MODULE – 3 THE MEMORY SYSTEM

5.1 BASIC CONCEPTS:

The maximum size of the Main Memory (MM) that can be used in any computer is determined by its addressing scheme. For example, a 16-bit computer that generates 16-bit addresses is capable of addressing upto 2^{16} =64K memory locations. If a machine generates 32-bit addresses, it can access upto 2^{32} = 4G memory locations. This number represents the size of address space of the computer.

If the smallest addressable unit of information is a memory word, the machine is called word-addressable. If individual memory bytes are assigned distinct addresses, the computer is called byte-addressable. Most of the commercial machines are byte-addressable. For example in a byte-addressable 32-bit computer, each memory word contains 4 bytes. A possible word-address assignment would be:

Word Address	**Byte Address**			
0	0	1	2	3
4	4	5	6	7
8	8	9	10	11
.		…..		
.		…..		
.		…..		

With the above structure a READ or WRITE may involve an entire memory word or it may involve only a byte. In the case of byte read, other bytes can also be read but ignored by the CPU. However, during a write cycle, the control circuitry of the MM must ensure that only the specified byte is altered. In this case, the higher-order 30 bits can specify the word and the lower-order 2 bits can specify the byte within the word.

CPU-Main Memory Connection – A block schematic: -

Data transfer between CPU and MM takes place through the use of two CPU registers, usually called MAR (Memory Address Register) and MDR (Memory Data Register). If MAR is K bits long and MDR is „n‟ bits long, then the MM unit may contain upto 2^k addressable locations and each location will be „n‟ bits wide, while the word length is equal to „n‟ bits. During a “memory cycle”, n bits of data may be transferred between the

MM and CPU. This transfer takes place over the processor bus, which has k address lines (address bus), n data lines (data bus) and control lines like Read, Write, Memory Function completed (MFC), Bytes specifiers etc (control bus). For a read operation, the CPU loads the address into MAR, set R/W to 1 and sets other control signals if required. The data from the MM is loaded into MDR and MFC is set to 1. For a write operation, MAR, MDR are suitably loaded by the CPU,R/W is set to 0 and other control signals are set suitably. The MM control circuitry loads the data into appropriate locations and sets MFC to 1. This organization is shown in the following block schematic.

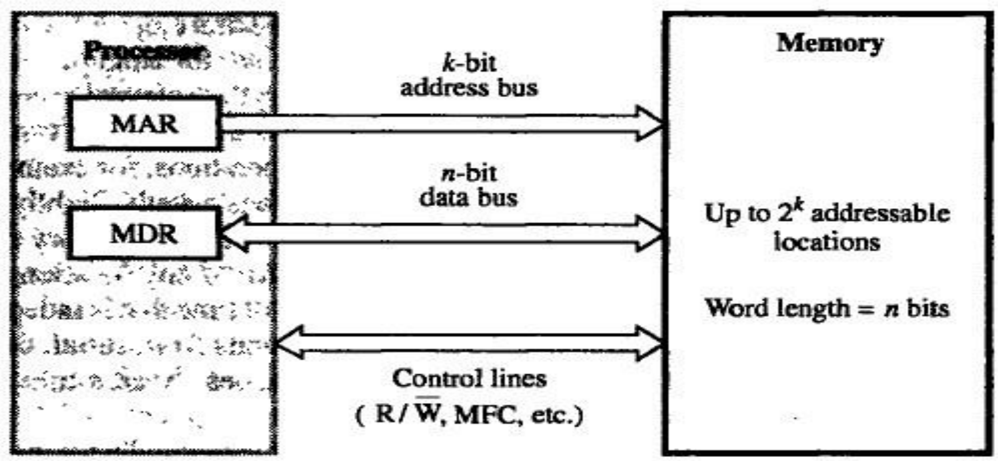


Figure 5.1 Connection of the memory to the processor.

Some Basic Concepts Memory Access Times: -

It is a useful measure of the speed of the memory unit. It is the time that elapses between the initiation of an operation and the completion of that operation (for example, the time between READ and MFC).

Memory Cycle Time :-

It is an important measure of the memory system. It is the minimum time delay required between the initiations of two successive memory operations (for example, the time between two successive READ operations). The cycle time is usually slightly longer than the access time.

RAM: A memory unit is called a Random Access Memory if any location can be accessed for a READ or WRITE operation in some fixed amount of time that is independent of the location‟s address. Main memory units are of this type. This distinguishes them from serial or partly serial access storage devices such as magnetic tapes and disks which are used as the secondary storage device.

Cache Memory:-

The CPU of a computer can usually process instructions and data faster than they can be fetched from compatibly priced main memory unit. Thus the memory cycle time becomes the bottleneck in the system. One way to reduce the memory access time is to use cache memory. This is a small and fast memory that is inserted between the larger, slower main memory and the CPU. This holds the currently active segments of a program and its data. Because of the locality of address references, the CPU can, most of the time, find the relevant information in the cache memory itself (cache hit) and infrequently needs access to the main memory (cache miss) with suitable size of the cache memory, cache hit rates of over 90% are possible leading to a cost-effective increase in the performance of the system.

Memory Interleaving: -

This technique divides the memory system into a number of memory modules and arranges addressing so that successive words in the address space are placed in different modules. When requests for memory access involve consecutive addresses, the access will be to different modules. Since parallel access to these modules is possible, the average rate of fetching words from the Main Memory can be increased.

Virtual Memory: -

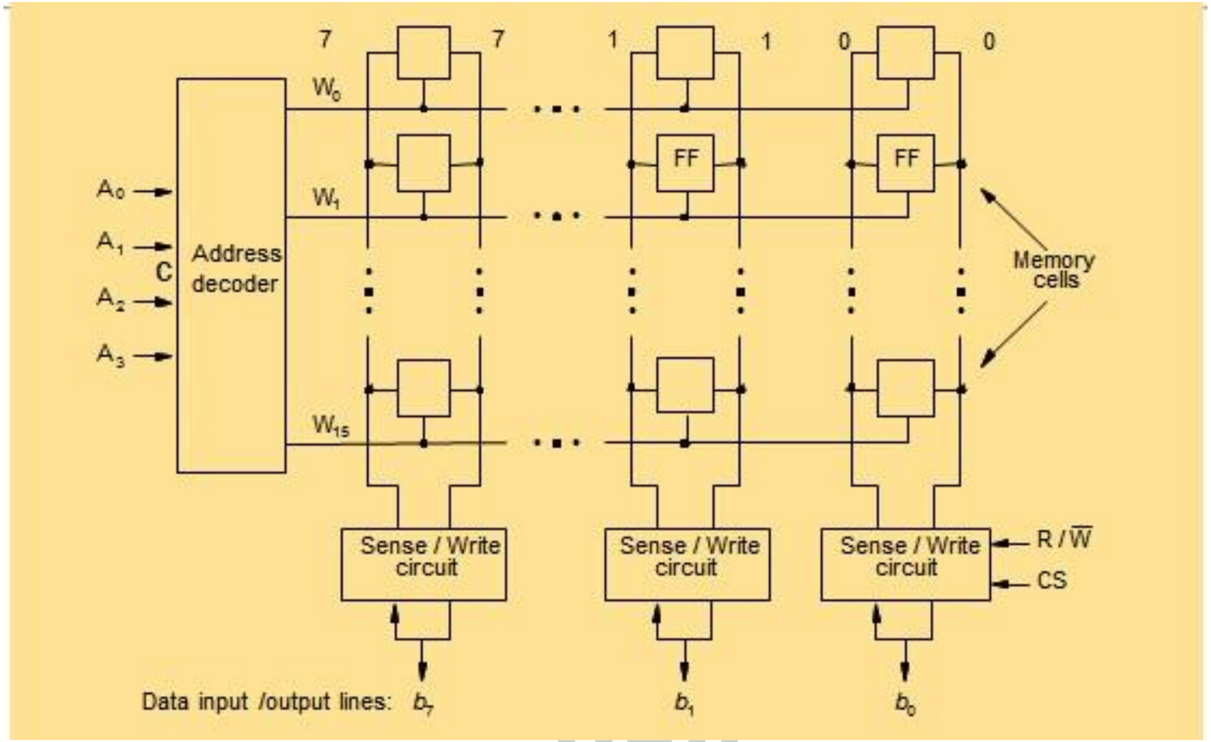
In a virtual memory System, the address generated by the CPU is referred to as a virtual or logical address. The corresponding physical address can be different and the required mapping is implemented by a special memory control unit, often called the memory management unit. The mapping function itself may be changed during program execution according to system requirements.

Because of the distinction made between the logical (virtual) address space and the physical address space; while the former can be as large as the addressing capability of the CPU, the actual physical memory can be much smaller. Only the active portion of the virtual address space is mapped onto the physical memory and the rest of the virtual address space is mapped onto the bulk storage device used. If the addressed information is in the Main Memory (MM), it is accessed and execution proceeds. Otherwise, an exception is generated, in response to which the memory management unit transfers a contigious block of words containing the desired word from the bulk storage unit to the MM, displacing some block that is currently inactive. If the memory is managed in such a way that, such transfers are required relatively infrequency (ie the CPU will generally find the required information in the MM), the virtual memory system can provide a reasonably good performance and succeed in creating an illusion of a large memory with a small, in expensive MM.

5.2 SEMICONDUCTOR RAM MEMORIES

5.2.1 Internal Organization of Memory Chips

Memory cells are usually organized in the form of an array, in which each cell is capable of storing on bit of information. Each row of cells constitutes a memory word, and all cells of a row are connected to a common line referred to as the word line, which is driven by the address decoder on the chip. The cells in each column are connected to a Sense/Write circuit by two bit lines. The Sense/Write circuits are connected to the data I/O lines of the chip. During the read operation, these circuits‟ sense, or read, the information stored in the cells selected by a word line and transmit this information to the output data lines. During the write operation, the Sense/Write circuits receive the input information and store in the cells of the selected word.



The above figure is an example of a very small memory chip consisting of 16 words of 8 bits each. This is referred to as a 16×8 organization. The data input and the data output of each Sense/Write circuit are connected to a single bidirectional data line that can be connected to the data bus of a computer. Two control lines, R/W (Read/ Write) input specifies the required operation, and the CS (Chip Select) input selects a given chip in a multichip memory system.

The memory circuit given above stores 128 and requires 14 external connections for address, data and control lines. Of course, it also needs two lines for power supply and ground connections. Consider now a slightly larger memory circuit, one that has a 1k (1024) memory cells. For a 1k×1 memory organization, the representation is given next. The required 10-bit address is divided into two groups of 5 bits each to form the row and column addresses for the cell array. A row address selects a row of 32 cells, all of which are accessed in parallel. However, according to the column address, only one of these cells is connected to the external data line by the output multiplexer and input demultiplexer.

