UNIT IV

5. THE MEMORY SYSTEM

5.1 Some Basic Concepts

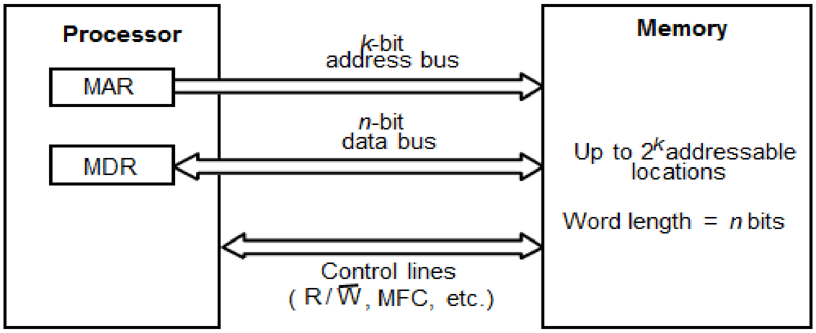
5.1.1 Size of the memory

Maximum size of the memory that can be used is dependent on the number of address lines that is present in the system. For example a computer with 16 bit address lines is capable of addressing 2^{16} locations which is equal to 64K locations. In general a computer with n bit address lines is capable of addressing 2^n locations of the memory.

5.1.2 Word Length

Number of bits present in a word is called as **word length.** Data from the memory system is accessed based on the word length. As an example a system with word length of 32 bit can access 32 bits of data on a single access.

5.1.3 Connection between memory and processor



One of the major functionality for the instruction execution is the communication between the memory unit and processor. Communication between memory unit and the processor is done through processor bus.

Processor bus consists of three different lines:

1. Address lines: Set of wires responsible for carrying the address information to be accessed from the memory unit.
2. Data lines: Set of wires that carry data information between the processor and the memory unit.
3. Control lines: Set of wires that carry the control signals from the processor required to control the memory unit.

When doing the communication, address is loaded every time into the register MAR and data to be sent or received from the memory unit is loaded into the register MDR. While doing the read operation, R/W signal is set to 1 and while doing the write operation, it is set to 0.

5.1.4 Memory access time

It is defined as the time between the initiation of the operation and the completion of the operation.

5.1.5 Memory Cycle Time

It is defined as the minimum time delay required between the initiation of two successive operations.

5.1.6 RAM (Random Access Memory)

It is the memory units were any location for read/write operation can be accessed within a fixed amount of time independent of the location address.

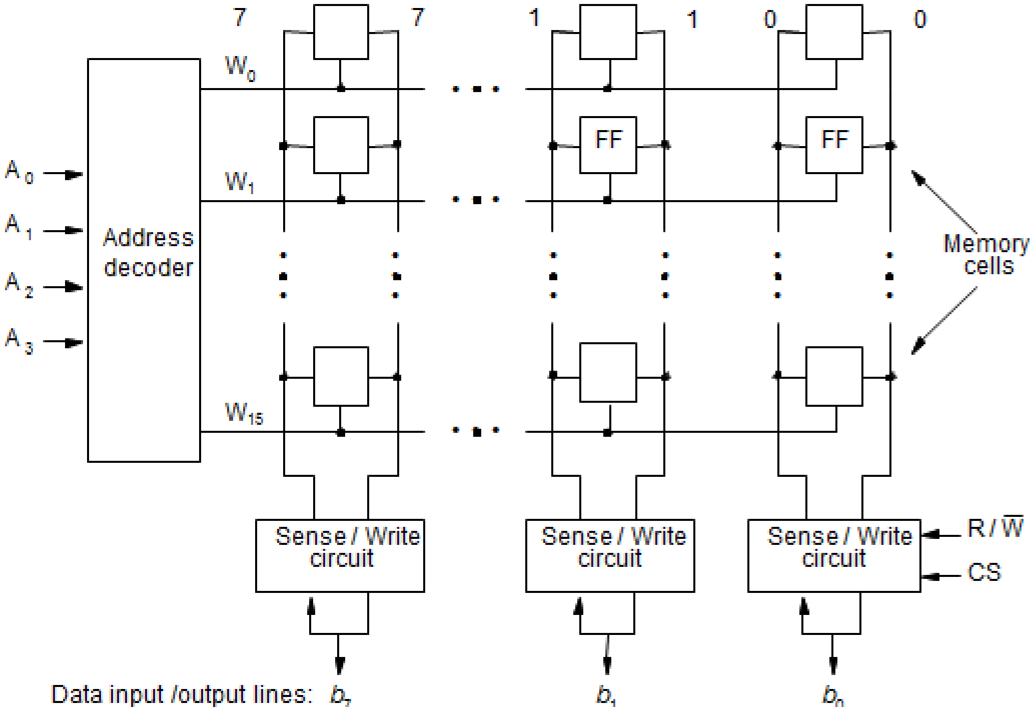
5.1.7 Cache Memory

It is the faster version of the RAM that is present on the processor chip.

5.2 Semiconductor RAM Memories

Semiconductor RAM memories are available in a very wide range of speeds. The cycle time for those memories range from 10ns to 100ns. First introduces around late 1960's and proved to be very expensive. Because of the invention of VLSI technology their cost dropped very dramatically.

5.2.1 Internal Organization of 16X8 Memory Chip



→ Memory cells of the semiconductor RAMs are organized in the form of a 2D array were each cell has the capacity of storing 1bit of data information.

→ Above shown figure is the organization of 16X8 memory cells were in there are 16 rows and 8 columns with each row having 8 memory cells arranged in a column structure. The organization can store 16 words with each word has the word length of 8 bits

→ When accessed, the organization accesses one complete row with all 8 cells arranged across it.

→ All the cells arranged across the row are connected to one common line called as **word line** which activates all the bits of the row when a particular row is accessed at a given time. For the circuit shown above, there are 16 word lines.

→ Word lines are controlled by address decoder which selects the required word from the address ranging between 0 to 15. First word line is given the address 0 and last word is given the address 15. To select required word, there is a 4 bit address input lines which ranges from 0000 to 1111. According to the input value, one out of 15 lines is activated at a given time.

