
**Module V (8 Hours)**

The Memory System – basic concepts, semiconductor RAM memories - organization – static and dynamic RAM, Structure of larger memories, semiconductor ROM memories, Speed, Size and cost, Cache memory – mapping functions – replacement algorithms , Virtual memory – paging and segmentation.


- Hamacher, Vranesic & Zaky, “Computer Organization” (5th Ed), McGraw Hill.

# Introduction

- Ideally, the memory would be fast, large, and inexpensive.
- Unfortunately, it is impossible to meet all three of these requirements simultaneously.
- Increased speed and size are achieved at increased cost.

# BASIC CONCEPTS

- The maximum size of the memory that can be used in any computer is determined by the addressing scheme.
- For example, a computer that generates 16-bit addresses is capable of addressing up to $2^{16}$ = 64K (kilo) memory locations.
- Machines whose instructions generate 32-bit addresses can utilize a memory that contains up to $2^{32}$ = 4G (giga) $\approx 4 \times 10^9$ locations, whereas machines with 64-bit addresses can access up to $2^{64}$ = 16E (exa) $\approx 16 \times 10^{18}$ locations.
- The number of locations represents the size of the address space of the computer.

Tera – $10^{12}$ Peta - $10^{15}$

# BASIC CONCEPTS

- 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.

# BASIC CONCEPTS

- Main Memory (MM) unit can be viewed as a "black box".
- 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.

# BASIC CONCEPTS

- 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 READ to 1 (R/W = 1) and sets other control signals if required. The data from the MM is loaded into MDR and MFC is set to 1. Upon receipt of the MFC signal, the processor loads the data on the datalines into the MDR register.

# BASIC CONCEPTS

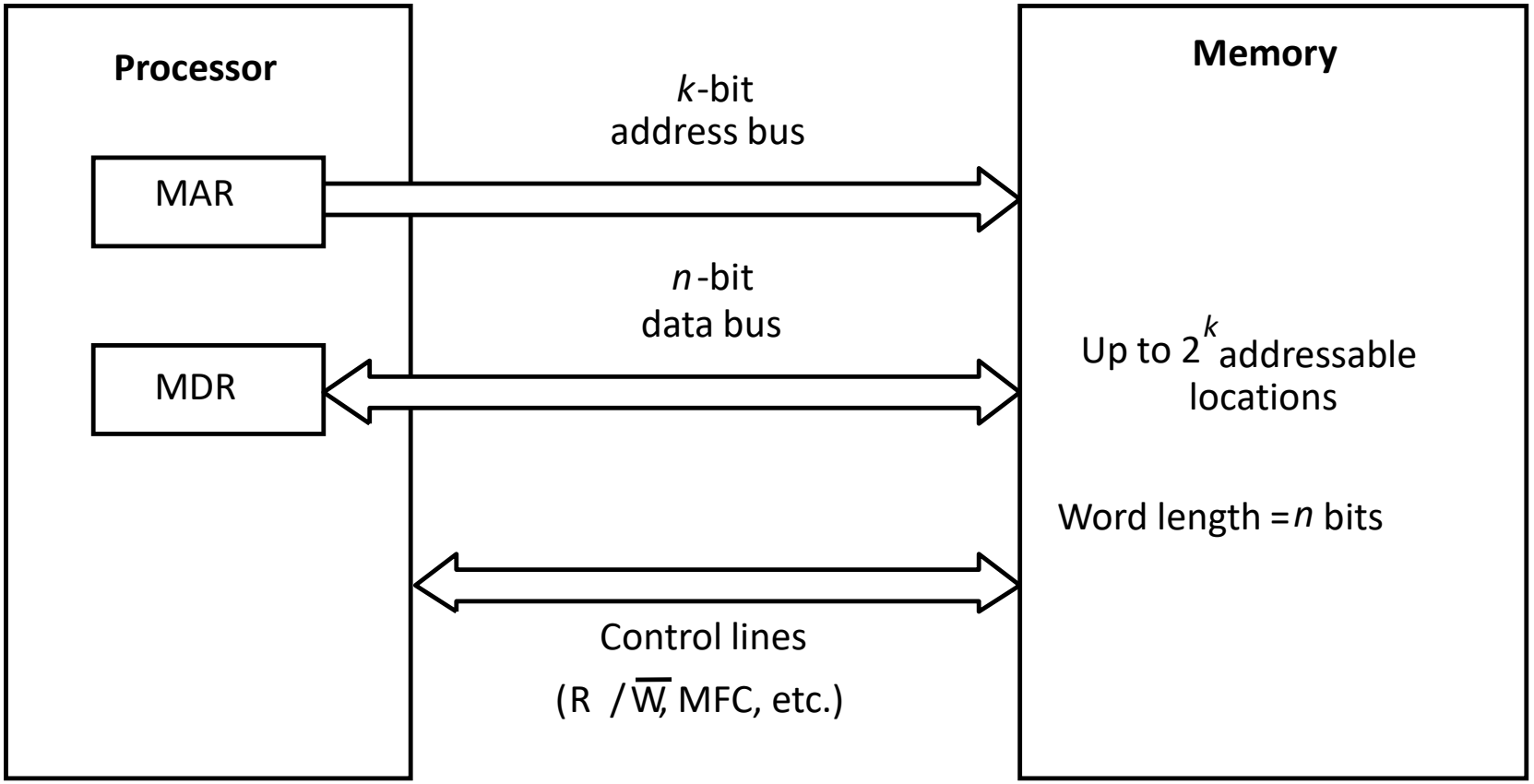



Figure 1 Connection of memory to the processor

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

# BASIC CONCEPTS

- For a write operation, MAR, MDR are suitably loaded by the CPU, write is set to 1 ($R/\overline{W} = 0$) and other control signals are set suitably. The MM control circuitry loads the data into appropriate locations and sets MFC to 1.
- The connection between the processor and its memory consists of address, data, and control lines.
- The processor uses the address lines to specify the memory location involved in a data transfer operation, and uses the data lines to transfer the data.
- At the same time, the control lines carry the command indicating a Read or a Write operation and whether a byte or a word is to be transferred.
- The control lines also provide the necessary timing information and are used by the memory to indicate when it has completed the requested operation.

# BASIC CONCEPTS

Measures for the speed of a memory:

- 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, depending on the implementation details of the memory unit.

# BASIC CONCEPTS

- Memory
  - Primary Memory (directly accessed by processor)
    - RAM(Random Access Memory) – Volatile(data stored is lost when power is turned off) – OS, applicn program, data currently using
    - ROM(Read Only Memory) – Non-volatile(data stored is retained even if when power is turned off) – BIOS(booting), firmware for other hardware devices
  - Secondary Memory

# BASIC CONCEPTS

- A memory unit is called a random-access memory (RAM) if the access time to any location is the same, independent of the location's address.
- This distinguishes such memory units from serial, or partly serial, access storage devices such as magnetic and optical disks.
- Access time of the latter devices depends on the address or position of the data.
- The technology for implementing computer memories uses semiconductor integrated circuits.

# BASIC CONCEPTS

- There are techniques used to increase the effective speed and size of the memory.
  - Cache Memory (to increase the effective speed)
  - Virtual memory (to increase the effective size)

- The processor of a computer can usually process instructions and data faster than they can be fetched from the main memory. Hence, the memory access time is the bottleneck in the system.
- One way to reduce the memory access time is to use a cache memory. This is a small, fast memory inserted between the larger, slower main memory and the processor. It holds the currently active portions of a program and their data.

