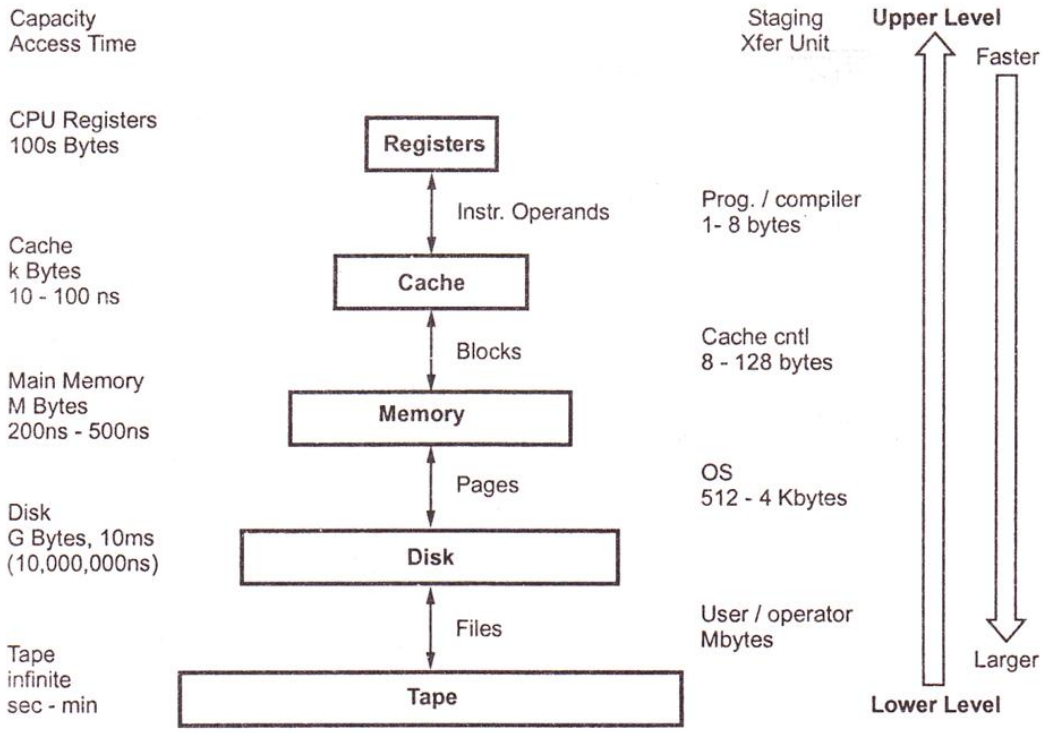
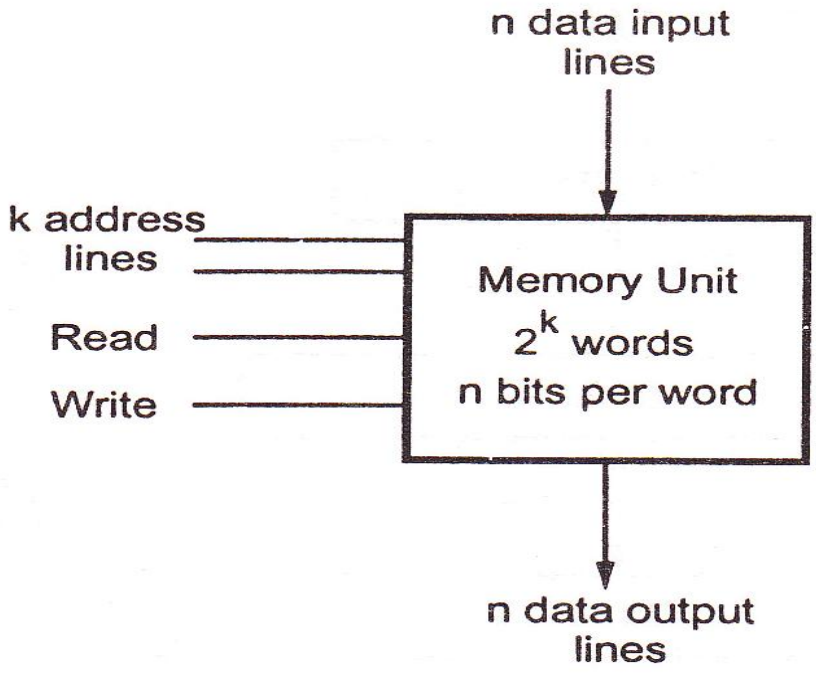
Memory Elements

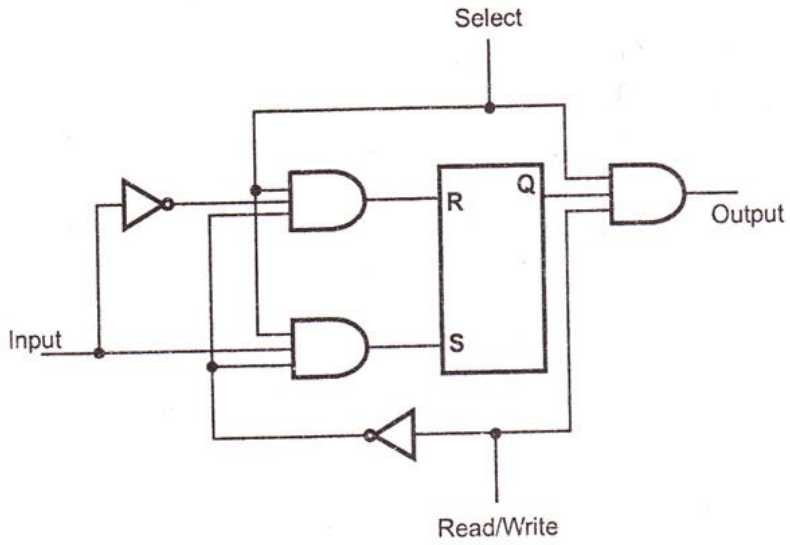
9.3.1 Levels of Memory Hierarchy

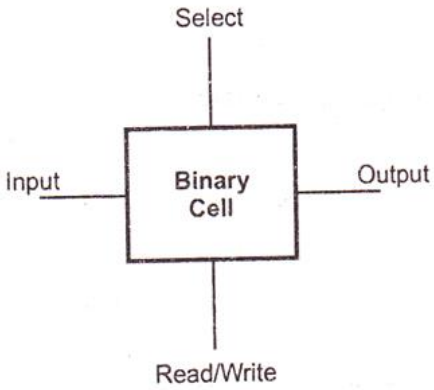
- Fig. 9.10 indicates different memory storage devices with their access time and the data transfer unit.



- The memory unit is specified by both :
  1. The number of words it contains
  2. The number of bits in each word.







- Implementation of a logic/binary cell for storing one-bit needs

i) 3 inputs and 1 output

ii) The *Data* input

iii) The *Select* input enables the cell for reading and writing.

iv) The *Read/Write* input determines the cell operation when it is selected.

v) Logic 1 in the Read/Write input provides the read operation by forming a path for the flip-flop to the output.

vi) Logic 0 in the Read/Write input provides the write operation by forming a path form the input terminal to the flip-flop.

Memory Basics

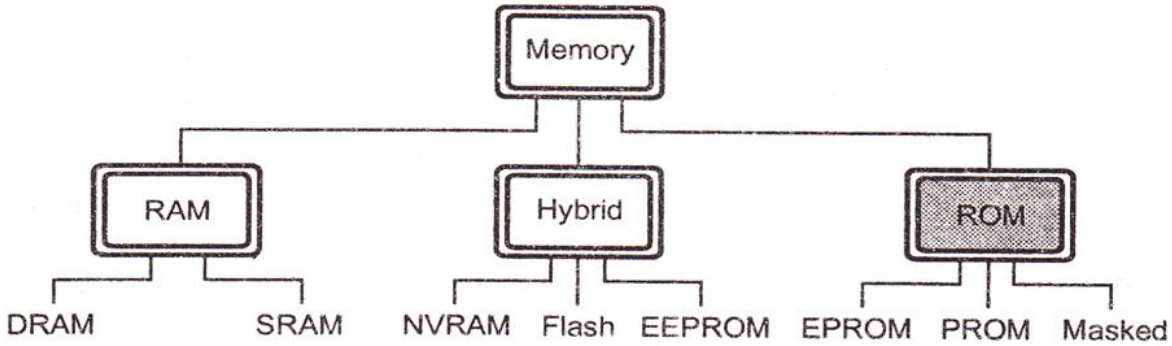
1. Memory elements stores a value as controlled by clock. It may have load signal, etc.
2. In CMOS, memory is created by :
   i) Capacitance (dynamic);
   ii) Feedback (static).