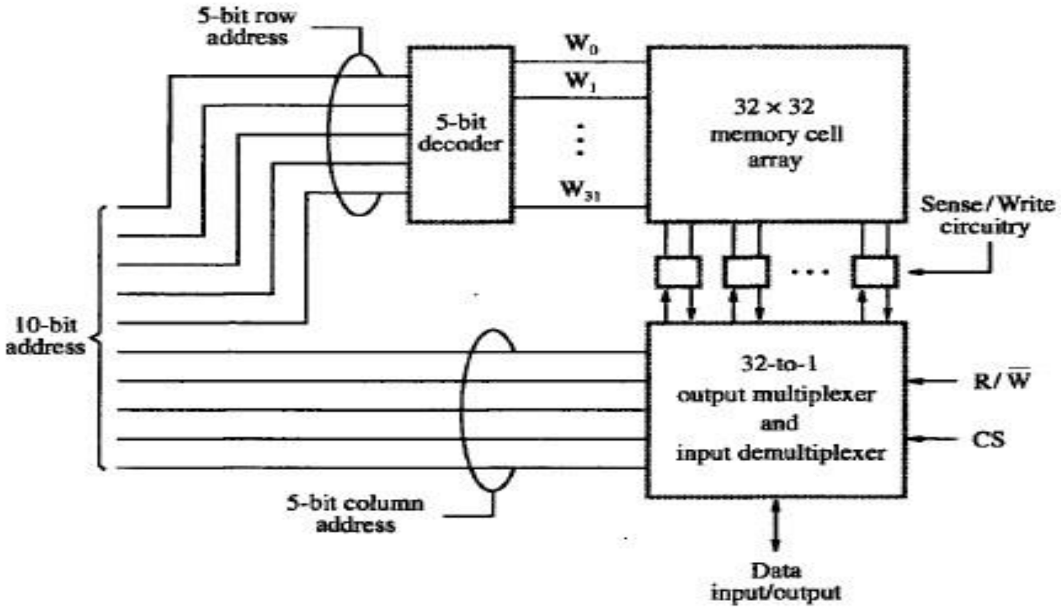
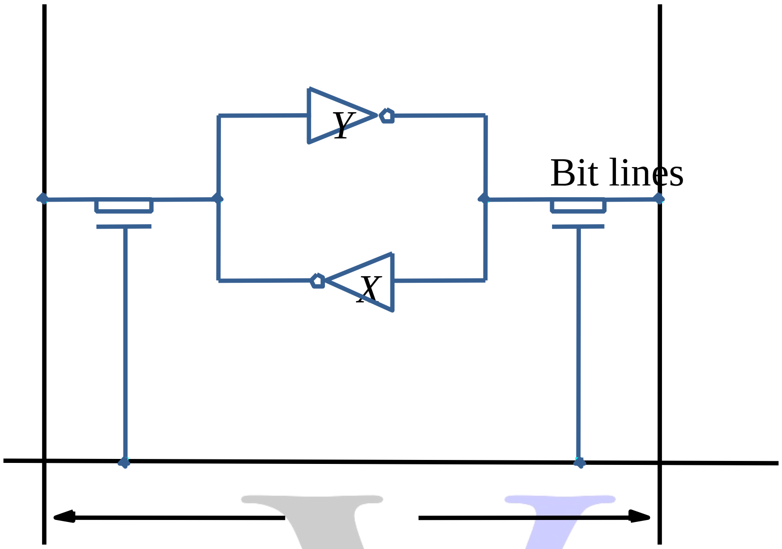


Figure 5.3 Organization of a 1K × 1 memory chip.

5.2.2 Static Memories

Memories that consist of circuits capable of retaining their state as long as power is applied are known as static memories.



The above figure illustrates how a static RAM (SRAM) cell may be implemented. Two inverters are cross- connected to form a latch. The latch is connected to two bit lines by transistors T_1 and T_2. These transistors act as switches that can be opened or closed under control of the word line. When the word line is at ground level, the transistors are turned off and the latch retains its state. For example, let us assume that the cell is in state 1 if the logic value at point X is 1 and at point Y is 0. This state is maintained as long as the signal on the word line is at ground level.

Read Operation

In order to read the state of the SRAM cell, the word line is activated to close switches T_1 and T_2. If the cell is in state 1, the signal on the bit line b is high and the signal on the bit line b‟ is low. The opposite is true if the cell is in state 0. Thus b and b‟ are compliments of each other. Sense/Write circuits at the end of the bit lines monitor the state of b and b‟ and set the output accordingly.

Write Operation

The state of the cell is set by placing the appropriate value on bit line b and its complement b‟, and then activating the word line. This forces the cell into the corresponding state. The required signals on the bit lines are generated by the Sense/Write circuit.

CMOS Cell

A CMOS realization of the static RAM cell is given below:

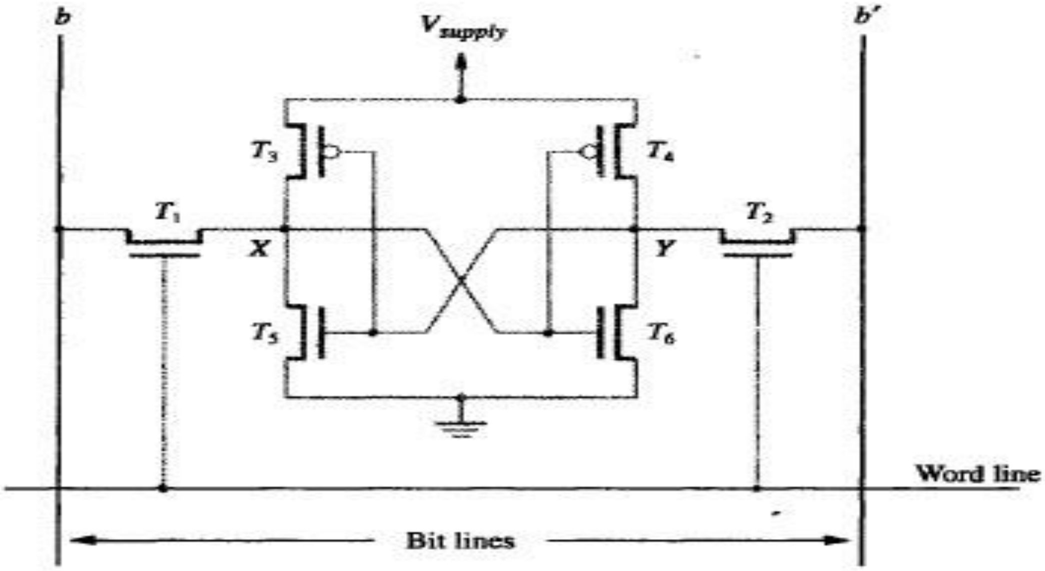


Figure 5.5 An example of a CMOS memory cell.

Transistor pairs (T_3, T_5) and (T_4, T_6) form the inverters in the latch (see Appendix A). The state of the cell is read or written as just explained. For example, in state 1, the voltage at point X is maintained high by having transistors T_3 and T_6 on, while T_4 and T_5 are off. Thus, if T_1 and T_2 are turned on (closed), bit lines b and b‟ will have high and low signals, respectively.

5.2.3 Asynchronous DRAMS

Information is stored in a dynamic memory cell in the form of a charge on a capacitor, and this charge can be maintained for only tens of milliseconds. Since the cell is required to store information for a much longer time, its contents must be periodically refreshed by restoring the capacitor charge to its full value. An example of a dynamic memory cell that consists of a capacitor, C, and a transistor, T, is shown below:

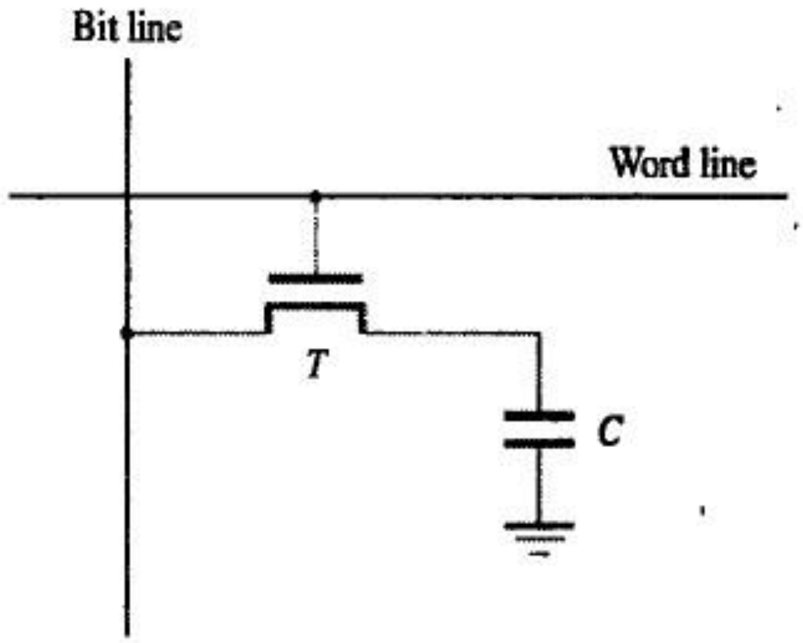


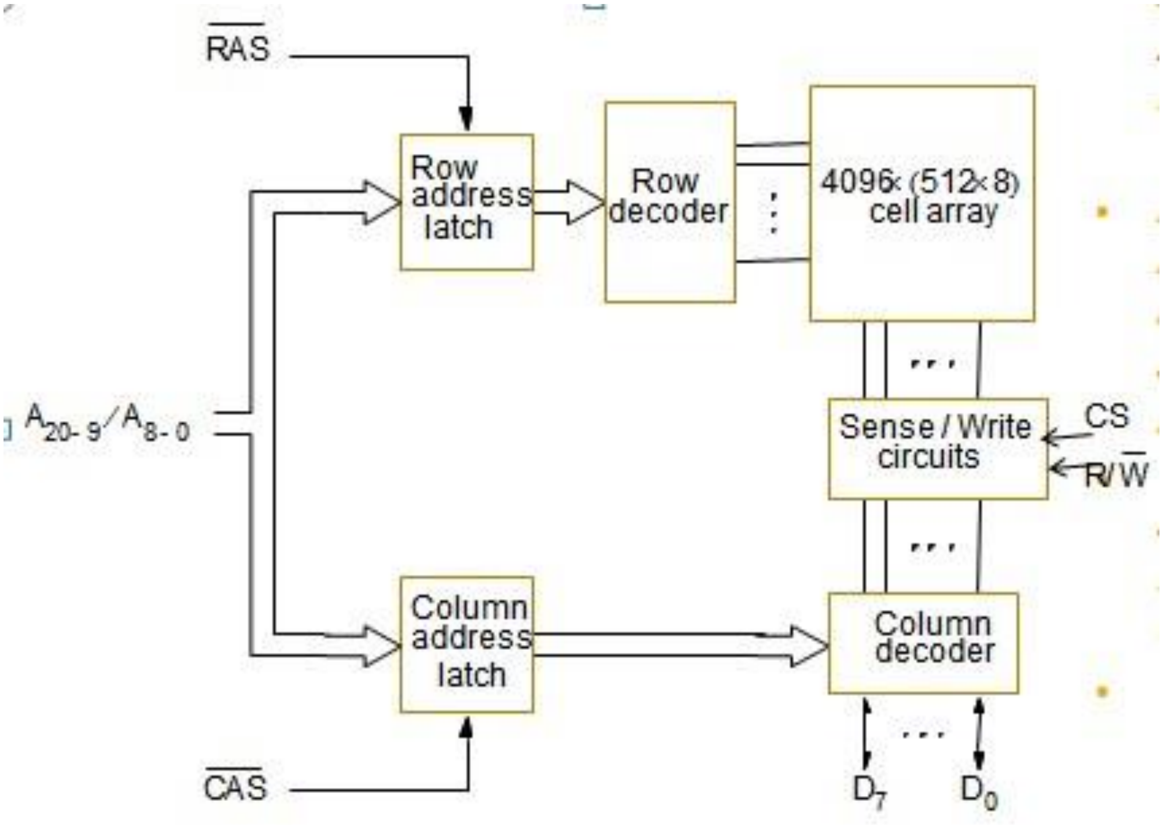
Figure 5.6 A single-transistor dynamic memory cell.

