Module 3
Memory system

Basic Concepts:

The maximum size of the memory that can be used in any computer is determined by its addressing scheme. For example, a 16-bit computer that generates 16-bit addresses can address up to 2^{16} =64K memory locations. If a machine generates 32-bit addresses, it can access up to 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.

Connection of the memory to the processor

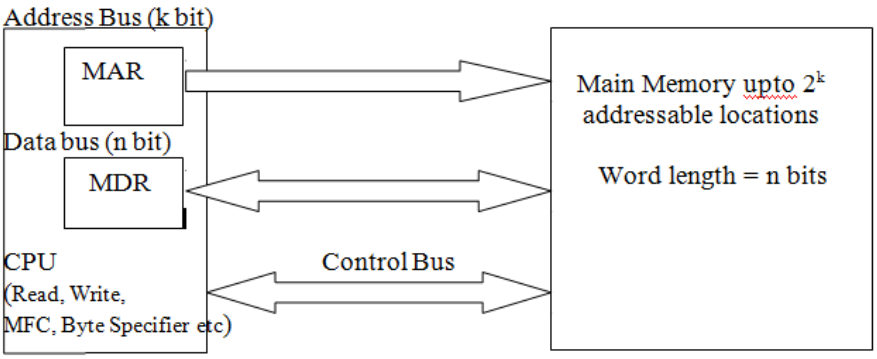
Data transfer between CPU and the memory takes place using two CPU registers, usually called MAR (Memory Address Register) and MDR (Memory Data Register) as shown in the below figure.

If MAR is k- bits long and MDR is 'n' bits long, then the memory unit may contain up to 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 memory 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 READ to 1 and sets other control signals if required. The data from the memory is loaded into MDR and MFC is set to 1.

For a write operation, MAR, MDR are suitably loaded by the CPU, write is set to 1 and other control signals are set suitably. The memory control circuitry loads the data into appropriate locations and sets MFC to 1. This organization is shown in the following block schematic



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.

Random Access Memory (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.

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. 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), cache hit rates of over 90% are possible leading to a cost-effective increase in the performance of the system.

SEMICONDUCTOR RAM MEMORIES:

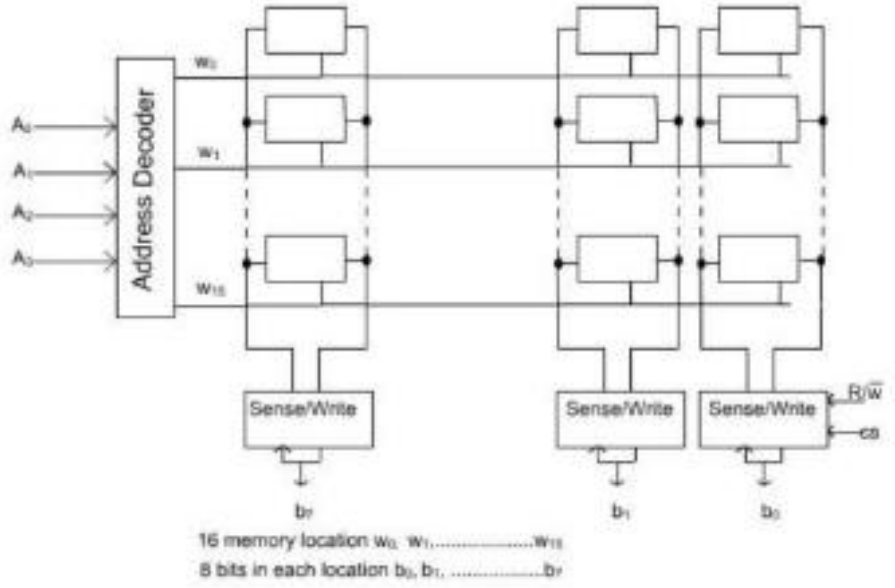
Internal Organization of Semiconductor Memory Chips: -

- Memory chips are usually organized in the form of an array of cells, in which each cell can store one bit of information.
- A row of cells constitutes a memory word, and the cells of a row are connected to a common line referred to as the word line, and this line is driven by the address decoder on the chip.
- The cells in each column are connected to a sense/write circuit by two lines known as bit lines.
- The sense/write circuits are connected to the data input/output lines of the chip.
- During a READ operation, the Sense/Write circuits sense, or read, the information stored in

the cells selected by a word line and transmit this information to the output lines.

- During a write operation, they receive input information and store it in the cells of the selected word.
- The following figure shows such an organization of a memory chip consisting of 16 words of 8 bits each, which is usually referred to as a 16 x 8 organization.
- The data input and the data output of each Sense/Write circuit are connected to a single bi-directional data line that can be connected to the data bus of a computer.
- One control line, the R/W (Read/Write) input is used a specify the required operation and another control line, the CS (Chip Select) input is used to select a given chip in a multi chip memory system.

Organization of bit cells in a memory chip

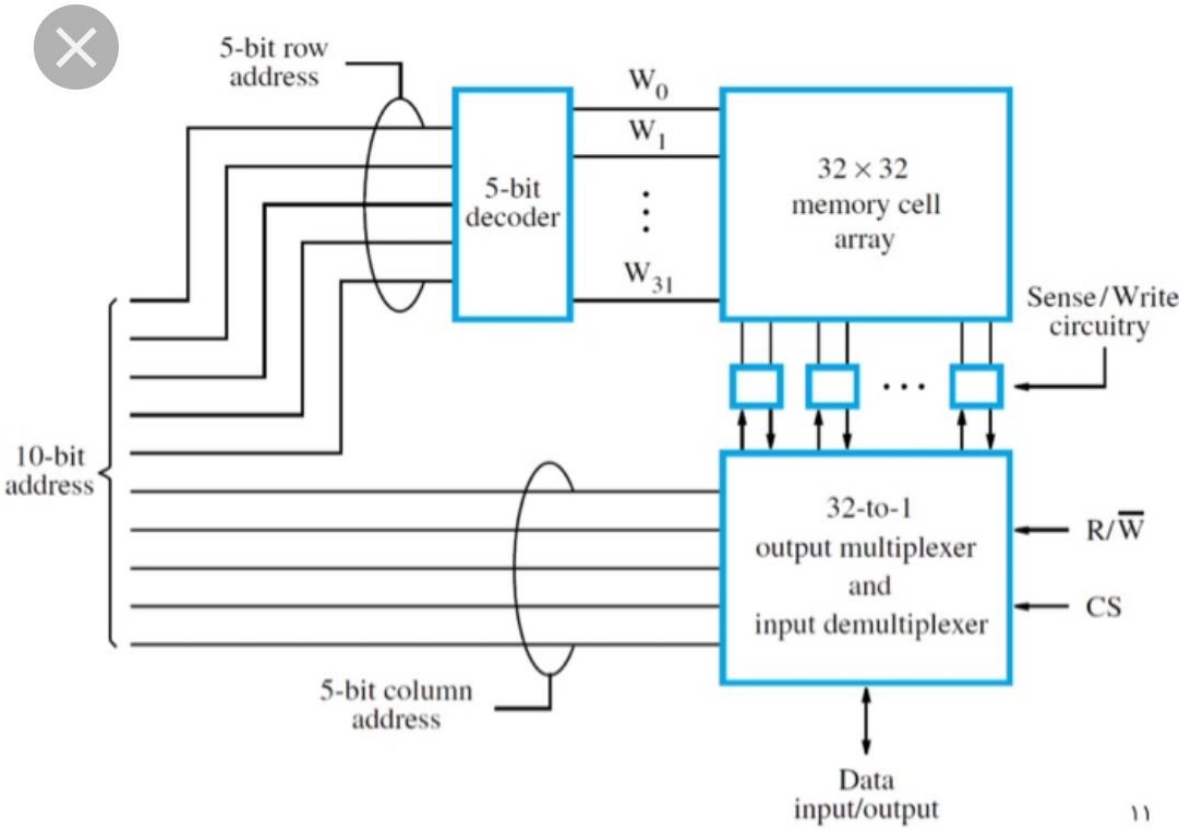


This memory circuit stores 128 bits and requires 14 connections (4 address lines, 8 data lines, 1 R/W, 1 CS), and allowing 2 pins for power supply and ground connections, can be manufactured in the form of a 16-pin chip.

