Unit 5 Memory System

Connection between processor and memory

- Address bus provides the address of the memory location to be accessed.
- Data bus transfers the data read from memory , or data to be written into memory.
  - Bidirectional
- Control bus provides various signals like READ, WRITE etc.

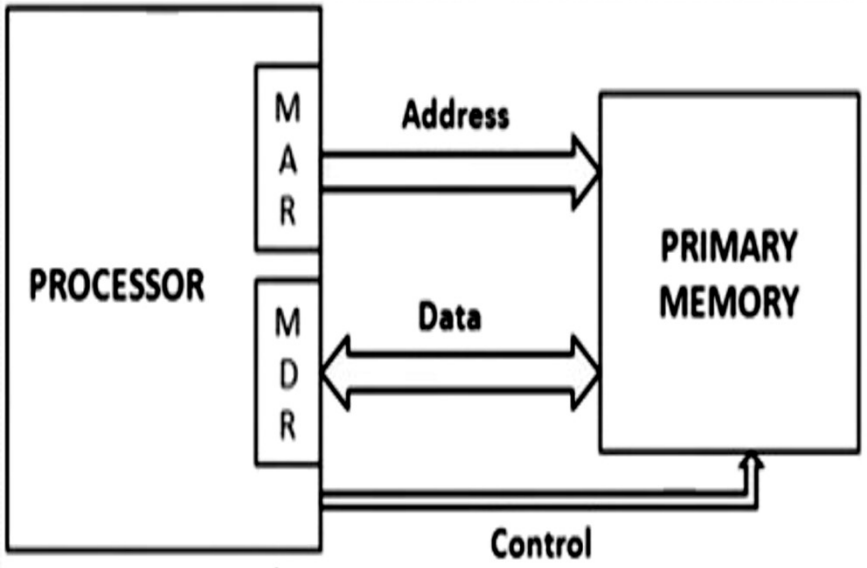


Figure 5.1 Connection between processor and memory

Memory System

- n-bit address bus can have a maximum of 2^n addressable memory locations.
- m-bit data at a time can be transferred to memory
- The RD/WR' control lines selects the memory for reading or writing (1:read, 0: write).
- The chip select line (CS') when active (=0) will enable the chip; otherwise, the data bus is in the **high impedance state**.

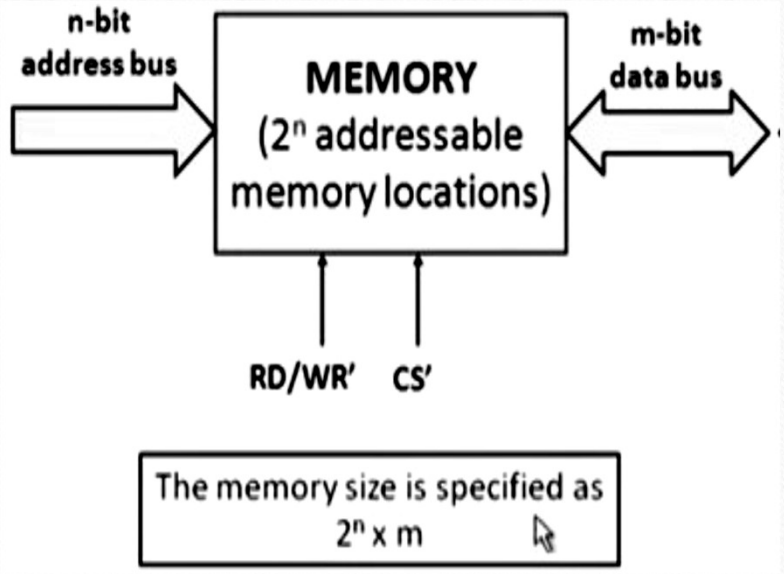


Figure 5.2 Memory System

Classification of Memory Systems

- **Volatile vs Non-volatile**
  - A volatile memory system is one where the stored data is lost when the power is switched off
    - Eg. : CMOS static memory and CMOS dynamic memory
    - Dynamic memory requires periodic refresh
- A non-volatile memory system is one where the stored data is retained even when the power is switched off.
  - Eg. : magnetic disk, CDROM, DVD, flash memory, and some resistive memories

Classification of Memory Systems

- **Random access versus direct or sequential access:**
  - A memory is said to be random-access when the read/write time is independent of the memory location being accessed.
    - Examples: CMOS memory (RAM and ROM).
  - A memory is said to be sequential access when the stored data can only be accessed sequentially in a particular order.
    - Examples: Magnetic tape, Punched paper tape.
  - A memory is said to be direct or semi-random access when part of the access is sequential and part is random.
    - Example: Magnetic disk.
    - We can directly go to a track after which access will be sequential.

Classification of Memory Systems

- **Read-only versus Random-access:**
  - Read-only Memory (ROM) is one where data once stored in permanent or semi-permanent.
    - Data written (programmed) during manufacture or in the laboratory.
    - Examples: ROM, PROM, EPROM, EEPROM.
  - Random Access Memory (RAM) is one where data access time is the same independent of the location (address).
    - Used in main/ cache memory systems.
    - Example: Static RAM (SRAM) -> data once written are retained as long as power is on.
    - Example: Dynamic RAM (DRAM) -> requires periodic refreshing even when power is on (data stored as charge on tiny capacitors).

Access time, latency and bandwidth

- Terminologies used to measure speed of the memory system.

a. **Memory Access Time**: Time between initiation of an operation (Read or Write) and completion of that operation.

b. **Latency:** Initial delay from the initiation of an operation to the time the first data is available.

c. **Bandwidth:** Maximum speed of the data transfer in bytes per second.

In modern memory organisation, every read request reads a block of the words into some high-speed registers (LATENCY), from where data are supplied to the processor one by one (ACCESS TIME).