A sense amplifier connected to the bit line detects whether the charge stored on the capacitor is above the threshold. If so, it drives the bit line to a full voltage that represents logic value 1. This voltage recharges the capacitor to full charge that corresponds to logic value 1. If the sense amplifier detects that the charge on the capacitor will have no charge, representing logic value 0.

A 16-megabit DRAM chip, configured as 2M×8, is shown below.

- ***Each row can store 512 bytes. 12 bits to select a row, and 9 bits to select a group in a row. Total of 21 bits.***
- ***First apply the row address; RAS signal latches the row address. Then apply the column address, CAS signal latches the address.***

- ***Timing of the memory unit is controlled by a specialized unit which generates RAS and CAS.***
- ***This is asynchronous DRAM***



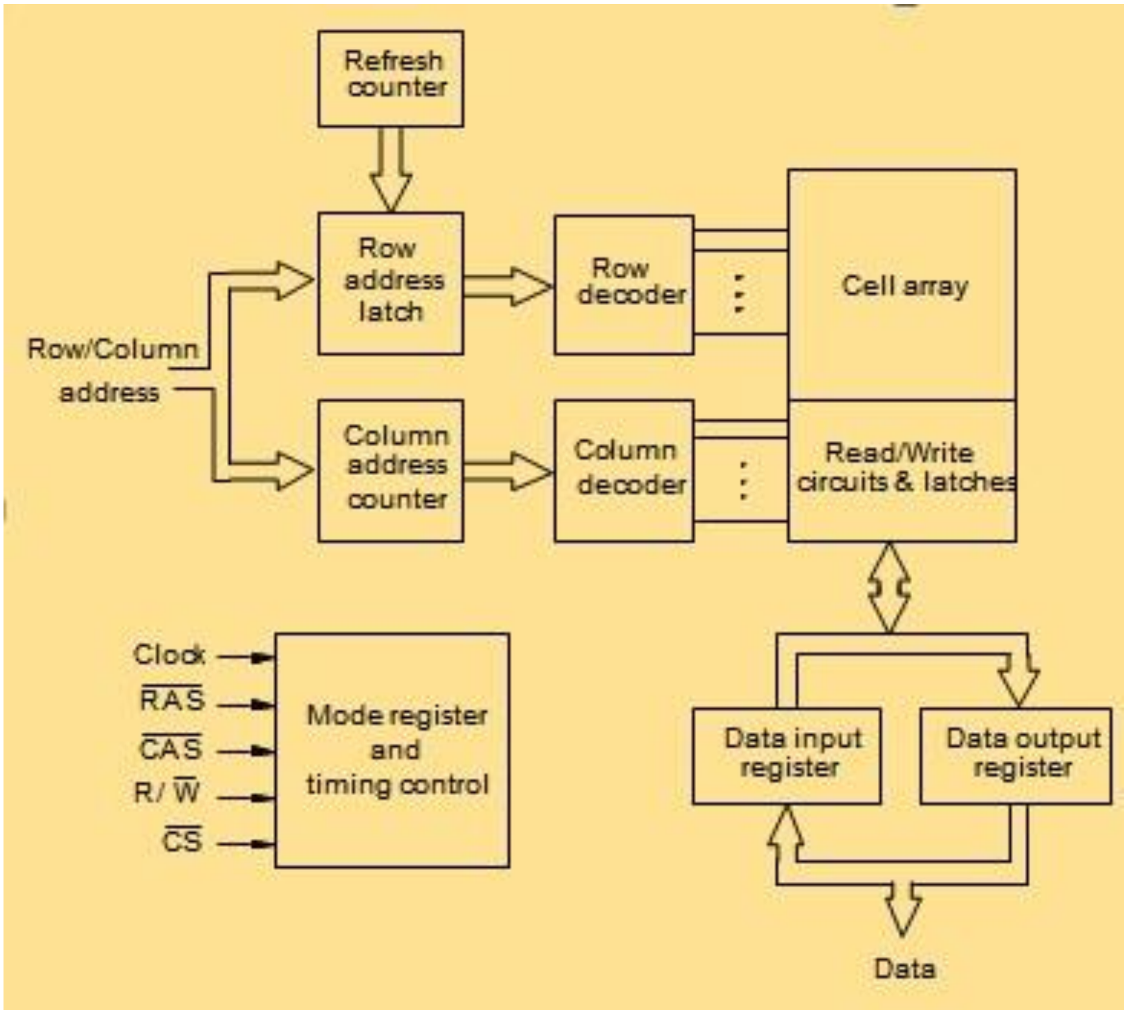
Fast Page Mode

- Suppose if we want to access the consecutive bytes in the selected row.
- This can be done without having to reselect the row.
  - Add a latch at the output of the sense circuits in each row.
  - All the latches are loaded when the row is selected.
  - Different column addresses can be applied to select and place different bytes on the data lines.
- Consecutive sequence of column addresses can be applied under the control signal CAS, without reselecting the row.

  - Allows a block of data to be transferred at a much faster rate than random accesses.
  - A small collection/group of bytes is usually referred to as a block.
- This transfer capability is referred to as the fast page mode feature.

5.2.4 Synchronous DRAMs

In these DRAMs, operation is directly synchronized with a clock signal. The below given figure indicates the structure of an SDRAM.



- The output of each sense amplifier is connected to a latch.
- A Read operation causes the contents of all cells in the selected row to be loaded into these latches.
- But, if an access is made for refreshing purpose only, it will not change the contents of these latches; it will merely refresh the contents of the cells.

- Data held in the latches that correspond to the selected column(s) are transferred into the output register, thus becoming available on the data output pins.

- SDRAMs have several different modes of operation, which can be selected by writing control information into a mode register. For example, burst operations of different lengths are specified.
- The burst operations use the block transfer capability described before as fast page mode feature.
- In SDRAMs, it is not necessary to provide externally generated pulses on the CAS line to select successive columns. The necessary control signals are provided internally using a column counter and the clock signal. New data can be placed on the data lines in each clock cycles. All actions are triggered by the rising edge of the clock.

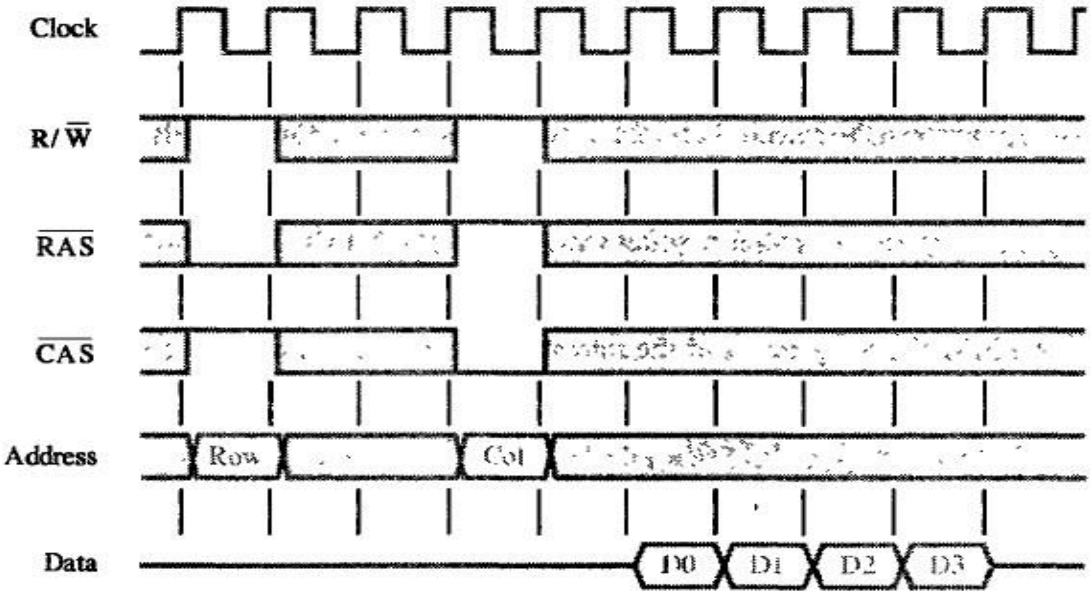


Figure 5.9 Burst read of length 4 in an SDRAM.

The above figure shows the timing diagram for a burst read of length 4.

- First, the row address is latched under control of the RAS signal.
- Then, the column address latched under control of the CAS signal.
- After a delay of one clock cycle, the first set of data bits is placed on the data lines.
- The SDRAM automatically increments the column address to access next three sets of the bits in the selected row, which are placed on the data lines in the next clock cycles.