Organization for 1k x 1 format is shown below:

- In this case a 10 bit address lines are needed, but there is only one data line, resulting in 15 external connections. Address lines=10; data line=1; R/W and CS = 2; Power supply & ground=2.
- 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 can be 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.



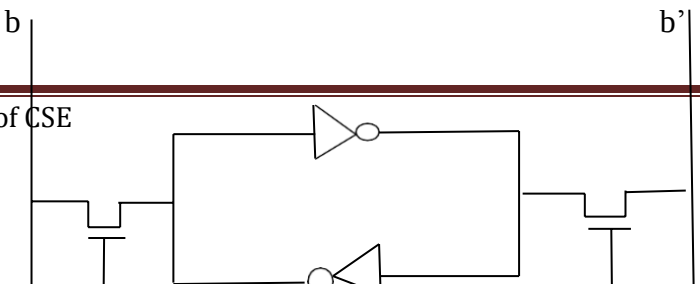
This structure can store 1024 bits, can be implemented in a 16-pin chip.

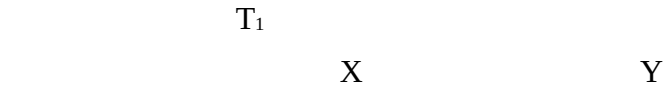
A larger memory cell array can have more external connections. For example, a 4M bit memory chip may have a 512K * 8 organization, in which 19 address and 8 data input/output pins are needed. Chips with a capacity of hundreds of megabits are now available.

Static Memories:

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

SRAM implementation:





- b and b’ are bit lines.
- 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 if the signal on the word line is at ground level.

Read Operation: -

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

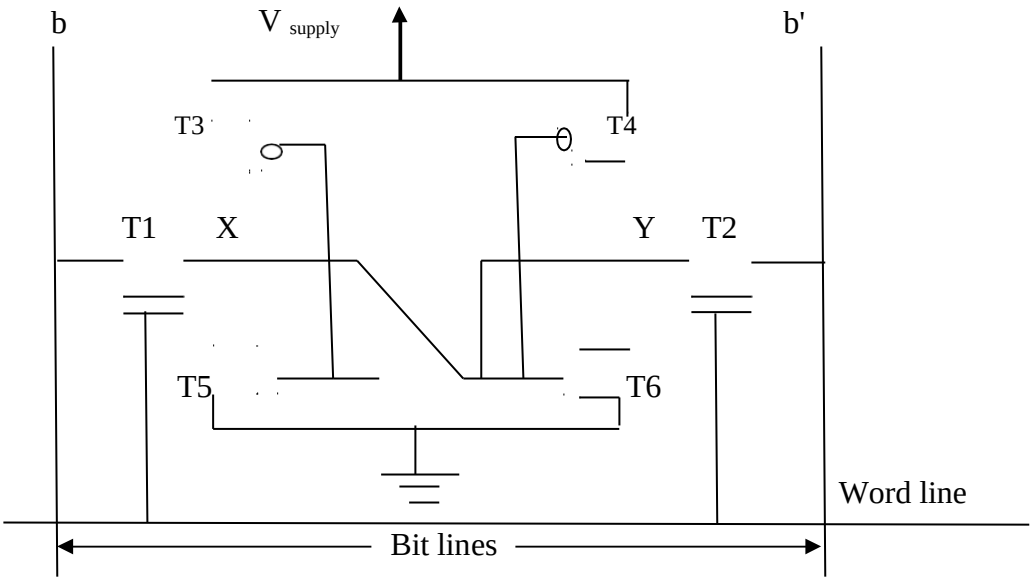
Write Operation: -

The state of the cell is set by placing the appropriate value on bit lines b and its complement on 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: -

Transistor pairs (T_3,T_5) and (T_4,T_6) form the inverters in the latch. The state of the cell is read or written as in SRAM. 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 T1 and T2 are turned on(closed), bit lines b and b' will have high and low signals, respectively.



An example of a CMOS memory cell.

Read Operation:-

The word line is activated to close switches T_1 and T_2. If the cell is in state 1, the signal on bit line b is high and the signal on bit line b' is low. The opposite is true if the cell is in state 0. Thus b and b' are complements of each other. The 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 lines b and its complement on 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.

The power supply voltage, V supply, is 5 V in older CMOS SRAMs or 3.3V in new low-voltage versions.

Note that continuous power is needed for the cell to retain its state. If power is interrupted, the cell's contents will be lost. When power is restored, the latch will settle into a stable state, but it will not necessarily be the same state the cell was in before the interruption.

Hence, SRAMs are said to be volatile memories because their contents are lost when power is interrupted.

Advantages:

CMOS SRAMs – very low power consumption because current flows in the cell only when the cell is being accessed.

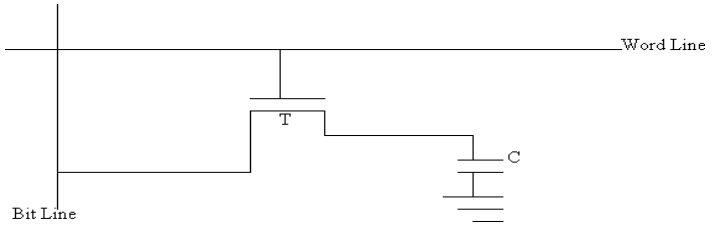
Static RAMs can be accessed very quickly. Access time – few nano seconds.

SRAMs are used in applications where speed is of critical concern.

Disadvantage: high cost because their cells require several transistors.

Dynamic Memories:-(DRAMs)

- DRAMs do not retain their state indefinitely.
- Information is stored in a dynamic memory cell in the form of a charge on the capacitor, and their 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.
- In order to store information in this cell, transistor T is turned on and an appropriate voltage is applied to the bit line. This causes a known amount of charge to be stored on the capacitor.
- After the transistor is turned off, the capacitor begins to discharge. This is caused by the capacitor's own leakage resistance. Hence the data is read correctly only if is read before the charge on the capacitor drops below some threshold value.



- During a Read operation, the transistor in a selected cell is turned on. A sense amplifier connected to the bit line detects whether the charge stored on the capacitor is above the threshold value. If so, it drives the bit line to a full voltage that represents logic value 1. this voltage recharges the capacitor to the full charge that corresponds to logic value 1.
- if the sense amplifier detects that the charge on the capacitor is below the threshold value, it pulls the bit line to ground level, which ensures that the capacitor will have no charge representing logic value 0.
- Thus, reading the contents of the cell automatically refreshes its contents. All cells in a selected row are read at the same time, which refreshes the contents of the entire row.

Asynchronous DRAMs:

Internal organization of a 2M * 8 dynamic memory chip:

- A 16-megabit DRAM chip is configured as 2M * 8 dynamic memory chip. The cells are organized in the form of a 4K * 4K array. The 4096 cells in each row are divided into 512 groups of 8, so that a row can store 512 bytes of data.
- 12 address bits are needed to select a row.
- 9 bits to specify a group of 8 bits in the selected row.
- Thus a 21-bit address is needed to access a byte in this memory. High order 12 bits and the

low order 9 bits of the address constitute the row and column addresses of the byte respectively. **To reduce the number of pins needed for external connections, the row and column addresses are multiplexed on 12 pins. If multiplexing is not used, the organization would need 21 pins to carry address.**