Design Issue of Memory System

- The most important issue is to bridge the processor-memory gap that has been widening with every passing year. — Advancements in memory technology are unable to cope with faster advancements in processor technology.
- How to make the memory system work faster?
- HOW to increase the data transfer rate between CPU and memory?
- HOW address the ever increasing storage needs Of applications?

Solution:

- Cache Memory: to increase the effective speed of the memory system
- Virtual Memory: to increase the effective size of the memory

Cache Memory

- A fast memory (possibly organized in several levels) that sits between processor and main memory
- Faster than main memory and relatively small.
- Frequently accessed data a instructions are stored here.
- Cache memory makes use of the fast SRAM technology.

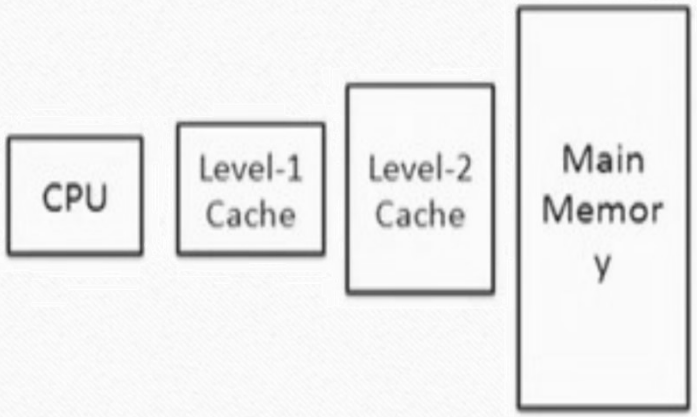


Figure 5.3 Cache Memory

Virtual Memory

- Technique used the operating system to provide an illusion Of very large memory to the processor
- Program and data are actually stored on secondary memory that is much larger.
- Transfer parts of program and data from secondary memory to main memory only when needed.

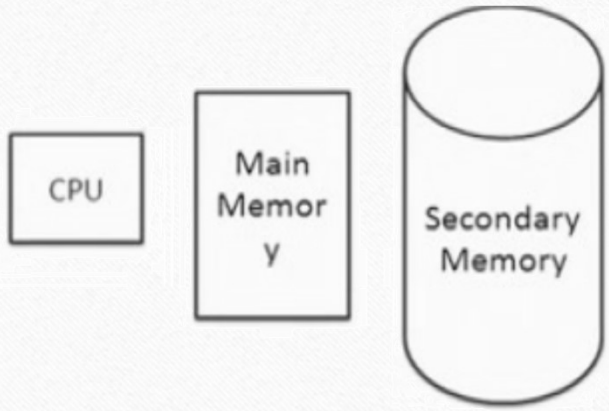


Figure 5.4 Virtual Memory

How a memory chip looks like?

- Memory cells are organized in the form of an array
- Every memory cell holds one bit Of data
- Present-day VLSI technology allows one to pack billions of bits per chip
- A memory module used in computers typically contains several such chips.