→ Cells arranged across the column are again connected to a sense/write circuit in common by two bit lines. There are 8 columns and altogether there 8 sense/write circuits connected in common.

→ Each sense/write circuit has one data input/output line connected across them which is a bidirectional line connected to a data bus and has one common control line R/W to select read or write operation and one common line CS which selects a particular chip.

→ Read operation: To do a read operation, a particular word line is activated by the 4 bit address and R/W is set to 1. All the cells in a specified row is activated and data from the row is sensed from each of the cell.

→ Write operation: To do a write operation, a particular word line is activated by the 4 bit address and R/W is set to 0 and the required data value is loaded on the data lines. All the cells in a specified row is activated and data from the data line is written into the cell.

→ Thus the above shown circuit has:

4 address lines

8 data lines

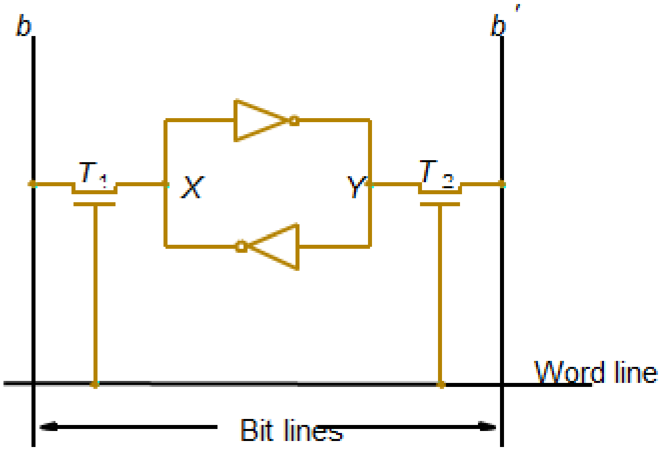
2 control lines

5.2.2 Static Memories

A memory that has the circuits which are capable of retaining their state as long as power is applied is called as static memories.

Even if there is a small interruption to continues power, it will wipe out all the data. That is why it is called as **volatile memory**. Type of cells used in this are called as static RAM cells

Organization of SRAM cell



→ To implement each RAM cell, two transistor inverters are cross connected to implement a basic flip-flop (latch).

→ This latch is connected to transistors T1 and T2 which acts as switch and these transistor lines are in turn connected to word line and a bit line b and $b^{|}$. Bit line b and $b^{|}$ are connected to each of the sense/write circuit.When word line is at ground level, the transistors are turned off and the latch retains its state.

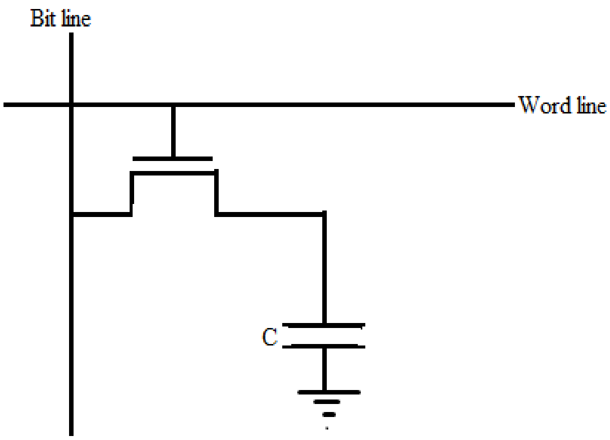
→**Read operation:** Read operation is performed by activating a particular word line to close the switches T1 and T2.Sense/Write circuits at the bottom monitor the state of b and $b^{|}$.

→**Write Operation:** It is performed by placing the required data on the line b and its compliment on line $b^{|}$ and particular word line is activated. Data from the sense/write circuit is stored inside the cell.

5.2.3 DRAMS (Dynamic RAM)

Static RAMs are fast but the major disadvantage of them is that they come at a very high cost because of too much involvement of the transistors in a single cell. It can be made to implement in a more simple way using another type of cells called as **Dynamic RAM** cells.

The cells are called as dynamic RAMs because they do not have the capacity to retain the state indefinitely for a long period of time when compared to static RAMs. Below figure shows the organization of DRAM cell.



Here the capacitor is used store the data in the form of a charge. The capacitor is connected to transistor and the transistor is connected to word line and the bit line. When word line is activated the switch is closed and required operation is done.

Read operation: Read operation is performed by activating a particular word line to close the switch T.Sense/Write circuits at the bottom senses the charge of the capacitor.

Write Operation: It is performed by placing the required data on the bit line and particular word line is activated. Data from the sense/write circuit is stored inside the cell in terms of charge.

In the DRAM cells, the capacitor gradually loses the charge with respect to time when switch is opened. So to retain the state, it is very much necessary to restore the charge periodically. The method of restoring the charge periodically in the DRAM cell is called as **Refresh Process.**

Refresh Process: After T is turned off, capacitor starts to discharge the charge. So the required cell is read time to time by closing T and applying the required voltage to retain the bit value. When closed, sense amplifier senses if the voltage of the cell is above or below the thresh hold voltage value. If above applies the charge and sets 1. If below pulls down to the ground voltage.

There are two types of DRAMs:

1. Asynchronous DRAM (DRAM)
2. Synchronous DRAM (SDRAM)

We will discuss each one of them in detail in the next sections.

5.2.4 Asynchronous DRAMs

One of the types of DRAM chips were the timing signals of the RAM circuit are controlled asynchronously. I.e timing signals are not in synchronization with processor clock and the memory chip derives the timing signals and control through a special circuit called **memory controller.**

To ensure the data attainment in the DRAM chip, there should be refresh process taking place periodically and this process is carried out by a dedicated circuit called **Refresh circuit**.