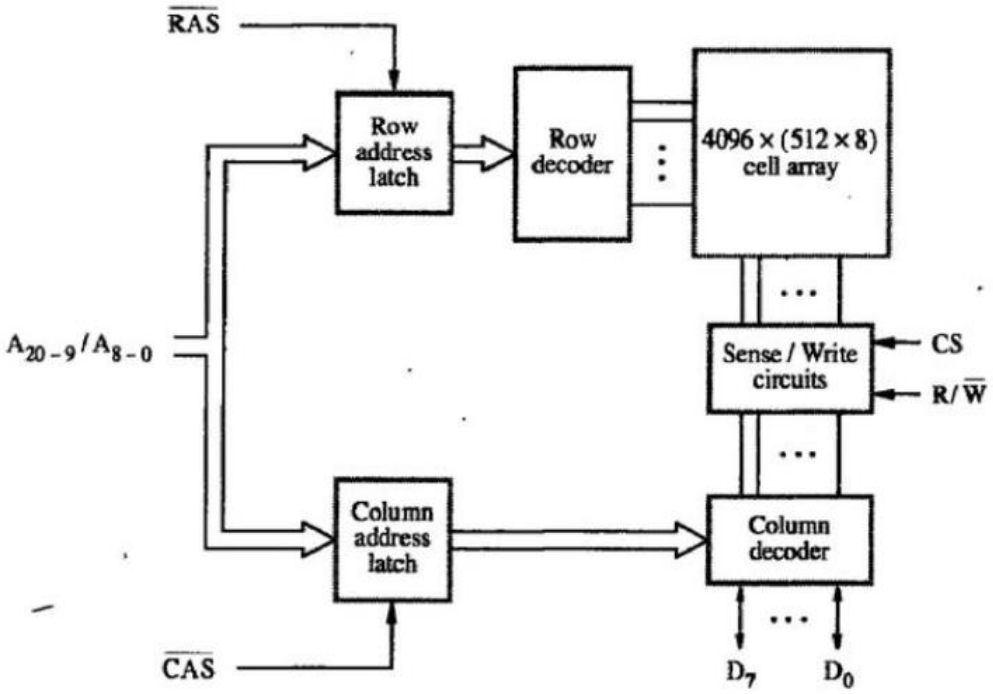


Figure: Internal organization of a 2M x 8 dynamic memory chip

Read or Write operation:

- The row address is applied first. It is loaded into the row address latch in response to a signal pulse on the Row Address Strobe (RAS) input of the chip. Then a read operation is initiated, in which all cells on the selected row are read and refreshed.
- Shortly after the row address is loaded, the column address is applied to the address pins and loaded into the column address latch under control of the Column Address Strobe (CAS) signal.
- The information in this latch is decoded and appropriate group of 8 Sense/Write circuits are selected.
- For read operation, the output values of the selected circuits are transferred to the data lines D7-0.
- For write operation, the information on the D7-0 lines are transferred to the selected circuits. This information is then used to overwrite the contents of the selected cells in the corresponding 8 columns.
- Applying a row address causes all cells on the corresponding row to be read and refreshed

during both read and write operations. A refresh circuit usually performs this function automatically.

- In DRAM, the timing of the memory device is controlled asynchronously. A specialized memory controller circuit provides the necessary control signals, RAS and CAS that govern the timing. The processor must consider the delay in the response of the memory.

Advantages:

----> high density, low cost, widely used in memory, Refreshing time not required.

Fast Page Mode:

- The contents of all 4096 cells in the selected row are sensed, but only 8 bits are placed on the data lines D7-0. This byte is selected by the column address bits A8-0.
- A simple modification can make it possible to access the other bytes in the same row without having to reselect the row.
- A latch can be added at the output of the sense amplifier in each column.
- The application of a row address will load the latch corresponding to all bits in the selected row.
- Then, it is only necessary to apply different bytes on the data lines.
- **The most useful arrangement is to transfer the bytes in sequential order, which is achieved by applying a consecutive sequence of column addresses under the control of successive CAS signals. This scheme allows a transferring a block of data at a much faster rate.**
- **The block transfer capability is referred to as the fast page mode feature.**

Synchronous DRAMs

DRAMs whose operation is directly synchronized with a clock signal are known as synchronous DRAMs (SDRAMs).

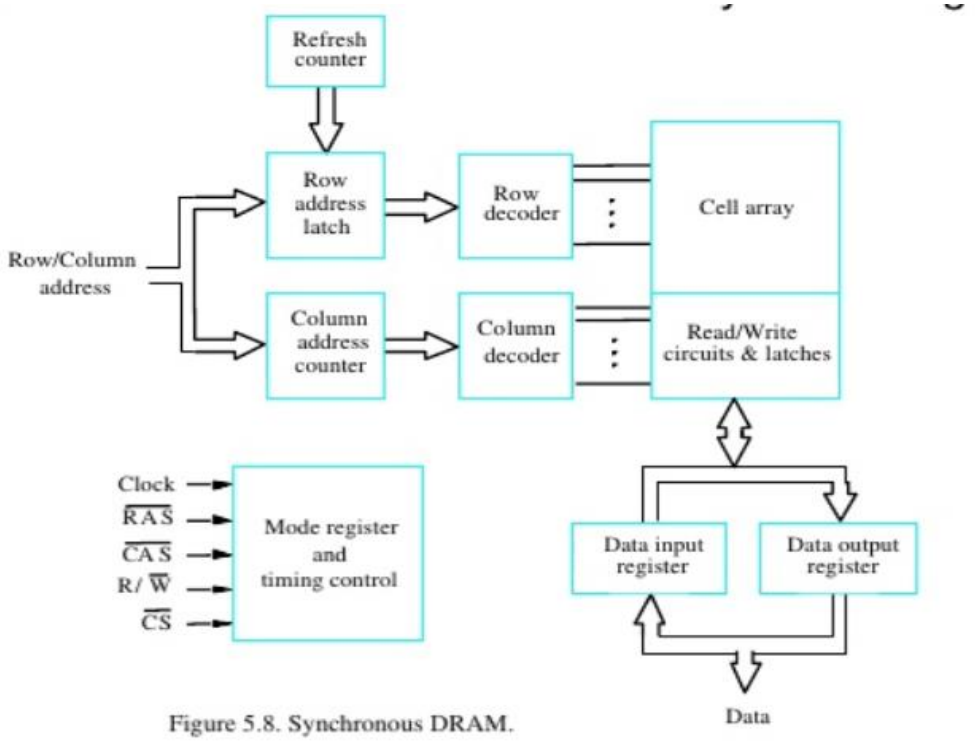


Figure 5.8. Synchronous DRAM.

The cell array is the same as in asynchronous DRAMS. The address and data connections are buffered by means of register. Output of each sense amplifier is connected to latch.

- Read operation causes the contents of all cells in the selected row to loaded into these latches. It will also refresh the contents of the cells.
- Data held in the latches that correspond to the selected column(s) are transferred into the data 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 can be specified.
- uses block transfer capability (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 cycle.
- **Timing diagram for a burst of length 4 in an SDRAM is shown below:**

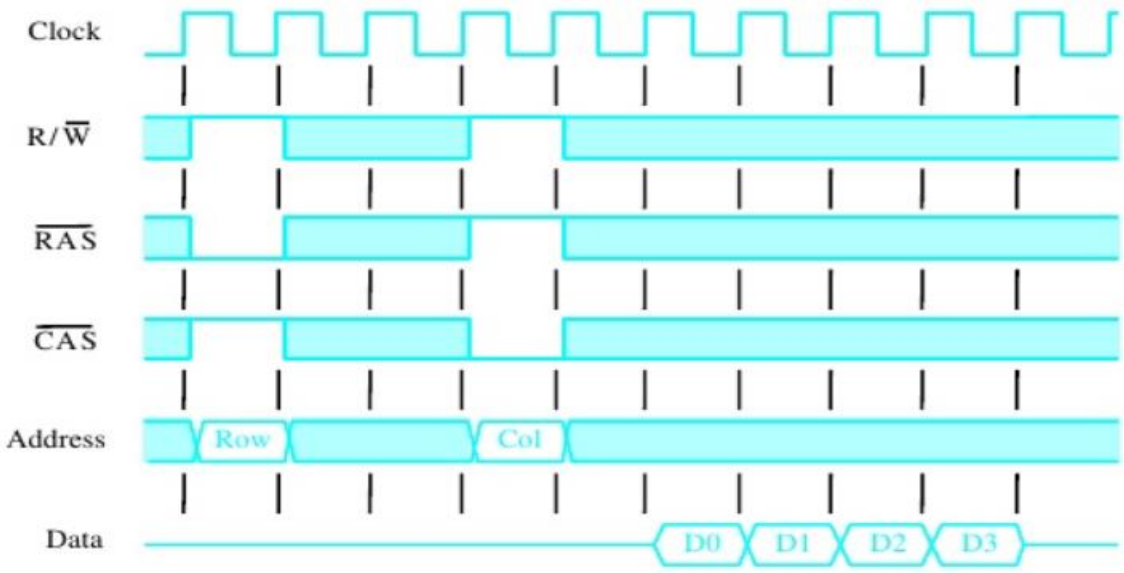


Figure 5.9. Burst read of length 4 in an SDRAM.