# BASIC CONCEPTS

- There are techniques used to increase the effective speed and size of the memory.
  - Cache Memory (to increase the effective speed)
  - Virtual memory (to increase the effective size)
- Virtual memory is another important concept related to memory organization.
- With this technique, only the active portions of a program are stored in the main memory, and the remainder is stored on the much larger secondary storage device.
- Sections of the program are transferred back and forth between the main memory and the secondary storage device in a manner that is transparent to the application program.
- As a result, the application program sees a memory that is much larger than the computer's physical main memory.

# SEMICONDUCTOR RAM MEMORIES

- Semiconductor random-access memories (RAMs) are available in a wide range of speeds.
- Their cycle times range from 100 ns to less than 10 ns.
- Previously very costly, but now it is dropped down. (rapid advances in VLSI)
- In this section, we discuss the main characteristics of these memories.
- We start by introducing the way that memory cells are organized inside a chip.

# INTERNAL ORGANIZATION OF MEMORY CHIPS

- Memory cells are usually organized in the form of an array, in which each cell is capable of storing one bit of information.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

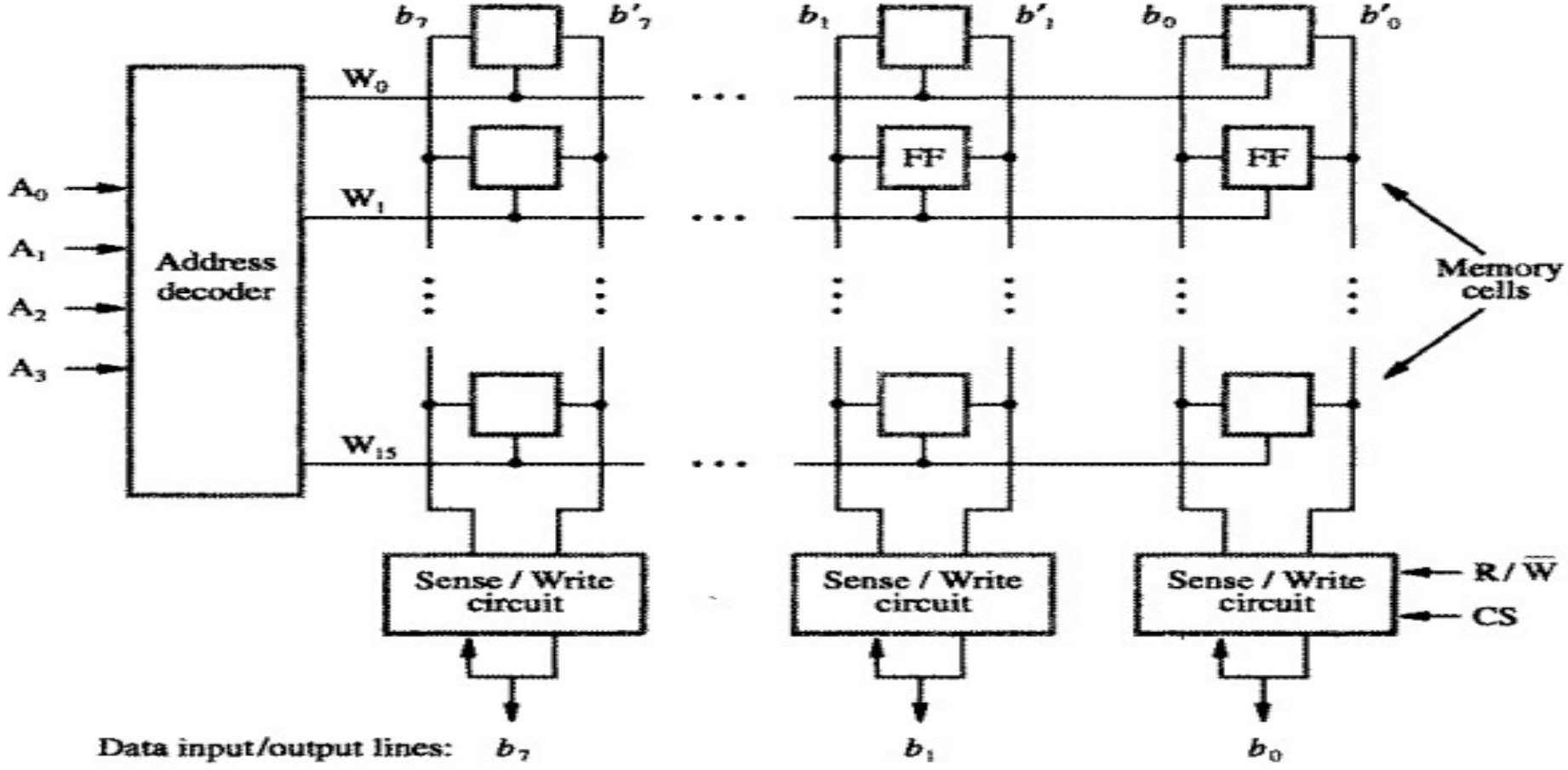



**Figure 2** Organization of bit cells in a memory chip.

# INTERNAL ORGANIZATION OF MEMORY CHIPS

- A possible organization is illustrated in Figure 2.
- 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, and the Sense/Write circuits are connected to the data input/output lines of the chip.