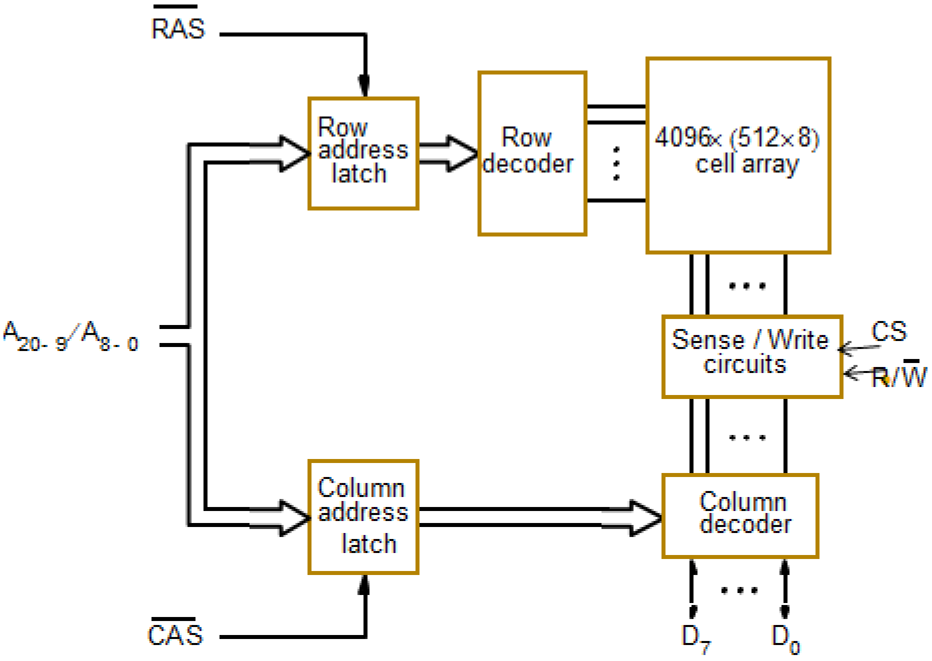
We will look into the detailed organization of asynchronous DRAM with a 16 MB DRAM. A 16MB DRAM is arranged as 4KX4K cells were in there are 4096 rows and each row has 4096 cells arranged in column wise fashion.

4096 cells arranged in each row are again divided into group of 8 . Thus 4096 cells are divided into 512 groups with each group is having 8 cells. Arrangement of the cell is shown below:

4KX4K = 4096X4096/8 = 4096X512X8

According to the above equation, it means there are 4096 rows and there are 512 columns and each column has 8 cells in it which forms a byte. When accessed, one out of 512 columns is accessed.

Therefore this configuration is called as 16MB DRAM chip configured as 2MX8 which means there are 2M different locations arranged in row and column wise structure. Circuit arrangement for asynchronous DRAM is shown below:



In the above circuit there are 4096 rows and in each row there are 512 groups having 8 bits each, which means each row has altogether 512 bytes of data. Therefore to access a data, it requires both row address and column address information. There are 4096 rows and to access row it requires 12 bits of information and there are 512 columns and to access a column it requires 9 bits of information. Altogether to access a particular byte in a column, it requires 21 bits of address information. The higher order 12 bits (A20-A9) are row address and low order 9(A8-A0) bits are column address. For a single access, it retrieves 8 bits of information. So there must be 8 data lines to access data to or from any location.

Following are the different units of the circuit and their functionality:

1. Row Address latch: This latch is used to strobe the row address to be accessed. The latch is controlled via signal RAS (Row Address Strobe) which is active low and when activated sends the row address to be accessed.

2. Column Address latch:This latch is used to strobe the column address to be accessed. The latch is controlled via signal CAS (Column Address Strobe) which is active low and when activated sends the column address to be accessed.

3. Row decoder: This unit receives the output of Row Address latch as its input and takes the responsibility of selecting one row out of 4096 rows.

4. Column decoder: Row decoder: This unit receives the output of Column Address latch as its input and takes the responsibility of selecting one column out of 512 columns. It also has bidirectional data lines which retrieves or sends the data information into a particular column.

5. Sense/Write Circuits: These are the circuits connected to columns of the DRAM chip in common. There are altogether 512X8 sense/write circuits which take the responsibility of reading/ writing operation.

Following are the important operations performed by the asynchronous DRAM chip:

1. Address multiplexing: One of the important features implemented in this DRAM chip. The technique of passing two different address information over the same line is called as **Address multiplexing.**

For example: We know that here 21 bits of address information are required to access any of the location. But at the same it is even known that Row address and Column address are accessed separately one after the other. So to reduce the number of pins row and column addressare multiplexed into same pins and the number of pins is chosen to be the highest valued information. In this case row address is 12 bits and column address is 9 bits. So column address is highest and number of pins is 12.

2. Read/Write Operation:

→During Read/Write operation, row address is applied first and is loaded onto the row address latch with signal $\overline{\text{RAS}}$. A complete row is selected

→ Next shorty after this, column address is loaded into column address latch with the signal $\overline{CAS}$.

→ Row address and column address are decoded and appropriate Read/Write signal is activated in the sense/Write ckt. If it is read operation data flows from the cell into the line and if write, it flows from the line into the cell.

3. Refresh Process: Required cells are read time to time to ensure the retaining of the data. Each and every row is read time to time and charge on them is ensured. This functionality in the circuit is controlled by a special circuit called memory controller circuit.

4. Fast Page mode: Another important feature that can be added to the DRAM chip. Fast page mode is a feature of having the block transfer capability. Fast page mode is nothing but the continuous transfer of data from the contiguous memory locations of the chip. Since this feature includes transfer of data from more than one location, it is often called as block transfer operation.

In this case, when the row address is applied all the cells a particular row gets activated. Then only the column address is generated to access the successive locations keeping the row address active till it completes the block transfer. This is implemented automatically by memory controller by keeping the row address active and only changing the column address.

