip. There is no need of read or write control signals because the unit can only read when CS1 = 1 and $\overline{CS2} = 0$.

For this reason the diagram specifies 512 byte ROM while the RAM is only 256 bytes.

Otherwise the data bus will be in a high-impedence state.

(For the same-size chip, it is possible to have more bits of ROM than of RAM, because the internal binary cells in ROM occupy less space than in RAM)

Cache

- A special very high speed memory called a cache is sometimes used to increase the speed of processing by making current programs and data available to the CPU at a rapid rate.

- The cache memory is employed in computer systems to compensate for the speed differential between main memory access time and processor logic.

- 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.

- A technique used to compensate for the mismatch in operating speeds is to employ an extremely fast, small cache between the CPU and main memory whose access time is close to processor logic clock cycle time

- The cache is used for storing segments of programs currently being executed in the CPU and temporary data frequently needed in the present calculations

- By making programs and data available at a rapid rate, it is possible to increase the performance rate of the computer

- The auxiliary memory has a large storage capacity, is relatively inexpensive, but has low access speed compared to main memory.

- The overall goal of memory hierarchy is to obtain the highest-possible average ~~access~~ access speed while minimizing the total cost of the entire memory system.

- Auxiliary and cache memories are used for different purposes. The cache holds those parts of the program and data that are most heavily used, while the auxiliary memory holds those parts that are not presently used by the CPU.

- Moreover the CPU has direct access to both cache and main memory but not to auxiliary memory. The transfer from auxiliary to main memory is usually done by means of direct memory access of large blocks of data

- The typical access time ratio between cache and main memory is about 1 to 7. For example, a typical cache memory may have an access time of 100 ns, while main memory access time maybe 700ns

- Auxiliary memory average access time is usually 1000 times that of main memory. Block size in auxiliary memory typically

ranges from 256 to 2048 words, while cache block size is typically from 1 to 16 words.

- Many operating systems are designed to enable the CPU to process a number of independent programs concurrently.

- This concept, called multiprogramming refers to the existence of two or more programs in different parts of the memory hierarchy at the same time. In this way it is possible to keep all parts of the computer busy by working with several programs in sequence.

- The CPU initiates the I/O processor to start executing the transfer. This leaves the CPU free to execute another program. In a multiprogramming system, when one program is waiting for input or output transfer, there is another program ready to utilize the CPU.

- With multiprogramming the need arises for running partial programs, for varying the amount of main memory in use by a given program, and for moving programs around the memory hierarchy.

- Computer programs are sometimes too long to be accommodated in the total space available in main memory.

- Moreover, a computer system uses many programs and all the programs cannot reside in main memory at all times. A program with its data normally resides in auxiliary memory.

- When the program or a segment of the program is to be executed, it is transferred to main memory to be executed by the CPU. Thus one may think of auxiliary memory as containing the totality of information stored in a computer system.

- It is the task of the operating system to maintain in main memory a portion of this information that is currently active.

- The part of the computer system that supervises the flow of information between auxiliary memory and main memory is called the memory management system.

- RAM Originally, RAM was used to refer to a random-access memory, but now it is used to designate a ~~wr~~ read/write memory to distinguish it from a read-only memory, although ROM is also random access.

- RAM is used for storing the bulk of the programs and data that are subject to change.

ROM

- ROM is used for storing programs that are permanently resident in the computer and for tables of constants that do not change in value once the production of the computer is completed.

- Among other things, the ROM portion of main memory is needed for storing an initial program called bootstrap loader.

Difference 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.

- Auxiliary memory holds programs and data for future use, and because it is nonvolatile (like ROM), it is used to store inactive programs and to archive data.