First, the row address is latched under the control of the RAS signal. The memory typically takes 2 or 3 clock cycles to activate the selected row. Then the column address is latched under the 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 the next three sets of bits in the selected row which are placed on the data lines in the next 3 clock cycles.

SDRAMs have built-in refresh circuitry. A part of this circuitry is a refresh counter, which provides the addresses of the rows that are selected for refreshing. Each row must be refreshed at least every 64 ms.

Latency and Bandwidth:

Memory latency – amount of time the memory takes to transfer a word of data to or from the memory.

In block transfers, latency is the time to transfer the first word of data. This time is usually longer than the time needed to transfer each subsequent word of a block.

Above timing diagram, latency is 5 clock cycle if the clock rate is 100 MHz, then the latency is 50ns.

Clock rate = 1 / P = 100 MHz where P = 1 /

100 MHz length of 1 clock cycle = 10 ns

latency = 5 * 10 = 50 ns

Memory Bandwidth- number of bits or bytes that can be transferred in one second.

it depends on the speed of access to the stored data and on the number of bits that can be accessed in parallel.

Speed of the memory also depends on the transfer capability of the links that connect the memory and the processor (ie speed of the bus).

Thus the bandwidth is the product of the rate at which data are transferred(accessed) and the width of the data bus.

Double-Data-Rate SDRAM (DDR SDRAM): (faster version)

Memory device which accesses the cell array in the same way as SDRAM, but transfers data on both edges of the clock.

Since it transfer data on both edges of the clock, their bandwidth is doubled for long burst transfers. Such devices are known as double-data-rate SDRAMs

Structure of Larger Memories:

Static Memory Systems:

Consider a memory consisting of 2M (2,097,152) words of 32 bits each. How we can implementation this memory using 512K * 8 memory chips.

→ Each column in the figure consists of four chips, which implement one-byte position. Four of these sets provide the required 2M * 32 memory.

→ Each chip has a control input called chip select. When this input is set to 1, it enables the chip to accept data from or to place data on its data lines.

→ The data output for each chip is of the three state type. Only the selected chip places data on the data output line, while all other outputs are in the high impedance state.

- Twenty-one address bits are needed to select a 32-bit word in this memory.
- the high order 2 bits of the address are decoded to determine which of the four chips select control signals should be activated, and the remaining 19 address bits are used to access specific byte locations inside each chip of the selected row.
- The R/W inputs of all chips are tied together to provide a common Read/Write control (not shown in the figure).

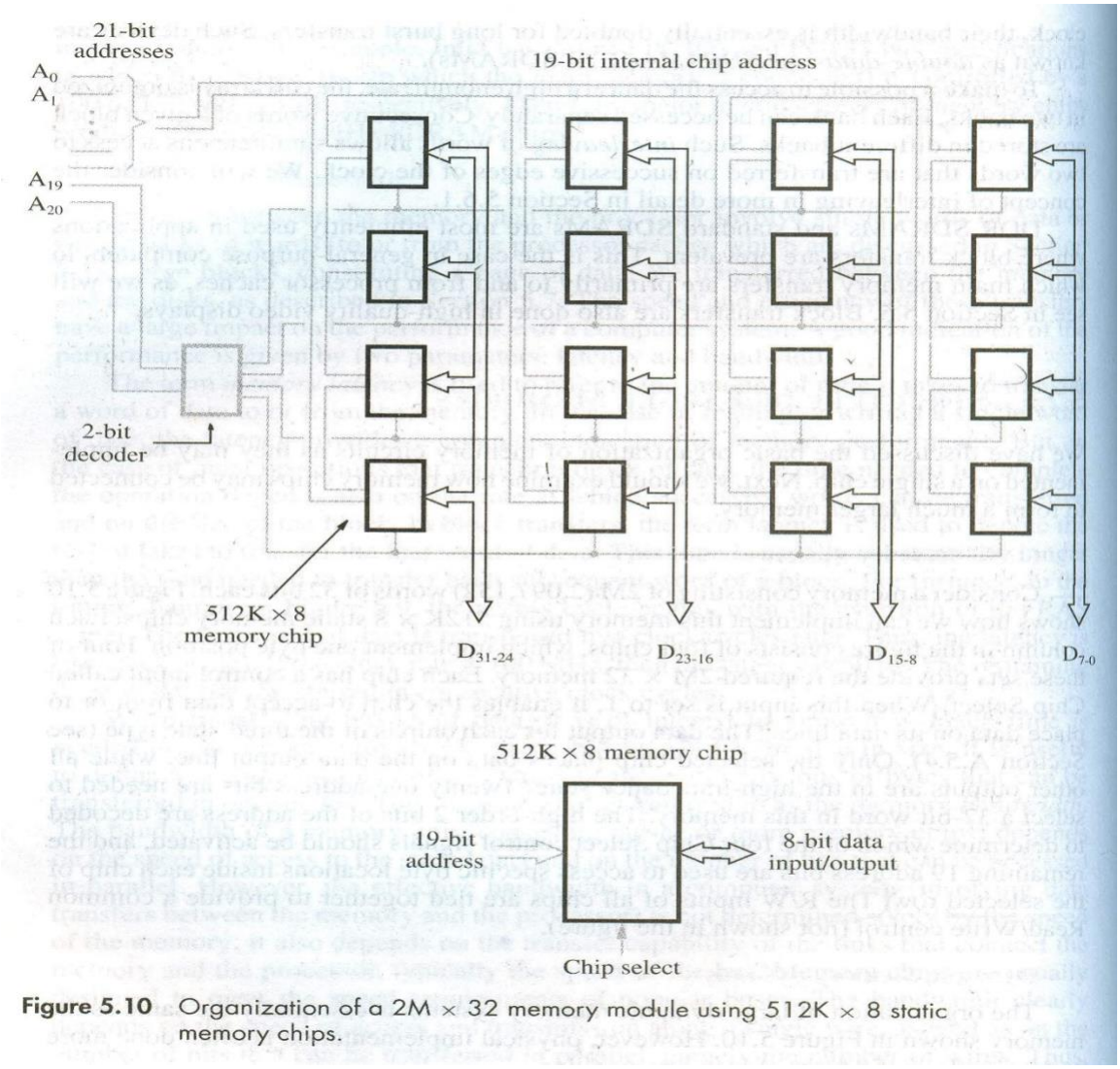


Figure 5.10 Organization of a 2M × 32 memory module using 512K × 8 static memory chips.

Total no. of chips = $\frac{2M * 32}{512K * 8} = \frac{2 * 1024 * 1024 * 32}{512 * 1024 * 8} = 16$

Dynamic Memory Systems:

The organization of large dynamic memory systems is essentially the same as the memory shown in the figure. However, physical implementation is often done more conveniently in the form of memory modules.

Large memories → better performance

Large memories are built by placing DRAM chips directly on the mother board, occupies an unacceptably large amount of space on the board. It is awkward to provide future expansion of the memory.

These 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).

Such a module is an assembly of several memory chips on a separate small board that plugs vertically into a single socket on the mother board.

Memory System Considerations:

RAM chip depends on the cost, speed, power dissipation and size of the chip.

Static RAMs - very fast operation (cost & size) - used in cache memories.

Dynamic RAMs – main memories.

Memory Controller:

To reduce the number of pins, the dynamic memory chips use multiplexed address inputs. High-order address bits, selects a row in the cell array, are provided first and latched into the memory chip under control of RAS signal. Low-order address bits, which select a column, are provided on the same pins and latched using CAS signal.

A typical processor issues all bits of an address at the same time. The required multiplexing of address bits is usually performed by a memory controller circuit, which is interposed between the processor and the dynamic memory.

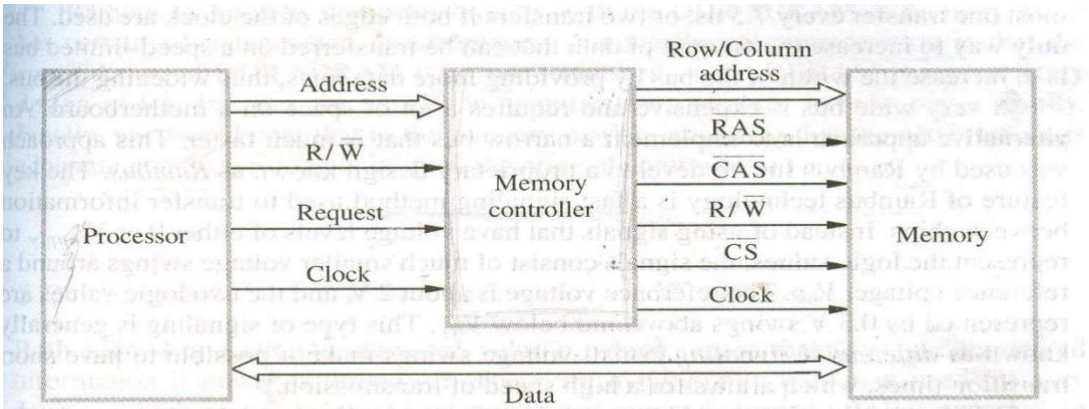


Figure 5.11 Use of a memory controller.