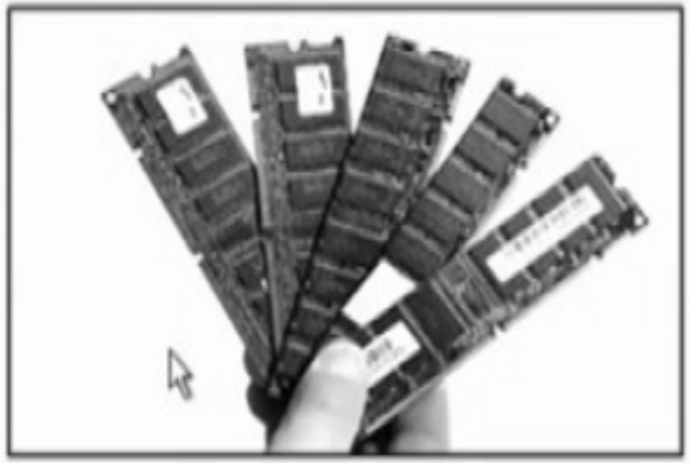


Figure 5.5 Memory Chip

Organization of Cells in 8x4 Memory Chip

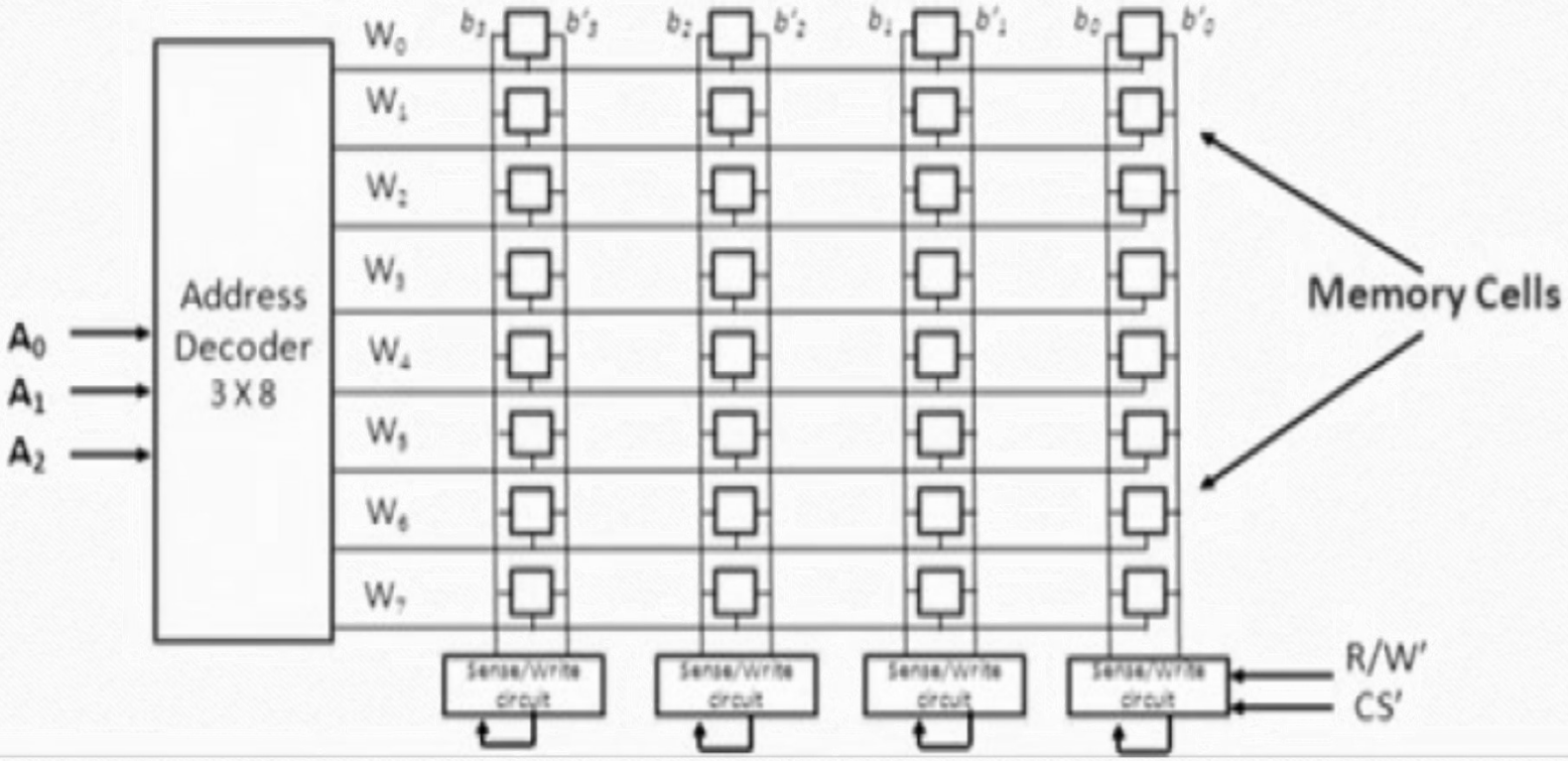


Figure 5.6 8x4 Memory Chip

Organization of Cells in 8x4 Memory Chip

- A 32 bit memory chip organized as 8 x 4 is shown in the figure 5.6
- Every row of the cell array constitutes a memory word
- A 3 x 8 decoder is required to access any one of the 8 rows
- The rows of the cells are connected to the word lines
- Individual cells are connected to two bit lines.
  - Bit b and its complement b'.
  - Required for reading and writing.
- Cells in each column are connected to a sense/write circuit by the two bit lines
- Other than address and data lines, there are two control lines: R/W' and CS' (Chip select)
  - CS is required to select One single Chip in a multi-chip memory system,

External Connection Requirements

- The 8 x 4 memory requires the following external connections:
- Address decoder of size: 3 x8
  - 3 external connections for address
- Data output : 4-bit
  - 4 external connections for data
- 2 external Connections for R/W' and CS’
- 2 external connections for power supply and ground
- Total of 3+4+2+2=11.

256x16 Memory System

Here the total number of external connections are estimated as follows.

- Address decoder size: 8 x 256
  - 8 external connections for address.
- Data output: 16 bit
  - 16 external connections for data
- 2 external connections for R/W' and CS'
- 2 external connections for power supply and ground.
- Total of 8+16 +2+2=28.

Static and Dynamic RAM

Broadly two types of semiconductor memory systems:

- a) Static Random Access Memory (SRAM)
- b) Dynamic Random Access Memory (DRAM)
  - i. Asynchronous DRAM
  - ii. Synchronous DRAM
- Vary in terms of speed, density, volatility properties, and cost.
- Present-day main memory systems are built using DRAM.
- Cache memory systems are built using SRAM

Static Random Access Memory(SRAM)

- SRAM consists of circuits which can store the data as long as power is applied
- It is a type of semiconductor memory that uses bistable latching circuitry (flip-flop) to store each bit
- SRAM memory arrays can be arranged in rows and columns of memory cells. — Called word line and bit line.

SRAM Technology

- Can be built using 4 or 6 MOS transistors
- Modern SRAM chips in the market uses 6-transistor implementations for CMOS compatibility
- Widely used in small-scale systems like microcontrollers and embedded systems
- Also used to implement cache memories in computer systems

A 1-bit SRAM Cell

- Two inverters are cross connected to form a latch
- The latch is connected to two bit lines with transistors T_1 and T_2
- Transistors behave like switches that can be opened (OFF) or closed (ON) under the control of the word line
- TO retain the State of the latch, the word line can be grounded which makes the transistors Off.

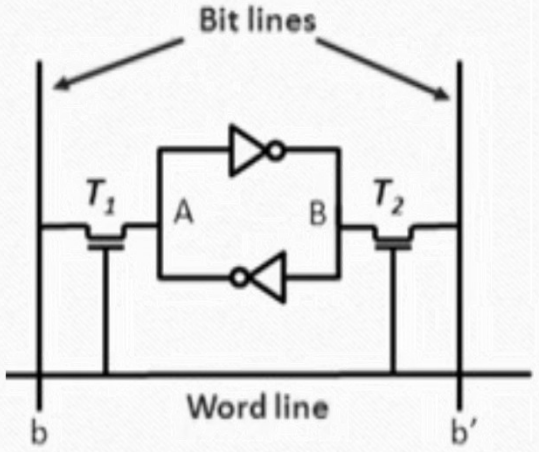
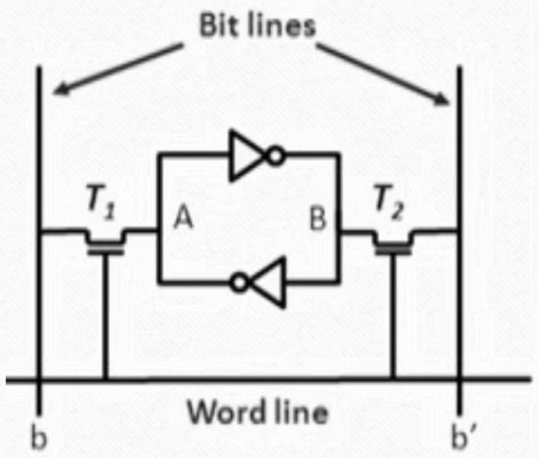