Variations in memory elements

- Memory elements can be distinguished on the basis of -

1. Form of required clock signal.
2. How behaviour of data input around clock affects the stored value.
3. When the stored value is presented to the output.
4. Whether there is ever a combinational path from input to output.

Types of Memory



RAM : Random Access Memory

- RAM is defined as memory array with individual bit access.
- It refers to memory with both Read and Write capabilities.
- There are two types of RAM :
  A) Static RAM (SRAM) and
  B) Dynamic RAM (DRAM)

ROM : Read Only Memory

- It has no capabilities for "online" memory Write operations.
- Write operation typically requires high voltages or erasing by UV light.
- There are 4 types of ROMs :
  a) Programmable ROM (PROM)
  b) Erasable Programmable ROM (EPROM)
  c) Electrically Erasable Programmable ROM (EEPROM)
  d) Flash ROM.

Memory Size Conversions

	Bytes			
Kilobyte (kB)	1,024	Kilobytes (kB)		
Megabyte (MB)	1,048,576	1,024	Megabytes (MB)	
Gigabyte (GB)	1,073,741,824	1,048,576	1,024	Gigabytes (GB)
Terabyte (TB)	1,099,511,627,776	1,073,741,824	1,048,576	1,024

Memory Element Parameters

1. **Setup time :** Time before clock during which data input must be stable.
2. **Hold time :** Time after clock event for which data input must remain stable.

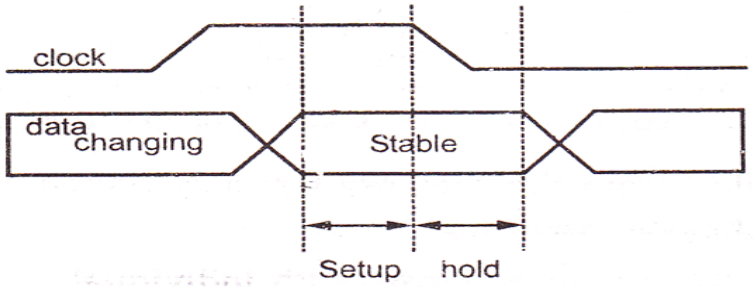


Fig. 9.14 Setup and hold times

4. Static Vs Dynamic Memory :

i) Static : Holds data as long as power is applied (SRAM).

ii) Dynamic : Must be refreshed periodically (DRAM).

Random Access Memory (RAM)

- Store data and instructions that are used by the CPU to perform some task. These instructions are usually loaded into RAM from a secondary storage device.
- RAM is also used to store instructions that tell the CPU how to work with its parts. These instructions are usually called drivers.
- The instructions in RAM are constantly changing, depending on the needs of the CPU.
- The instructions in RAM are volatile. When the computer is turned-off the information in RAM disappears.
- The information in RAM needs to be saved to secondary storage before the computer is turned-off.

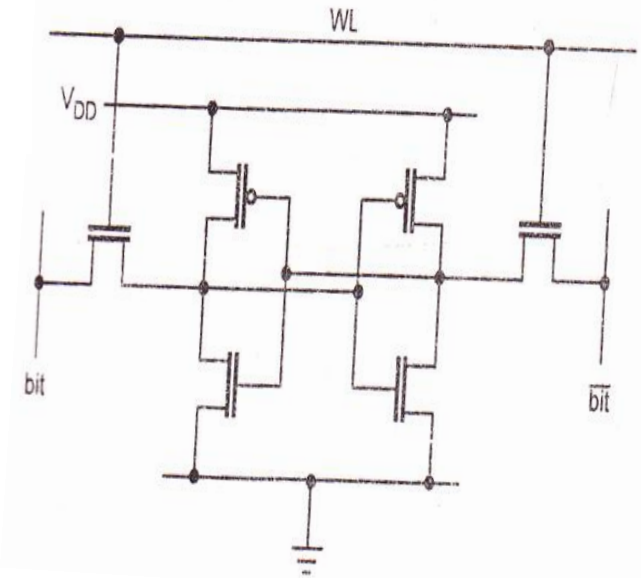
SRAM

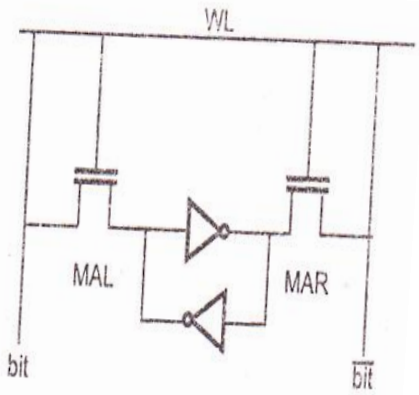
- SRAM means Static Random Access Memory. Random access means that locations in the memory can be written to or read from in any order, regardless of the memory location that was last accessed.
  1) Static : Holds data as long as power is applied.
  2) Volatile : Cannot hold data if power is removed.
- Static Random Access Memory (SRAM) is widely used in integrated circuits. All of the popular microprocessors have elaborate on-chip caching schemes for improved performance. All of these caches use SRAM.
- A SRAM cell has three different states it can be in :
  1) Standby or hold where the circuit is idle,
  2) Reading when the data has been requested and
  3) Writing when updating the data

Advantages of SRAM

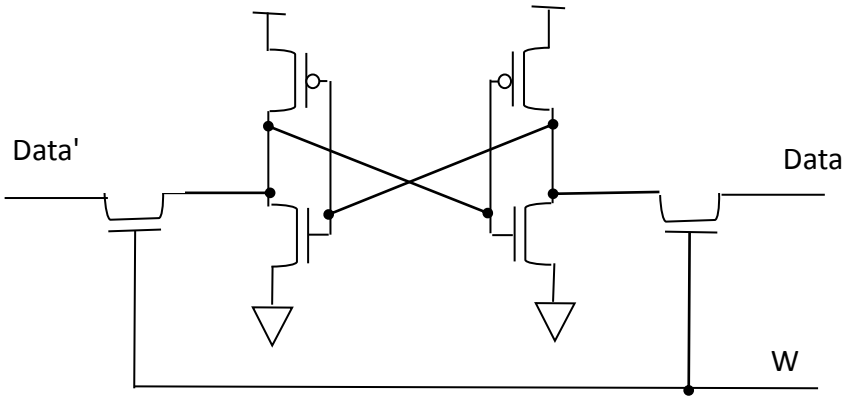
1. Compatibility with modern logic - optimized CMOS manufacturing processes.
2. High speed - fastest of all semiconductor memories.
3. Ease of use - no refresh necessary.
4. High density - high bits per area can be achieved.
5. Versatility :
   i) Embedded SRAM size can range from less than 1 kB to 10's of MB.
   ii) Basic cell structure can be modified to support multi-port access.