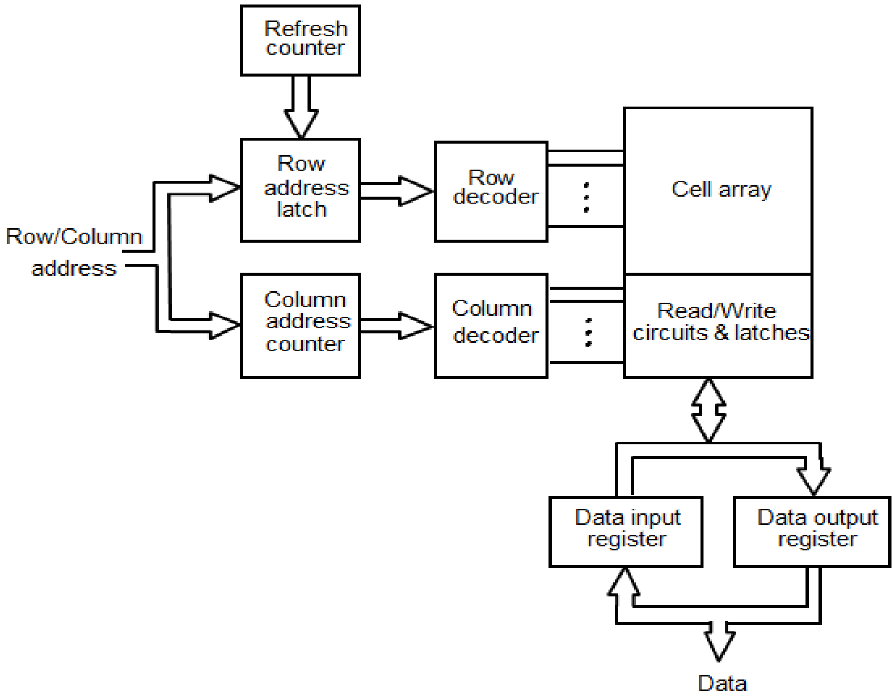
5.2.5 Synchronous DRAM (SDRAM)

In this type of DRAM chip, operations are directly controlled by processor. I.e timing signals are in synchronization with processor clock. When it comes to the working procedure, it is very similar to the asynchronous DRAM chip. But has few changes made to the implementation of the memory chip.

To ensure the data attainment in the DRAM chip, there should be refresh process taking place periodically and this process is carried out by a dedicated circuit called **Refresh circuit**.

To ensure the data attainment in the SDRAM chip, there should be refresh process taking place periodically and to carry out this process, a circuit has a built in component called **Refresh counter.** This circuit automatically refreshes the cells of RAM periodically.

Another important change is that, this circuit doesn’t need a memory controller and processor itself controls all the operations.



Following are the different units of the circuit and their functionality:

1. Row Address latch: This latch is used to strobe the row address to be accessed.

2. Column Address counter: This component is one of the new feature added and latch is used to strobe the column address and it also has the capability of implementing block transfer were in it automatically increments the column address. Hence the unit is called as column address counter.

3. Row decoder: This unit receives the output of Row Address latch as its input and takes the responsibility of selecting one row out of m rows.

4. Column decoder: Row decoder: This unit receives the output of Column Address latch as its input and takes the responsibility of selecting one column out of n columns. It also has bidirectional data lines which retrieves or sends the data information into a particular column.

5. Sense/Write Circuits: These are the circuits connected to columns of the DRAM chip in common.

6. Data input register: This register is used to hold the data value that needs to be written in to the memory location.

7. Data output register: This register is used to hold the data value that is read from the memory location.

Following are the important operations performed by the asynchronous SDRAM chip:

1. Address multiplexing: One of the important features implemented in this DRAM chip. The technique of passing two different address information over the same line is called as **Address multiplexing.**

For example: If 21 bits of address information are required to access any of the location. But at the same it is even known that Row address and Column address are accessed separately one after the other. So to reduce the number of pins row and column address are multiplexed into same pins and the number of pins is chosen to be the highest valued information. In this case row address is 12 bits and column address is 9 bits. So column address is highest and number of pins is 12.

2. Read/Write Operation:

→ During Read/Write operation, row address is applied first and is loaded onto the row address latch with signal $\overline{RAS}$. A complete row is selected

→ Next shorty after this, column address is loaded into column address latch with the signal $\overline{CAS}$.

→ Row address and column address are decoded and appropriate Read/Write signal is activated in the sense/Write ckt. If it is read operation data flows from the cell into the line and if write, it flows from the line into the cell.

3. Refresh Process: Required cells are read time to time to ensure the retaining of the data. Each and every row is read time to time and charge on them is ensured. This functionality in the circuit is controlled automatically by a refresh counter circuit.

4. Fast Page mode: Another important feature that can be added to the DRAM chip. Fast page mode is a feature of having the block transfer capability. Fast page mode is nothing but the continuous transfer of data from the contiguous memory locations of the chip. Since this feature includes transfer of data from more than one location, it is often called as block transfer operation.

In this case, when the row address is applied all the cells a particular row gets activated. Then only the column address is generated to access the successive locations keeping the row address active till it completes the block transfer.

5.3 Structure of larger memories

Here we will see how a larger memory system is built using smaller memory chips discussed. Larger memory element constructed using smaller memory chips is called as **memory module.**