The controller accepts a complete address and the R/W signal from the processor, under control of a Request signal which indicates that a memory access operation is needed. The controller then forwards the row and column portions of the address to the memory and generates the RAS and CAS signals. Thus the controller provides the RAS and CAS timing, in addition to its address multiplexing function.

It also sends the R/W and CS signals to the memory. Data lines are connected directly between the processor and the memory. Clock signal – SDRAM chip. DRAM chips, which do not have self refreshing capability, the memory controller must provide all the information needed to control the refreshing process. It contains a refresh counter that provides successive row addresses. Its function is to cause the refreshing of all rows to be done within the period specified for a device.

Refresh overhead:

All dynamic memories must be refreshed. DRAMs, a typical period for refreshing all rows was 16 ms. In typical SDRAMs, a typical period is 64 ms. Consider, an SDRAM whose cells are arranged in 8K = 8192 rows. Suppose that it takes 4 clock cycles to access (read) each row. Then, it takes 8192 * 4 = 32768 cycles to refresh all rows.

Clock rate of 133 MHz, the required to refresh all rows = $1 / 133 * 10^6 * 32768 = 246 * 10^{-6}$ = 0.246 ms in each 64 ms time interval.

Refresh overhead = 0.246 / 64 = 0.0038 which is less than 0.4 percent of the total time available for accessing the memory.

Read Only Memories:

Both SRAM and DRAM chips are volatile, means that they lose the stored information if power is turned off.
When a computer is turned on, the operating system software has to be loaded from the disk to the memory. This requires execution of a program that boots the operating system. This program is quite large and it is stored on the disk. So, the processor must execute some instructions that load the boot program into the memory and the boot program loads the operating system.

- ROM provides a small amount of non volatile memory that holds the instructions whose execution results in loading the boot program from the disk. Non volatile memory is used extensively in embedded systems .
- Normal operation involves only reading of stored data, a memory of this type is called read only memory (ROM).

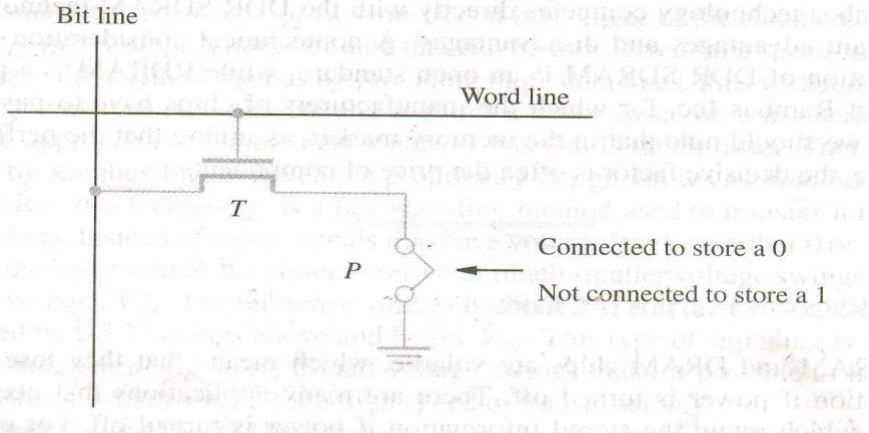


Figure 5.12 A ROM cell.

A logic 0 is stored in the cell if the transistor is connected to ground at point p; otherwise, a 1 is stored. The bit line is connected through a resistor to the power supply. Read word line is activated thus the transistor switch is closed and the voltage on the bit line drops to near zero if there is a connection between the transistor to ground. If there is no connection to ground, the bit line remains at the high voltage indicating a 1.

A sense circuit at the end of the bit line generates the proper output value. Data are written into a ROM when it is manufactured.

PROM: Programmable ROM

It allows the data to be loaded by the user. Programmability is achieved by inserting a fuse at point p. The memory contains all 0's before it is programmed. The user can insert 1's at the required locations by burning out the fuses at their locations using high-current pulses. Of course, this process is irreversible.

Cost is high (small numbers) high volumes (less expensive) PROMs provide a faster and considerably less expensive approach because they can be programmed directly by the user.

EPROM: Erasable PROM

Stored data can be erased and new data to be loaded. EPROMs are capable of retaining stored information for a long time. They can be used in place of ROMs. While software is being developed. In this way, memory changes and updates can be easily made.

An EPROM cell has a structure like ROM cell. Special transistor is used, which can function either as a normal transistor or as a disabled transistor that is always turned off. Thus, transistor can be programmed to behave as a permanent open switch, by injecting charge into it that becomes trapped inside. Thus, an EPROM cell can be used to construct a memory in the same way as the ROM cell.

Advantages:

- Contents can be erased and reprogrammed.
- Erasure requires dissipating the charges trapped in the transistor of memory cells; this can be done by exposing the chips to ultraviolet light.
- For this reason EPROM chips are mounted in packed that have transparent windows.

Disadvantage:

Chip must be physically removed from the circuit for reprogramming that its entire contents are erased by the ultraviolet light programmed and erased electrically.

EEPROM:

EEPROM do not have to be removed for erasure. Possible to erase the call contents selectively.
Disadvantage - different voltages are needed for erasing-writing and reading the stored data.

Flash memory: (similar to EEPROM technology)

- A flash cell is based on a single transistor controlled by trapped charge, just like an EEPROM cell. In EEPROM it is possible to read and write the contents of a single cell. In

flash device it is possible to read the contents of a single cell, but it is only possible to write an entire block of cells.

- Before writing the previous contents of the block are erased.
- Flash drives hence greater density, which leads to higher capacity and a lower cost per bit.
- They require a single power supply voltage and consume less power in this operation.
- Low power consumption of flash memory makes it use in portable equipment that is battery driven.

Applications:

1. Hand-held computer, cell phones- holds the s/w needed.
2. Digital camera- store picture image data.
3. Mp3 music players- store the data that represent sound.
4. Single flash chips do not provide sufficient storage capacity for the applications mentioned above.

Larger memory modules consisting of a number of chips are needed. Two popular choices for the implementation of such modules

1. flash cards 2. flash drives.

Flash cards:

Mount flash chips on a small card. Such flash cards have a standard interface that makes them usable in a variety of products. A card is simply plugged into a conveniently accessible slot. Flash cards come in a variety of memory sizes. 8, 32, and 64 Mbytes. 1 min of music- 1 Mbyte of memory using the mp3 encoding format. 64 MB flash card can store an hour of music.

