Unit 9 Memory Unit – Part I

Structure:

9.1 Introduction

Memory unit is used for storage, and retrieval of data and instructions. A typical computer system is equipped with a hierarchy of memory subsystems, some internal to the system and some external. Internal memory systems are accessible by the CPU directly and external memory systems are accessible by the CPU by an I/O module.

Objectives:

By the end of Unit 9, you should be able to:

1. Define the different units of transfer of data.
2. Explain the various accessing methods
3. Explain with neat sketches various memory organization.
4. Discuss the design of memory subsystem using dynamic memory chips.

9.2 Characteristics of Memory Systems

Memory systems are classified according to their key characteristics. The most important are listed below:

Location

The classification of memory is done according to the location of the memory as:

- ***CPU*:** The CPU requires its own local memory in the form of registers and also the control unit requires local memories which are fast accessible. We have already studied this in detail in our earlier discussions.
- ***Internal (main)*:** It is often equated with the main memory. There are other forms of internal memory. We will be discussing the internal memory in the next coming sections of this unit.

  ***External (secondary)*:** It consists of peripheral storage devices like hard disks, magnetic disks, magnetic tapes, CDs etc.

Capacity

Capacity is one of the important aspects of the memory.

Word size: Word size is the natural unit of organization of memory. The size of the word is typically equal to the number of bits used to represent a number and is equal to the instruction length. But there are many exceptions. Common word lengths are 8, 16 and 32 bits.

Number of words: The addressable unit is the word in many systems. However external memory capacity is generally expressed in terms of bytes.

Unit of Transfer

Unit of transfer for internal memory is equal to the number of data lines into and out of memory module.

- **Word:** For internal memory, unit of transfer is equal to the number of data lines into and out of the memory module. Need not be equal to a word or addressable unit.
- **Block:** For external memory, data are often transferred in much larger units than a word, and these are referred to as blocks.

Access Method

- **Sequential:** Tape units have sequential access. Data are generally stored in units called ***"records".*** Data is accessed sequentially; the records may be passed (or rejected) until the record that is searched for is found. The access time to a certain record is highly variable.
- **Direct:** Individual blocks or records have a unique address based on physical location. A block may contain a group of data. Access is accomplished by direct address to reach general vicinity, plus sequential searching, counting or waiting to reach the final location. Disk units have direct access.
- **Random:** Each addressable location in memory has a unique, physically wired-in addressing mechanism. The time to access a given location is independent of the sequence of prior accesses and constant. Any location can be selected at random and directly addressed and accessed. Main memory and some cache systems are random access.
- **Associative:** This is a random-access type of memory that enables one to make a comparison of desired bit locations within a word for a specified match, and to do this for all words simultaneously. Thus, a word is retrieved based on a portion of its contents rather than its address. Some cache memories may employ associative access.

Performance

- **Access time:** For random-access memory, this is the time it takes to perform a read or write operation. That is, the time from the instant that

an address is presented to the memory to the instant that data have been stored or made available for use. For non-random-access memory, access time is the time it takes to position the read-write mechanism at the desired location.

- **Cycle time:** Applied to random-access memory. It consists of the access time plus any additional time required before a second access can commence.
- **Transfer rate:** This is the rate at which data can be transferred into or out of a memory unit. For random-access memory, it is equal to (1/<cycle-time>).

For non-random-access memory, the following relationship holds:

Tn = Ta + N/R

where Tn = Average time to read or write N bits;

Ta = Average access time,

N = Number of bits

R = Transfer rate, in bits per second (bps).

Physical Type

- **Semiconductor:** Main memory, cache. RAM, ROM.
- **Magnetic:** Magnetic disks (hard disks), magnetic tape units.
- **Optical:** CD-ROM, CD-RW.
- **Magneto-Optical:** The recording technology is fundamentally magnetic. However an optical laser is used. The read operation is purely optical.

Physical Characteristics

- **Volatile/Non-volatile:** In a volatile memory, information decays naturally or is lost when electrical power is switched off. In a non-volatile memory, information once recorded remains without deterioration until deliberately changed; no electrical power is needed to retain information.

Magnetic-surface memories are nonvolatile. Semiconductor memories may be either volatile or non-volatile.

- **Erasable/Non-erasable:** Non-erasable memory cannot be altered (except by destroying the storage unit). ROMs are non-erasable.

Memory Hierarchy

Design constraints: How much? How fast? How expensive?

- Faster access time, greater cost per bit
- Greater capacity, smaller cost per bit,
- Greater capacity, slower access time.

9.3 Main Memory

The main memory stores data and instructions. Main memories are usually built from dynamic IC's known as dynamic RAMs. These semiconductor ICs can also implement static memories referred to as static RAMs (SRAMs). SRAMs are faster but cost per bit is higher. These are often used to build caches.

Types of Random-Access Semiconductor Memory:

Dynamic RAM (DRAM): Example: Charge in capacitor. It requires periodic refreshing.

Static RAM (SRAM): Example: Flip-flop logic-gates. Applying power is enough (no need for refreshing). Dynamic RAM is simpler and hence smaller than the static RAM. Therefore they are denser and less expensive. But it requires supporting refresh circuitry. Static RAMs are faster than dynamic RAMs.