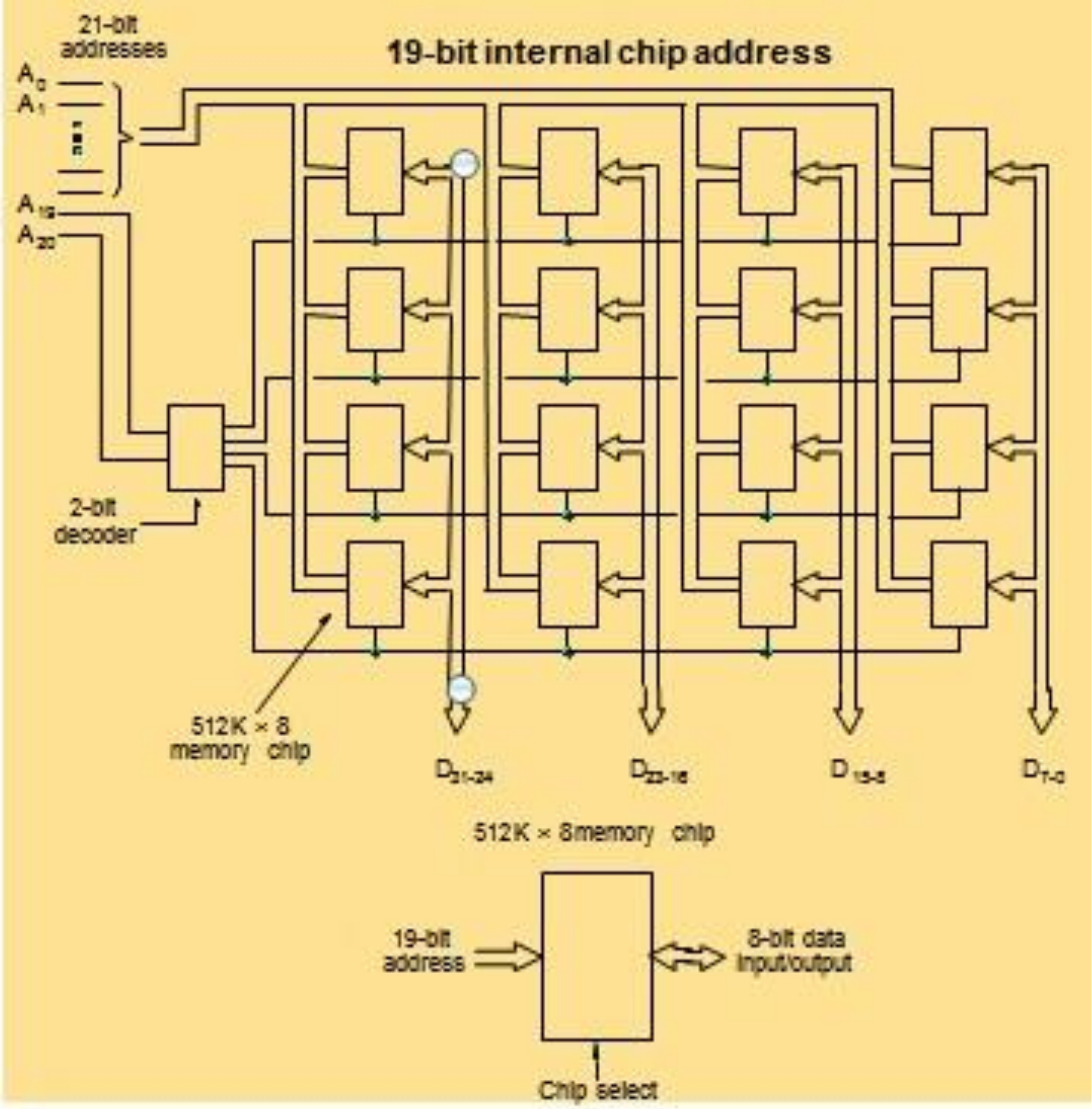
Latency and Bandwidth

- Memory latency is the time it takes to transfer a word of data to or from memory

- Memory bandwidth is the number of bits or bytes that can be transferred in one second.
- DDRSDRAMs- Cell array is organized in two banks.

Double Data Rate- Synchronous DRAMs (DDR- SDRAMs)

To assist the processor in accessing data at high enough rate, the cell array is organized in two banks. Each bank can be accessed separately. Consecutive words of a given block are stored in different banks. Such interleaving of words allows simultaneous access to two words that are transferred on the successive edges of the clock. This type of SDRAM is called Double Data Rate SDRAM (DDR- SDRAM). **5.2.5 Structure of larger memories**



- *Implementing a memory unit of 2M words of 32 bits each.*
- *Using 512x8 static memory chips. Each column consists of 4 chips. Each chip implements one byte position.*
- *A chip is selected by setting its chip select control line to 1. Selected chip places its data on the data output line, outputs of other chips are in high impedance state.*
- *21 bits to address a 32-bit word. High order 2 bits are needed to select the row, by activating the four Chip Select signals.*
- *19 bits are used to access specific byte locations inside the selected chip.*

Dynamic Memory System

- Large dynamic memory systems can be implemented using DRAM chips in a similar way to static memory systems.
- Placing large memory systems directly on the motherboard will occupy a large amount of space.
  - Also, this arrangement is inflexible since the memory system cannot be expanded easily.
- Packaging considerations have led to the development of larger memory units known as SIMMs (Single In-line Memory Modules) and DIMMs (Dual In-line Memory Modules).
- Memory modules are an assembly of memory chips on a small board that plugs vertically onto a single socket on the motherboard.
  - Occupy less space on the motherboard.
  - Allows for easy expansion by replacement.

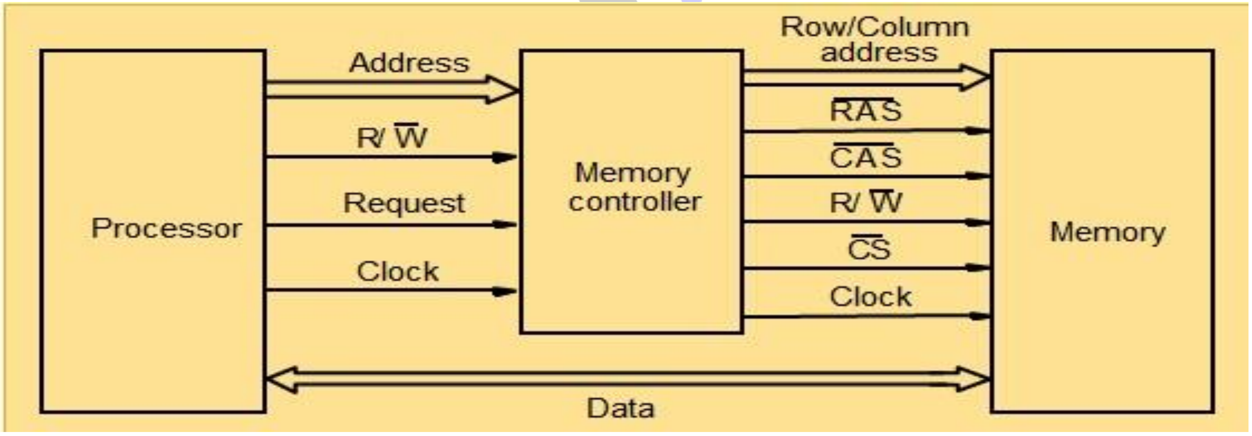
Recall that in a dynamic memory chip, to reduce the number of pins, multiplexed addresses are used.

- Address is divided into two parts:
  - High-order address bits select a row in the array.
  - They are provided first, and latched using RAS signal.
  - Low-order address bits select a column in the row.
  - They are provided later, and latched using CAS signal.
- However, a processor issues all address bits at the same time.

- In order to achieve the multiplexing, memory controller circuit is inserted between the processor and memory.

5.2.6 Memory System Considerations

- Recall that in a dynamic memory chip, to reduce the number of pins, multiplexed addresses are used.
- Address is divided into two parts:
  - High-order address bits select a row in the array.
  - They are provided first, and latched using RAS signal.
  - Low-order address bits select a column in the row.
  - They are provided later, and latched using CAS signal.
- However, a processor issues all address bits at the same time.
- In order to achieve the multiplexing, memory controller circuit is inserted between the processor and memory.



Refresh Operation:-

The Refresh control block periodically generates Refresh, requests, causing the access control block to start a memory cycle in the normal way. This block allows the refresh operation by activating the Refresh Grant line. The access control block arbitrates between Memory Access requests and Refresh requests, with priority to refresh requests in the case of a tie to ensure the integrity of the stored data.

As soon as the Refresh control block receives the Refresh Grant signal, it activates the Refresh line. This causes the address multiplexer to select the Refresh counter as the source and its contents are thus loaded into the row address latches of all memory chips when the RAS signal is activated.

5.3 Semi-Conductor Rom Memories: -

Semiconductor read-only memory (ROM) units are well suited as the control store components in micro programmed processors and also as the parts of the main memory that contain fixed programs or data. The following figure shows a possible configuration for a bipolar ROM cell.

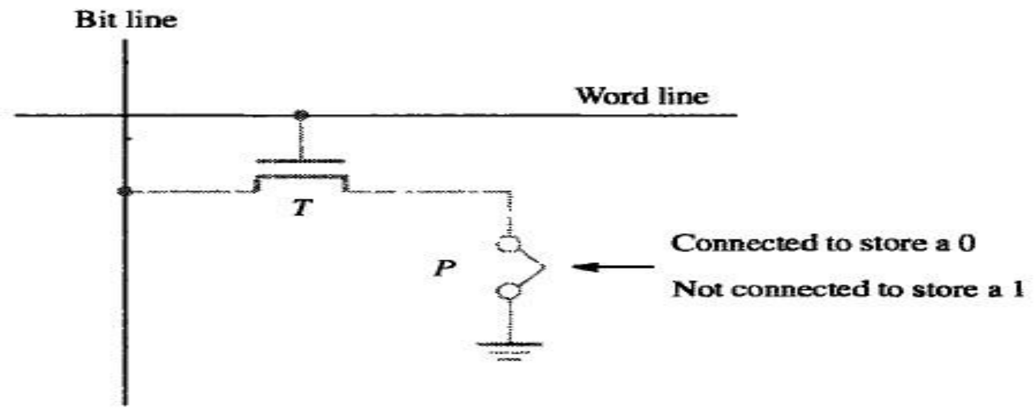


Figure 5.12 A ROM cell.

The word line is normally held at a low voltage. If a word is to be selected, the voltage of the corresponding word line is momentarily raised, which causes all transistors whose emitters are connected to their corresponding bit lines to be turned on. The current that flows from the voltage supply to the bit line can be detected by a sense circuit. The bit positions in which current is detected are read as 1s, and the remaining bits are read as O_s. Therefore, the contents of a given word are determined by the pattern of emitter to bitline connections similar configurations are possible in MOS technology.

Data are written into a ROM at the time of manufacture programmable ROM (PROM) devices allow the data to be loaded by the user. Programmability is achieved by connecting a fuse between the emitter and the bit line. Thus, prior to programming, the memory contains all 1s. The user can inserts O_s at the required locations by burning out the fuses at these locations using high-current pulses. This process is irreversible.

ROMs are attractive when high production volumes are involved. For smaller numbers, PROMs provide a faster and considerably less expensive approach. Chips which allow the stored data to be erased and new data to be loaded. Such a chip is an erasable, programmable ROM, usually called an EPROM. It provides considerable flexibility during the development phase. An EPROM cell bears considerable resemblance to the dynamic memory cell. As in the case of dynamic memory, information is stored in the form of a charge on a capacitor. The main difference is that the capacitor in an EPROM cell is very well insulated. Its rate of discharge is so low that it retains the stored information for very long periods. To write information, allowing charge to be stored on the capacitor.