Figure 5.7 1-bit SRAM Cell

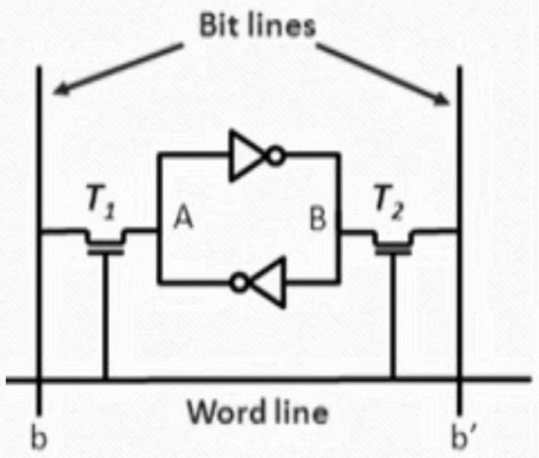
Read Operation in SRAM

- To read the content of the cell, the word line is activated (=1) to make the transistors T_1 and T_2 on
- The value stored in latch is available on bit line b and its complement on b'
- Sense/write circuits connected to the bit lines monitor the states of b and b'



Write Operation in SRAM

- To write 1: The bit line b is set with 1 and bit line b' is set with 0. Then the word line is activated and the data is written to the latch
- TO write 0: The bit line b is set with 0 and bit line b' is set with 1. Then the word line is activated and the data is written to the latch. The required signals (either 1 or 0) are generated by the sense/write circuit.



6 Transistor Static Memory Cell

- 1-bit SRAM cell with 6-transistors are used in modern-day SRAM implementations
- Transistors(T3&T5) and (T4&T6)form the CMOS inverters in the latch
- The data can be read or written in the same way as explained.

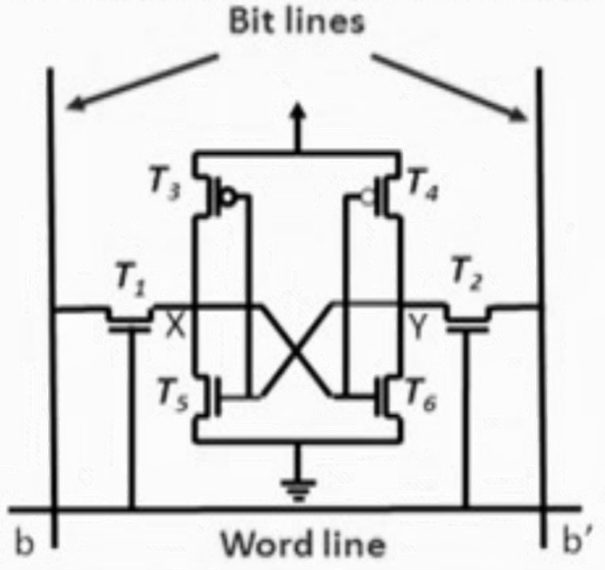
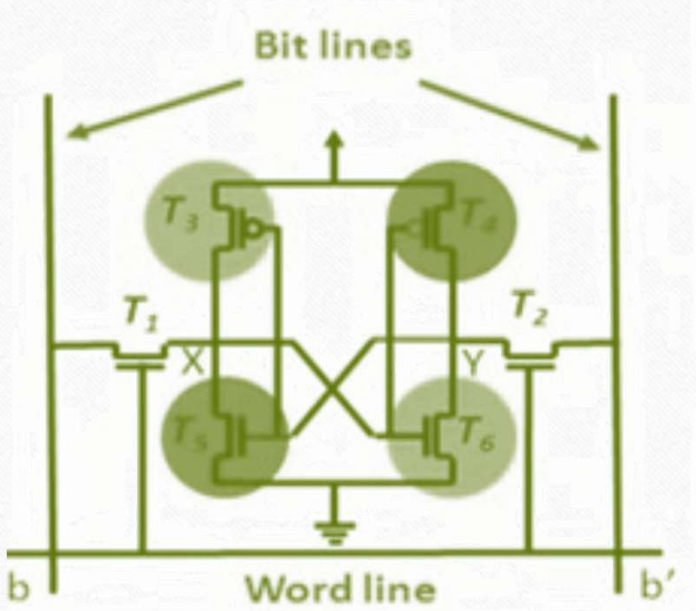


Figure 5.8 6 Transistor SRAM Cell

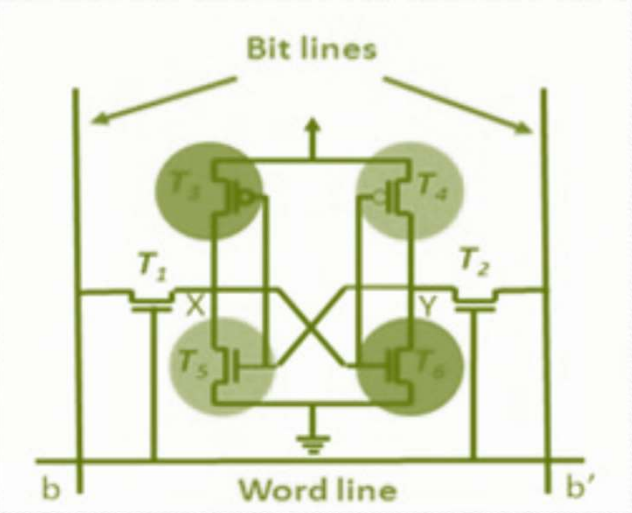
In State 0

- In state 0 the voltage at X is low and the voltage at Y is high
- When the voltage at X is low, transistors (T4 & T5) are on while (T3& T6) are off
- When word line is activated, and T2 are turned on and the bit lines b Will have 0 and b' Will have 1.