ROM: The data is actually wired in the factory. It can never be altered.

PROM: Programmable ROM. It can only be programmed once after its fabrication. It requires special device to program.

EPROM: Erasable Programmable ROM. It can be programmed multiple times. Whole capacity need to be erased by ultraviolet radiation before a new programming activity. It cannot be partially programmed.

EEPROM: Electrically Erasable Programmable ROM. Erased and programmed electrically. It can be partially programmed. Write operation takes considerably longer time compared to read operation.

Each more functional ROM is more expensive to build, and has smaller capacity than less functional ROM's.

Organization

Basic element of semiconductor memory is the memory cell. All semiconductor memory cells have certain properties:

- Have two stable states that represent binary 0 and 1.
- Capable of being written into (at least once), to set the state.
- Capable of being read to sense the state.
  Individual cells can be selected for reading and writing operations.

The cell has three functional terminals which are shown in figure 9.1. The select terminals selects a cell for read or write operation, the control indicates read or write, and the third terminal is used to write into the cell, that is to set the state of the cell to 0 or 1. Similarly for read operation the third terminal is used to output the state of the cell.

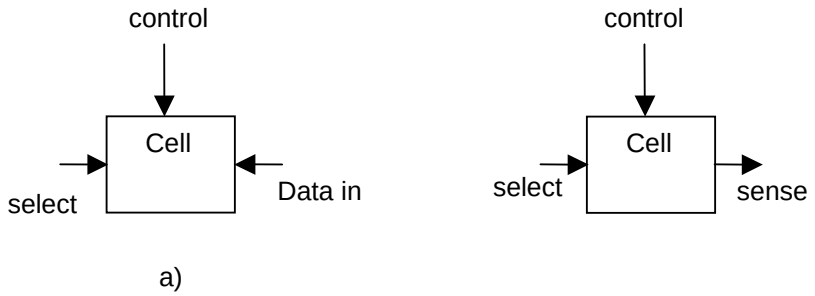

Figure 9.1: Memory cell organization (a) : write (b) for read operation

Chip Logic

Semiconductor memory comes in package chips. Each chip contains an array of memory cells. Two organizational approaches have been used 2D and 2½D.

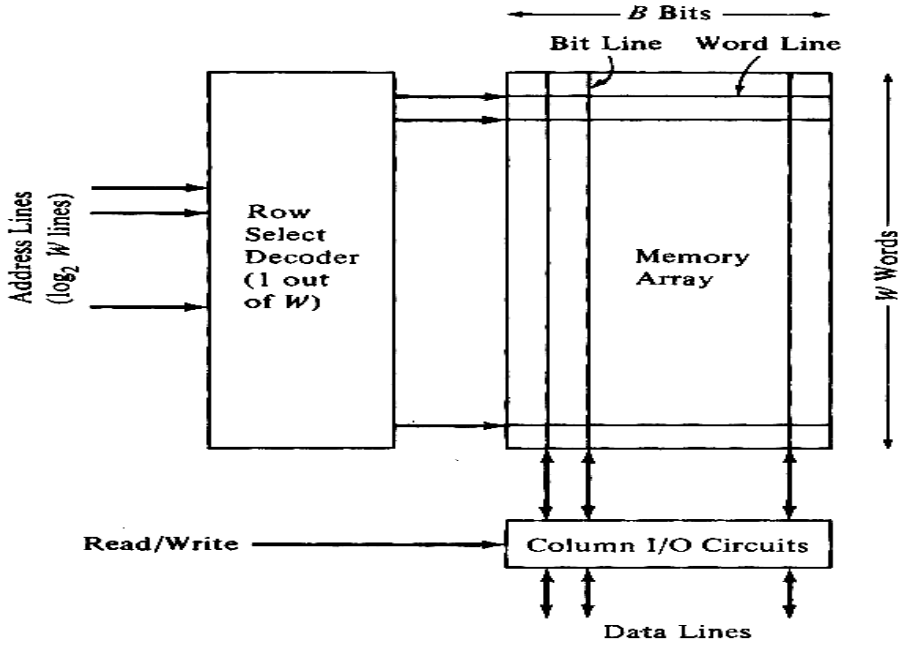

Figure 9.2: 2D memory organizations

A) 2D memory organization

In 2D memory organization the physical arrangements of a memory chip (memory cells in an array) is same as logical words in memory. Figure 9.2 shows the 2D memory organization. A memory chip has W words of B bits each. Hence its capacity is (W x B) bits.

E.g.: A 16-Mbit memory chip can be organized as 1M 16-bit words or as 4M 4-bit words. The 1st memory chip will have 20 address and 16 data lines, the 2nd memory chip will have 22 address and 4 data lines. Remember that the lines may be multiplexed. So less address lines can be used in a specific memory chip.

Number of Address Lines = $\mathbf{\log_2 W}$.

The figure also depicts the additional circuitry. Address lines supply the address of the word to be selected. A total of log2 W lines are needed are input to the decoder and the output is to activate a single output based on bit pattern out of W words.

Example: Address lines 0101 input to a decoder activates the 6^{th} output line (address starts from 0). This output is then used to select one of the word lines. Data lines are used for input and output of B bits from/to the selected word line chip simultaneously to/from the data buffer. The disadvantage of 2D organization is that all bits of any given word are on the same chip.