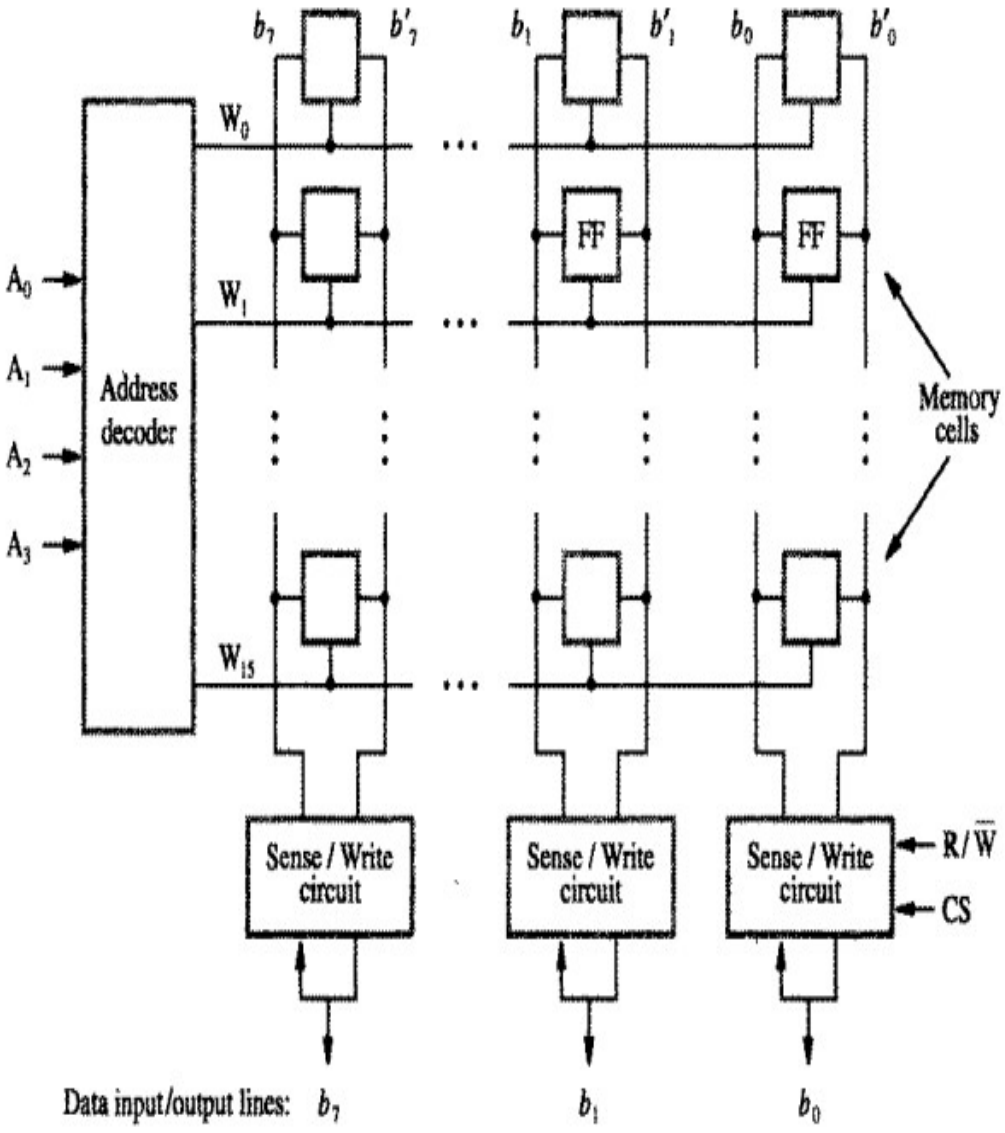



Figure 2 Organization of bit cells in a memory chip.

# INTERNAL ORGANIZATION OF MEMORY CHIPS

- During a Read operation, these circuits sense, or read, the information stored in the cells selected by a word line and place this information on the output data lines.
- During a Write operation, the Sense/Write circuits receive input data and store them in the cells of the selected word.

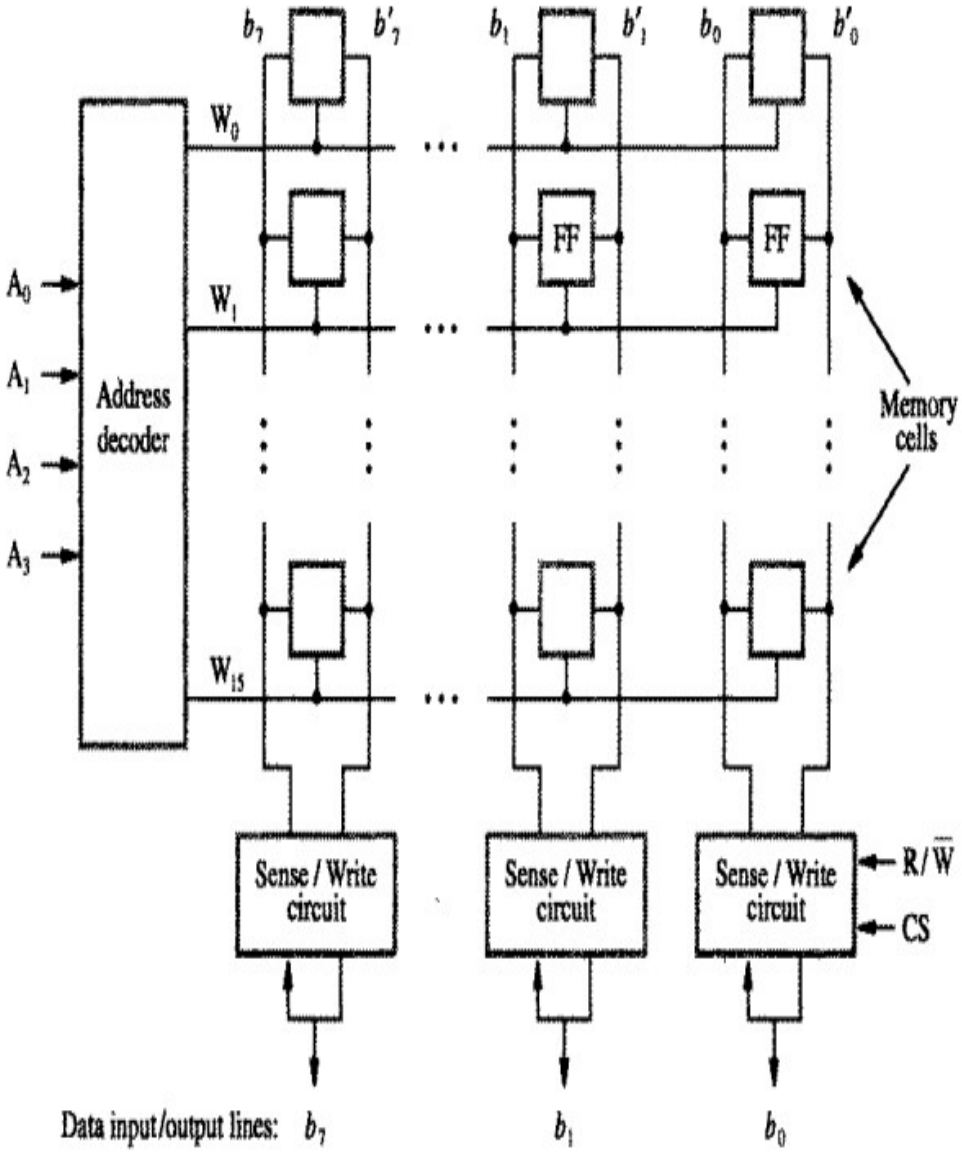



Figure 2 Organization of bit cells in a memory chip.

# INTERNAL ORGANIZATION OF MEMORY CHIPS

- Figure 2 is an example of a very small memory circuit consisting of 16 words of 8 bits each (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 lines of a computer.
- Two control lines, R/W and CS, are provided. The R/W (Read/Write) input specifies the required operation, and the CS (Chip Select) input selects a given chip in a multichip memory system.