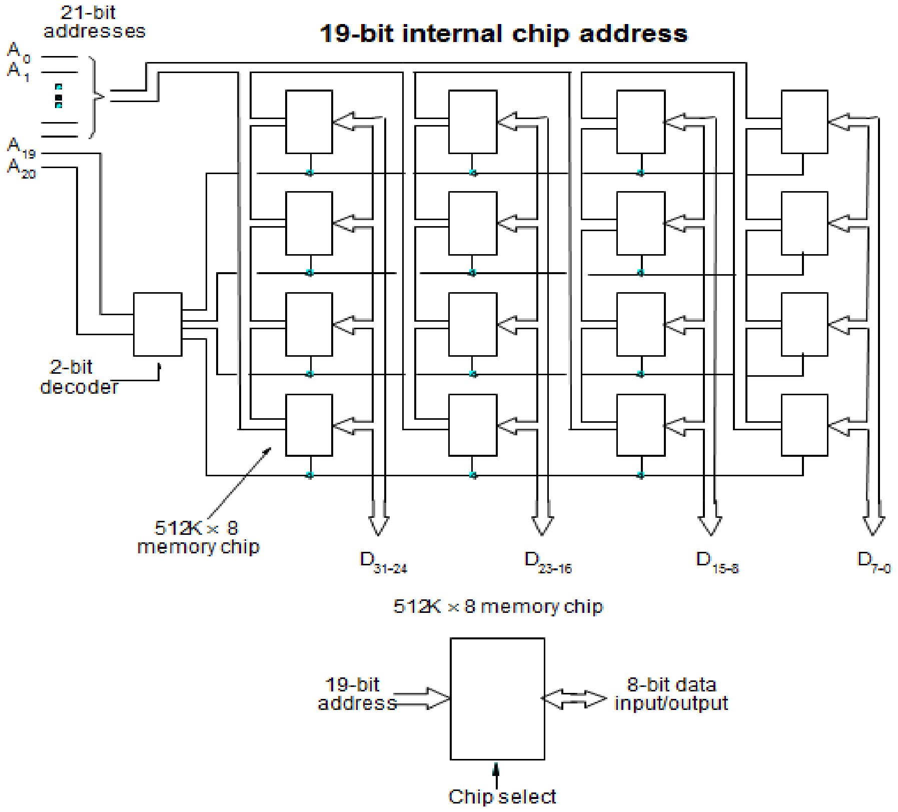
Organization of a 2MX32 memory module using 512K X 8 static memory chips

To know the implementation details of this we consider the memory configuration mentioned above. Here in this we have 512K X 8 memory chips available using which we have to design a 2M X 32 memory module. So in the first step we have to determine how many chips are required in order to design a module. This is done in a very simple way as shown below:

Divide the memory module value with a chip value separately:

$$\frac{\text{memory module}}{\text{chip size}} = \frac{2M \text{ X } 32}{512K \text{ X } 8} = \frac{(2 \text{ X } \overset{2}{\cancel{1024 \text{ X}}} \; \cancel{1024)} \text{ X } \overset{4}{\cancel{(32)}}}{(\cancel{512 \text{ X}} \; \cancel{1024)} \text{ X } \cancel{8}} = 4 \text{ X } 4$$

This suggests that there should be 16 512K X 8 memory chips arrange in 4 different rows and each row having 4 chips arranged in column wise to design 2M X 32 memory module. The typical arrangement of this configuration is designed as shown below:



We know that memory consists of 2M words and each word access 32 bits data. So to access 2M different words, it needs 21 bits of address lines and to access 32 bits data 32 data lines are required.

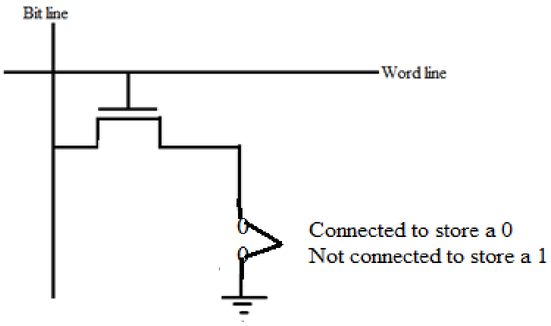
But when u consider single chip it has only 512K words and access 8 bits of data. So the higher order 2 bits is used to select the row of chip connected to chip select CS show above and remaining to select word from each chip. When a row is accessed it access 8 bit from each chip and altogether access 32 bits of data in total.

(Solve the problems discussed in the class)

5.4 Read Only Memories (ROM)

We know that both SRAM and SDRAM chips are volatile which means that it loses the contents when power is switched off. But some of the applications need to retain data even if the power is switched off. For example, computer is turned on, the operating system must be loaded from the disk into the memory. Hence we need a non-volatile memory system to store these kind of applications and one such memory system used in computer is ROM. Non-volatile memory is read in the same manner as volatile memory. Separate writing process is needed to place information in this memory.

The diagram shown below is the hardware configuration of ROM cell.



Here in this is case data are written when it is manufactured. Were ever the data bit value 1 is required, the wire is burnt to disconnect the connection.

Types of ROM

5.4.1 Programmable Read-Only Memory (PROM)

A variation of ROM which allows the programmer to store the data. Before programming ROM chip consists of all 0’s and wherever the data 1 is required, it is stored by burning the fuse using the PROM programmer to disconnect the connection.

5.4.2 Erasable PROM (EPROM)

This type of ROM allows stored data to be erased and new data to be loaded. Hence it is also called as reprogrammable ROM. Erasure process is carried out by exposing the ROM to UV light.

The main disadvantage of this type of ROM is that the ROM chip should be physically removed in order to erase the contents.

5.4.3 Electrically Erasable PROM (EEPROM)

In this method the contents of ROM is erased using the electric voltage and the ROM chip need not be removed. The only disadvantage of this method is that, it requires different voltages for erasing, writing and reading the data.

5.4.4 Flash memory

Has similar approach to EEPROM. Only difference in this technology is that, it reads the contents of a single cell, but writes the contents of an entire block of cells. Flash devices have greater density. Higher capacity and low storage cost per bit.

Main feature of this technology is that power consumption of flash memory is very lowmaking it attractive for use in equipment that is battery-driven. Single flash chips are not sufficiently large, so larger memory modules are implemented using flash cards and flash drives.

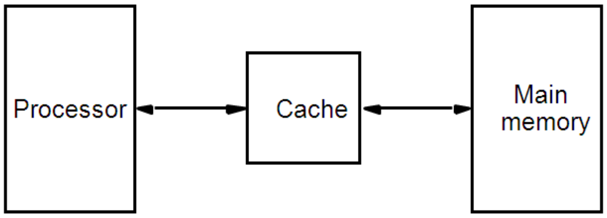
Flash cards basically act as removable memory devices and flash drives acts a mounted memory devices in the battery operated devices.

5.5 Cache Memories

Fastest RAM technology is called as cache memory.