A) **2½ D memory organization**

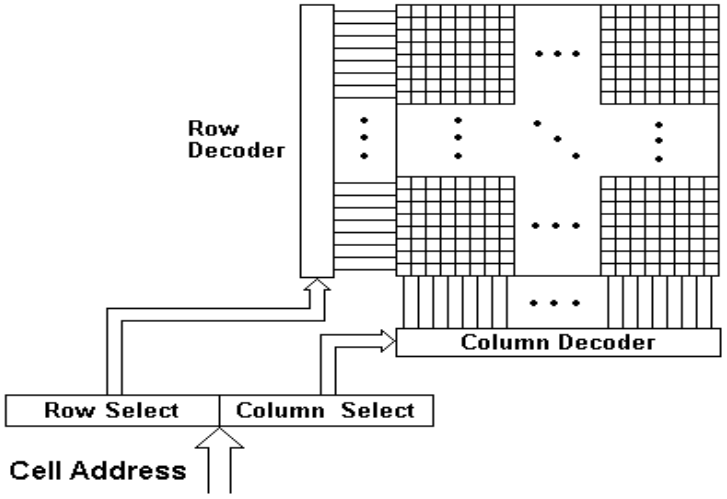

Figure 9.3: 2½ D memory organization

The 2½D memory organization is as shown in figure 9.3. In 2½D memory organization the bits of a particular word are spread across multiple chips. The most extreme and most common organization is to allow only 1 bit of a given word on a chip. Now the array itself is like a matrix, each cell is connected to a row line and a column line. For any operation to select a bit of a particular word, the word address is split into two. One part is fed to the decoder to select one row and second part of address is fed to another decoder to select one of the columns.

From the system standpoint, the memory unit can be viewed as a “black box”. Data transfer between the main memory and the CPU register takes place through two registers namely **MAR** (memory address register) and **MDR** (memory data register). If MAR is k bits long and MDR is n bits long, the main memory unit can contain up to 2^k addressable locations. During a ‘memory cycle’ n bits of data are transferred between main memory and

CPU. This transfer takes place over the *processor bus*, which has k address lines and n data lines.

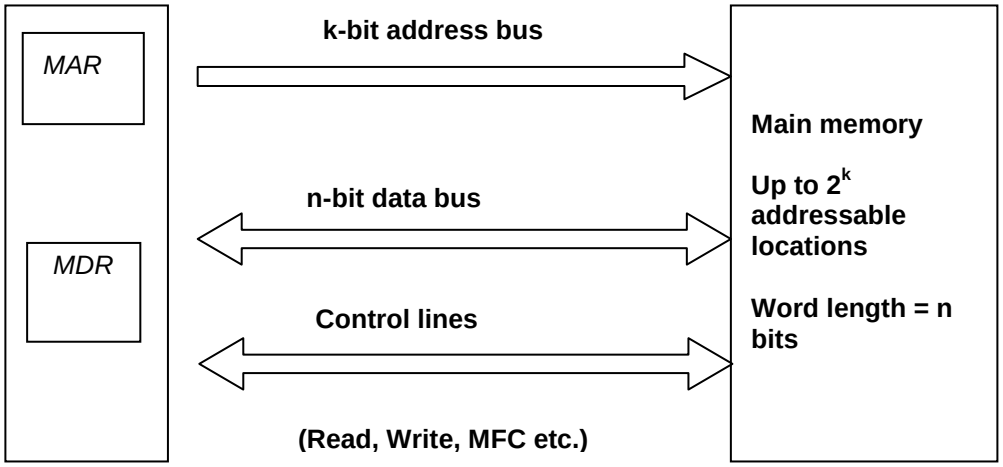

Figure 9.4: Connection of main memory to CPU

The bus also includes control lines *Read, Write, and Memory Function Completed* (**MFC**) for coordinating the data transfers. In the case of byte addressable computers, another control line is added to indicate when only a byte, rather than a full word of n bits, is transferred. Fig. 9.4 shows the connection between the CPU and main memory. The CPU initiates a memory operation by loading the appropriate data into registers MDR and MAR, and setting either Read or Write memory control line to 1. When the required operation is completed the memory control circuitry sends MFC signal to CPU.

The time that elapses between the initiation of an operation and completion of that operation is called *memory access time*. The minimum time delay between two successive memory operations is called *memory cycle time.* The cycle time is usually slightly lower than the access time. The main memory is *Random Access Memory*.

Therefore any location can be accessed for a Read or Write operation in some fixed amount of time. Semiconductor Integrated Circuits are used for the implementation of main memories.

Some of the techniques used to increase the effective speed and size of the main memory are discussed in the following sections.

Static and Dynamic Memories

Semiconductor memories in which the storage cells are small transistor circuits are used for high speed CPU registers. Single chip RAMs can be manufactured in sizes ranging from a few hundred bits to 1 GB or more. Semiconductor memories fall into two categories, SRAMs (static RAMs) and DRAMs (dynamic RAMs). Both bipolar transistor and MOS ((Metal Oxide Semiconductor) transistor are used for designing RAMs but MOS is dominant technology for designing large RAMs.