In State 1

- In state 1 the voltage at X is high and the voltage at Y is LOW
- When the voltage at X is high, transistors (T3& T6) are on while (T4& T5) are off
- When word line is activated,T1 and T2 are turned on and the bit lines b will and b' will have 0.



Features of SRAM

- Moderate/ High power consumption
  - Current flows in the cells only when the cell is accessed
  - Because of latch operation. power consumption is higher than DRAM
- Simplicity — refresh circuitry is not needed.
  - Volatile :: continuous power supply is required
- Fast operation
  - Access time is very fast; fast memories (cache) are built using SRAM
- High cost.
  - 6 transistors per cell.
- Limited capacity - Not economical to manufacture high-capacity SRAM chips.

Dynamic Random Access Memory(DRAM)

- Dynamic RAM do not retain its state even if power supply is on,
  - Data stored in the form of charge stored on a capacitor.
- Requires periodic refresh.
  - The charge stored cannot be retained over long time (due to leakage)
- Less expensive that SRAM
  - Requires less hardware lone transistor and one capacitor per cell)
- Address lines are multiplexed.

Dynamic Random Access Memory(DRAM)

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

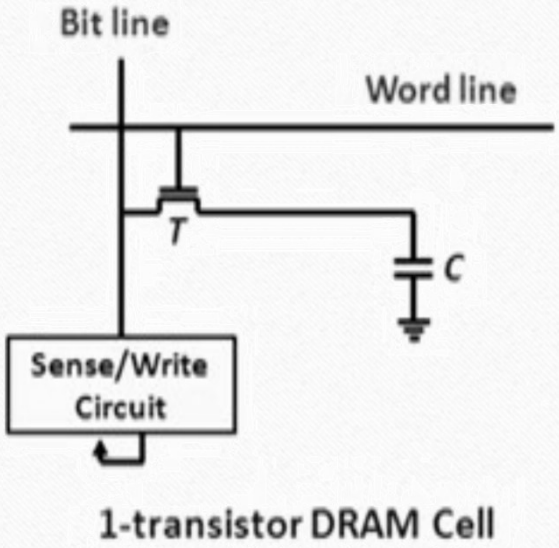
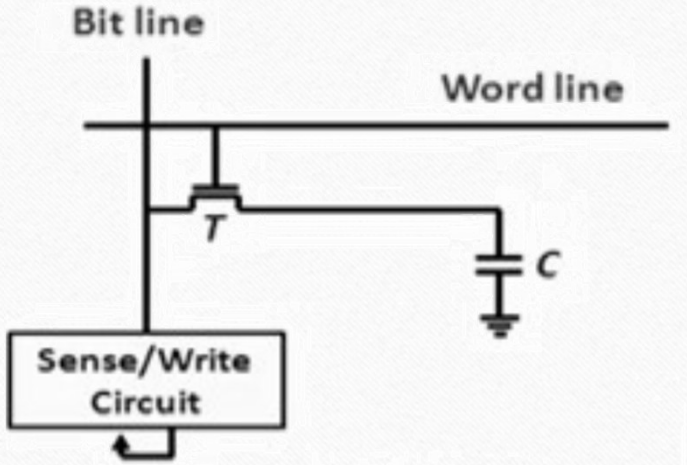


Figure 5.9

Read Operation in DRAM

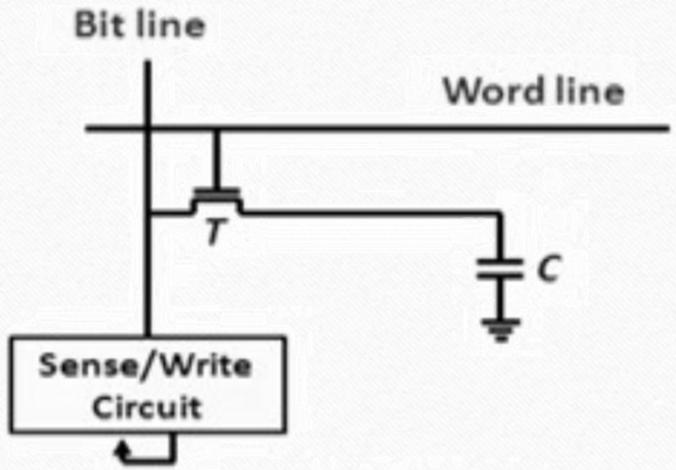
- The transistor of the particular cell is turned on by activating the word line
- A sense amplifier connected to bit line senses the charge stored in the capacitor
- If the charge is above threshold, the bit line is maintained at high voltage, which represents logic l
- If the charge is below threshold, the bit line is grounded, which represent logic 0



1-transistor DRAM Cell

Write Operation in DRAM

- The transistor of the particular cell is turned on by activating the word line
- Depending on the value to be written (0 or 1), an appropriate voltage is applied to the bit line
- The capacitor gets charged to the required voltage state
- Refreshing Of the capacitor requires periodic READ-WRITE cycles (every few m sec).



1-transistor DRAM Cell

Types of DRAM

Asynchronous DRAM (ADRAM)

- Timing of the memory device is handled asynchronouslv
- A special memory controller circuit generates the signals asynchronously.
- DRAM chips produced between the early 1970s to mid•1990s used asynchronous DRAM.

Synchronous DRAM(SDRAM)

- Memory operations are synchronized by a clock.
- Concept of SDRAM Came in the 1993 by Samsung. – by 2000
- SDRAM replaced almost all types Of DRAMs in the market.
- Performance of SDRAM is much higher compared to all other existing DRAM,

Asynchronous DRAM

- The timing of the memory device is controlled asynchronously
- The device connected to this memory is responsible for the delay.
- Address lines are divided into two parts and multiplexed.
  - — Upper half Of address: Loaded into ROW Address Latch using ROW Address Strobe (RAS).
  - — Lower half Of address: Loaded into Column Address Latch using Column Address Strobe (CAS).