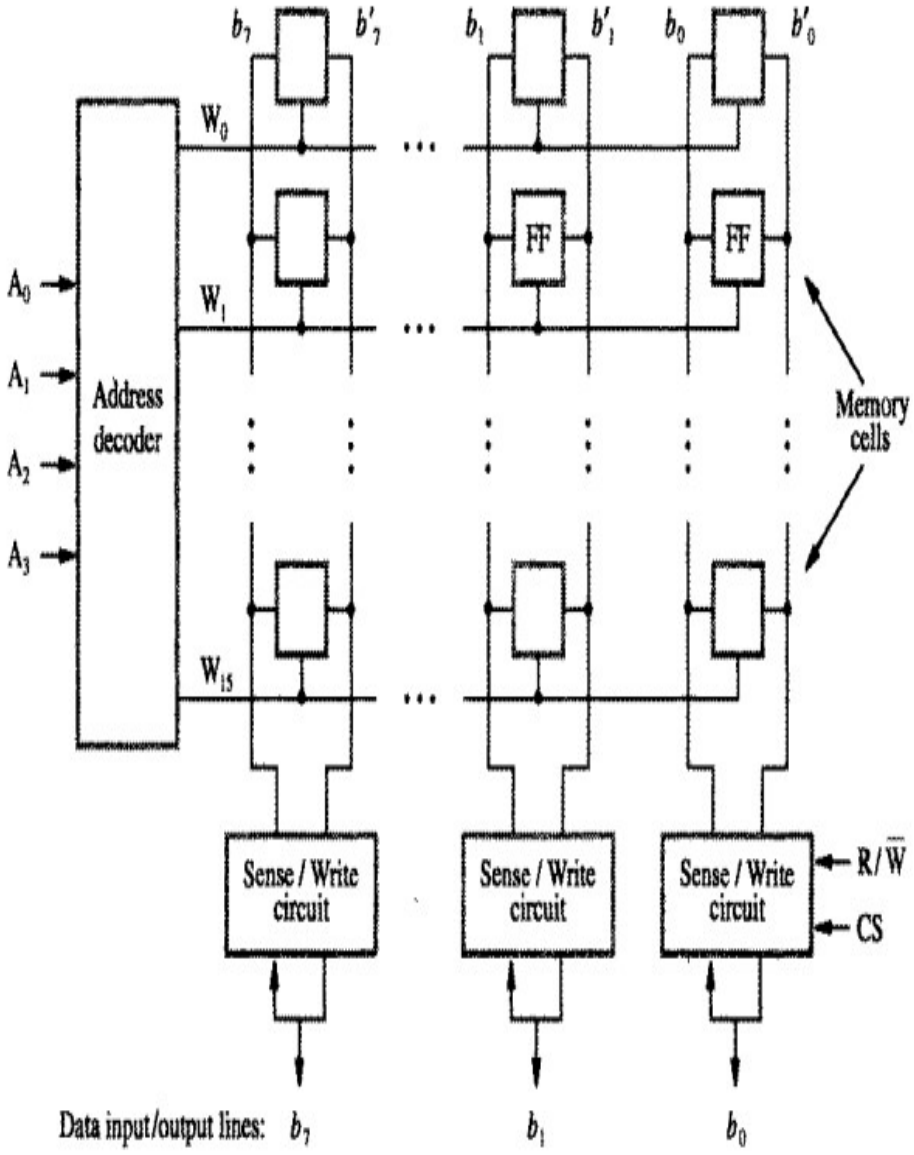
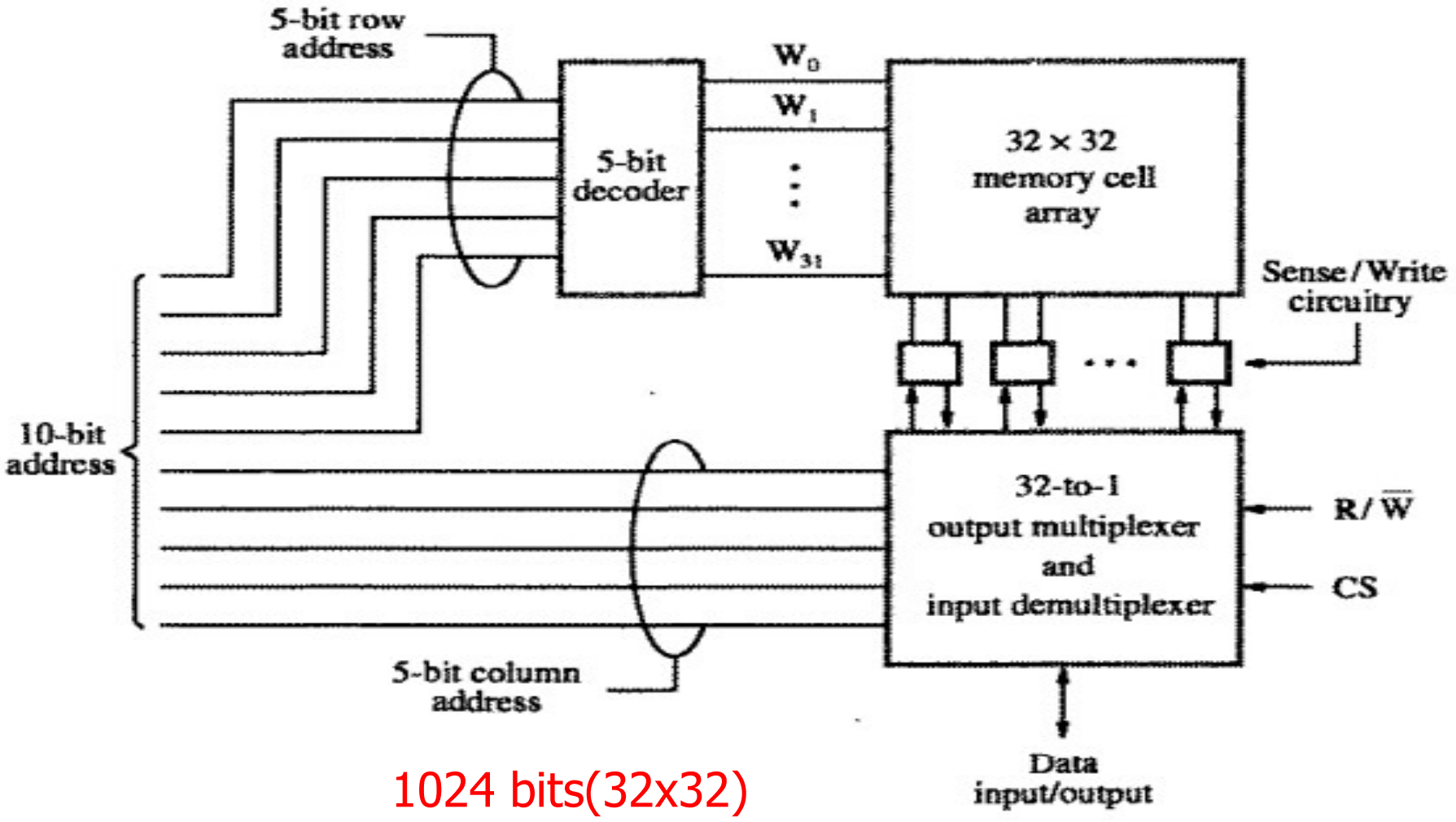



Figure 2 Organization of bit cells in a memory chip.

# INTERNAL ORGANIZATION OF MEMORY CHIPS

- The memory circuit in Figure 2 stores 128 bits(16 x 8) and requires 14 external connections for address(4), data(8), and control lines(2).
- It also needs two lines for power supply and ground connections.
- Consider now a slightly larger memory circuit, one that has 1K (1024) memory cells. This circuit can be organized as a 128 × 8 memory, requiring a total of 19 external connections.(7+8+2+2(power and gnd))
- Alternatively, the same number of cells can be organized into a 1K×1 format. In this case, a 10-bit address($1K \approx 2^{10}$) is needed, but there is only one data line, resulting in 15 external connections. (10+1+2+2)

# INTERNAL ORGANIZATION OF MEMORY CHIPS




1024 bits(32x32)

**Figure .3** Organization of a 1K × 1 memory chip.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

# INTERNAL ORGANIZATION OF MEMORY CHIPS

- The 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.
- One of these, selected by the column address, is connected to the external data lines by the input and output multiplexers.
- This structure can store 1024 bits(32x32), can be implemented in a 16-pin chip.

# STATIC MEMORIES

- Memories that consist of circuits capable of retaining their state as long as power is applied are known as static memories.
- Figure illustrates how a static RAM (SRAM) cell may be implemented.