A semiconductor memory constructed using bipolar transistors or MOS transistor stores information in the form of a flip-flop voltage levels. These voltage levels are not likely to be discharged. Such memories are called *static memories*. In these memories information remains constant for longer period of time. Semiconductor memory designed using a MOS with a capacitor stores the information in the form of charge on a capacitor. The stored charge has a tendency to leak away. A stored ‘1’ will become ‘0’ if no precautions are taken. These types of memory are called *dynamic memories*.

A) Static memories

Static RAM cells use 4 – 6 transistors to store a single bit of data. This provides faster access times at the expense of lower bit densities. A processor's internal memory (registers and cache) is fabricated using static RAM. SRAMs resemble the flip-flops used in the processor design. SRAM

cells differ from the flip-flops primarily in methods used to address the cells and transfer data to and from them.

The six-transistor SRAM is shown in Fig. 9.5. A signal applied to the *word line* (also called as *address line)* by the address decoder selects the cell either for Read or Write operation. The two *bit lines* (also called as *data lines*) are used to transfer stored data and its complement between the cell and data drivers.

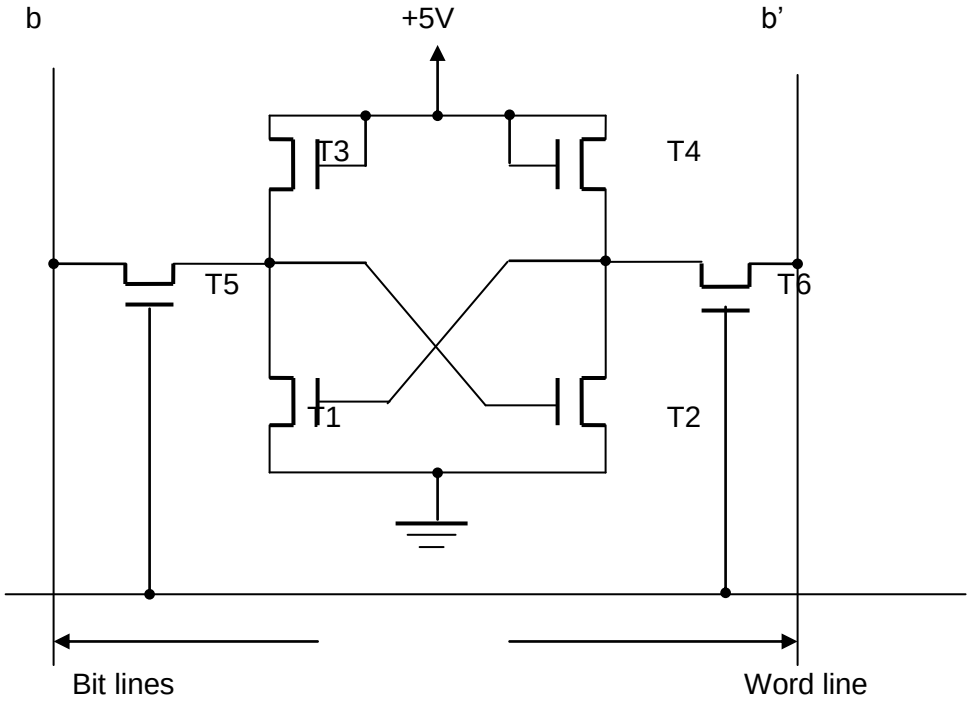

Fig. 9.5: An n-channel MOS memory cell

Static RAM is used extensively for second level cache memory, where its speed is needed and a relatively small memory will lead to a significant increase in performance.

B) Dynamic memories

The bulk of a modern processor's memory is composed of dynamic RAM (DRAM) chips. A DRAM memory cell uses a single transistor and a capacitor to store a bit of data. In a DRAM cell, the '1' and '0' states correspond to the presence or absence of a stored charge in a capacitor controlled by the transistor switching circuit. Since a DRAM can be constructed by a single transistor, the storage density is higher. Since the charge stored in DRAM may leak with time, the cell must be periodically refreshed.

Fig. 9.6 illustrates one-transistor DRAM cell. Transistor used is a MOS transistor which acts as a switch and a capacitor to store a data bit. To write information into the cell, a voltage signal is applied to the data line. Voltage signal can either be high or low representing 1 and 0 respectively. A signal is applied to the word line to switch on T. Now the capacitor charges if the data line is 1. When the transistor is off, the capacitor begins to discharge due to capacitor's own leakage resistance and due to the fact that the transistor continues to conduct a very small amount of current after it is turned off. Hence the information stored in the cell can be retrieved correctly only if it is read before the charge on the capacitor drops below some threshold value. The memory cell is therefore refreshed every time its contents are read. When a DRAM is being refreshed, other accesses must be "held off". This increases the complexity of DRAM controllers.