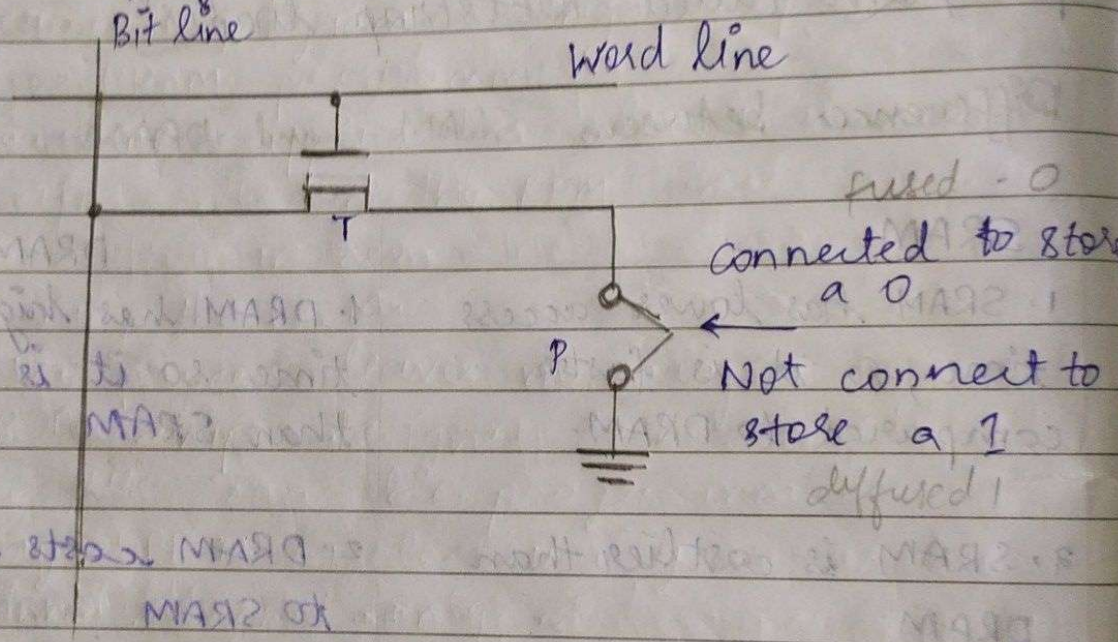


fig of ROM cell

Classifications of ROM :-

Programmable ROM - PROM

PROM or programmable ROM (programmable read-only memory) is a computer memory chip that can be programmed once after it is created.
Once the PROM is programmed, the information written is ~~perfo~~ permanent and cannot be erased or delated. PROM was first developed by Wen Tsing Chow in 1956.
An example of a PROM is a computer BIOS in early computers.

OR

PROM is read-only memory that can be modified once by a user.

PROM is first prepared as blank memory, and then it is programmed to store the information

To program the PROM, a PROM programmer or PROM burner is used. The process of programming the PROM is called as burning the PROM. Also, the data stored in it cannot be modified, so it is called as one-time programmable device.

When the PROM is created, all bits read as 1. During the programming, any bit needing to be changed to a "0" is etched or burned into the chip using a gang programmer.
If a PROM is programmed with an error or needs ~~up~~ to be updated, the chip is discarded

and a new PROM is created, replacing the old chip. A variation of the PROM is an EPROM, which is a PROM that can be erased and reprogrammed without being replaced. Today, PROM in computers has been replaced by EEPROM.

- PROM is a way of allowing a user to tailor a microcode program using a special machine called a PROM programmer.
This machine supplies an electrical current to specific cells in the ROM that effectively blows a fuse in them. The process is known as burning the PROM.
Since this process leaves no margin for error, most ROM chips designed to be modified by users use erasable programmable read-only memory (EPROM) or electrically erasable programmable read-only memory (EEPROM).

Uses:
They have several different applications, including cell phones, video game consoles, RFID tags, medical devices, and other electronics.

- Some ROM allow the data to be loaded by the user, thus providing a programmable ROM.
- Programmability is achieved by inserting a fuse at point P in the above ROM figure.

EPROM - Erasable Programmable ROM

EPROM stands for erasable programmable read only memory. It overcomes the disadvantage of PROM that once programmed, the fixed pattern is permanent and cannot be altered.

If a bit pattern has been established, the PROM becomes unusable, if the bit pattern has to be changed.

This problem has been overcome by the EPROM, as when the EPROM is placed under a special ultraviolet light for a length of time, the shortwave radiation makes the EPROM return to its initial state, which then can be programmed accordingly. Again for erasing the content, PROM programmer or PROM burner is used.

Uses:-

Before the advent of EEPROMs, some micro-controllers like some versions of intel 8048, the Freescale 68HC11 used EPROM to store their program.

IV. In EPROM we can erase the data and new data can be stored (or data is rewritten).

- Disadvantage is that if we have to erase the data we are compelled to disconnect the ROM cell from the mother board or unmount

- EEPROM has window like cells.