Flash drives:

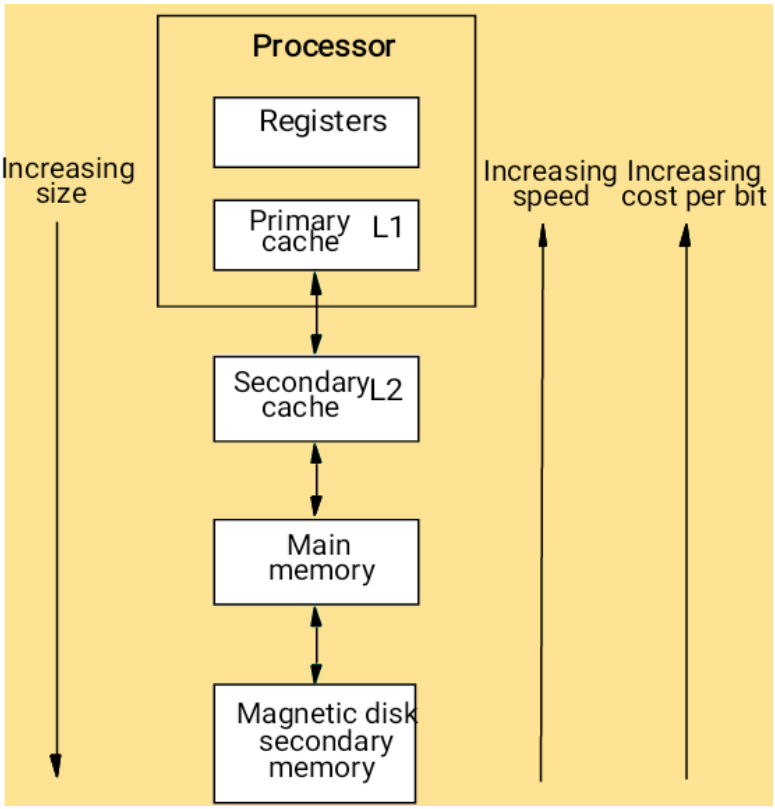
Larger flash memory modules have been developed to replace hard disk drives. However, the storage capacity of flash drives is significantly lower (1 giga byte). Hard disk can store many gigabytes. Flash drives are solid state electronic devices that have no movable parts provides some important **advantages.**

- Short seek and access times=faster response.
- Lower power consumption (for battery driven applications).
- Insensitive to vibration.

Disadvantages:

- Smaller capacity and higher cost per bit while disk are extremely low cost per bit.
- Flash memory will deteriorate after it has been written several times. (at least one million times).

Memory hierarchy



- Ideal memory is the one which is fast, large and inexpensive. Fast memory can be implemented if SRAM chips are used. But it is expensive, because its basic cell contains 6 transistors. So, it is impractical to build large memory using SRAM chips.
- Use of dynamic RAM chips is less expensive and significantly slower.
- Secondary storage- magnetic disk, large memory spaces available at a reasonable price, they are used extensively in computer systems.
- Main memory can be built with dynamic RAM.
- Cache memories: A smaller unit where speed is of the essence uses SRAMs.
- The fastest access is to data held in **processor registers**. This is the top in terms of the speed of access.
- The next level of the hierarchy is a relatively small amount of memory that can be implemented directly on the processor chip, called a processor cache, holds copies of instructions and data stored in a much larger memory that is provided externally. The primary cache is referred to as level 1 (L1) cache.
- A larger, secondary cache is placed between the primary cache and the rest of the memory. It

is referred to as level 2(L2) cache. It is implemented using SRAM chips.

- It is possible not to have a cache on the processor chip at all. Also, it is possible to have L1 and L2 caches on the processor chip.

- The next level hierarchy is called the main memory in the form of SIMMs, DIMMs or RIMMs (Rambus inline memory modules). This is slower than the cache memory and the access time for the main memory is about ten times larger than the access for the L1 cache.

- Disk devices provide a huge amount of inexpensive storage. They are very slow compared to the semiconductor devices used to implement the main memory.

Cache Memory:

The speed of the main memory is very low in comparison with the speed of modern processors.

- Processor cannot spend much of its time waiting to access instructions and data in main memory.
- Speed of main memory unit is limited by electronic and packaging constraints

Solution is to use a fast cache memory which essentially makes the main memory appear to the processor to be faster than it really is.

- The effectiveness of the cache mechanism is based on a property of computer programs called **locality of reference.**
- Many instructions in localized areas of the program are executed repeatedly during same time period and the remainder of the program is accessed relatively infrequently. This is referred to as locality of reference.

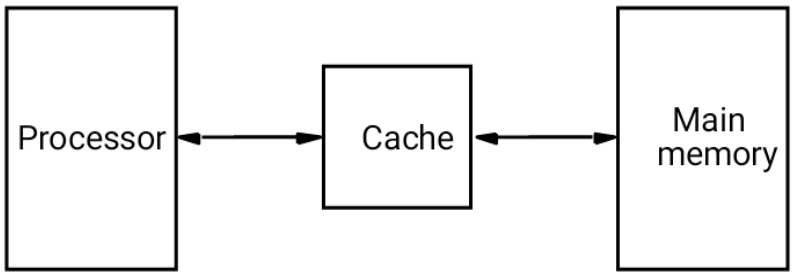
  - Locality of reference is done it two ways: - Temporal and Spatial

Temporal: - Recently executed instruction is likely to be executed again very soon.

Spatial: - Instructions in close proximity to a recently executed instruction (with respect to the instruction addresses) are also likely to be executed soon.

Block refers to a set of contiguous address locations of same size.

Use of a cache memory:



When a read request is received from the processor, the contents of a block of memory words containing the location specified are transferred into the cache one word at a time.

Cache hit: the fraction of memory accesses found in cache memory

Cache miss:

- When the address word in a read operation is not in the cache, a **read miss** occurs.
- During read miss, the block of words that contains the requested word is copied from the main memory into the cache.
- After the entire block is loaded into the cache, the particular word requested is forwarded to the processor.
- Alternatively, the word may be sent to the processor as soon as it is read form the main memory. This approach, which is called **load-through or early restart**, reduces processor's waiting period somewhat, but at the expense of more complex circuitry.
- During a write operation, if the address word is not in the cache, a **write miss** occurs. Then if write-through is used, the block containing the addressed word is first brought into the cache, and then the desired word in the cache is overwritten.

For a write operation, the system can proceed in two ways:

- **Write through protocol**: Cache location and the main memory location are updated simultaneously.

- **Write back or copy back protocol**: Update only the cache location and to mark it as updated with an associated flag bit called 'dirty' or 'modified' bit. The main memory location of the word is updated when the block containing this marked word is to be removed from the cache to make room for a new block.

Mapping functions:

Consider a cache consisting of 128 blocks of 16 words each, total of 128x16=2048(2k) words. Main memory 16-bit address, therefore the main memory has 64k words, which is viewed as 4k

blocks of 16 words each. (Consecutive addresses refer to consecutive words.)

Direct mapping:

- Simplest way to determine cache locations in which to store memory blocks is the direct mapping technique.
- In this technique, block j of the memory maps onto block ***j modulo 128*** of the cache.
- Whenever one of the main memory blocks 0,128,256... is loaded in the cache, it is stored in cache block 0.
- Block 1,129,257... are stored in cache block 1 and so on.
- Since more than one memory block is mapped onto a given cache block position contention may arise for that position even when the cache is not full.

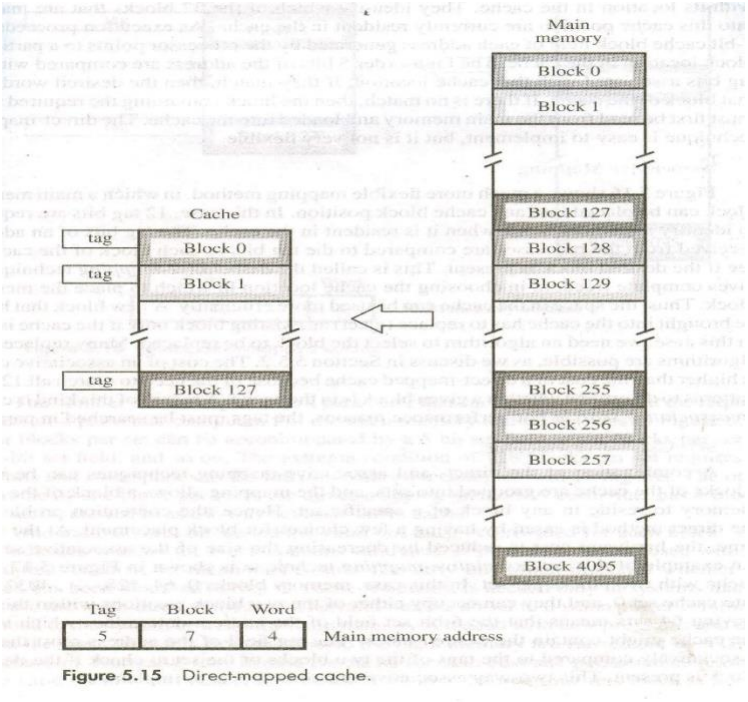


Figure 5.15 Direct-mapped cache.

For example, instructions of a program are in block 1 and block 129. As this program is executed, both of these blocks must be transferred to the block 1 position in the cache. (1%128=1 and 129%128=1). Contention is resolved by allowing the new block to overwrite the currently resident block.