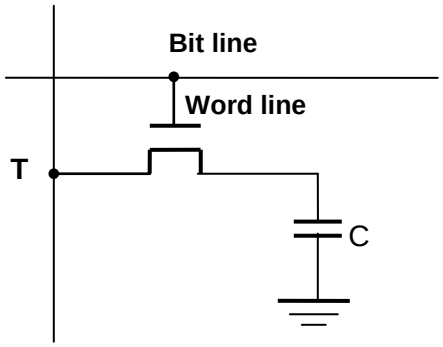

Fig. 9.6: A single-transistor dynamic memory cell

To read the cell, the word line is activated. Now the charge stored in capacitor is transferred to the bit line where it is detected. Almost all DRAMs fabricated require the address applied to the device to be asserted in two parts: a *row address* and a *column address*. This has a deleterious effect on the access time, but enables devices with large numbers of bits to be fabricated with fewer pins (enabling higher densities).

Let us study the internal organization of a 64K $\times$ 1 dynamic memory chip. The cells are arranged in the form of a square array. The higher order 8 bits of the 16 bits constitute the row address of the cell and the lower order 8 bits constitute the column address of a cell. The row and column addresses are multiplexed on eight pins. This is done to reduce the number of pins needed for external connections. During a Read or Write operation the row address is applied first. It is loaded into the row address latch under the control of Row Address Strobe (RAS) input of the chip.

Then a Read operation is initiated in which all cells in a selected row are read and refreshed. Soon after the row address is loaded, the column address is applied to the address pins and loaded into the column address latch under the control of Column Address Strobe (CAS) signal. The information is decoded in the column decoder and the appropriate $\text{sense}/\overline{\text{write}}$ circuit is selected. If the $R/\overline{W}$ control signal initiates a Read operation, output of the selected circuit is transferred to the data output D_0.

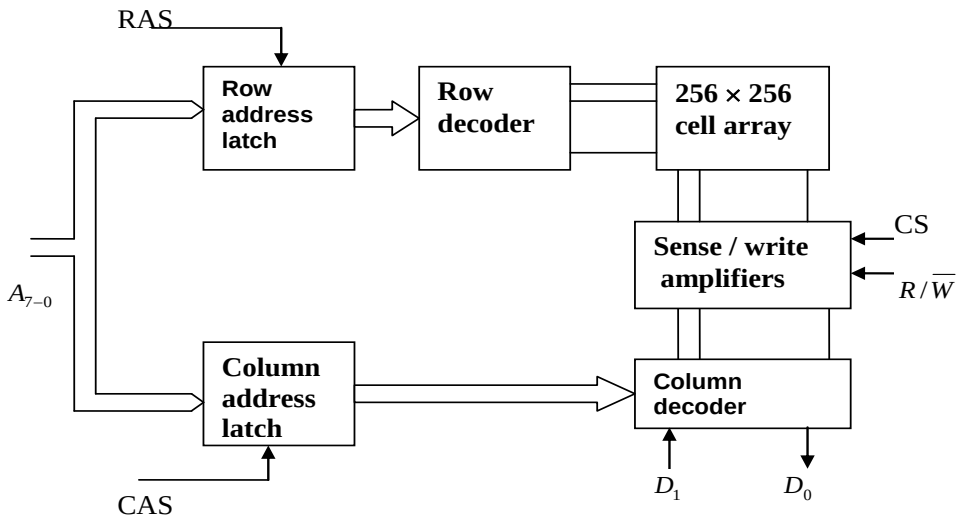

Fig. 9.7: Internal organization of a 64K × 1 dynamic memory chip

To perform a Write operation, the information at the data input DI is given to the column decoder. The information is then used to overwrite the contents of the selected cell in the corresponding column.

Application of a row address causes cells on the corresponding row to be read and refreshed during both Read and Write operations. To ensure that the contents of dynamic memory are maintained, each row of cells must be addressed periodically. A refresh circuit usually performs this function. In some dynamic memory chips the refresh facility is available within the chip itself. In such cases dynamic nature of these chips are completely invisible to the user. Such chips are known as *pseudo static.*

Advantages

1. High bit density
2. Available chips range from 1K to 4M bits, and even larger chips are being developed.
3. Low power dissipation

Disadvantages

Slower speed of operation

9.4 Memory System Considerations

While choosing a RAM chip for a given application several factors have to be taken into account. The most important factors are the speed, power dissipation, and size of the chip. In some situations other features such as availability of block transfers may be important. Because of high power dissipation in bipolar circuits, high bit density cannot be achieved. Hence a memory constructed using bipolar chips needs relatively large number of chips. Bipolar memories (static memories) are used only when very fast operation is required.

Static MOS memory chips have higher densities and slightly longer access time than bipolar memories. They have lower densities than dynamic memories, but are easier to use because they do not require refreshing circuits. Since it doesn't need refresh, static RAMs consume less power than dynamic RAMs. SRAMs will be found in battery-powered systems. The absence of refresh circuitry leads to slightly simpler systems, so SRAM will also be found in very small systems, where the simplicity of the circuitry compensates for the cost of the memory devices themselves.

Dynamic MOS memory is the ruling technology used in computer main memories. Since higher density is achieved with this technology, it is economical to implement large memories using dynamic MOS memories. Let us study how a memory subsystem is designed using static and dynamic memory chips.

Design of memory subsystem using Static Memory Chips

Consider a small memory consisting of 64K words; each word is of 16 bits. Figure 9.8 shows organization of this memory using 16K $\times$ 1 static memory

chips. A single static memory chip has a control input called Chip Select (CS). When CS is set to 1, it enables the chip to accept data input or to place data on the output line. The data output for each chip is of three state type. Only the selected chip places the data on the output line while all the other outputs are in high impedance state.