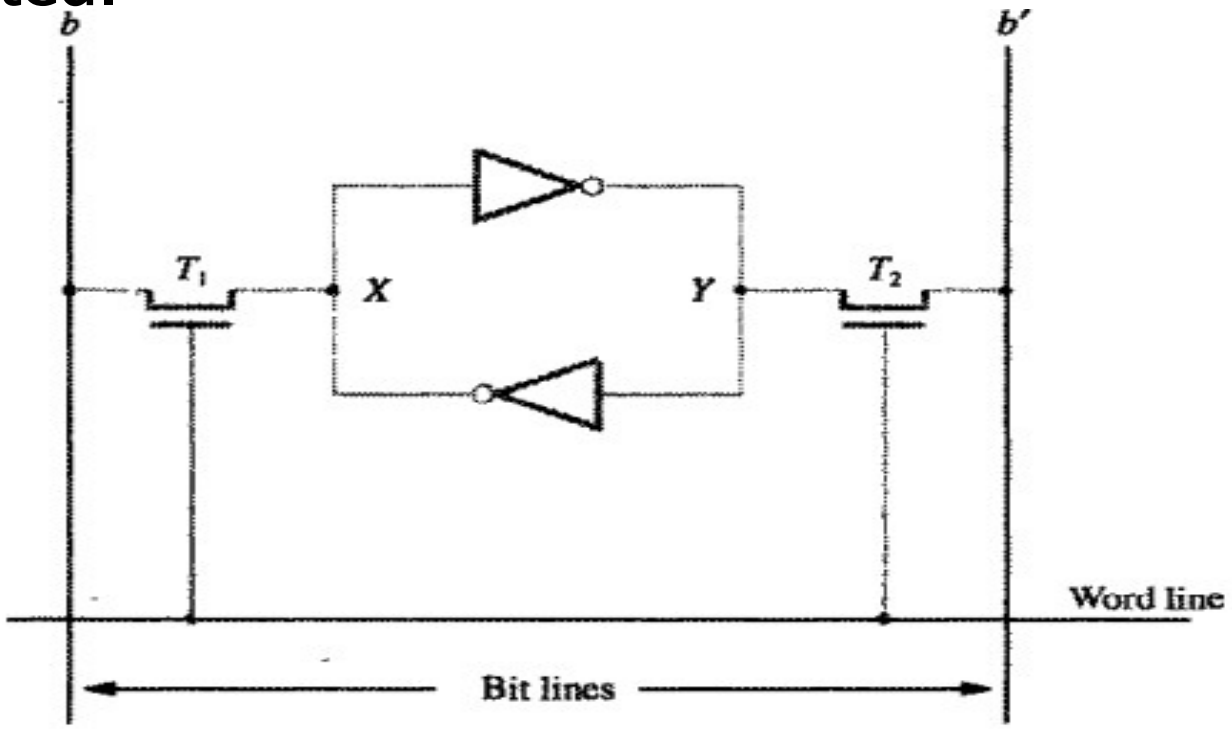



**Figure 4** A static RAM cell.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

# STATIC MEMORIES

- Two inverters are cross-connected to form a latch.
- The latch is connected to two bit lines by transistors T1 and T2.
- 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, 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.

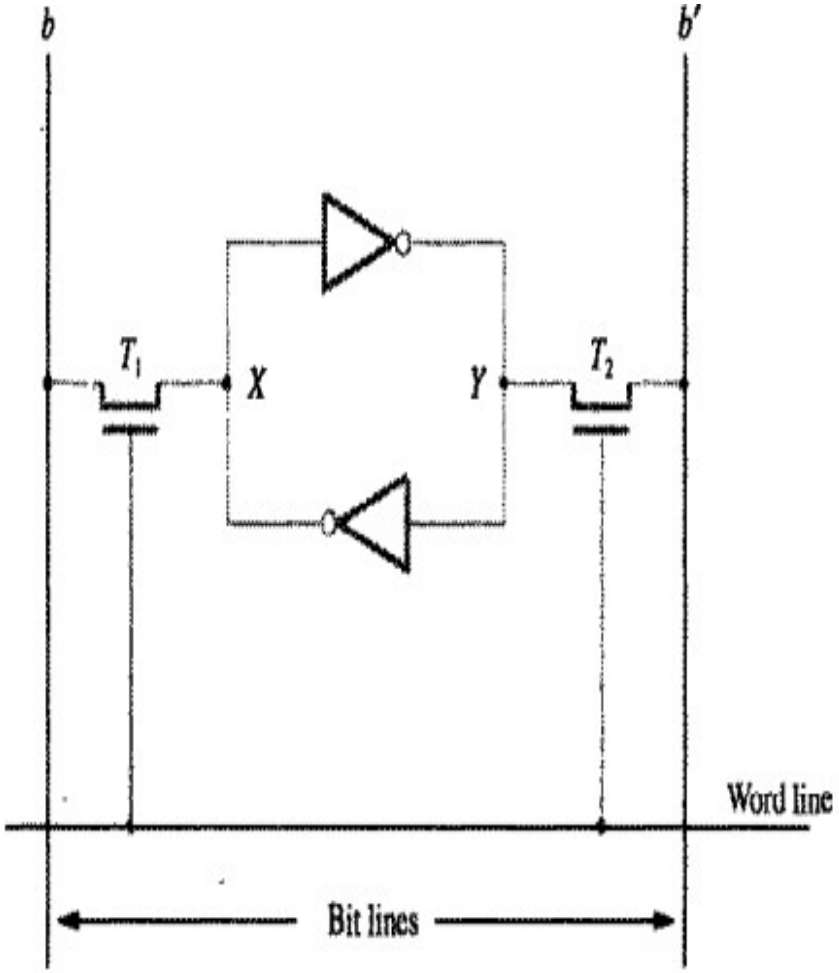



Figure 4 A static RAM cell.

# STATIC MEMORIES

Read Operation

- In order to read the state of the SRAM cell, the word line is activated to close switches T1 and T2.
- 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 always complements of each other.
- The Sense/Write circuit at the end of the two bit lines monitors their state and sets the corresponding output accordingly.

# STATIC MEMORIES

## Write Operation

- During a Write operation, the Sense/Write circuit drives bit lines b and b', instead of sensing their state.
- It places the appropriate value on bit line b and its complement on b' and activates the word line.
- This forces the cell into the corresponding state.

# CMOS Cell

- A CMOS realization of the cell is in Figure5.
- Transistor pairs (T3, T5) and (T4, T6) form the inverters in the latch.
- 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 T3 and T6 on, while T4 and T5 are off.
- If T1 and T2 are turned on, bit lines b and b' will have high and low signals, respectively.