Basic 6T (Transistor) SRAM Cell



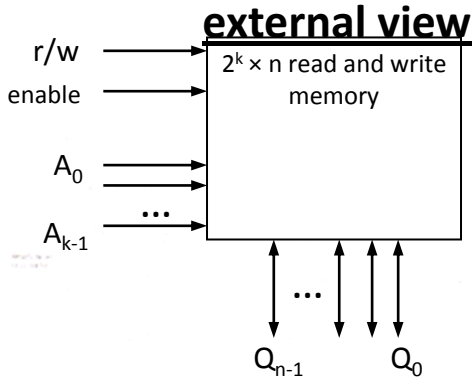


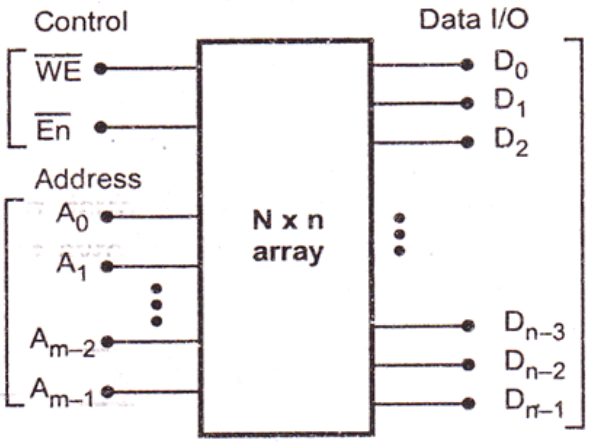
SRAM

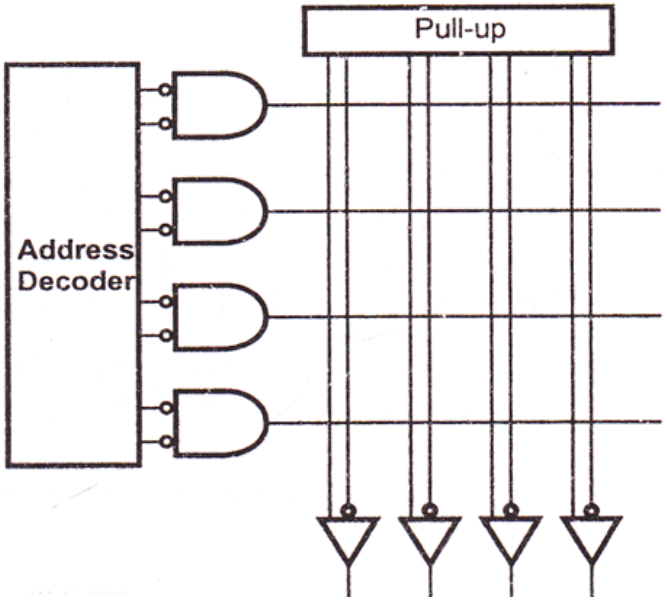


SRAM Array

- N × n array of 1-bit cells
  1) n = Byte width; 8, 16, 32 etc.
  2) N = Number of bytes
  3) m = Number of address bits
  4) max N = 2 m

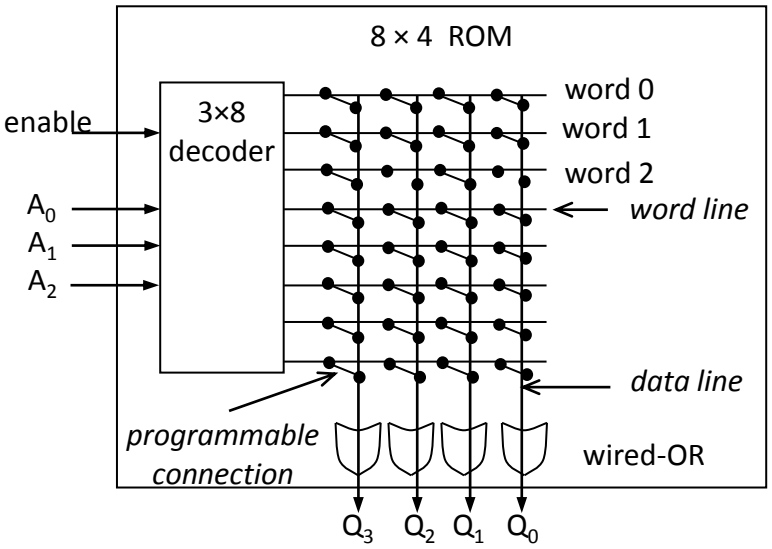




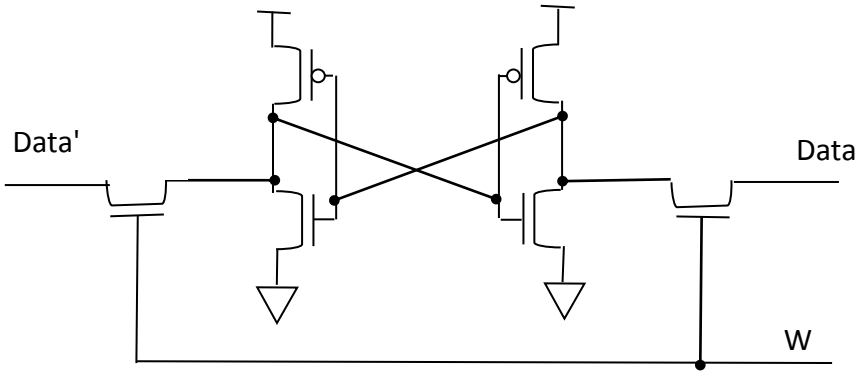


- Horizontal lines = words
- Vertical lines = data
- Lines connected only at circles
- Decoder sets word 2's line to 1 if address input is 010
- Data lines Q_3 and Q_1 are set to 1 because there is a "programmed" connection with word 2's line
- Word 2 is not connected with data lines Q_2 and Q_0
- Output is 1010

Internal view



SRAM



SRAM Write Operation

- The SRAM writes into the latch through the NMOS transistors on the left and right sides of the cell. The problem is that the coupled inverters will resist state changes. Their resistance is lowered by making them small and by overpowering them with a write driver in the column amplifier.

SRAM Read Operation

- The SRAM reads the value stored in the latch using a differential amplifier. The transistors in the SRAM cell do not produce good logic levels because 1) they have small sizes (so we can write into them), 2) they have to pass through NMOS transistors which have non-ideal (resistive) behaviour when passing a "one" value and 3) they may be driving a significant capacitive load. The poor output levels that they produce must be translated into valid logic levels and a special amplifier is required to do the translation.

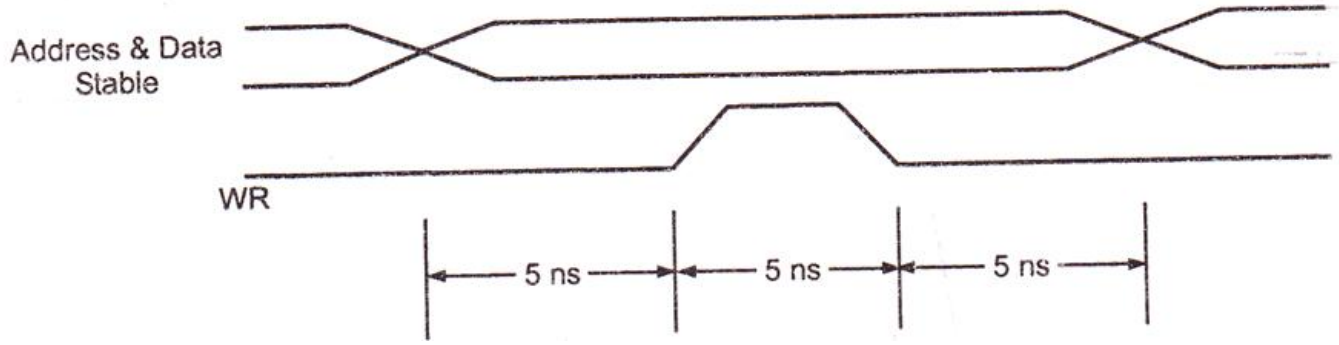
Read Timing

- Read timing in the SRAM is similar to the ROM. A valid address is presented and a short time later the contents of the SRAM cell will appear at the Q output. The primary difference is that the SRAM is more sensitive to the number of rows because the NMOS transistors associated with Bit and nBit are not designed to drive those signals poorly.

SRAM configuration (rows × columns)	Read timing Address to output
4 × 16	2.2 ns
4 × 32	3.2 ns
16 × 16	2.7 ns
4 × 1024	37.9 ns

Write Timing

- Writing an SRAM requires a little care. The column write driver must not drive the Bit and nBit lines before the addresses stabilize. Fig. 9.17 shows the waveform required for writing into the SRAM.



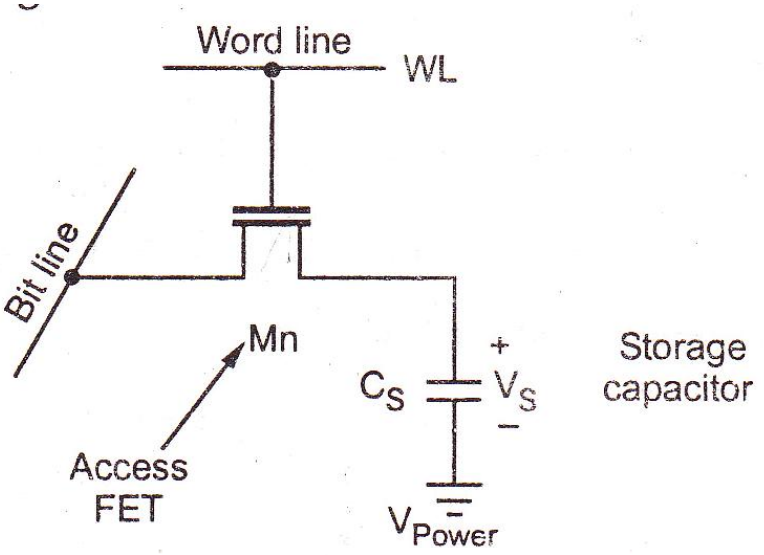
DRAM

- The term **dynamic** indicates that the memory must be constantly **refreshed** (re-energized) or it will lose its contents. **RAM** (Random Access Memory) is sometimes referred to as **DRAM** to distinguish it from **static RAM (SRAM)**. Static RAM is faster and less volatile than dynamic RAM, but it requires more power and is more expensive.