There are 4 chips in each column. Each column represents one bit position. There are 16 such columns to build a 64K $\times$ 16 memory. The address bus required for this memory is of 16 bits wide. The high-order 2 bits of the address are decoded to obtain the 4 chip select control signals. The remaining 14 address bits are used to access specified bit locations inside the chip of the selected row. The R / W inputs are given to all chips which provide common Read / Write control.

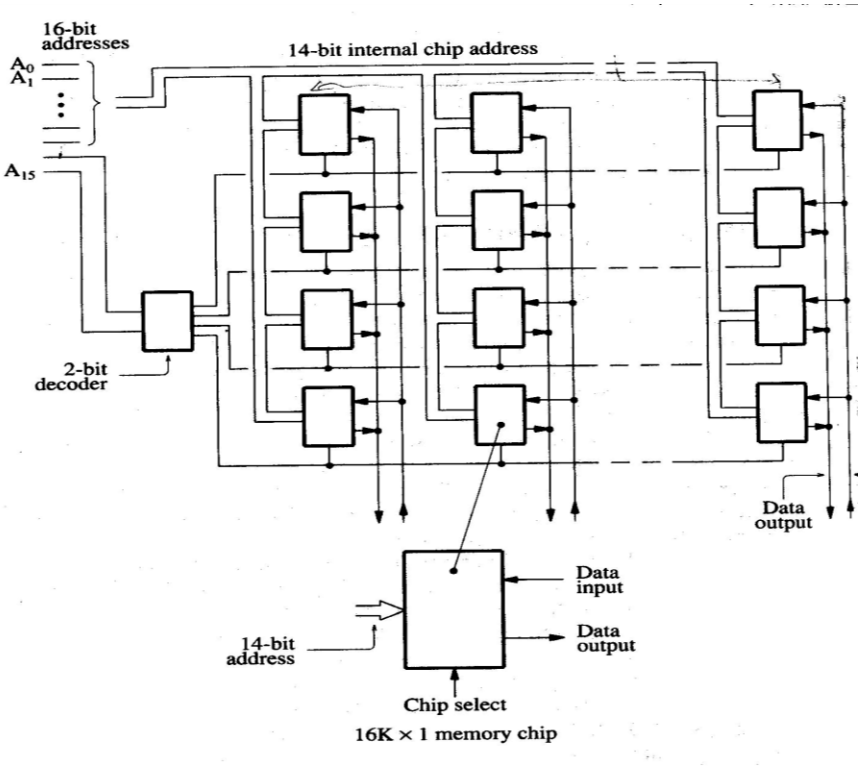

Fig. 9.8: Organization of a 64K $\times$ 16 memory using static memory chips

Design of memory subsystem using Dynamic Memory Chips

Now consider a memory implemented using dynamic memory chips. The organization of this memory is same as that of static memory. However control circuitry of this memory differs from that of static memory in three respects. The block diagram of a 256 × 16 dynamic memory chip is as shown in figure 9.9. First, the row and column parts of the address for each chip have to be multiplexed. Second, a refresh circuit is needed. Finally the timing of various steps of the memory cycle must be carefully controlled.

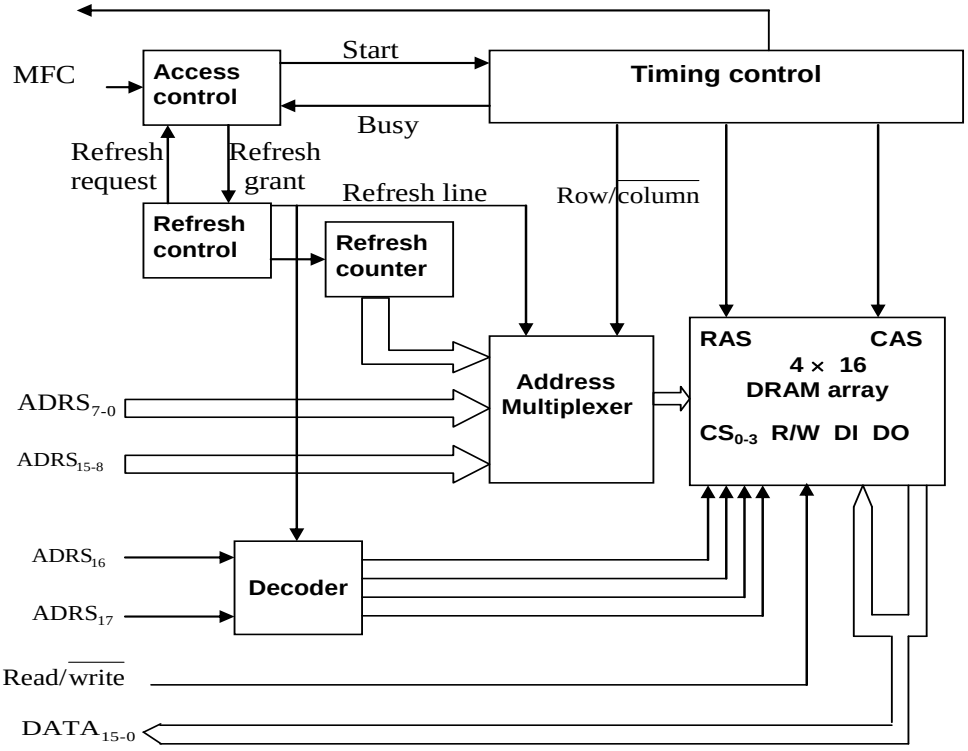

Figure 9.9: A block diagram of a 256 × 16 dynamic memory chip

Block diagram consists of dynamic memory chips and a control circuitry. Dynamic memory chips are arranged in a 4 × 16 array. If the individual chips have 64K × 1 organization, then the total storage capacity of array is