The memory address can be divided into three fields.

- The low order 4 bits select one of 16 words in a block.
- When a new block enters the cache, the 7-bits are used to identify the cache block. (128 cache blocks= 2^7)
- 4096/128 =32=2^5, hence 5 bits are needed to represent the tag field. This field specifies how many main memory blocks can be mapped to one cache block.

Associative mapping:

- More flexible mapping method, in which a main memory block can be placed anywhere in the cache block.
- 12 tag bits are required to identify a memory block when it is resident in the cache. (2^{12}=4096)
- The tag bits of an address received form the processor are compared to the tag bits of each block of the cache to see if the desired block is present.
- It gives complete freedom in choosing the cache location in which to place the memory block.
- A new block that has to be brought into the cache has to replace (eject) an existing block only if the cache is full. It needs an algorithm to select the block to be replaced.
- Cost is higher compared to direct mapped cache, because need to search all 128 tag patterns to determine whether a given block is in the cache.

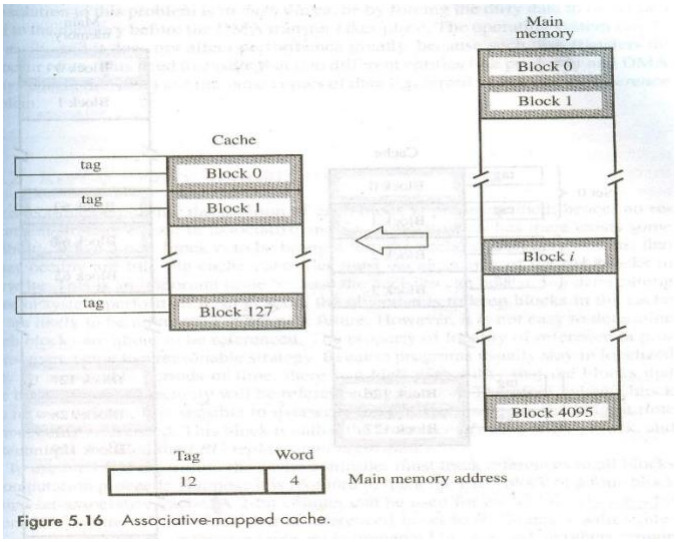


Figure 5.16 Associative-mapped cache.

Set Associative mapping:

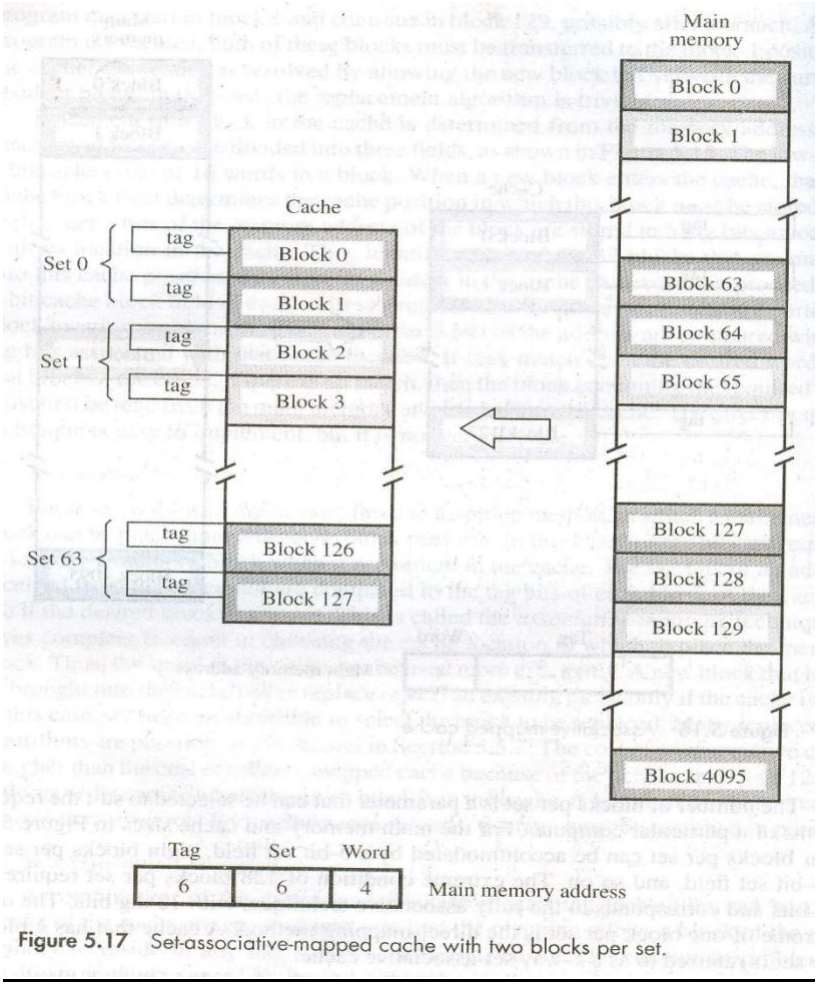


Figure 5.17 Set-associative-mapped cache with two blocks per set.

- A combination of the direct and associative mapping techniques can be used. Blocks of cache are grouped into sets, and the mapping allows a block of specified set.
- Hence, the contention problem of the direct method is eased by having a few choices for block placement. At the same time, the hardware cost is reduced by decreasing the size of the associative search.
- Example: cache with two blocks per set memory block 0,64,128...4032 map into cache set 0 and they can occupy either of the two block positions within this set.
- Total address lines contains 16 bits, divided into:
  Word: 4 bits (2^4 =16 words per block)
  Set : 64-sets are possible out of 128 cache blocks; hence 6-bit set field of the address determines which set of the cache might contain the desired blocks.
  Tag : The tag field of the address must then be associatively compared to the tags of the two blocks of the set to check if the desired block is present. 64 main memory blocks can map to one cache block i.e. 4096/64=64=2^6 , hence 6 bits are needed.

If the number of blocks per set is changed, then set field changes accordingly.

- 4 blocks per set Set field 5 bit (32 sets)
- 8 blocks per set set field 4 bit (16 sets)
- Extreme conditions 128 blocks per set, then no set bit and full association, 12 tag

bits. 1 block per set- direct mapping.

Cache that has **k** blocks per set referred to as a **k-way** set associative cache.

Control bit valid bit must be provided for each block indicates whether the block contains valid data.

Replacement Algorithms:

when a block is to be overwritten, 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 the technique is called the LRU replacement algorithm.

To use LRU algorithm the cache controller must track references to all blocks as computation proceeds. Suppose it is required to track the LRU block of a four-block set in a set-associative cache.

Interleaving:

If the main memory of a computer is structured as a collection of physically separate modules, each with its own address buffer register (ABR) and data buffer register (DBR), memory access operations may proceed in more than one module at the same time. Thus the aggregate rate transmission of words to and from the main memory system can be increased.

How individual addresses are distributed over the modules determine the average number of modules that can be kept busy as computation proceed.
Two methods of address layouts are present.

1) The memory address generated by the processor is decoded as high order **'k- bits'** to name one of the n modules and the low order **'m- bits'** name a particular word in that module.

   When consecutive locations are accessed or when a block of data is transferred to a cache, only one module is involved and at the same time devices with DMA ability may be accessing information in other memory modules. Hence processor accesses a single module and DMA device simultaneously accessing other memory modules. (fig a)

2) Second and more effective way to address the modules is called **memory interleaving.** The lower order **'k- bits'** of the memory address select a module and the higher order **'m-bits'** name a location within that module.

   Here consecutive addresses are located in successive modules.