  1. **Dynamic :** Must be refreshed periodically.
  2. **Volatile** : Loses data when power is removed.

- DRAM is manufactured using a similar process to how processors are : a silicon substrate is etched with the patterns that make the transistors and capacitors (and support structures) that comprise each bit. DRAM costs much less than a processor because it is a series of simple, repeated structures, so there is not the complexity of making a single chip with several million individually-located transistors.

1-Transistor DRAM Cell

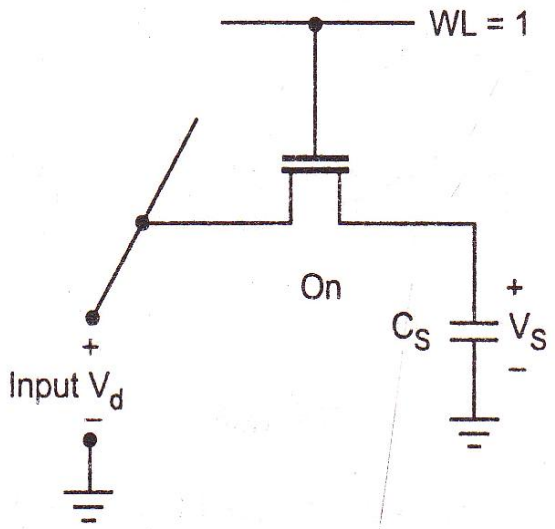


DRAM Operation

- RAM data is held on the storage capacitor.
  i) Temporary - due to leakage currents which drain charge.
- Charge storage
  i) If C_S is charged to V_S
  ii) $Q_S = C_S\ V_S$
- If $V_S = 0$, then $Q_S = 0$: LOGIC 0.
- If V_S = large, then $Q_S > 0$: LOGIC 1.

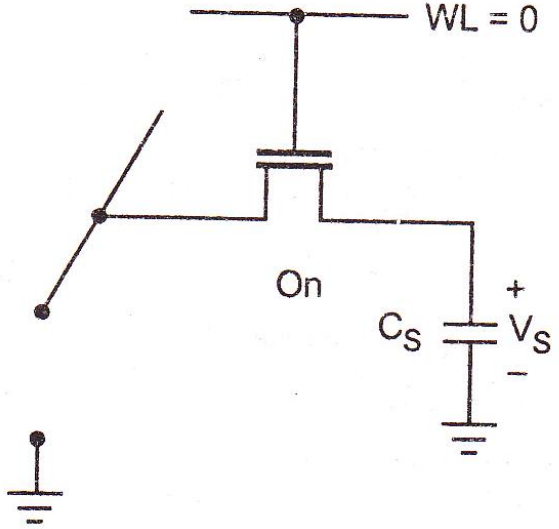
1. Write Operation

- Turn-on access transistor : WL = V_{DD}.
  - i) Apply voltage, V_d (high or low), to bit line.
  - ii) C_S is charged (or discharged).
  - iii) If $V_d = 0$.
- $V_S = 0$, $Q_S = 0$, store logic 0.
  - i) If $V_d = V_{DD}$.
- $V_S = V_{DD} - V_{th}$, $Q_S = C_S (V_{DD} = V_{th})$, logic 1.



2. Hold Operation

- Turn-off access transistor : WL = 0.
- Charge held on C_S.
- During hold, leakage currents will slowly discharge C_S.
  - i) Due to leakage in the access transistor when it is OFF.
- If I_L is known, can determine discharge time.



3. Hold Time, t_h

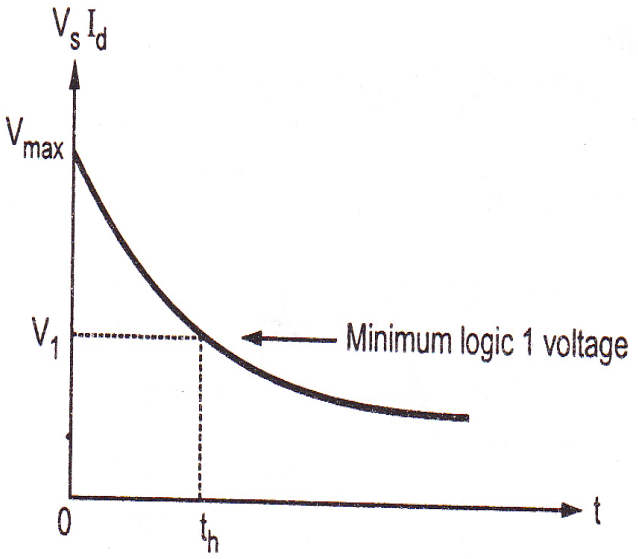
i) Max time voltage on C_S is high enough to be a logic 1.

- Time to discharge from V_{max} to V_1 (in Fig. 9.20).

ii) $t_h = (C_S / I_L)\ (V_S)$, if we estimate I_L as a constant.

- Desire large hold time.
- t_h increases with larger C_S and lower I_L.
- Typical value, $t_h = 0.5$ sec.

iii) With $I_L = 1$ nA, $C_S = 50$ pF, and $V_S = 1$ V.



4. Refresh Operation

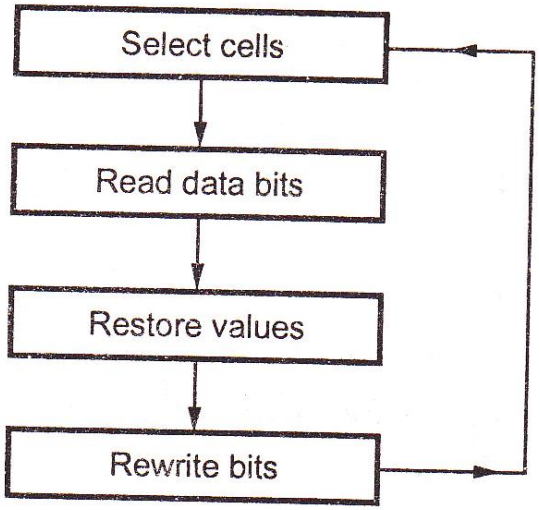
- To hold data as long as power is applied, data must be refreshed.
- Periodically read every cell.

i) Amplify cell data.

ii) Rewrite data to cell.

iii) Refresh rate, $f_{refresh}$.

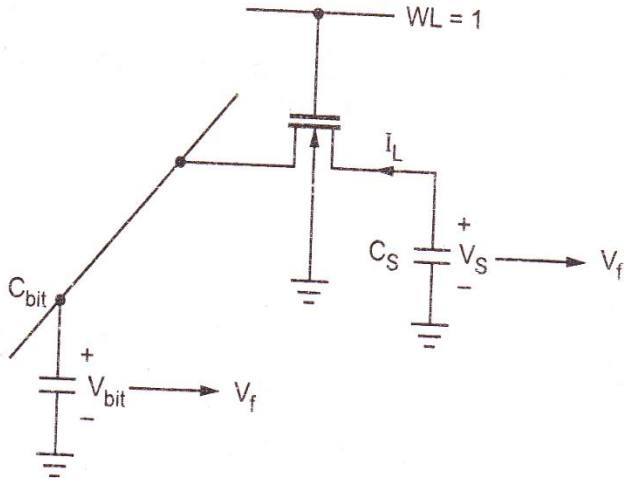
- Frequency at which cells must be refreshed to maintain data.
- $f_{refresh} = 1/2^{th}$.
- Must include refresh circuitry in a DRAM circuit.