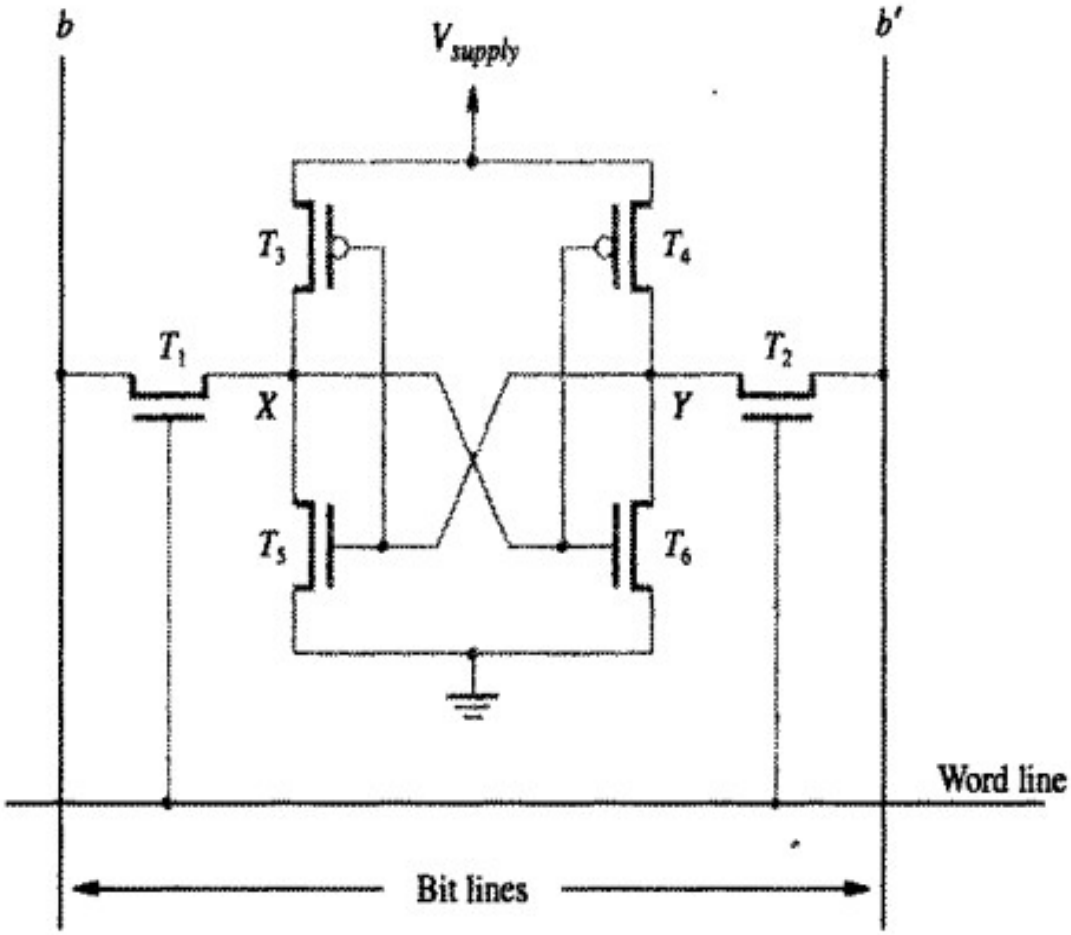



**Figure 5** An example of a CMOS memory cell.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

# CMOS Cell

- Continuous power is needed for the cell to retain its state.
- If power is interrupted, the cell's contents are lost.
- When power is restored, the latch settles into a stable state, but not necessarily 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.

- A major advantage of CMOS SRAMs is their very low power consumption, because current flows in the cell only when the cell is being accessed.
- Otherwise, T1, T2, and one transistor in each inverter are turned off, ensuring that there is no continuous electrical path between Vsupply and ground.

# Static RAM (SRAM)

- Static RAMs can be accessed very quickly. Access times on the order of a few nanoseconds are found in commercially available chips.
- SRAMs are used in applications where speed is of critical concern.

# RAM

- Static RAMs (SRAMs):
  - Consist of circuits that are capable of retaining their state as long as the power is applied.
  - Volatile memories, because their contents are lost when power is interrupted.
  - Access times of static RAMs are in the range of few nanoseconds.
  - However, the cost is usually high.
- Dynamic RAMs (DRAMs):
  - Do not retain their state indefinitely.
  - Contents must be periodically refreshed.
  - Contents may be refreshed while accessing them for reading.

# SRAM VS DRAM

## SRAM

- Very fast
- Very Expensive
- Used in Cache memory and CPU register

## DRAM

- Slower than SRAM
- Cheaper than SRAM
- Used in most computer as main memory
- Need to be refreshed periodically

# ASYNCHRONOUS DYNAMIC RAMS

- Static RAMs are fast, but their cells require several transistors.
- Less expensive and higher density RAMs can be implemented with simpler cells. Such cells do not retain their state indefinitely; hence, they are called dynamic RAMs (DRAMs).

- Information is stored in a dynamic memory cell in the form of a charge on a capacitor, but 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.

# ASYNCHRONOUS DYNAMIC RAMS

- An example of a dynamic memory cell that consists of a capacitor, C, and a transistor, T.
- 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 in the capacitor.

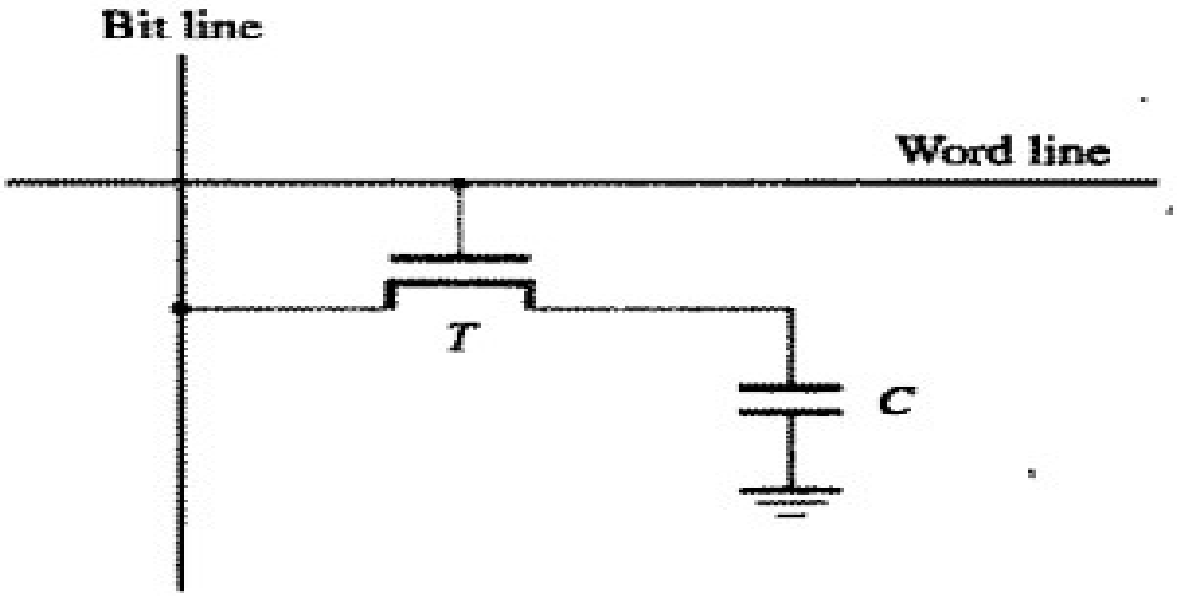



Figure .6 A single-transistor dynamic memory cell.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

# ASYNCHRONOUS DYNAMIC RAMS

- After the transistor is turned off, the charge remains stored in the capacitor, but not for long.
- The capacitor begins to discharge. This is because the transistor continues to conduct a tiny amount of current, measured in picoamperes, after it is turned off.
- Hence, the information stored in the cell can be retrieved correctly only if it is read before the charge in the capacitor drops below some threshold value.

# ASYNCHRONOUS DYNAMIC RAMS

- 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 in the capacitor is above or below the threshold value.
- If the charge is above the threshold, the sense amplifier drives the bit line to the full voltage representing the logic value 1. As a result, the capacitor is recharged to the full charge corresponding to the logic value 1.
- If the sense amplifier detects that the charge in the capacitor is below the threshold value, it pulls the bit line to ground level to discharge the capacitor fully.
- Thus, reading the contents of a cell automatically refreshes its contents.
- Since the word line is common to all cells in a row, all cells in a selected row are read and refreshed at the same time.