The contents of EPROM cells can be erased by increasing the discharge rate of the storage capacitor by several orders of magnitude. This can be accomplished by allowing ultraviolet light into the chip through a window provided for that purpose, or by the application of a high voltage similar to that used in a write operation. If ultraviolet light is used, all cells in the chip are erased at the same time. When electrical erasure is used, however, the process can be made selective. An electrically erasable EPROM, often referred to as EEPROM. However, the circuit must now include high voltage generation.

Some EEPROM chips incorporate the circuitry for generating these voltages o the chip itself. Depending on the requirements, suitable device can be selected.

Flash memory:

- Has similar approach to EEPROM.
- Read the contents of a single cell, but write the contents of an entire block of cells.
- Flash devices have greater density.
    - Higher capacity and low storage cost per bit.
- Power consumption of flash memory is very low, making 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.

(REFER slides for point wise notes on RoM and types of ROM)

5.4 Speed, Size and Cost

A big challenge in the design of a computer system is to provide a sufficiently large memory, with a reasonable speed at an affordable cost.

Static RAM: Very fast, but expensive, because a basic SRAM cell has a complex circuit making it impossible to pack a large number of cells onto a single chip.

Dynamic RAM: Simpler basic cell circuit, hence are much less expensive, but significantly slower than SRAMs.

Magnetic disks: Storage provided by DRAMs is higher than SRAMs, but is still less than what is necessary. Secondary storage such as magnetic disks provides a large amount of storage, but is much slower than DRAMs**.**

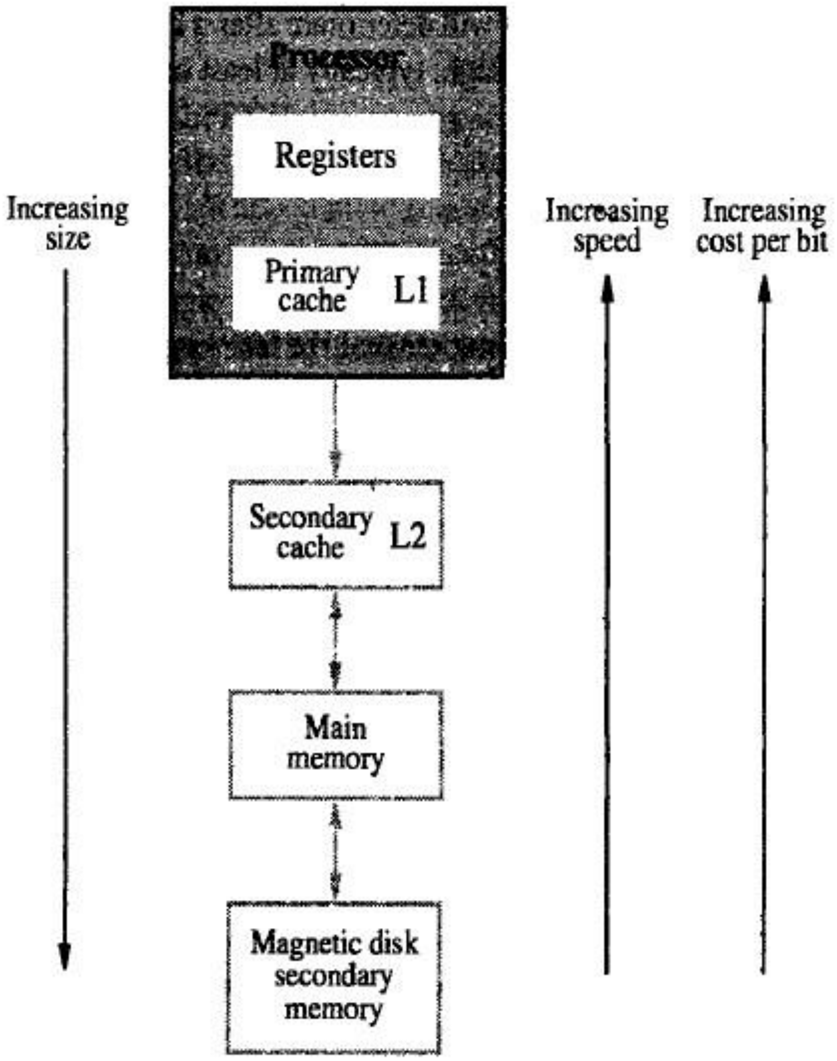


Figure 5.13 Memory hierarchy.

Fastest access is to the data held in processor registers. Registers are at the top of the memory hierarchy. Relatively small amount of memory that can be implemented on the processor chip. This is processor cache. Two levels of cache. Level 1 (L1) cache is on the processor chip. Level 2 (L2) cache is in between main memory and processor. Next level is main memory, implemented as SIMMs. Much larger, but much slower than cache memory. Next level is magnetic disks. Huge amount of inexpensive storage. Speed of memory access is critical, the idea is to bring instructions and data that will be used in the near future as close to the processor as possible.

5.5 Cache memories

Processor is much faster than the main memory. As a result, the processor has to spend much of its time waiting while instructions and data are being fetched from the main memory. This serves as a major obstacle towards achieving good performance. Speed of the main memory cannot be increased beyond a certain point. So we use Cache memories. Cache memory is an architectural arrangement which makes the main memory appear faster to the processor than it really is. Cache memory is based on the property of computer programs known as “locality of reference”.

Analysis of programs indicates that many instructions in localized areas of a program are executed repeatedly during some period of time, while the others are accessed relatively less frequently. These instructions may be the ones in a loop, nested loop or few procedures calling each other repeatedly. This is called “locality of reference”. Its types are:

Temporal locality of reference: Recently executed instruction is likely to be executed again very soon.

Spatial locality of reference: Instructions with addresses close to a recently instruction are likely to be executed soon.

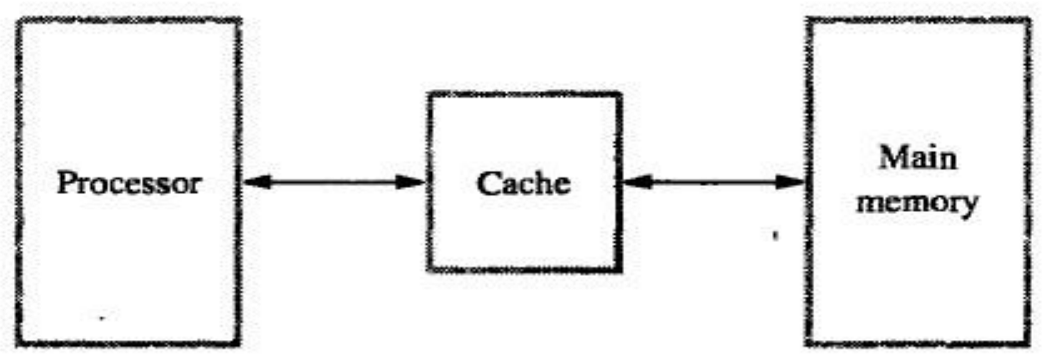


Figure 5.14 Use of a cache memory.

A simple arrangement of cache memory is as shown above.

- Processor issues a Read request, a block of words is transferred from the main memory to the cache, one word at a time.
- Subsequent references to the data in this block of words are found in the cache.

- At any given time, only some blocks in the main memory are held in the cache. Which blocks in the main memory are in the cache is determined by a "mapping function".

- 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".

Cache hit:

Existence of a cache is transparent to the processor. The processor issues Read and Write requests in the same manner. If the data is in the cache it is called a Read or Write hit.

Read hit: The data is obtained from the cache.

Write hit: Cache has a replica of the contents of the main memory. Contents of the cache and the main memory may be updated simultaneously. This is the write-through protocol.

Update the contents of the cache, and mark it as updated by setting a bit known as the dirty bit or modified bit. The contents of the main memory are updated when this block is replaced. This is write-back or copy-back protocol.

Cache miss:

- If the data is not present in the cache, then a Read miss or Write miss occurs.

- Read miss: Block of words containing this requested word is transferred from the memory. After the block is transferred, the desired word is forwarded to the processor. The desired word may also be forwarded to the processor as soon as it is transferred without waiting for the entire block to be transferred. This is called load-through or early-restart.

- Write-miss: Write-through protocol is used, then the contents of the main memory are updated directly. If write-back protocol is used, the block containing the addressed word is first brought into the cache. The desired word is overwritten with new information.

Cache Coherence Problem:

A bit called as "valid bit" is provided for each block. If the block contains valid data, then the bit is set to 1, else it is 0. Valid bits are set to 0, when the power is just turned on.

When a block is loaded into the cache for the first time, the valid bit is set to 1. Data transfers between main memory and disk occur directly bypassing the cache. When the data on a disk changes, the main memory block is also updated. However, if the data is also resident in the cache, then the valid bit is set to 0.

The copies of the data in the cache, and the main memory are different. This is called the cache coherence problem

Mapping functions: Mapping functions determine how memory blocks are placed in the cache.

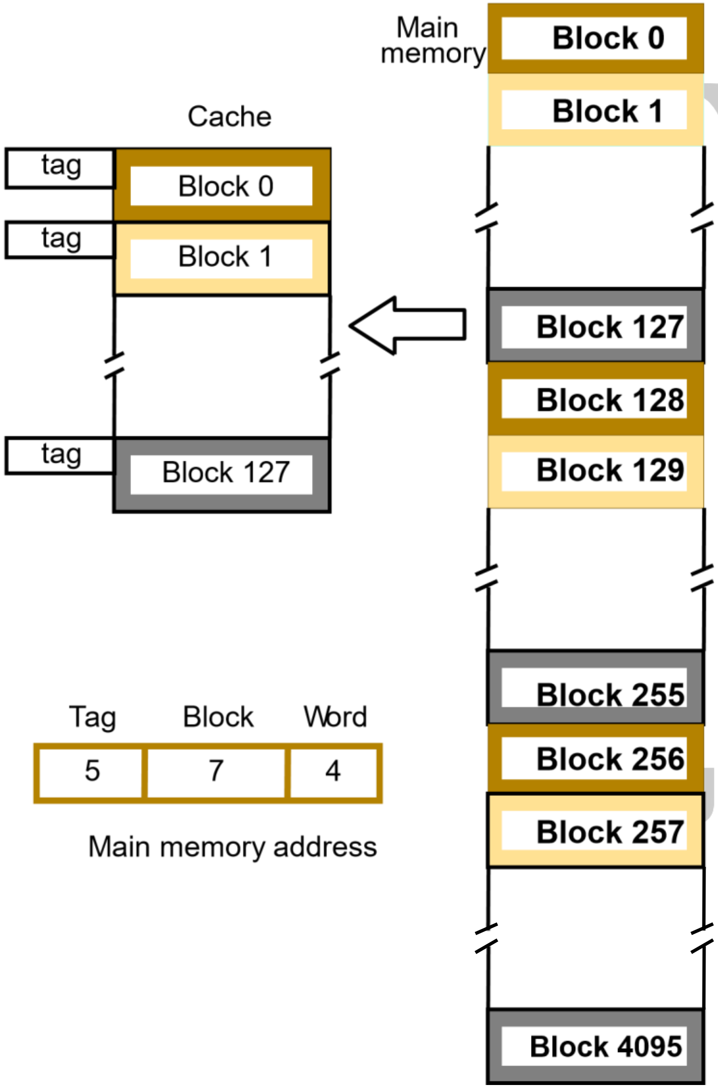
A simple processor example:

- Cache consisting of 128 blocks of 16 words each.
- Total size of cache is 2048 (2K) words.
- Main memory is addressable by a 16-bit address.
- Main memory has 64K words.
- Main memory has 4K blocks of 16 words each.