5. DRAM Read Operation

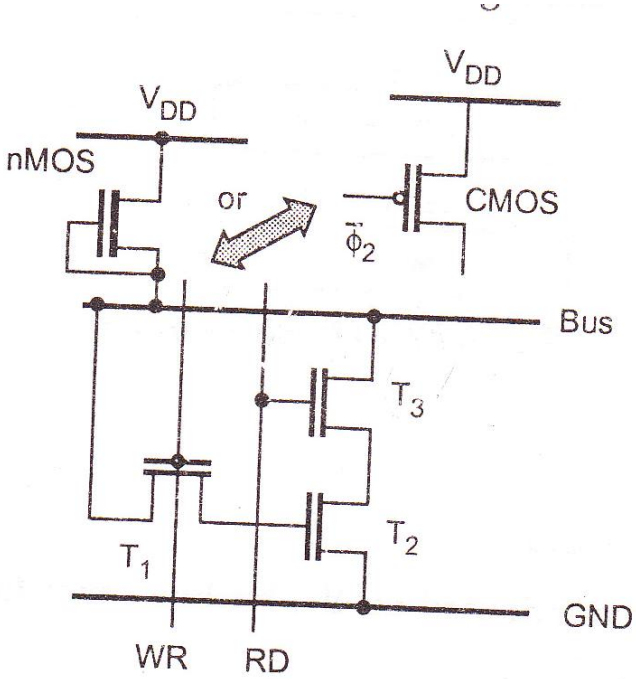
Read Operation

- Turn on access transistor.
- Charge on C_S is redistributed on the bit line capacitance, C_{bit}.
- This will change the bit line voltage, V_{bit}.
- Which is amplified to read a 1 or 0.



3-Transistor DRAM Cell

- When control line RD is LOW, then a bit may be read from bus through T_1 by making WR to HIGH. This in turns WR to LOW through T_2. The bit value is stored by gate capacitance (C_g) of T_2 while RD and WR are LOW.
- For reading stored bit RD is mode HIGH, is it pulls down bus to ground through T_3 and T_2 if a logic '1' was stored. Otherwise T_2 will not conduct and bus will remain HIGH because of pull-up arrangements.



Difference between SRAM and DRAM

- Static - SRAM

i) Data is stored as long as supply is applied.

ii) Large cells (6 FETs/cell) - so fewer bits/chip.

iii) Fast - so used where speed is important (e.g., caches).

iv) Differential outputs (output BL and !BL).

v) Use sense amps for performance.

vi) Compatible with CMOS technology.

Dynamic - DRAM

i) Periodic refresh required (every 1 to 4 ms) to compensate for the charge loss caused by leakage.

ii) Small cells (1 to 3 FETs/cell) - so more bits/chip.

iii) Slower - so used for main memories.

iv) Single ended output (output BL only).

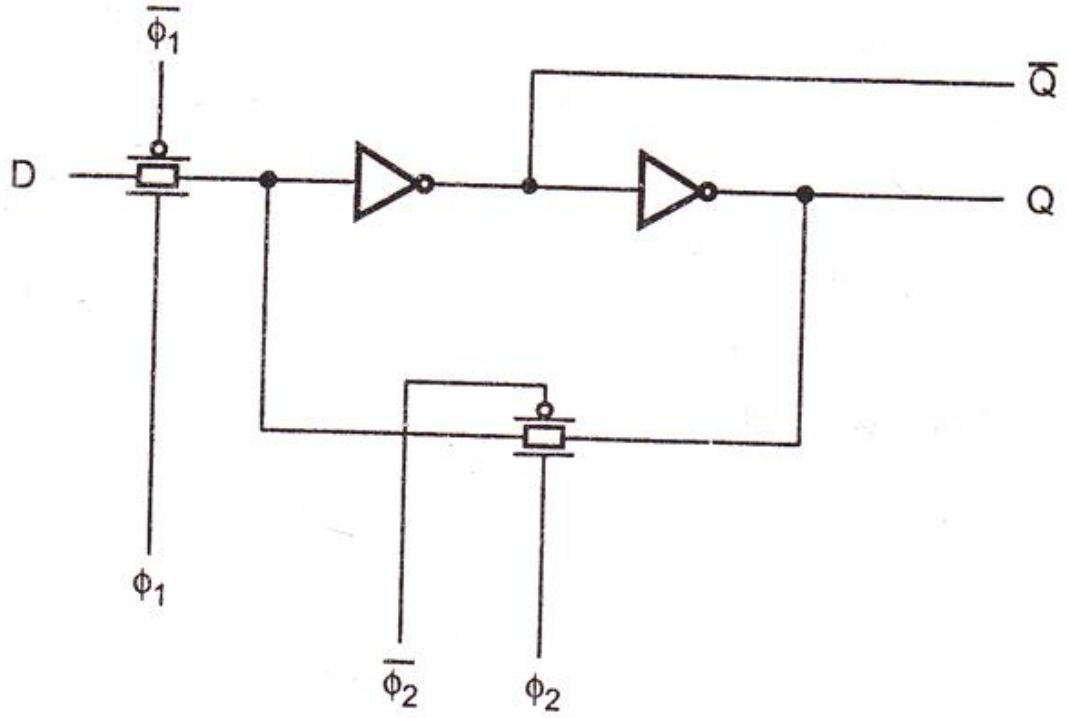
v) Meed sense amps for correct operation.

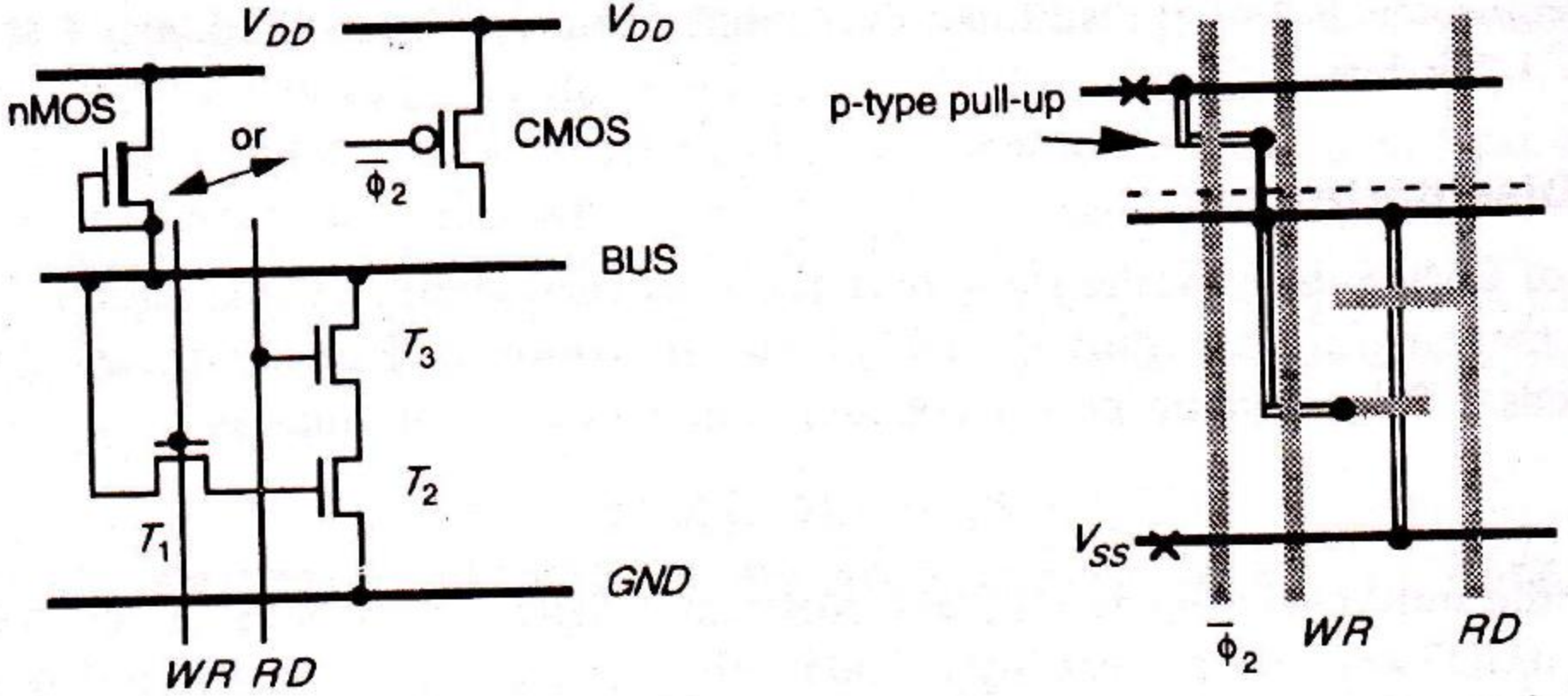
vi) Not typically compatible with CMOS technology.