Internal Organization of DRAM Chip

- Cells are organized in the form of an array, in rows and columns.
- Cells of each row are divided into fixed number of columns, m bits each.
- m is 8, 16. 32 or 64.

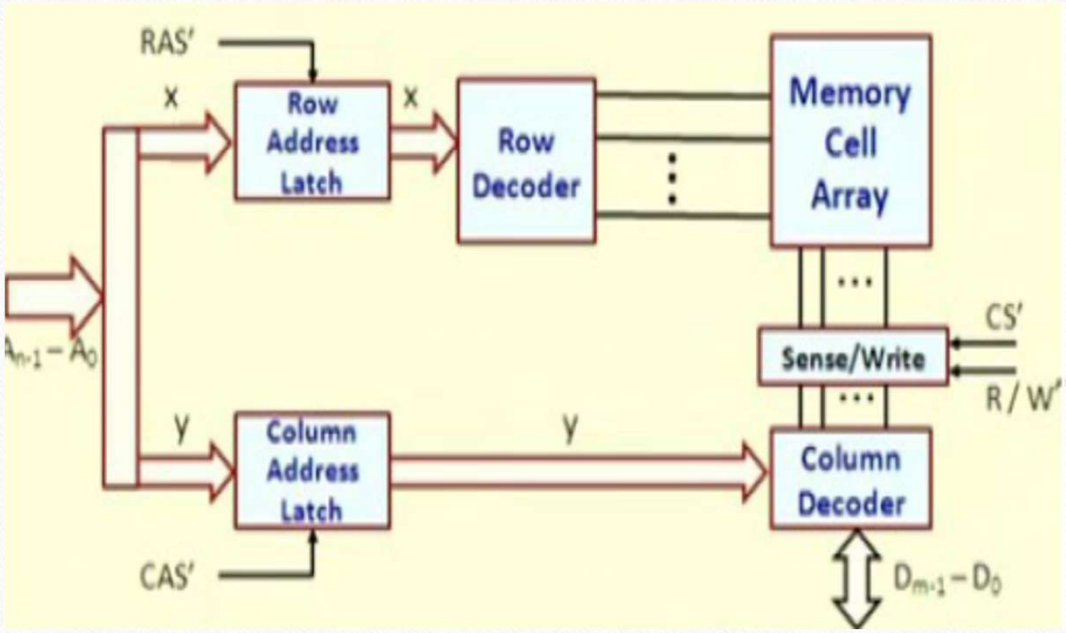


Figure 5.10

Internal Organization of DRAM Chip

- Suppose that the memory cell array is organized as r x c. — r rows and c columns,
- An x-bit address is required to select a row r, where $x = \log_2 r$.
- An y-bit address is required to select a column c, where $y = \log_2 c$.
- Total address bits: n x (high order) + y (low order)

READ or WRITE Operation

- For a read operation, the x-bit row address is applied first.
  - It is loaded into Row Address latch in response to the signal RAS'.
  - The read operation is performed in which all the cells of the selected row are read and refreshed.
  - After loading of row address, the column address is selected.
  - In response to CAS' the column address is loaded into Column Address latch.
  - Then the column decoder selects a particular column from c columns and an appropriate group of m sense/write circuits are selected.