Three mapping functions can be used.

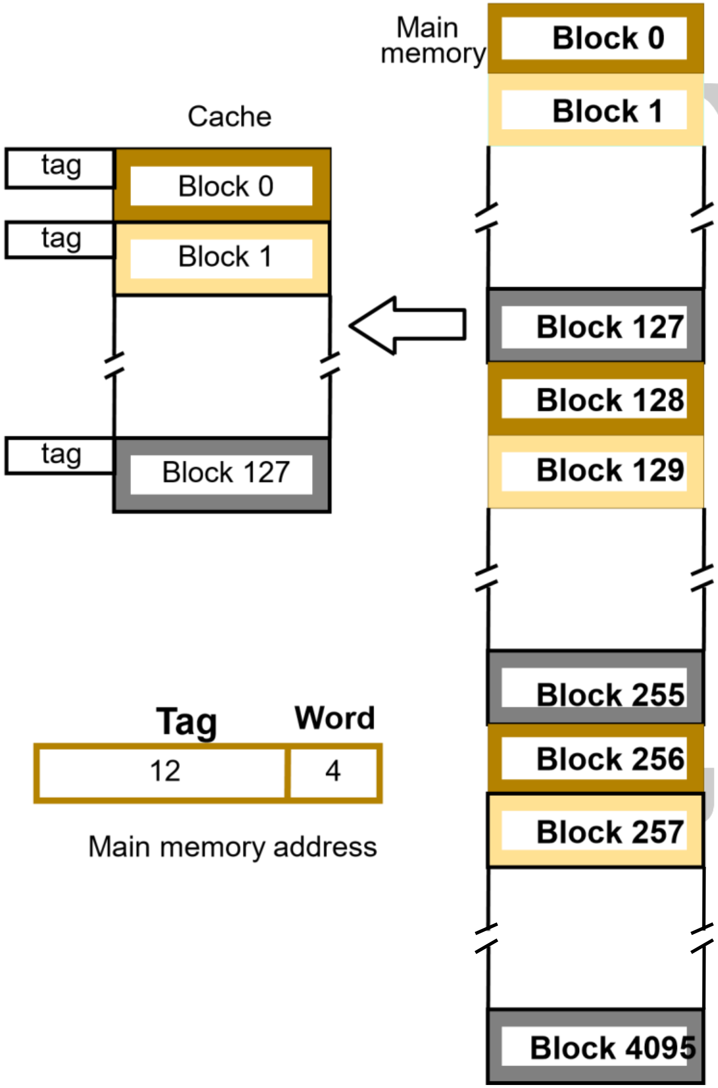
1. Direct mapping
2. Associative mapping 3. Set-associative mapping.

Direct mapping



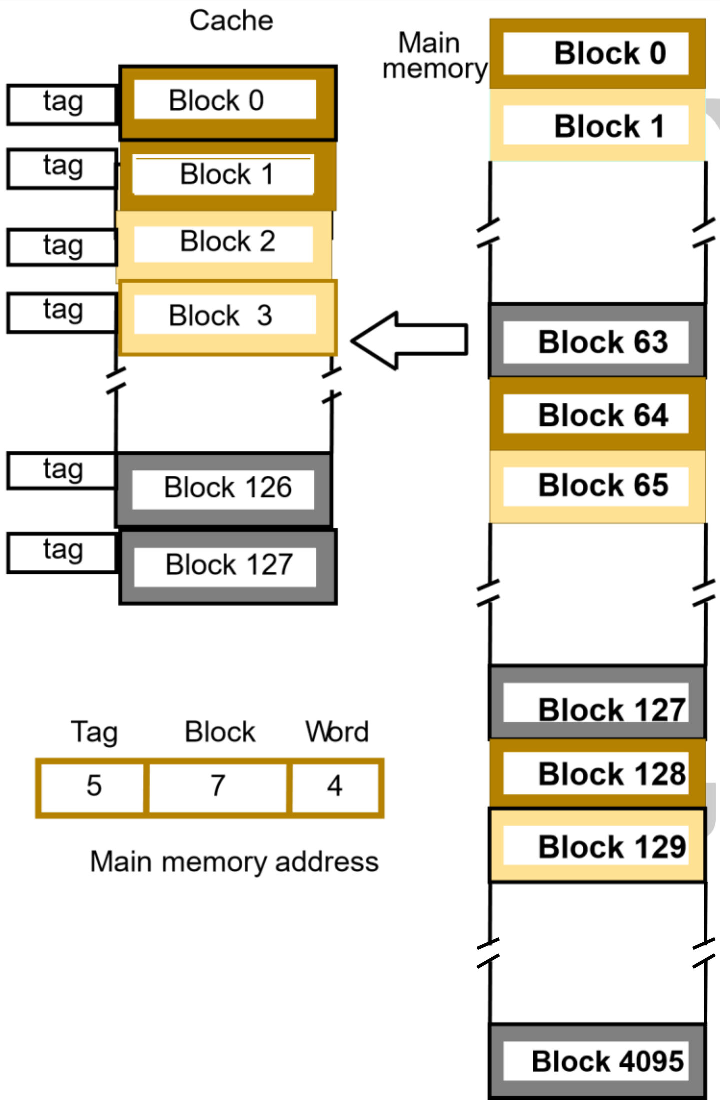
- Block j of the main memory maps to j modulo 128 of the cache. 0 maps to 0, 129 maps to 1.
- More than one memory block is mapped onto the same position in the cache.
- May lead to contention for cache blocks even if the cache is not full.
- Resolve the contention by allowing new block to replace the old block, leading to a trivial replacement algorithm.
- 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 the next 7 bits determine which cache block this new block is placed in.
  - High order 5 bits determine which of the possible 32 blocks is currently present in the cache. These are tag bits.
- Simple to implement but not very flexible.

Associative mapping



- Main memory block can be placed into any cache position.
- 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.
- Flexible, and uses cache space efficiently.
- Replacement algorithms can be used to replace an existing block in the cache when the cache is full.
- 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.

Set-Associative mapping



Blocks of cache are grouped into sets.

Mapping function allows a block of the main memory to reside in any block of a specific set.

Divide the cache into 64 sets, with two blocks per set.

Memory block 0, 64, 128 etc. map to block 0, and they can occupy either of the two positions.

Memory address is divided into three fields:

- 6 bit field determines the set number.
- High order 6 bit fields are compared to the tag fields of the two blocks in a set.

Set-associative mapping combination of direct and associative mapping.

Number of blocks per set is a design parameter.

- One extreme is to have all the blocks in one set, requiring no set bits (fully associative mapping).
- Other extreme is to have one block per set, is the same as direct mapping.

Replacement Algorithm

In a direct-mapped cache, the position of each block is fixed, hence no replacement strategy exists. In associative and set-associative caches, when a new block is to be brought into the cache and all the Positions that it may occupy are full, the cache controller must decide which of the old blocks to overwrite. This is important issue because the decision can be factor in system performance. The objective is to keep blocks in the cache that are likely to be referenced in the near future. Its not easy to determine which blocks are about to be referenced. The property of locality of reference gives a clue to a reasonable strategy. When a block is to be over written, it is sensible to overwrite the one that has gone the longest time without being referenced. This block is called the least recently used (LRU) block, and technique is called the LRU Replacement algorithm. The LRU algorithm has been used extensively for many access patterns, but it can lead to poor performance in some cases. For example, it produces disappointing results when accesses are made to sequential elements of an array that is slightly too large to fit into the cache. Performance of LRU algorithm can be improved by introducing a small amount of randomness in deciding which block to replace.

Solved Problems:-

1. A block set associative cache consists of a total of 64 blocks divided into 4 block sets. The MM contains 4096 blocks each containing 128 words.

a) How many bits are there in MM address?

b) How many bits are there in each of the TAG, SET & word fields

Solution:- Number of sets = 64/4 = 16

Set bits = 4(2^4 = 16)

Number of words = 128

Word bits = 7 bits (2^7 = 128)

MM capacity : 4096 x 128 (2^{12} x 2^7 = 2^{19})

a) Number of bits in memory address = 19 bits

b)

8	4	7
TAG	SET	WORD

TAG bits = 19 – (7+4) = 8 bits.

2. A computer system has a MM capacity of a total of 1M 16 bits words. It also has a 4K words cache organized in the block set associative manner, with 4 blocks per set & 64 words per block. Calculate the number of bits in each of the TAG, SET & WORD fields of MM address format. **Solution:** Capacity: 1M (2^{20} = 1M)

Number of words per block = 64

Number of blocks in cache = 4k/64 = 64

Number of sets = 64/4 = 16 Set

bits = 4 (2^4 = 16)

Word bits = 6 bits (2^6 = 64)

Tag bits = 20-(6+4) = 10 bits

MM address format: 10 tag bits, 6 word bits and 4 set bits.

5.6 PERFORMANCE CONSIDERATIONS

A key design objective of a computer system is to achieve the best possible performance at the lowest possible cost. Price/performance ratio is a common measure of success.

Performance of a processor depends on: How fast machine instructions can be brought into the processor for execution. How fast the instructions can be executed.

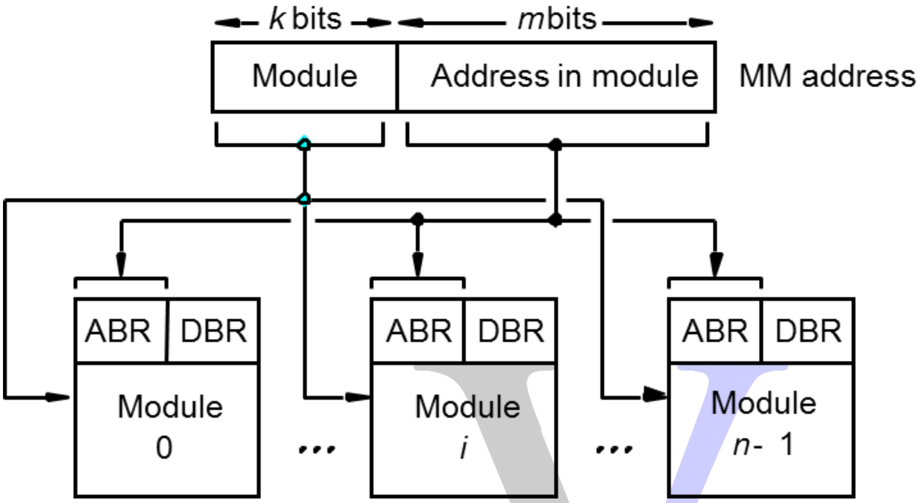
Interleaving

Divides the memory system into a number of memory modules. Each module has its own address buffer register (ABR) and data buffer register (DBR). Arranges addressing so that successive words in the address space are placed in different modules. When requests for memory access involve consecutive addresses, the access will be to different modules.