Type	Volatile	Writeable	Erase Size	Max Erase Cycles	Cost (per byte)	Speed
SRAM	Yes	Yes	Byte	Unlimited	Expensive	Fast
DRAM	Yes	Yes	Byte	Unlimited	Moderate	Moderate
Masked ROM	No	No	n/a	n/a	Inexpensive	Fast
PROM	No	Once, with a device programmer	n/a	n/a	Moderate	Fast
EPROM	No	Yes, with a device programmer	Entire chip	Limited (consult datasheet)	Moderate	Fast
EEPROM	No	Yes	Byte	Limited (consult datasheet)	Expensive	Fast to read, slow to erase/write.
Flash	No	Yes	Sector	Limited (consult datasheet)	Moderate	Fast to read, slow to erase/write.

CMOS Static Pseudo-static D Flip-Flop

- A simple version of CMOS D flip-flop with pseudo-static approach is shown in Fig. 9.25.



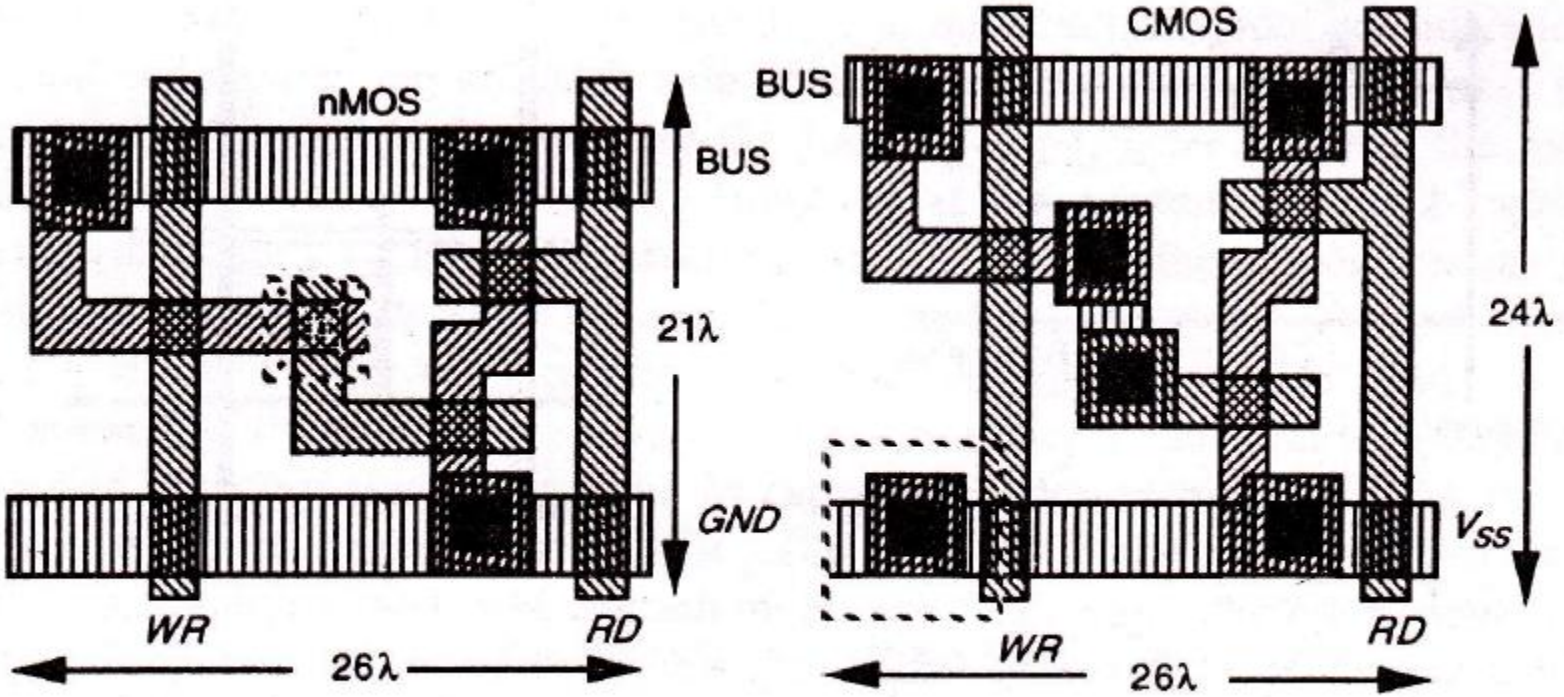


(a) Circuit

(b) CMOS stick diagram

Note: *WR* and *RD* are coincident with ϕ_1.

FIGURE 9.1 Three-transistor dynamic memory cell.



URE 9.2 Mask layouts* for three-transistor (nMOS and CMOS) memory cell *(pull-ups not shown).

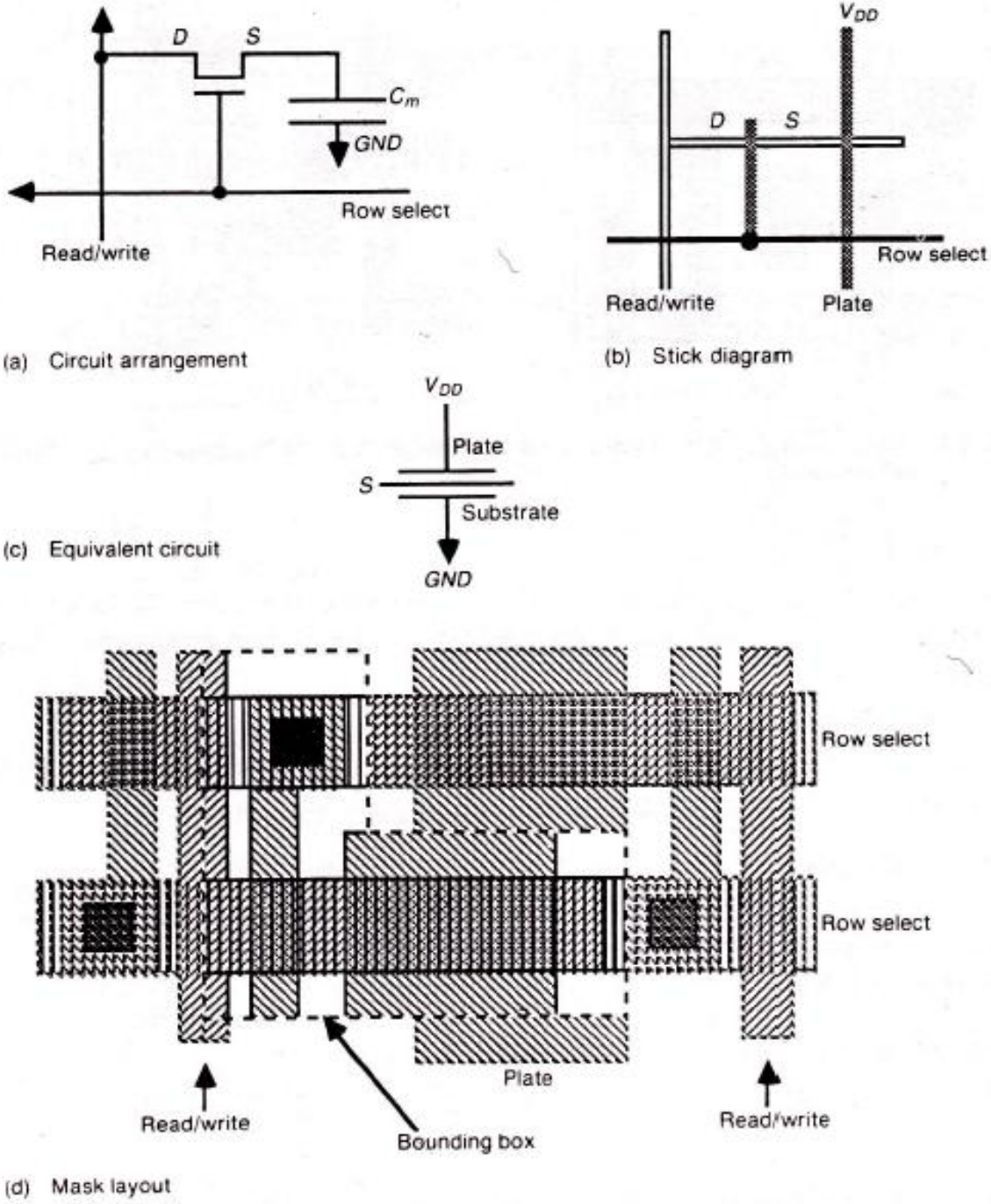
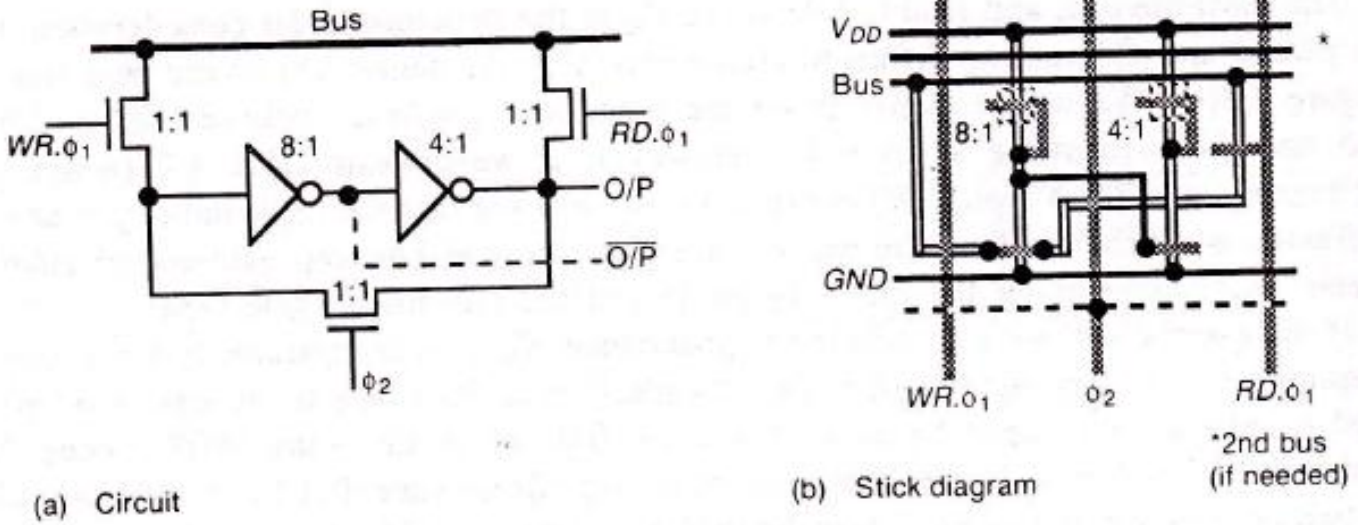