We know that speed of main memory is very less when compared to the speed of the processor. Hence addressing the main memory unit for fetching the data takes lot of time. So to reduce the access time and to increase the performance, cache memory is used which makes main memory appear to the processor closer and faster.

Typical memory unit arrangement with cache is shown below:



Cache memory is very small compared to main memory since its cost is high. Basically with the implementation of cache memory data that is required for the processor are first bought into cache from the main memory before executing.

Working principle of the cache memory:
Cache memory works with a property called as **locality of reference.**
Locality of reference: It is a principle which states that many instructions in the localized area of program are executed repeatedly while remaining are executed infrequently.
It supported by two other aspects:

1. Temporal locality of reference
2. Spatial locality of reference

Temporal aspect of locality of reference means that recently executed instruction is likely to be executed very soon. Cache exhibits this property by keeping the fetched data even after using it

Spatial aspect of locality of reference means that instructions in the close proximity of recently executed instruction are likely to be executed very soon. This property is exhibited by cache by fetching contiguous instructions rather than fetching a single word. Implementing this type transfer is called as **block transfer.**

Some important definition:
1. Mapping function: It is function that decides the placement of memory block inside the cache block.

2. Replacement algorithm: When the cache is full, and a block of words needs to be transferred from the main memory, some block of words in the cache must be replaced. This is determined by a “replacement algorithm”.

3. Cache hit/ cache miss: If the requested information is present inside the cache, it is called cache hit or else it is called cache miss.
If the requested information is present inside the cache and if the operation is read, it is called read hit or else read miss.
If the requested information is present inside the cache and if the operation is write, it is called write hit or else write miss.

4. Write through protocol: one that is in reference to write operation. In this technique, when location of cache is updated, simultaneously location of the main memory is also updated.

5. Write back protocol: In this technique, when location of cache is updated, location of the main memory is updated when the block needs to be replaced back. Here when the contents of cache is updated, it sets **dirty bit or modified bit** to 1. If the block replaced has this bit set to 1, then the contents in the main memory are updated.

6. Load through protocol: If there is a read miss and instead of transferring the block of data into the main memory and then to the processor, the data is first loaded in to the processor and then transferred into the main memory for later usage. This concept is called **Load through protocol.**

5.5.1 Mapping Functions

It is function that decides the placement of memory block inside the cache block. There are three mapping functions:

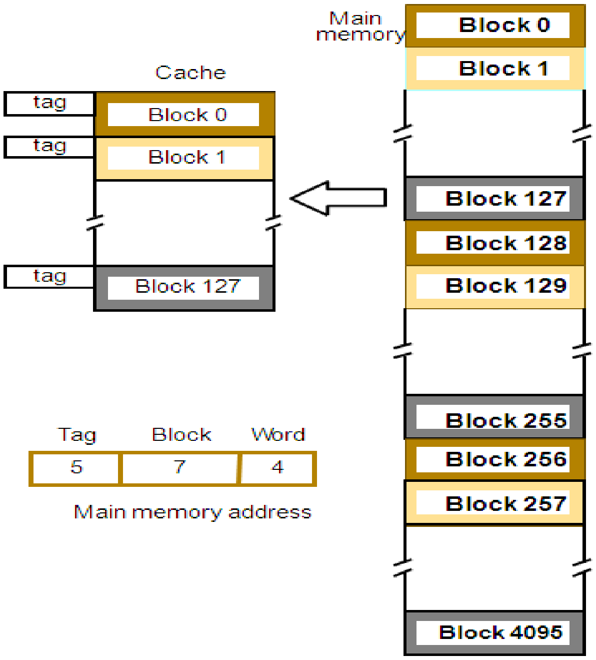
1. Direct mapping
2. Associative mapping
3. Set-associative mapping.

To discuss all these mapping functions technique, we will consider a common processor configuration problem:

A system has a Cache consisting of 128 blocks of 16 words each. So total size of the cache is 2048 (2K) words. Assume that Main memory is addressable by a 16-bit address. So main memory has 64K words which mean that it has 4K blocks of 16 words each.

5.5.1.1 Direct Mapping

- This mapping technique works according to the fixed formula.
- Placement of main memory block into the cache block is fixed all time and is predetermined.
- Here block j of main memory maps to the block j modulo number of block of the cache, which is given by a formula:
- **j % number of cache blocks**
- In this example, it is: **j % 128**
- **For example:**
  Block 0 maps to 0 % 128 = 0
  Block 1 maps to 1 % 128 = 1
  Block 32 maps to 32 % 128 = 32
  Block 129 maps to 129 % 128 = 1
- Altogether 4096/128 = 32 different blocks are mapped into a particular cache block.
- Main disadvantage is that it leads to a contention problem: Even if other locations are full, it will not be acquired as it works according to the formula.
- Following is the arrangement of direct mapping technique :

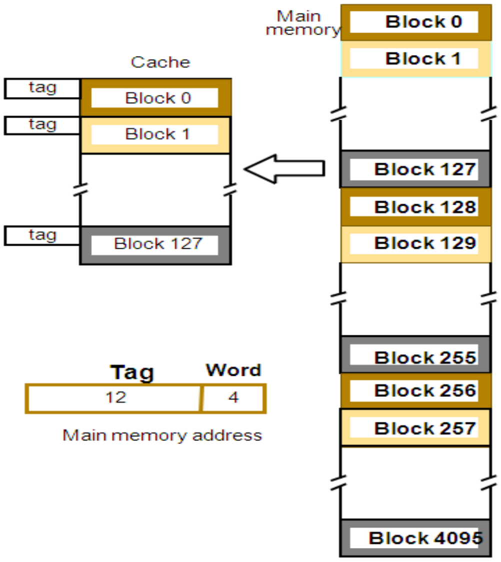


Memory address is divided into three fields:

- ✓ Low order 4 bits determine one of the 16 words in a block. When a new block is brought into the cache,
- ✓ The next 7 bits determine which cache block this new block is placed in.
- ✓ High order 5 bits determine which of the possible32 blocks is currently present in the cache. These are tag bits.