   Thus any component of the system that generates requests for access to consecutive memory locations can keep several modules busy at any one time. (fig b)

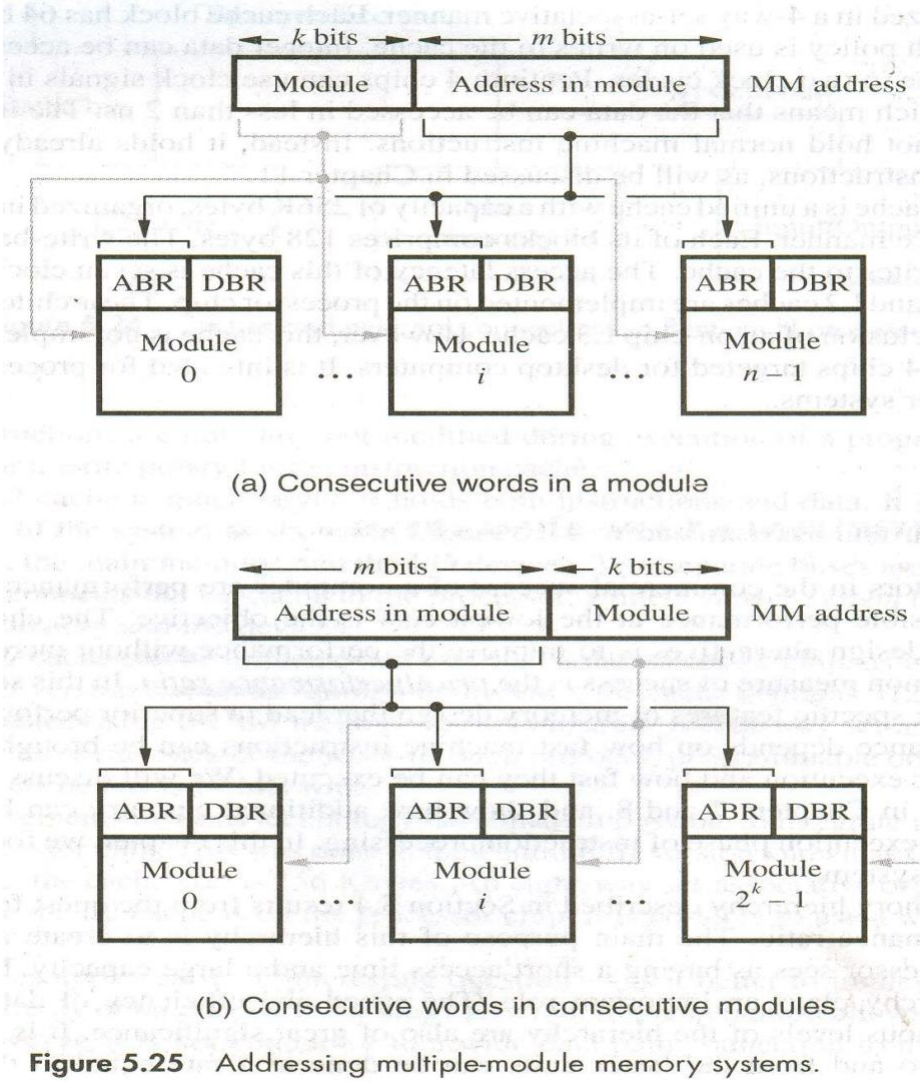


Figure 5.25 Addressing multiple-module memory systems.

This results in both faster accesses to a block of data and higher average utilization of the memory system as a whole. To implement the interleaved structure, there must be 2k modules; otherwise there will be gaps of nonexistent locations in the memory address space.

Hardware properties:

It takes 1 clock cycle to send an address to the main memory which is built with DRAM chip. First word can be accessed in 8 cycles; subsequent words of the block are accessed in 4 clock cycles per word. 1 cycle is needed to send one word to the cache.

If a single memory module is used, then the time needed to load the desired block into the cache is 1 + 8 + (7 * 4) + 1 = 38 cycles

Suppose now that the memory is constructed as 4 interleaved modules.

When the starting address of the block arrives at the memory, all four modules begin accessing the required data, using the high-order bits of the address. After 8 clock cycles, each module has one word of data in its DBR. These words are transferred to the cache, one word at a time, during the next 4 clock cycles. During this time the next word in each module is accessed. Then , it takes another 4 cycles to transfer these words to the cache.

Total time needed to load the block from the interleaved memory is 1 + 8 + 4+ 4 = 17 cycles.

Thus interleaving reduces the block transfer time by more than a factor of 2.

Hit Rate and Miss Penalty:

- Successful access to data in a cache is called **hit**.
- The no. of hits stated as a fraction of all attempted accesses is called **hit rate** and the
- **miss rate** is the no. of misses stated as a fraction of attempted accesses.
- The extra time needed to bring the desired information into the cache is called **miss penalty**.

Let 'h' be the hit rate, M be the miss penalty ,that is the time to access the information in the MM(main memory) and C the time to access information in the cache. Then the average access time experienced by the processor is

$$t_{avg} = hC + (1-h) M$$

If the computer has no cache,then using a fast processor and typical DRAM main memory, it takes 10 clock cycles for each memory read access.

Suppose the computer has a cache that holds 8-word blocks and an interleaved main memory. Then 17 clock cycles are needed to load a block into the cache

- Assume that 30% of the instructions in a typical program perform a read and write operation,which means that there are 130 memory accesses for every 100 instructions executed. Assume that the hit rates in the cache are 0.95 for instruction and 0.9 for the data. Assume that the miss penalty is the same for both read and write accesses.

Then,a rough estimate of the improvement in performfance that results from using the cache can be obtained as follows

$$\frac{\text{Time without cache}}{\text{Time with cache}} = \frac{(130*10)}{100 (95 * 1 + 0.05 * 17) + 30 (0.9 * 1 + 0.1 * 17)} = 5.04$$

- This result suggests that the computer with the cache performs 5 times better.

Comparison to an ideal cache that has a hit rate of 100% (in which case ,all memory references

take one cycle). Rough estimateof relative performance of the cache is

$$\frac{100\,(0.95 * 1 + 0.05 * 17) + 30\,(0.9 * 1 + 0.1 * 17)}{130} = 1.98$$

This means the actual cache provides an environment in which the processor effectively works with a large DRAM based main memory that appears to be only 2 times slower than the circuit in the cache.

Caches on the processor chip:

From the speed point of view, the optimal place for the cache is on the processor chip. Unfortunately, space on the processor chip is needed for many other functions; this limits the size of the cache that can be accommodated.

1. Separate caches, one for instructions another for data.(More expenses complex circuit-possible to access both the caches at the same time).
2. Combined cache: better hit rate, greater flexibility in mapping new information into the cache.

L1 cache(s) is on the proceesor chip. L1 is designed for very fast access by the processor and a smaller L2 cache can be on the processor chip. L1-tens of kilobytes.

L2 cache (SRAM chip) can be slower, larger to ensure a high hit rate. It only affects the miss penalty of the L1 cache. L2-several megabytes.

The average access time experienced by the proc in a system with 2 levels of cache is

tavg = h1C 1 + (1-h1) h2C2 + (1-h1) (1-h2) M

where h1-hit rate in the L1 cache

h2-hit rate in the L2 cache

C1-time to access information in the L1 cache

C2-time to access information in the L2 cache

M-time to access information in the MM memory

Other enhancements:

1. Write buffer
2. Prefetching
3. Lock up free cache

1. Write buffer:

In **Write through protocol** processor has to wait for the write request to complete. To improve performance, a write buffer can be included for temporary storage of write requests. Processor places each write request into this buffer and continues execution of the next instruction. Write requests stored in the write buffer are sent to the Main memory whenever the memory is not responding to the read requests.

It is important that the read request be serviced immediately because the processor usually cannot proceed without the data that are to be read from the memory.

Read request may refer to data that are still in the write buffer address of the data to be read from the memory are compared with the addresses of the data in the write buffer. In case of the match, the data in the write buffer are used.

In Write-back protocol: Write operation is simply performed on the corresponding word in the cache. New block as the result of read miss, which replaces an existing block that has some dirty data.

Dirty block has to be written into the main memory. Write-back performed first, then the processor will have to wait longer for the new block to be used into the cache.

2) Prefetcing:

When miss occurs, the processor has to pause until the new data arrive, which is the effect of the miss penalty. To avoid stalling the processor, it is possible to prefetch the data into the cache before they are needed. The simplest way to do this is through software. A prefetch instruction is inserted in the program to cause the data to be loaded in the cache by the time they are needed in the program, the compiler inserts the instructions.

3) **Lock up free cache**:

If the action of prefetching stops other accesses to the cache until the prefetch is completed, this type of cache is said to be locked while it serves a miss. It should allow the processor to access the cache while a miss is being serviced. It is desirable that more than one outstanding miss can be supported.

A cache that can support multiple outstanding misses is called lock-up free. Since it can service only one miss at a time it must include circuitry that keeps track of all outstanding misses. This may be done with special registers that hold pertinent information about these misses.