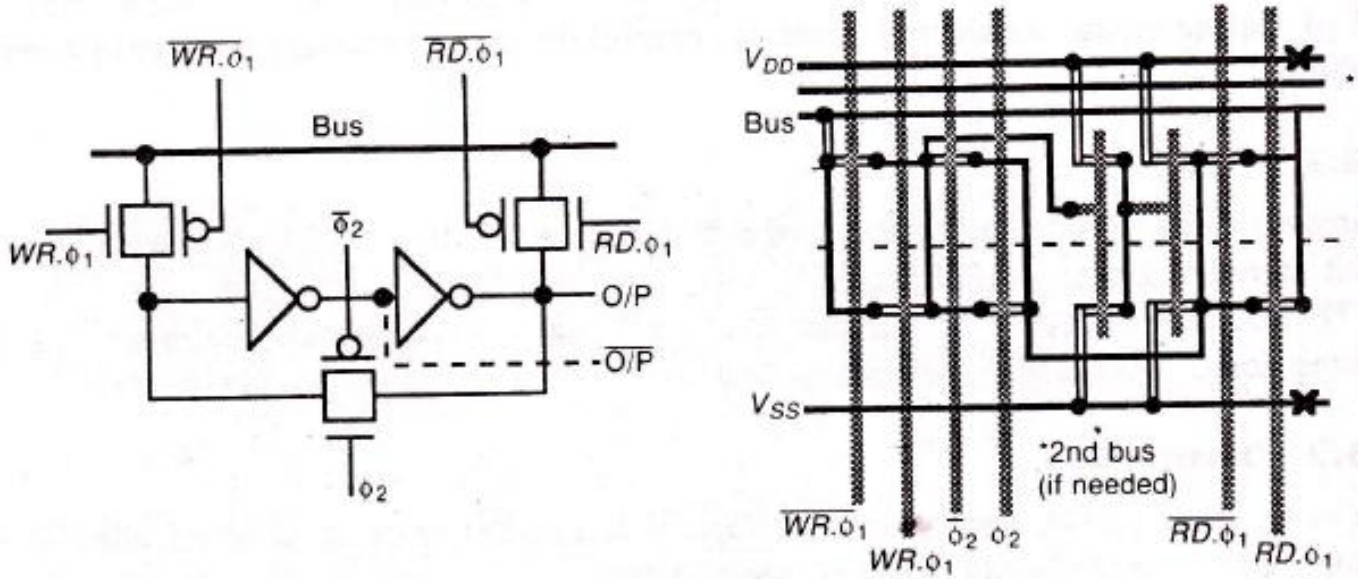
FIGURE 9.3 One-transistor memory cell.



(a) Circuit

(b) Stick diagram

FIGURE 9.4 nMOS pseudo-static memory cell.



(a) CMOS circuit

(b) Stick diagram

FIGURE 9.5 CMOS pseudo-static memory cell.

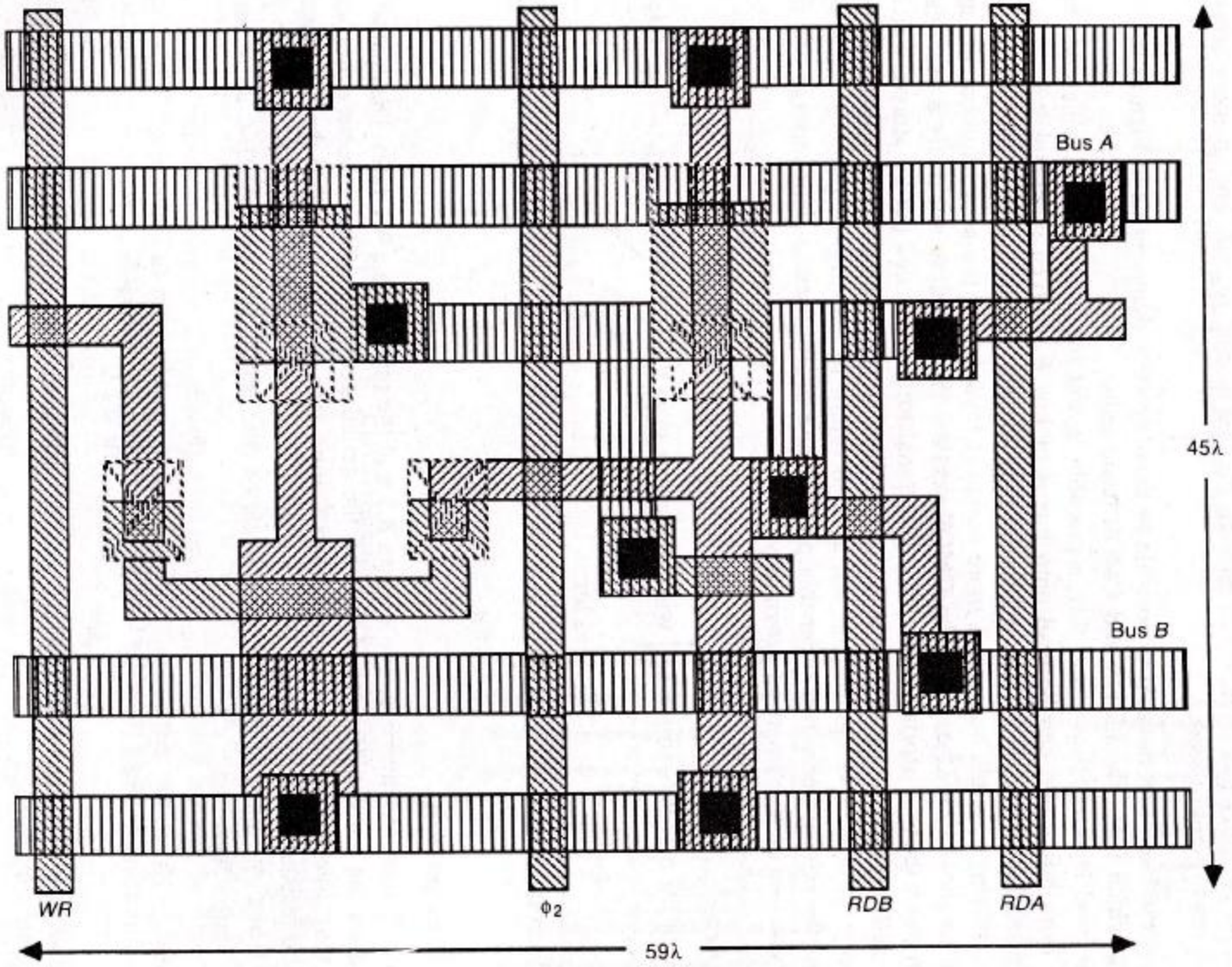


FIGURE 9.6 nMOS pseudo-static memory cell with read to either of two buses.