256K words, each word is of 16 bits. The control circuitry provides the multiplexed address, chip select, row and column address strobe signals (RAS and CAS) to the memory chip array. The memory unit is assumed to be connected to an asynchronous memory bus that has 18 address lines ($ADRS_{17-0}$), 16 data lines ($DATA_{15-0}$), two handshake signals (MFC and Memory Request) and a Read / $\overline{\text{Write}}$ line to indicate the type of the memory cycle required.

Operation of the control circuitry for a memory read cycle

1. The cycle starts when the CPU activates the address, the Read / $\overline{\text{Write}}$ and the memory request lines.
2. When the memory request signal becomes active, the access control block recognizes the request and it sets start signal to 1.
3. The timing and control circuit sends a busy signal in order to prevent the access control box from accepting the new requests until the current cycle ends.
4. The timing and control block then loads the row and column address into the memory chips by activating RAS and CAS. First it uses Row / $\overline{\text{column}}$ line to select the row address (ADD_{15-8}), followed by the column address (ADD_{7-0}).
5. The decoder block decodes two most significant bits of the address and generates 4 chip select signals CS_{0-3}.
6. After obtaining the row and column parts of the address, the selected memory chips place the contents of the requested bit cells on their data outputs.
7. This data is then transferred to the data lines of the memory bus through appropriate drivers.
8. Now timing and control circuit send MFC signal to CPU indicating that the requested data is available on the memory bus.

9. Finally the *busy* signal is deactivated so that the access control unit is now free to accept new requests.

The timing unit is responsible throughout the process to ensure that various signals are activated and deactivated in accordance with the specifications of the particular type of memory chips used.

Control sequence for refresh operation

The main purpose of the refresh circuit is to maintain the integrity of the stored contents of the cell. A priority is given to the refresh over the CPU bus to ensure that no information is lost, during simultaneous requests. Sequence of refresh operation is as follows:

1. The refresh control block periodically generates refresh requests, causing the access control block to start a memory cycle in the normal way.
2. The access control block arbitrates between memory access requests and refresh requests. If two requests are activated simultaneously, refresh requests are given priority in order to ensure that no information is lost.
3. As soon as the refresh control block receives the refresh grant signal, it activates the refresh line.
4. The address multiplexer selects the refresh counter as the source for the row address, instead of $ADD_{15\text{-}8}$ and the contents of the counter will be loaded into the row address latches of all the memory chips when the RAS signal is activated.
5. During this time, the Read / $\overline{\text{Write}}$ line may indicate a write operation. It is important to ensure that this does not inadvertently cause new information to be loaded into some of the cells that are being refreshed. This needed protection can be provided in several ways. One way is to have the decoder block deactivate all CS lines to prevent the memory

chips from responding to the Read / $\overline{\text{Write}}$ line. The remainder of the refresh cycle is then same as in a normal cycle.

6. At the end, the refresh control block increments the refresh counter in preparation for the next refresh cycle.

Although it requires an 8-bit row address, the refresh counter needs only seven bits wide because of the cell organization inside its memory chips. In fact the 256 $\times$ 256 array in Fig. 9.9, consists of two 128 $\times$ 256 array each having its own set of sense / $\overline{\text{Write}}$ circuits. One row of each of the two arrays is accessed during any memory cycle, depending on the lower order 7 bits of the row address. The most significant bit is used only in a normal Read / $\overline{\text{Write}}$ cycle to select one of the two groups of 256 columns. Because of this organization, the frequency of refresh operation can be reduced to half of what would be needed if the memory cells had been organized in a single 256 $\times$ 256 array.

The main purpose of the refresh circuit is to maintain the integrity of the stored information. Its existence should be invisible to the remainder of the computer system. Also other parts of the system should not be affected by the operation of the refresh circuit. If memory access and refresh requests occur simultaneously, refresh circuit is given the priority so that no stored information is lost. Thus the response of the memory to a request from CPU may be delayed if refresh operation is in progress. The amount of delay depends on the mode of refresh operation. During a refresh operation, all memory rows may be refreshed in succession before the memory is returned to normal use. A more common scheme, however, interleaves refresh operations on successive rows with accesses from the memory bus, which results in shorter, but more frequent refresh periods.

In case of a synchronous, it may be possible to hide a refresh cycle within the early part of a bus cycle if sufficient time remains after a refresh cycle to carry out a Read or Write access.