5.5.1.2 Associative mapping

- ➢ To rule out the contention problem of the direct mapping, this technique is used. Here any main memory block can be placed anywhere in the cache block which is free.
- ➢ When the new block needs to bought into cache, it does linear search for the free block and inserts the new block into that area.
- ➢ When cache is full, requires replacement algorithm to bring the new block into the cache.
- ➢ Advantage is that, flexible, and uses cache space efficiently.
- ➢ Only disadvantage is that, it Cost is higher than direct-mapped cache because of the need to search all 128 patterns to determine whether a given block is in the cache.
- ➢ Following is the arrangement of associative mapping technique :

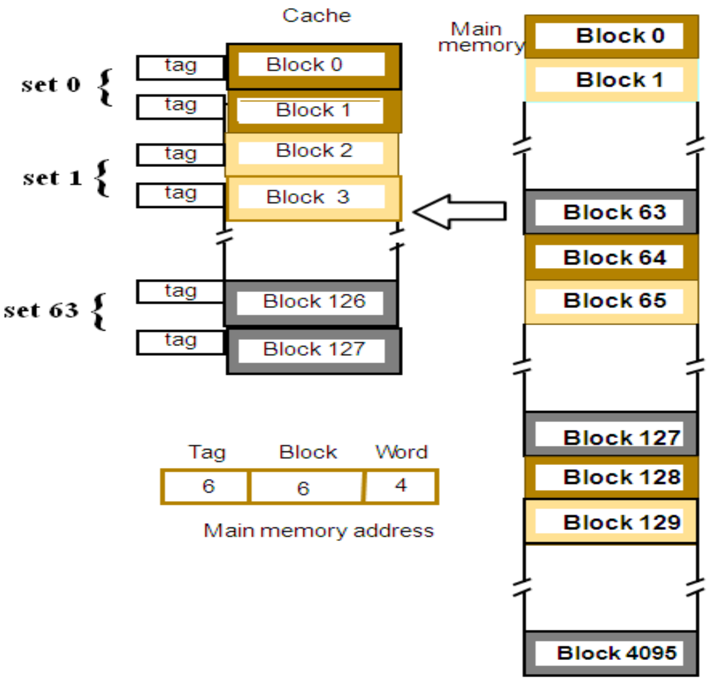


Memory address is divided into two fields:

- ✓ Low order 4 bits identify the word within a block.
- ✓ High order 12 bits or tag bits identify a memory block when it is resident in the cache.

5.5.1.3 Set Associative mapping

- ➢ This technique is a combination of direct mapping and the associative mapping which rules out both contention problem and cost of searching.
- ➢ Here in this technique, blocks are again grouped into entities called sets.
- ➢ Here block j of main memory maps to the block j modulo number of sets of the cache, which is given by a formula:
- ➢ **j % number of cache sets**
- ➢ Within the set, it can be placed anywhere according to the associative mapping technique.
- ➢ Here we will divide the cache into 64 sets, with two blocks per set. I.e.
- ➢ **j % 64**
- ➢ **For example:**
  Block 0 maps to 0 % 164 = 0 set
  Block 1 maps to 1 % 64 = 1 set
  Block 32 maps to 32 % 64 = 32 set
  Block 129 maps to 129 % 64 = 1 set
- ➢ Altogether 4096/64 = 64 different blocks are mapped into a particular cache set.
- ➢ Following is the arrangement of set associative mapping technique:



Memory address is divided into three fields:

- ✓ Low order 4 bits determine one of the 16 words in a block. When a new block is brought into the cache,
- ✓ The next 6 bits determine which cache set this new block is placed in.
- ✓ High order 6 bits determine which of the possible 64 blocks is currently present in the cache. These are tag bits.

5.5.2 Cache Replacement Algorithm

When the cache is full, and a block of words needs to be transferred from the main memory, some block of words in the cache must be replaced. This is determined by a “replacement algorithm”.

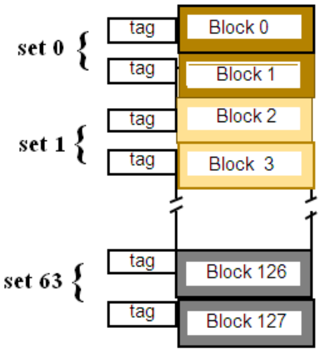
There are three different replacement algorithms:

1. Random technique
2. First In First Out (FIFO) Technique
3. Least Recently Used (LRU) Technique.

Random technique: As the name indicates, in this technique replacement is done randomly based on some random number generation.

First In First Out (FIFO) Technique: In this technique the oldest cache block entered is the one to be replaced first which works according to the circular queue technique.

Least Recently Used (LRU) Technique: Cache block that is not referred for a very long time is removed. To know how it works let's take the example of typical cache block with 64 sets were each set has 2 blocks which uses associative mapping technique.



Here it maintains two bit information called counter for the blocks in each set to determine the least recently used block. The technique works according to the following steps:

1. When a cache hit occurs, set the counter of referred block to 0 and all other blocks counter which was originally less than the referred block counter value is incremented by 1.
2. When miss occurs and the set is not full, counter associated with new block is set to 0 and all other blocks counter is incremented by 1.
3. When a miss occurs and the set is full, the block with counter value 3 is removed and the new block value is set to 0 and all other blocks within the set is incremented by 1.

5.6 Virtual memories

What are virtual memories?

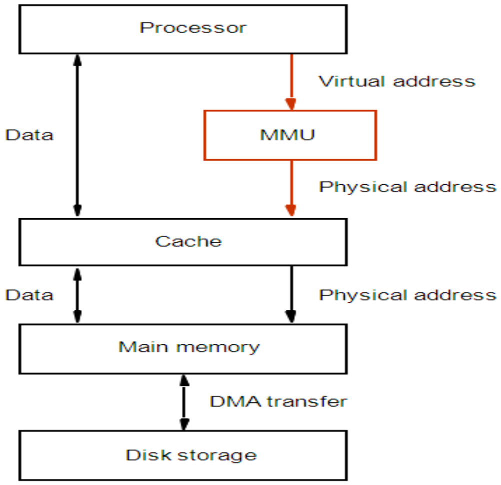
We know that if there is bit address space, it can only address 2^n locations. That means it can address the program of 2^n words only and if there are n bit address space, physical memory may not be 2^n. But even if there are no 2^n locations processor will address all those non existing locations.