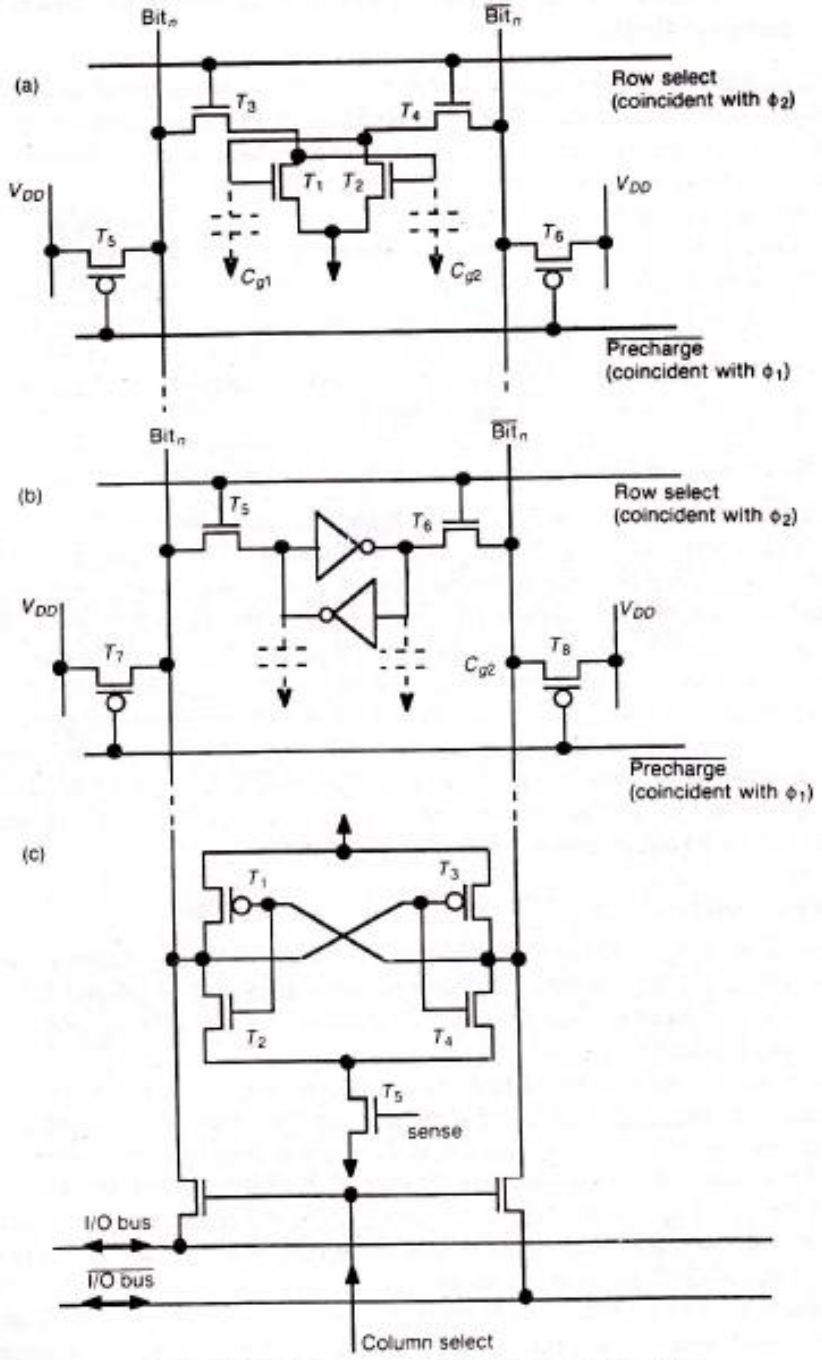


FIGURE 9.7 Dynamic and static memory cells.

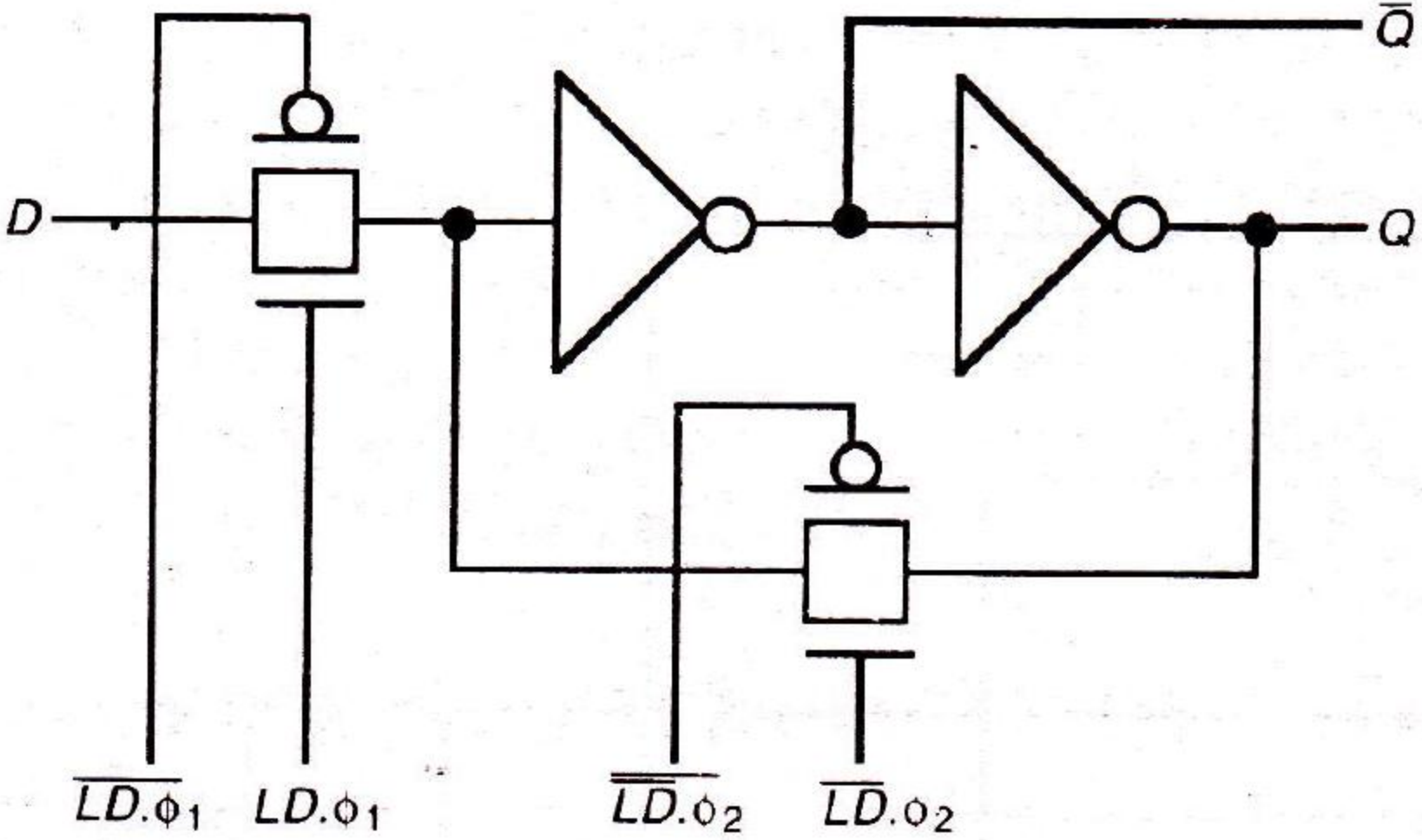


FIGURE 9.12 A CMOS pseudo-static D flip-flop.