# ASYNCHRONOUS DYNAMIC RAMS

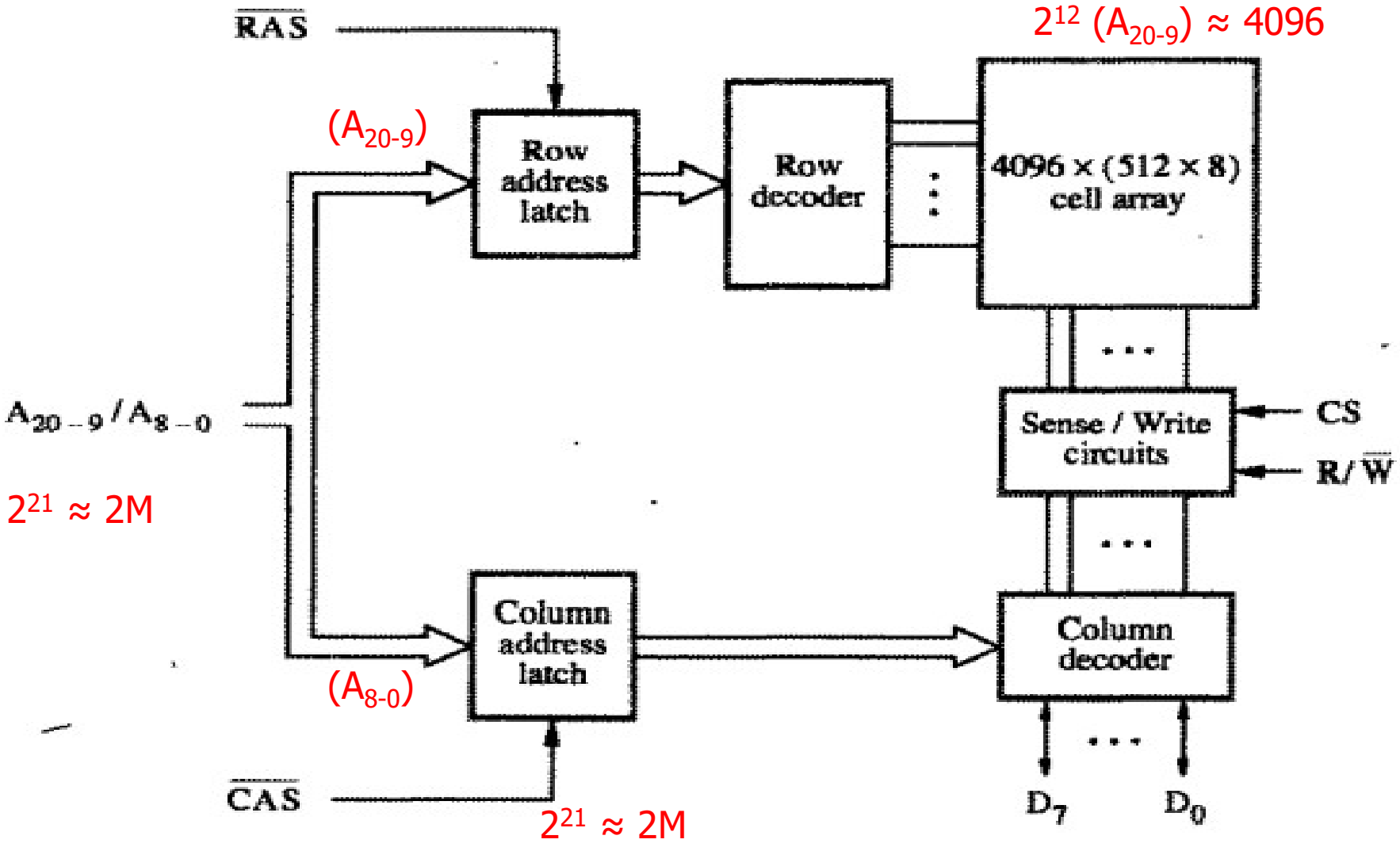



**Figure 7** Internal organization of a 2M x 8 dynamic memory chip.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

# ASYNCHRONOUS DYNAMIC RAMS

- During a Read or a Write operation, the row address is applied first.
- It is loaded into the row address latch in response to a signal pulse on an input control line called the Row Address Strobe (RAS). This causes a Read operation to be initiated, in which all cells in 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 a second control line called the Column Address Strobe (CAS).

# ASYNCHRONOUS DYNAMIC RAMS

- The timing of the operation of the DRAM is controlled by the RAS and CAS signals.
- These signals are generated by a memory controller circuit external to the chip when the processor issues a Read or a Write command.
- During a Read operation, the output data are transferred to the processor after a delay equivalent to the memory's access time. Such memories are referred to as asynchronous DRAMs.
- The memory controller is also responsible for refreshing the data stored in the memory chips.

# Fast Page Mode

- When the DRAM in Figure 7 is accessed, the contents of all 16,384 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 A10–0. A simple addition to the circuit makes it possible to access the other bytes in the same row without having to reselect the row. Each sense amplifier also acts as a latch. When a row address is applied, the contents of all cells in the selected row are loaded into the corresponding latches. Then, it is only necessary to apply different column addresses to place the different bytes on the data lines.

# Fast Page Mode

- All bytes in the selected row can be transferred in sequential order by applying a consecutive sequence of column addresses under the control of successive CAS signals.
- Thus, a block of data can be transferred at a much faster rate than can be achieved for transfers involving random addresses.
- The block transfer capability is referred to as the fast page mode feature. (A large block of data is often called a page.)

# SYNCHRONOUS DRAMS

- Developments in memory technology resulted in DRAMs whose operation is synchronized with a clock signal. Such memories are known as synchronous DRAMs (SDRAMs).

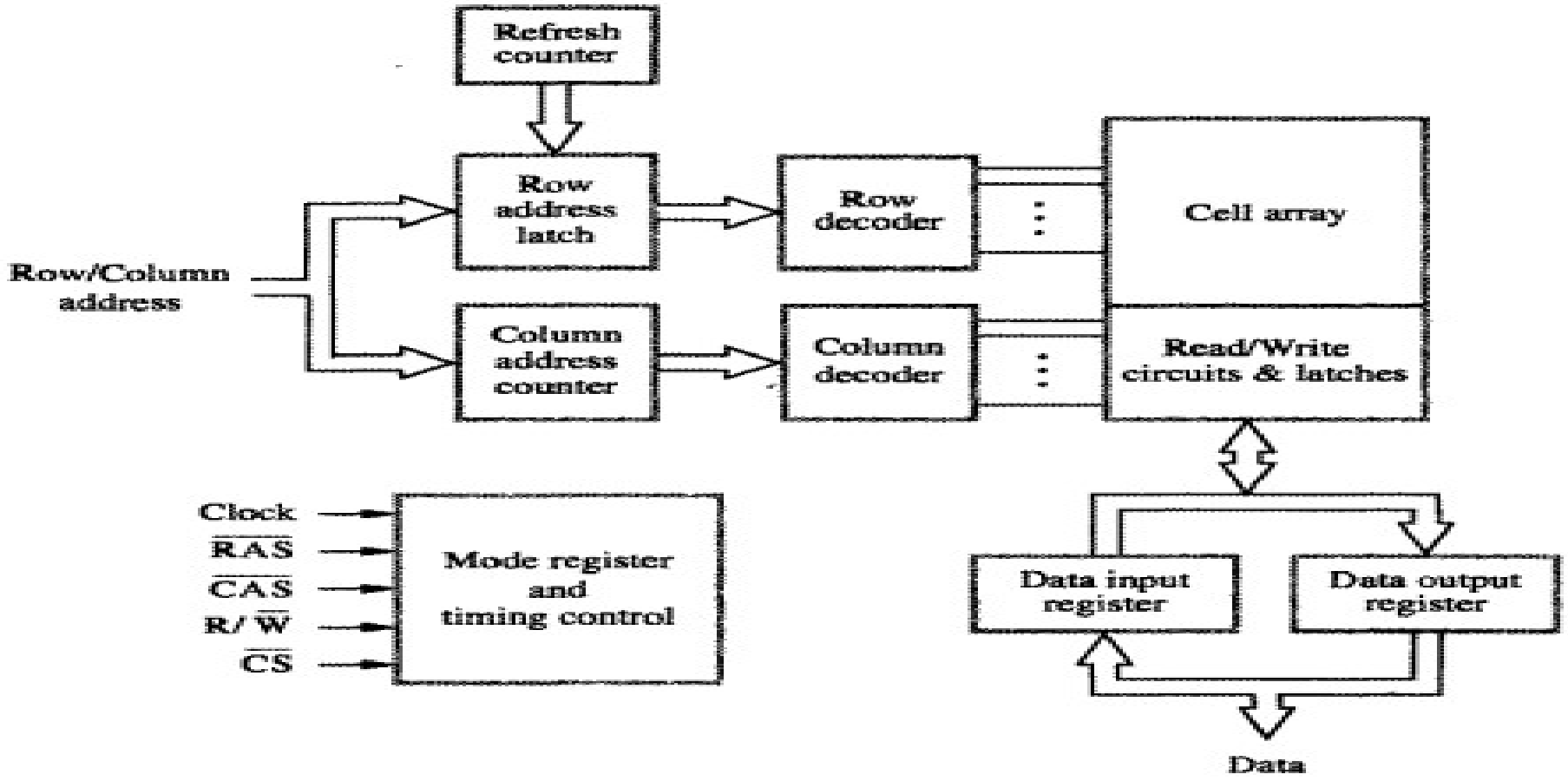



Figure 8 Synchronous DRAM.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]