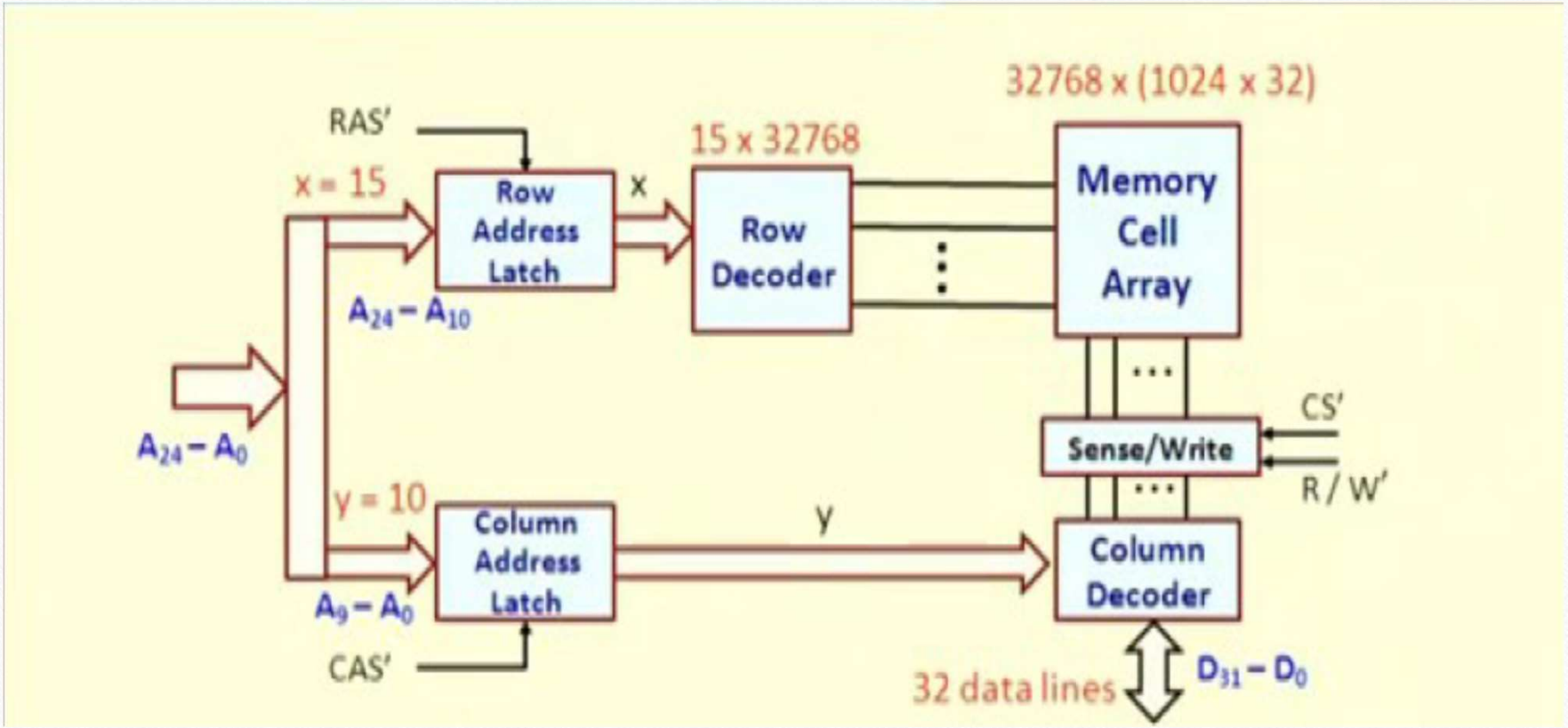
READ or WRITE Operation

- For a READ operation, the output values of the selected circuits are transferred to data lines D_{m-1} to D_0.
- For a WRITE operation, the data available on the data lines D_{m-1} to D_0 is transferred to the selected circuits.
  - This information is stored in the selected cell.
- Both *RAS'* and *CAS'* are active low signals. That is they cause latching the addresses when they move from high to low.
- Each row of the cell array must be periodically refreshed to prevent data loss.
- Cost is low but access time is high compared to SRAM.
- Very high packing density (few billion cells per chip).
- Widely used in the main memory of modern computer systems.

An Example: 1 Gbit ADRAM Chip

- We assume that the 1 Gbit memory cells are organized as 32768 (2^{15}) rows and 32768 (2^{15}) columns.
- Let us assume that data bus is 32-bit long.
- So, the memory can be organized as (2^{15}) x (2^{10} x 2^{5}).
  - Total number of address lines is 25 bits.
- High order 15 bits of the address is used to select a row.
  - Requires a 15 x 32768 row-address decoder.
- Low order 10 bits of the address is used to select a column.
  - Requires a 10 x 1024 column decoder.

Example-1 Gbit ADRAM Chip



Example-1 Gbit ADRAM Chip

- Operation: 15-bit row address is selected (i.e., x =15)
- With the help of RAS control signal the row address is latched. The 15 x 32768 Row Decoder selects a particular row.
- Then the 10-bit column address is applied and with the help of CAS the address is latched. The 10 x 1024 column decoder selects a particular column.
- A group of 32 bits are selected as the 32-bit word to be accessed

Fast Page Mode

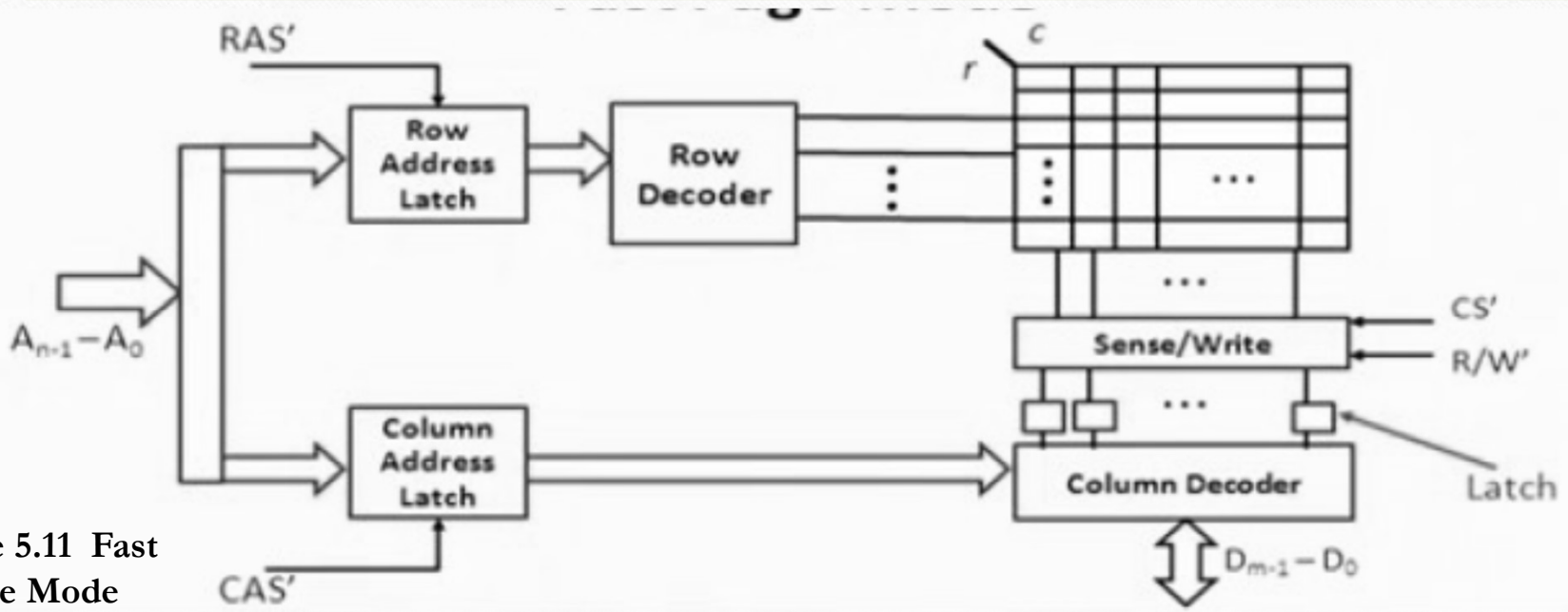


Figure 5.11 Fast Page Mode

Fast Page Mode

Operation:

- When the DRAM cell is accessed only an m bit word (data bus width) is transferred
- But when we select a row, we can select not only the data of a single column but multiple columns as well
- A latch can be connected at the output of the sense amplifier in each column
- Once we apply a row address. the row get selected
  - Different column addresses are required to place different bytes on the data lines
- Hence consecutive bytes can be transferred by applying consecutive column addresses under the control of successive CAS signals.
- This also helps in faster transfer of blocks of data.
- This block transfer capability is termed as *Fast Page Mode access.*

Synchronous DRAM

- SDRAM is the commonly used name for various kinds Of dynamic RAM that are synchronized with clock
- The Structure Of this memory is same as asynchronous DRAM
- The concept of SDRAM were known from 70's but it is first developed by Samsung in the year 1993 (KM48SL2000).
  - By 2000 all kinds of DRAM were replaced by SDRAM.

Internal Organization of SDRAM chip

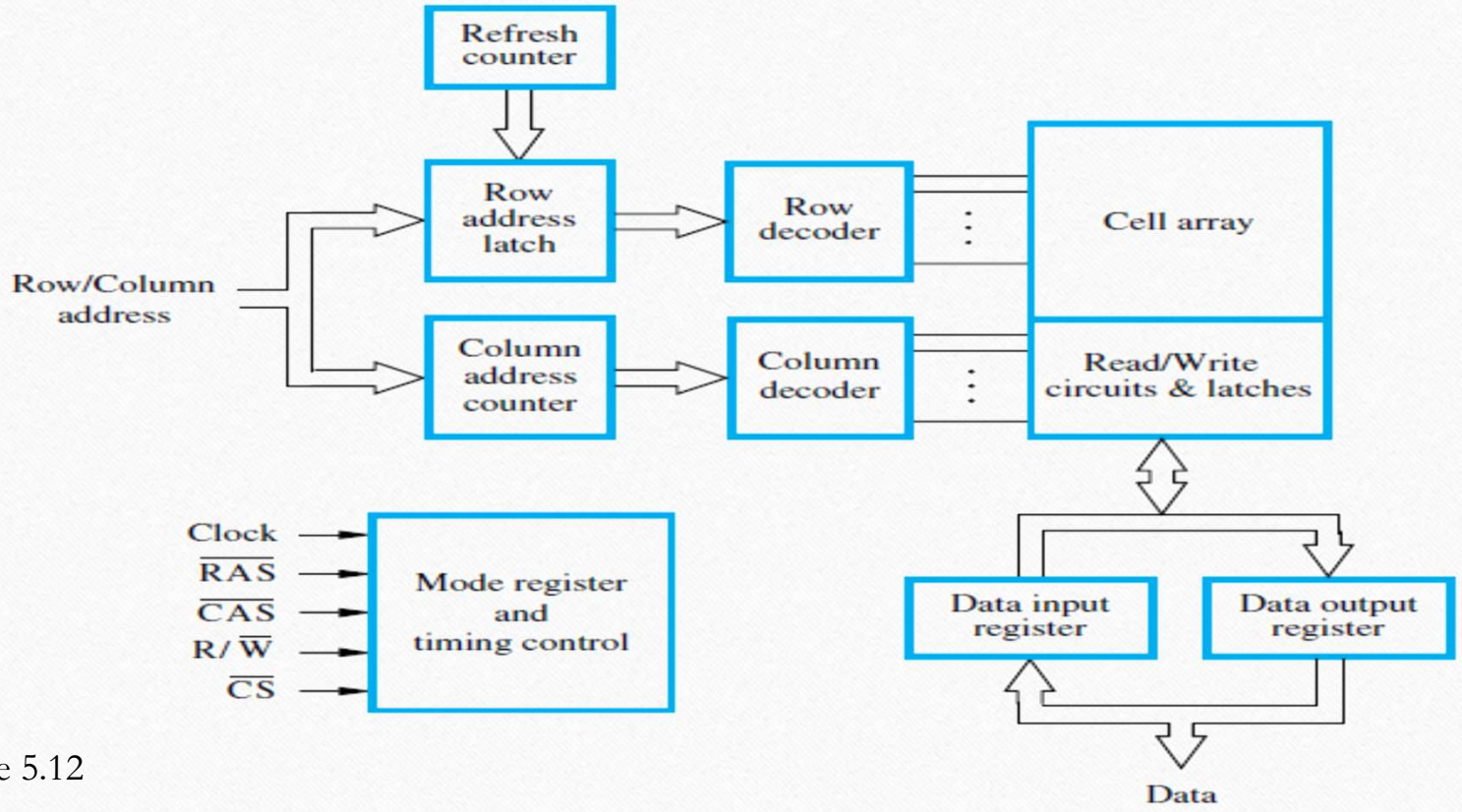


Figure 5.12

Internal Organization of SDRAM chip

- In SDRAM, address and data connections are buffered by registers
- The output of individual sense amplifier is connected to a latch
- Mode register is present which can be set to operate the memory chip in different modes
- To select successive columns it is not required to provide externally generated pulses on CAS line
- A column counter is used internally to generate the required signals.

Read and Write Operation

- For READ operation, the row address is applied first, and in response to the column address, the data present in the latches for the selected columns are transferred to the data output register.
  - Then the data is available on the data bus.
- For WRITE operation, the row address is applied first, and in response to the column address, the data present in the data bus is made available to the latches through data input register.
  - The data is then written to the particular cell.

Burst Mode Transfer

- As in fast page mode, after the column address is applied the data from successive column addresses are read out or written into
- In same way burst operation of different lengths can be specified
- This uses the same block transfer capability of fast page mode.
- Here the control signal required are provided internally by column counter and clock signals
- New set of data are available in the data lines after every clock cycle
- All the operations are triggered at the rising edge of the clock
- It has built-in refresh circuitry, and refresh counter is part of it.