Since parallel access to these modules is possible, the average rate of fetching words from the Main Memory can be increased.

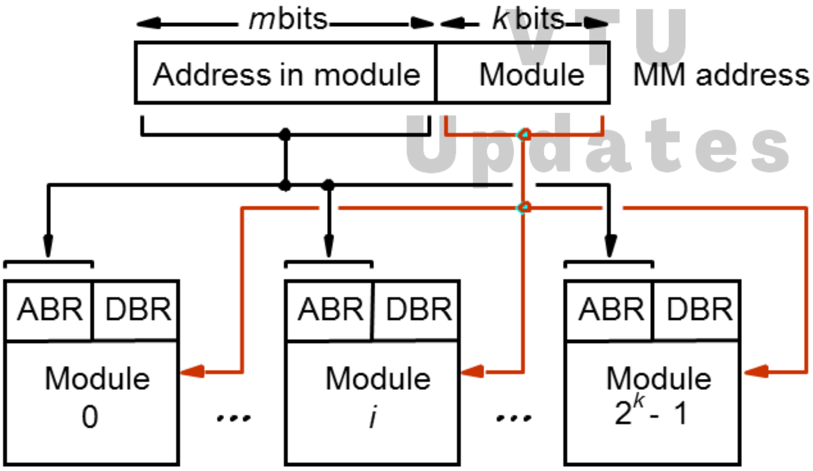
Methods of address layout:

1)



Consecutive words are placed in a module. High-order k bits of a memory address determine the module. Low-order m bits of a memory address determine the word within a module. When a block of words is transferred from main memory to cache, only one module is busy at a time.

2)



Consecutive words are located in consecutive modules. Consecutive addresses can be located in consecutive modules. While transferring a block of data, several memory modules can be kept busy at the same time.

Hit rate and miss penalty

- The number of hits stated as fraction of all attempted accesses is called the **hit rate** and the number of misses stated as a fraction of all attempted accesses is called the **miss rate.**
- The extra time needed to bring the desired information into cache is called as miss penalty.
- Hit rate can be improved by increasing block size, while keeping cache size constant ▪ Block sizes that are neither very small nor very large give best results.
- Miss penalty can be reduced if load-through approach is used when loading new blocks into cache.
- Avg access time experienced by processor is

$tave = hC + (1-h)M$ h- hitrate, M – miss

penalty, C- time to access info from cache.

Caches on the processor chips

- In high performance processors 2 levels of caches are normally used.
- Avg access time in a system with 2 levels of caches is
- $T\ ave = h1c1+(1-h1)h2c2+(1-h1)(1-h2)M$

h1- hitrate of Cache 1, h2- hit rate of Cache 2, M – miss penalty, c1- time to access info from cache1, c2- time to access info from cache2.

Write buffer

Write-through: Each write operation involves writing to the main memory. If the processor has to wait for the write operation to be complete, it slows down the processor. Processor does not depend on the results of the write operation. Write buffer can be included for temporary storage of write requests. Processor places each write request into the buffer and continues execution. If a subsequent Read request references data which is still in the write buffer, then this data is referenced in the write buffer.

Write-back: Block is written back to the main memory when it is replaced. If the processor waits for this write to complete, before reading the new block, it is slowed down.Fast write buffer can hold the block to be written, and the new block can be read first.

In addition to these prefetching and lock up free cache can also be used to improve performance of the processor.

5.7 Virtual Memory

An important challenge in the design of a computer system is to provide a large, fast memory system at an affordable cost. Cache memories were developed to increase the effective speed of the memory system. Virtual memory is an architectural solution to increase the effective size of the memory system.

The addressable memory space depends on the number of address bits in a computer. For example, if a computer issues 32-bit addresses, the addressable memory space is 4G bytes. Physical main memory in a computer is generally not as large as the entire possible addressable space. Physical memory typically ranges from a few hundred megabytes to 1G bytes. Large programs that cannot fit completely into the main memory have their parts stored on secondary storage devices such as magnetic disks. Pieces of programs must be transferred to the main memory from secondary storage before they can be executed.

Techniques that automatically move program and data between main memory and secondary storage when they are required for execution are called virtual-memory techniques. Programs and processors reference an instruction or data independent of the size of the main memory. Processor issues binary addresses for instructions and data. These binary addresses are called logical or virtual addresses. Virtual addresses are translated into physical addresses by a combination of hardware and software subsystems. If virtual address refers to a part of the program that is currently in the main memory, it is accessed immediately. If the address refers to a part of the program that is not currently in the main memory, it is first transferred to the main memory before it can be used.

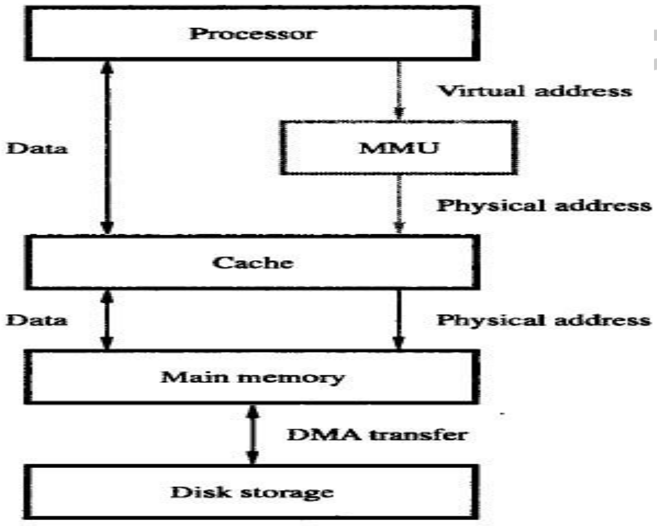


Figure 5.26 Virtual memory organization.

Memory management unit (MMU) translates virtual addresses into physical addresses. If the desired data or instructions are in the main memory they are fetched as described previously. If the desired data or instructions are not in the main memory, they must be transferred from secondary storage to the main memory. MMU causes the operating system to bring the data from the secondary storage into the main memory.

Address Translation

Assume that program and data are composed of fixed-length units called pages. A page consists of a block of words that occupy contiguous locations in the main memory. Page is a basic unit of information that is transferred between secondary storage and main memory. Size of a page commonly ranges from 2K to 16K bytes. Pages should not be too small, because the access time of a secondary storage device is much larger than the main memory. Pages should not be too large, else a large portion of the page may not be used, and it will occupy valuable space in the main memory.

Concepts of virtual memory are similar to the concepts of cache memory.

Cache memory: Introduced to bridge the speed gap between the processor and the main memory. Implemented in hardware.

Virtual memory: Introduced to bridge the speed gap between the main memory and secondary storage. Implemented in part by software.

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. Main memory address where the page is stored. Current status of the page. Area of the main memory that can hold a page is called as page frame. Starting address of the page table is kept in a page table base register.

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.

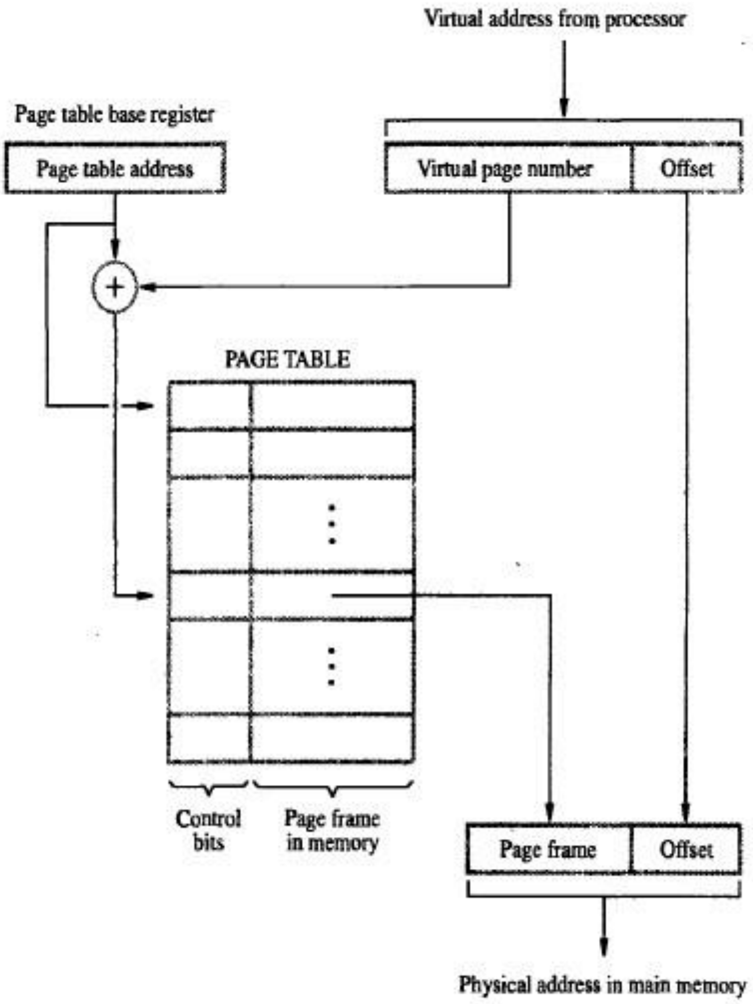


Figure 5.27 Virtual-memory address translation.

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. One bit indicates the validity of the page. Indicates whether the page is actually loaded into the main memory. Allows the operating system to invalidate the page without actually removing it. One bit indicates whether the page has been modified during its residency in the main memory. This bit determines whether the page should be written back to the disk when it is removed from the main memory. Similar to the dirty or modified bit in case of cache memory. Other control bits for various other types of restrictions that may be imposed. For example, a program may only have read permission for a page, but not write or modify permissions.

The page table is used by the MMU for every read and write access to the memory. Ideal location for the page table is within the MMU. Page table is quite large. MMU is implemented as part of the processor chip. Impossible to include a complete page table on the chip. Page table is kept in the main memory. A copy of a small portion of the page table can be accommodated within the MMU. Portion consists of page table entries that correspond to the most recently accessed pages.

A small cache called as Translation Lookaside Buffer (TLB) is included in the MMU. TLB holds page table entries of the most recently accessed pages. The cache memory holds most recently accessed blocks from the main memory. Operation of the TLB and page table in the main memory is similar to the operation of the cache and main memory. Page table entry for a page includes: Address of the page frame where the page resides in the main memory. Some control bits. In addition to the above for each page, TLB must hold the virtual page number for each page.