9.5 Memory interleaving

Another technique called *memory interleaving* divides the system into a number of modules and arranges them so that successive words in the address space are placed in different modules. If memory access requests are made for consecutive addresses, then the access will be made for different modules. Since parallel access to these modules is possible, the average rate of fetching words from the main memory can be increased.

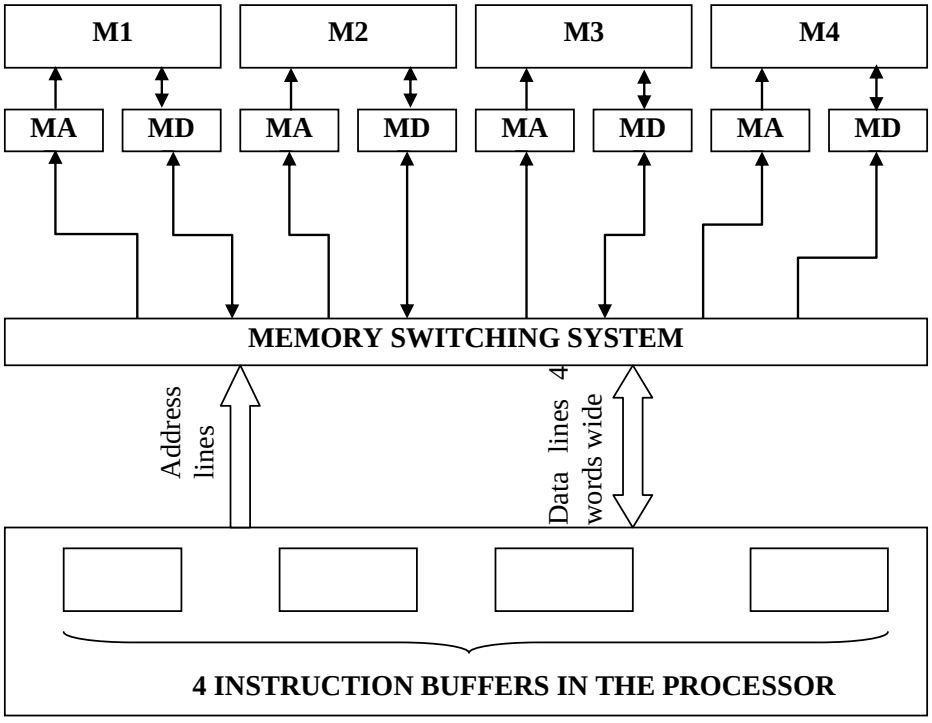

Figure 9.10: Memory interleaving

Fig. 9.10 illustrates interleaved memory where main memory is divided into 4 different modules. When an instruction FETCH is issued by the processor, a memory access circuit creates four consecutive addresses and places them in four MARs. A memory read command reads all the four modules simultaneously and retrieves four instructions. These are sent to the processor. Thus each FETCH instruction fetches 4 consecutive instructions.

One may combine interleaving and cache to reduce the speed mismatch between the cache memory and main memory. To illustrate this consider the time required for transferring a block of data from the main memory to the cache when a read miss occurs. Assume that a cache with 8-word blocks is used. When cache miss occurs, the block that contains desired word must be copied from the main memory into the cache.

Assume that the hardware has the following properties,

- It takes one clock cycle to send an address to the main memory.
- The memory is build with DRAM chips that allow the first word to be accessed in 8 clock cycles, but subsequent words of the block are accessed in 4 clock cycles per word.
- One clock cycle is needed to send one word to the cache.

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

Similarly now assume that the main memory is divided into 4 interleaved modules using interleaving technique. When the starting address of the block arrives at the memory all the four modules start accessing the required data, using higher order bits of the address. After 8 clock cycles, each module has one word of data in its MDR. These words are transferred to the cache, one word at a time, during the next 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. Therefore the total time required to load the block from the interleaved memory is 1 + 8 + 4 + 4 = 17 cycles. Thus interleaving reduces the block transfer by more than a factor of 2.

9.6 Summary

A computer system is equipped with a hierarchy of memory subsystems. There are several memory types with very different physical properties. The important characteristics of memory devices are cost per bit and access time data transfer rate, alterability and compatibility with processor technologies.

The main memory is a major component in any computer system. Its characteristics in terms of size and speed play an important role in determining the performance of a given computer. There are several techniques available to increase the effective speed and size of the memory. An intermediate memory called cache memory can be introduced between main memory and CPU. We have introduced the reader the principle, structure of cache memory along with the mapping functions and replacement algorithms. Finally we touched upon few external memory elements and another method to increase the effective size of the memory is to introduce virtual memory.

Self Assessment Questions

1. _____ requires its own local memory in the form of registers.
2. _____ is often equated with the main memory.
3. Disk units have _____.
4. _____ is dominant technology for designing large RAMs.
5. Static RAM cells use _____ transistors to store a single bit of data.

9.7 Terminal Questions

1. Explain the characteristics of memory system.
2. Discuss the organization of main memory.
3. Explain the concept of memory interleaving.

9.8 Answers

Self Assessment Questions:

1. CPU
2. internal
3. direct access
4. MOS
5. 4 to 6

Terminal Questions:

1. Refer Section 9.2
2. Refer Section 9.3
3. Refer Section 9.5