# SYNCHRONOUS DRAMS

- The cell array is the same as in asynchronous DRAMs.
- The distinguishing feature of an SDRAM is the use of a clock signal, the availability of which makes it possible to incorporate control circuitry on the chip that provides many useful features.
- For example, SDRAMs have built-in refresh circuitry, with a refresh counter to provide the addresses of the rows to be selected for refreshing.
- The address and data connections of an SDRAM may be buffered by means of registers.
- Internally, the Sense/Write amplifiers function as latches, as in asynchronous DRAMs.

# SYNCHRONOUS DRAMS

- A Read operation causes the contents of all cells in the selected row to be loaded into these latches.
- The data in the latches of the selected column are transferred into the data register and thus becoming available on the data output pins.
- The buffer registers are useful when transferring large blocks of data at very high speed.

# SYNCHRONOUS DRAMS

- SDRAMs have several different modes of operation, which can be selected by writing control information into a mode register.
- It is not necessary to provide externally-generated pulses on the CAS line to select successive columns.
- The necessary control signals are generated internally using a column counter and the clock signal.
- New data are placed on the data lines at the rising edge of each clock pulse.
- Synchronous DRAMs can deliver data at a very high rate, because all the control signals needed are generated inside the chip.

# SYNCHRONOUS DRAMS

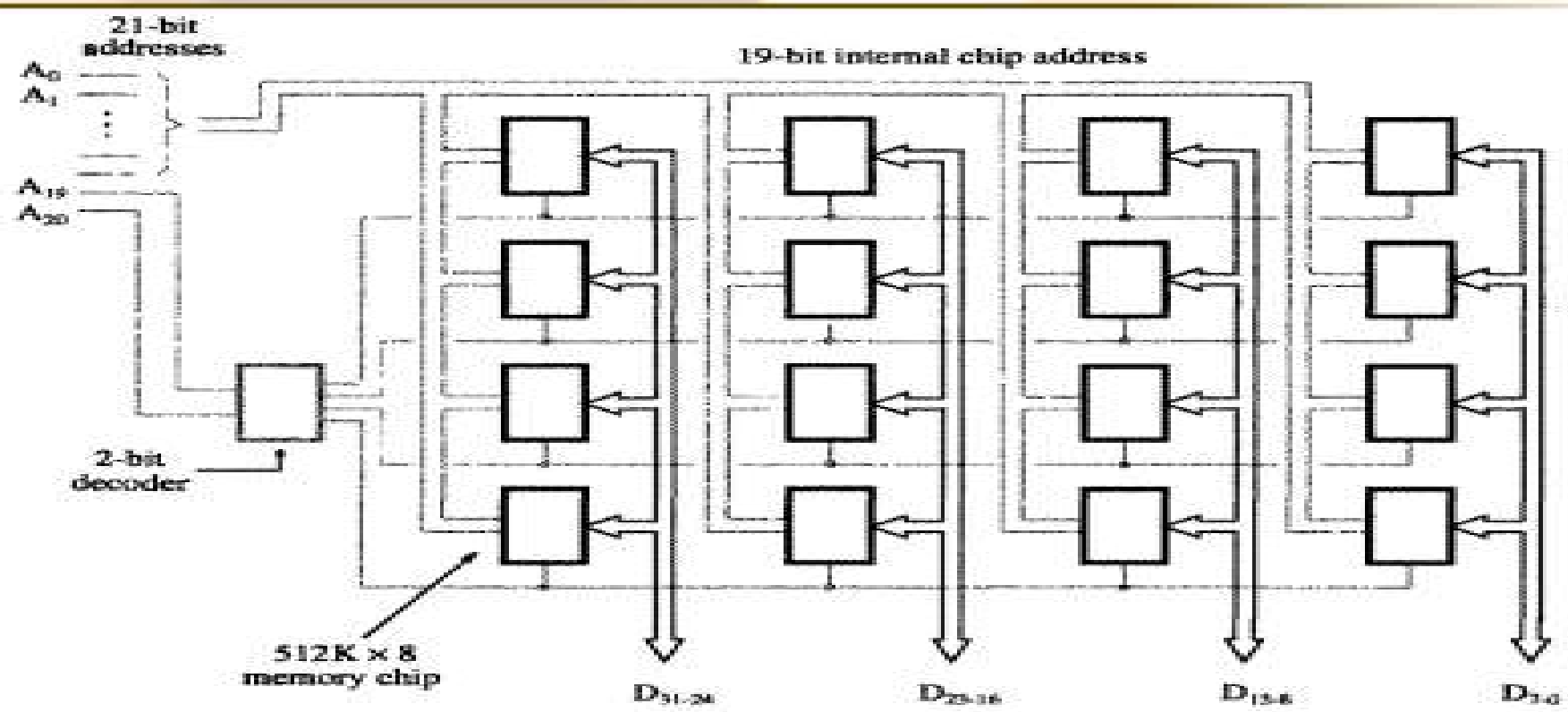
- Indication of performance : Latency and bandwidth
  Memory latency is the time it takes to transfer a word of data to or from memory
  Memory bandwidth is the number of bits or bytes that can be transferred in one second.

- Various techniques are used to transfer the bits quickly to the pins of the chip.
- To make the best use of the available clock speed, data are transferred externally on both the rising and falling edges of the clock. The memories that use this technique are called double-data-rate SDRAMs (DDR SDRAMs).

# Structure of larger memories

- Static Memory Systems
- Dynamic Memory Systems

# Structure of larger memories




[512 x 8] * 4 units = 2M x 32

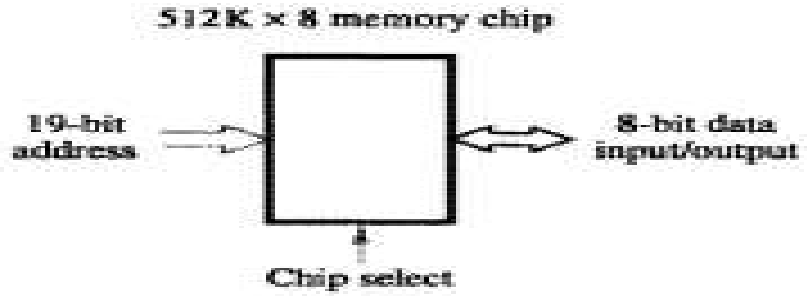



**Figure 8a** Organization of a 2M × 32 memory module using 512K × 8 static memory chips.

[Hamacher, Vranesic & Zaky, "Computer Organization" (5th Ed), McGraw Hill]