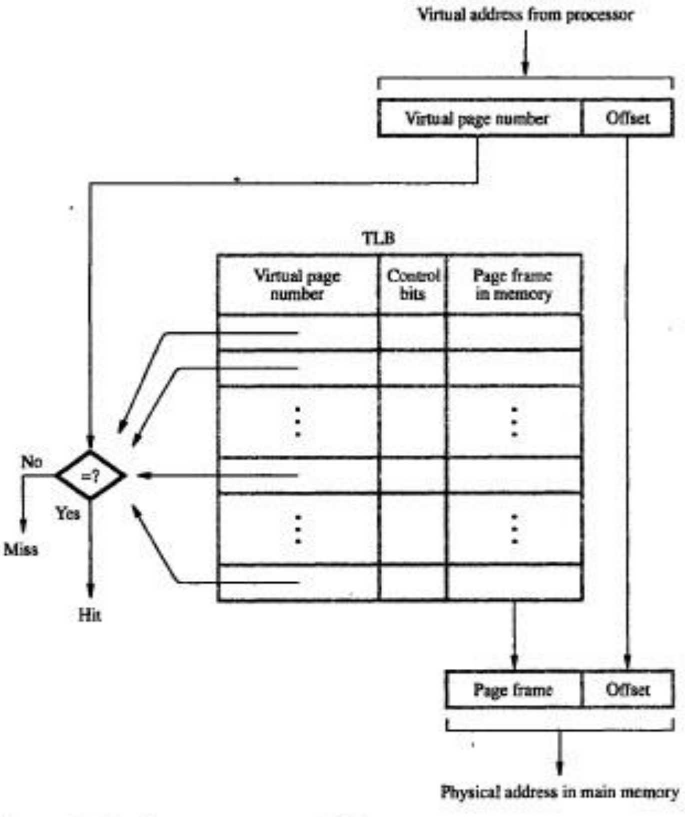


Figure 5.28 Use of an associative-mapped TLB.

Associative-mapped TLB: 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.
Set-associative mapped TLBs are found in commercial processors.

If a program generates an access to a page that is not in the main memory a page fault is said to occur. Whole page must be brought into the main memory from the disk, before the execution can proceed. Upon detecting a page fault by the MMU, following actions occur: MMU asks the operating system to intervene by raising an exception. Processing of the active task which caused the page fault is interrupted. Control is transferred to the operating system. Operating system copies the requested page from secondary storage to the main memory. Once the page is copied, control is returned to the task which was interrupted.

5.8 Secondary Storage

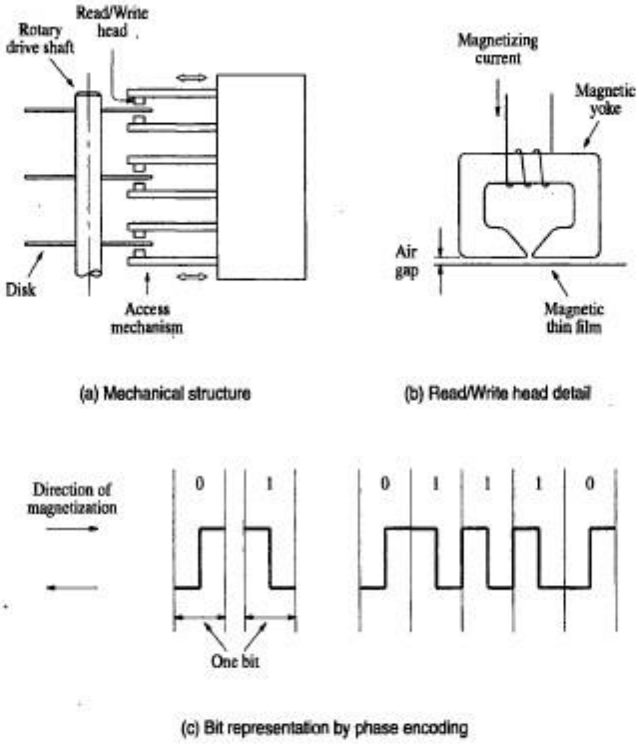
1. Magnetic Disk Drives: **Hard disk Drive organization:**

The modern hard disk drive is a system in itself. It contains not only the disks that are used as the storage medium and the read write heads that access the raw data encoded on them, but also the signal conditioning circuitry and the interface electronics that separate the system user from the details & getting bits on and off the magnetic surface. The drive has 4 platters with read/write heads on the top and bottom of each platter. The drive rotates at a constant 3600rpm.

Platters and Read/Write Heads: - The heart of the disk drive is the stack of rotating platters that contain the encoded data, and the read and write heads that access that data. The drive contains five or more platters. There are read/write heads on the top and bottom of each platter, so information can be recorded on both surfaces. All heads move together across the platters. The platters rotate at constant speed usually 3600 rpm.

Drive Electronics: - The disk drive electronics are located on a printed circuit board attached to the disk drive. After a read request, the electronics must seek out and find the block requested, stream is off of the surface, error check and correct it, assembly into bytes, store it in an on-board buffer and signal the processor that the task is complete. To assist in the task, the drive electronics include a disk controller, a special purpose processor.

Data organization on the Disk:- The drive needs to know where the data to be accessed is located on the disk. In order to provide that location information, data is organized on the disk platters by tracks and sectors. Fig below shows simplified view of the organization of tracks and sectors on a disk. The fig. shows a disk with 1024 tracks, each of which has 64 sectors. The head can determine which track it is on by counting tracks from a known location and sector identities are encoded in a header written on the disk at the front of each sector. The number of bytes per sector is fixed for a given disk drive, varying in size from 512 bytes to 2KB. All tracks with the same number, but as different surfaces, form a cylinder. The information is recorded on the disk surface 1 bit at a time by magnetizing a small area on the track with the write head. That bit is detected by sending the direction of that magnetization as the magnetized area passes under the read head as shown in fig below.



(a) Mechanical structure

(b) Read/Write head detail

(c) Bit representation by phase encoding

Figure 5.29 Magnetic disk principles.

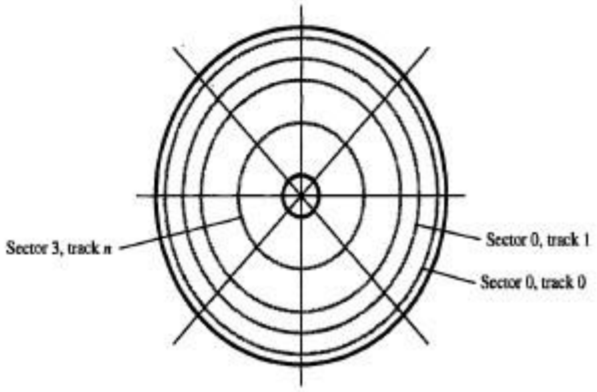


Figure 5.30 Organization of one surface of a disk.

The header usually contains both synchronization and location information. The synchronization information allows the head positioning circuitry to keep the heads centered on the track and the location information allows the disk controller to determine the sectors & identifies as the header passes, so that the data can be captured if it is read or stored, if it is a write. The 12 bytes of ECC (Error Correcting Code) information are used to detect and correct errors in the 512 byte data field.

Disk Drive Dynamic Properties: -

Dynamic properties are those that deal with the access time for the reading and writing of data. The calculation of data access time is not simple. It depends not only as the rotational speed of the disk, but also the location of the read/write head when it begins the access. There are several measures of data access times.

1. **Seek time: -** Is the average time required to move the read/write head to the desired track. Actual seek time which depend on where the head is when the request is received and how far it has to travel, but since there is no way to know what these values will be when an access request is made, the average figure is used. Average seek time must be determined by measurement. It will depend on the physical size of the drive components and how fast the heads can be accelerated and decelerated. Seek times are generally in the range of 8-20 m sec and have not changed much in recent years.

2. **Track to track access time: -** Is the time required to move the head from one track to adjoining one. This time is in the range of 1-2 m sec.

3. **Rotational latency: -** Is the average time required for the needed sector to pass under head once and head has been positioned once at the correct track. Since on the average the desired sector will be half way around the track from where the head is when the head first arrives at the track, rotational latency is taken to be ½ the rotation time. Current rotation speeds are from 3600 to 7200 rpm, which yield rotational latencies in the 4-8 ms range.

4. **Average Access time:-** Is equal to seek time plus rotational latency.

5. **Burst rate: -** Is the maximum rate at which the drive produces or accepts data once the head reaches the desired sector, It is equal to the rate at which data bits stream by the head, provided that the rest of the system can produce or accept data at that rate

Burst rate (byte/sec) = rows/sec * sector/row * bytes/sector

6. **Sustained data rate: -** Is the rate at which data can be accessed over a sustained period of time.

Optical Disks

Compact Disk (CD) Technology:- The optical technology that is used for CD system is based on laser light source. A laser beam is directed onto the surface of the spinning disk. Physical indentations in the surface are arranged along the tracks of the disk. They reflect the focused beam towards a photo detector, which detects the stored binary patterns.

The laser emits a coherent light beam that is sharply focused on the surface of the disk. Coherent light consists of Synchronized waves that have the same wavelength. If a coherent light beam is combined with another beam of the same kind, and the two beams are in phase, then the result will be brighter beam. But, if a photo detector is used to detect the beams, it will detect a bright spot in the first case and a dark spot in the second case.A cross section of a small portion of a CD shown in fig. below. The bottom payer is Polycarbonate plastic, which functions as a clear glass base. The surface of this plastic is Programmed to store data by indenting it with pits. The unindented parts are called lands. A thin layer of reflecting aluminium material is placed on top of a programmed disk. The aluminium is then covered by a protective acrylic. Finally the topmost layer is deposited and stamped with a label.

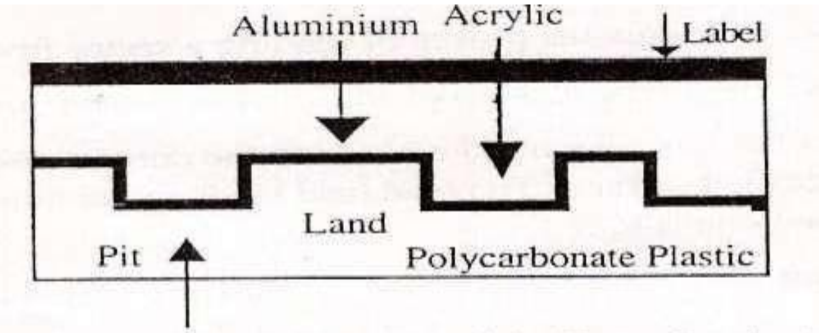


Fig. 16 Optical disk (Cross Section)

The laser source and the Photo detector are positioned below the polycarbonate plastic. The emitted bean travels through this plastic, reflects off the aluminium layer and travels back toward photo detector.

Some important optical disks are listed below

1. CD-ROM
2. CD-RWs (CD-re writables)
3. DVD technology (Digital Versatile disk)