EEPROM - Electrically Erasable Programmable ROM

It stands for electrically erasable programmable read only memory. It is similar to EPROM, except that in this, the EEPROM is returned to its initial state by application of an electrical signal, in place of ultraviolet light. Thus, it provides the ease of erasing as this can be done, even if the memory is positioned in the computer.

It erases or writes one byte of data at a time.

Uses:-
It is used for storing the computer system BIOS

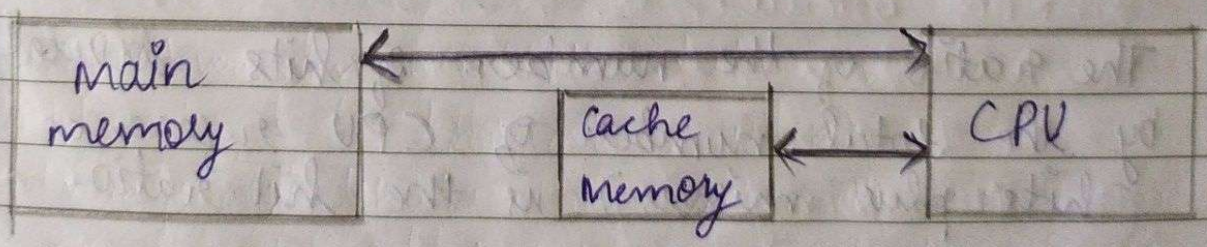
- We donot have to disconnect / unmount the ROM from motherboard
- EEPROM erases data using electrical charges and hence we donot need to unmount the ROM.
  Hence this overcomes the disadvantage of EPROM - ie the chip must be physically removed from the circuit for reprogramming in the case of EPROM

Cache Memory

Analysis of a large number of typical programs has shown that the references to memory at any given interval of time tend to be ~~continued~~ confined within a few localized areas in memory.
This phenomenon is known as the property of locality of reference.

eg:- When a program loop is executed, the CPU repeatedly refers to the set of instructions in memory that constitute the loop.

* [If the active portions of the program and data are placed in a fast small memory, the average memory access time can be reduced, thus reducing the total execution time of the program.
Such a fast small memory is referred to as a "cache memory". It is placed between the CPU and the main memory]



The fundamental idea of cache organization is that by keeping the most frequently accessed instructions and data in the

fast cache memory.

The basic operation of the cache is as follows :

When the CPU needs to access memory, the cache is examined. If the word is found in the cache it is read from the fast memory. If the word is not found in the cache, the main memory is accessed to read the word.

The performance of cache memory is frequently measured in terms of quantity called hit ratio.

When the CPU refers to memory and finds the word in the cache, it is said to produce a hit.

If the word is not found in cache, it is in main memory and it count as a miss.

The ratio of the number of hits divided by the total number of CPU reference (hits plus misses) is the hit ratio.

The transformation of data from main memory to cache memory is referred to as a mapping process.

There are three types of mapping procedures

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

Associative Mapping

The fast and most flexible cache organization uses an associative memory. The associative memory stores both the address and content of the word.

This permits any location in cache to store any word from main memory. The following diagram shows three words presently stored in the cache.

CPU address (15 bits)

↓

Argument Register

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

A CPU address of 15 bits is placed in the argument register and the associative memory is searched for a matching address.
If the address is found, the corresponding 12-bit data is read and sent to the CPU. If no match occurs, the main memory is accessed for the word.