For example, in 32 bit machine it addresses up to 4G programs but the RAM it may have will be 1G. So when a complete program doesn't fit inside main memory, parts of it are bought into main memory when required and remaining are kept inside the hard disk. This technique of moving the programs and data automatically between hard disk and main memory is called as **virtual memory technique.**

When a new piece of a program is to be transferred to the main memory, and the main memory is full, then some other piece in the main memory must be replaced.

Programs and processors reference an instruction or data independent of the size of the main memory according to addressing capability which may or may not exist and this address is called as logical address (virtual address).

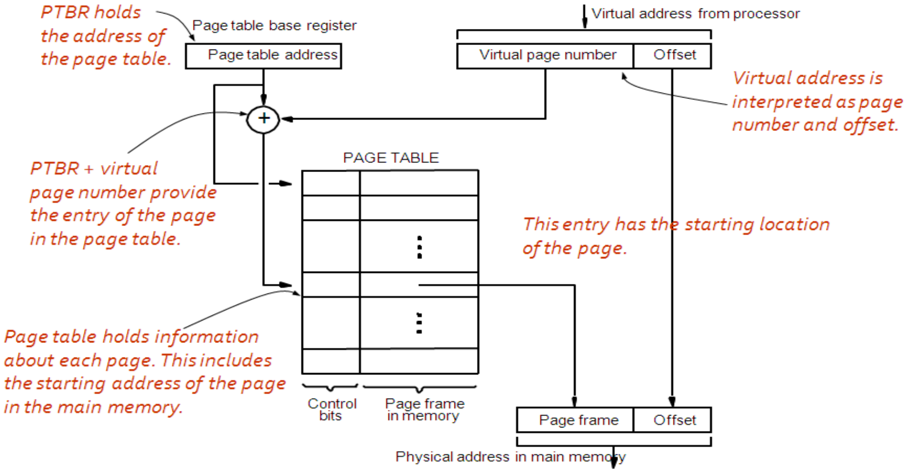
If the address generated by the processor is existing, then it will be accessed accordingly or else it needs to be converted into physical address by a hardware unit called **Main Memory Management Unit (MMU).** Translation of logical address into physical address is always done by MMU. Following is the memory configuration with virtual memory technique.



5.6.1 Address Translation

Data stored in the hard disk is assumed to be composed of fixed number of units called **pages (similar to the blocks in cache).** Each page is again a collection of words called. In virtual memory transfer between main memory and hard disk is done in terms of **page.**

Hardware configuration for the implementation of the address translation is shown below:

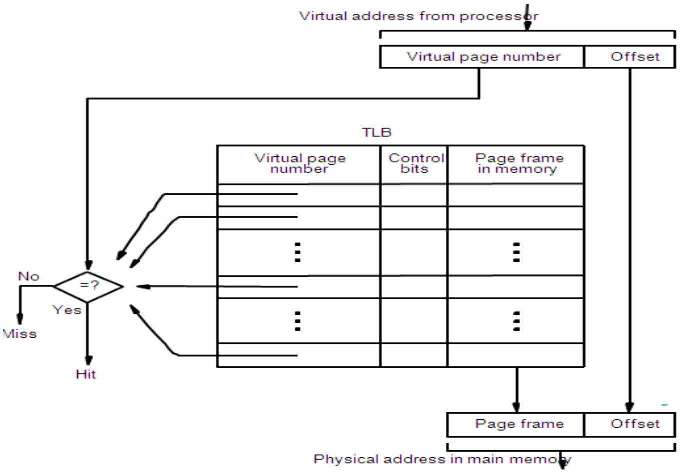

PTBR holds the address of the page table.
Page table base register
Page table address
Virtual address from processor
Virtual page number
Offset
Virtual address is interpreted as page number and offset.
+
PTBR + virtual page number provide the entry of the page in the page table.
PAGE TABLE
This entry has the starting location of the page.
Page table holds information about each page. This includes the starting address of the page in the main memory.
Control bits
Page frame in memory
Page frame
Offset
Physical address in main memory

- ➢ Each virtual or logical address generated by a processor is interpreted as a virtual page number (high-order bits) plus an offset (low-order bits) that specifies the location of a particular byte within that page.
- ➢ Information about the main memory location of each page is kept in the page table. It has following information:
  1. Main memory address were the page is stored
  2. Current status of the page (valid, modified etc)
- ➢ Area of the main memory that can hold a page is called as **page frame** and starting address of the page table in the memory location is kept in a special register called **page table base register.**
- ➢ Page frames of the program are kept in a contiguous manner in a page table were the first page information starts from the top of the table.
- ➢ Virtual page number generated by the processor is added to the contents of the page table base register. This provides the address of the corresponding entry in the page table. The contents of this location in the page table give the starting address of the page if the page is currently in the main memory.
- ➢ Page table entry for a page also includes some control bits which describe the status of the page while it is in the main memory.

5.6.3 Associative mapped Translation Look aside Buffer (TLB)

Page table in virtual memory is used for every read and write operation. Hence it takes good amount of time to access the page table resided inside the main memory. So there has to be some means were in the page table has to be accessed fast. As page table is quiet large, it is impossible to reside the complete page table inside MMU which is a part of processor chip. So a copy of a small portion of the page table can be accommodated within the MMU which is called as **TLB.**

This TLB includes the information of few pages that are accessed very recently. Implementation of TLB is shown below:



High-order bits of the virtual address generated by the processor select the virtual page. These bits are compared to the virtual page numbers in the TLB. If there is a match, a hit occurs and the corresponding address of the page frame is read. If there is no match, a miss occurs and the page table within the main memory must be consulted.