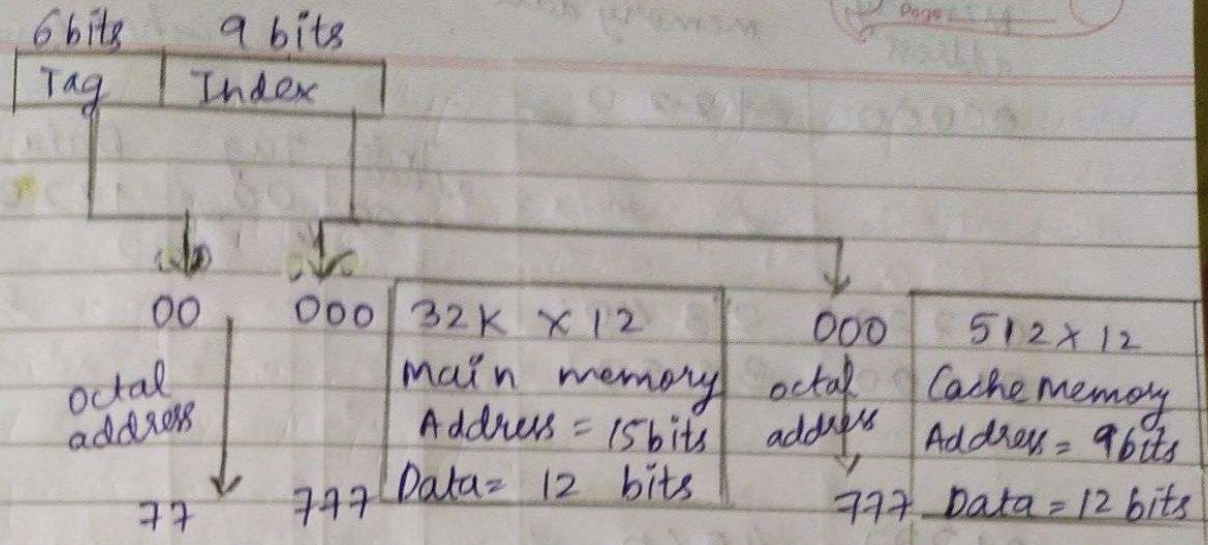
The address data pair is then transferred to the associative cache memory. If the cache is full, an address data pair must be displaced to make room for a pair that is needed and not presently in the cache.

The replacement is determined from replacement algorithm.

Direct Mapping

Associative memories are expensive compared to RAM because of the added logic associated with each cell.

The following figure shows addressing relationship between main memory and cache.



The CPU address of 15 bits is divided into two fields. The nine least significant bits (LSB's) constitute the index field and the remaining 6 bits form the tag field

Main memory needs an address that includes both the tag & the index bits.

The number of bits in the index field is equal to the number of address bits required to access the cache memory

In general case there are 2^k 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 n − k bits for the tag field.

Following figure illustrates direct mapping organization.

Main memory

Memory address	memory data
00000	1220
00777	2340
01000	3450
01777	4560
02000	5670
02777	6710

Cache Memory

Index address	Tag	Data
000	00	1220
777	02	6710

Cache memory

Each word in the cache consists of the data word and associated tag. When a new word is first brought into the cache the tag bits are stored alongside ~~wit~~ the data bits.

When the CPU generates a memory request, the index field is used for the address to access the cache.

The tag field of the CPU address is compared with the tag in the word read from the cache. If the two tags match, 'there is a hit' and the desired data word is in cache. If there is no match 'there is a miss' and the required word is read from main memory. It is then stored in the cache together with the new tag, replacing the previous value.

eg:- The word at address zero is presently stored in the cache (index = 000, tag = 00, data = 1220) Suppose that the CPU now ~~wo~~ wants to access the word at address 02000.

The index address is 000, so it is used to access the cache. The two tags are then compared. The cache tag is 00 but the address tag is 02, which does not produce a match. ~~Th~~

Therefore the main memory is accessed and the data word 5670 is transferred to the CPU. The cache word at index address 000 is then replaced with a tag of 02 and data of 5670.

Set-Associative Mapping

The disadvantage of direct mapping is that two words with the same index in their address but with different tag values cannot reside in cache memory at the same time.

A third type of cache organization, is set-associative mapping. In this mapping, each word of cache can store two or more words of memory under the same index address. 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.

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

When the CPU generates a memory request, the index value of the address is used to access the cache. The tag field of the CPU address is then compared with both tags in the cache to determine if a match occurs.

The comparison logic is done by an associative search of the tags in the set similar to an associative memory search, thus the name "set-associative".

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

The most common replacement algorithms are: random replacement, first-in-first-

out (FIFO) and least recently used (LRU). With the random replacement policy, the control chooses one tag-data item for replacement at random. The FIFO ~~proce~~ procedure selects for replacement the item that has been in the set the longest.

The LRU algorithm selects for replacement the item that has been least recently used by the CPU.

Writing into Cache

When the CPU finds a word in cache during a read operation, the main memory is not involved in the transfer. If the operation is a write, there are two ways that the system can proceed.

(i) Write through method

The simplest and most commonly used procedure is to update main memory with every memory write operation, with cache memory being updated in parallel, if it contains the word at the specified address.

This is called write through method. This method has the advantage that main memory always contain the same data

as the cache.

(ii) <u>Write-back method</u>

In this method, only the cache location is ~~updt~~ updated during a write operation.

The location is then marked by a flag so that later when the word is removed from the cache it is copied into main memory.