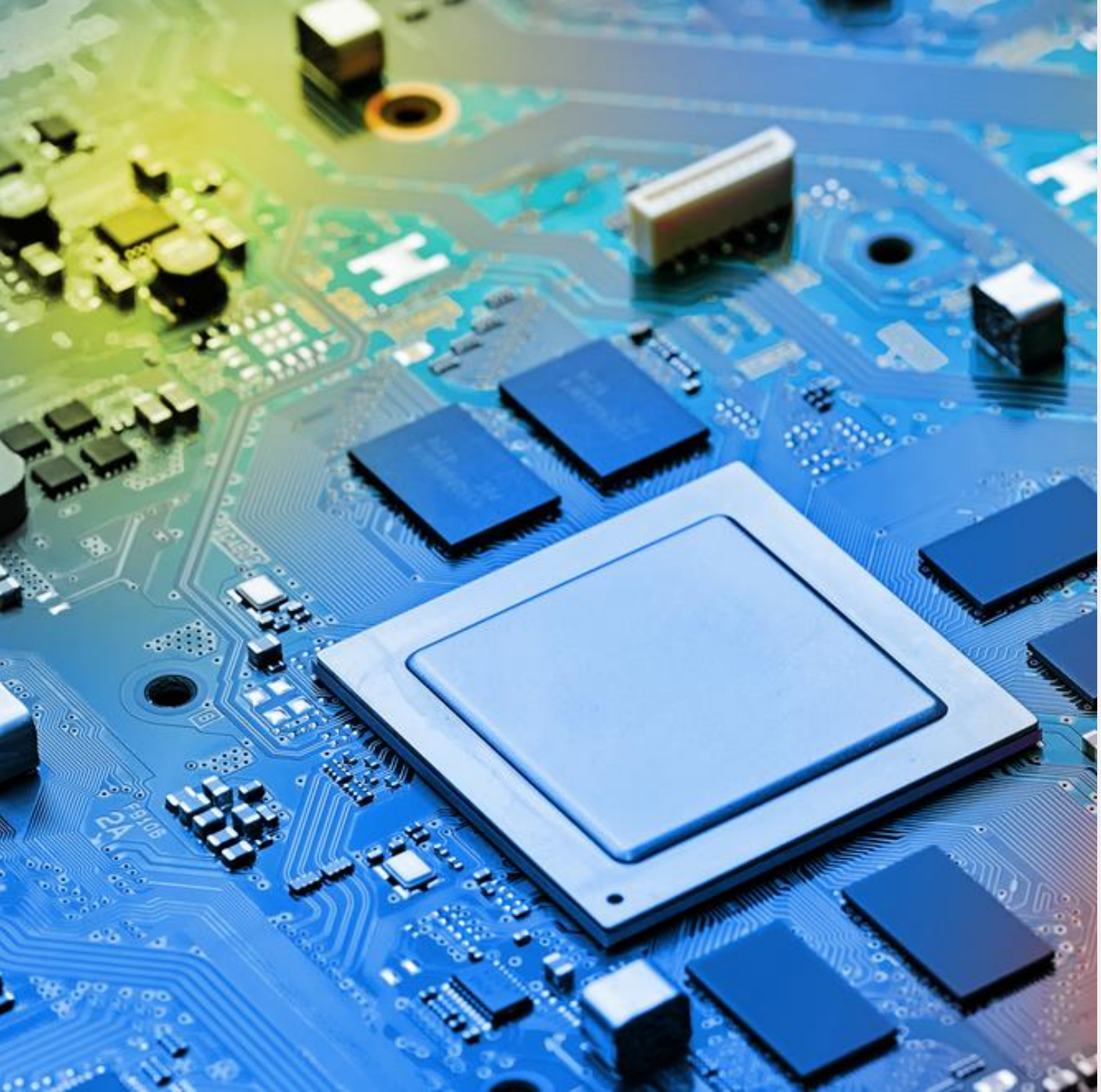


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

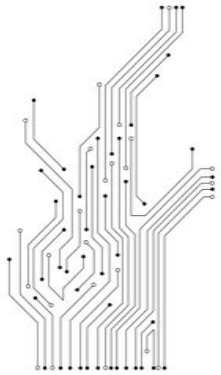
EE3019 – Integrated Electronics

Semiconductor Memory

Learning Objectives

By the end of this lesson, you should be able to:

- Discuss the different types of semiconductor memories and memory array organisation.
- Discuss the read, write and refresh operations of Dynamic Random Access Memory (DRAM).
- Explain the read and write operations of Static Random Access Memory (SRAM).
- Explain the operation of Read-only Memory (ROM) and the different types of implementation.
- Discuss the operation of Row and Column decoders.
- Calculate the row access time of a memory array.



Overview of Semiconductor Memory Types

← Portability

- Area (cost) ↓
- Speed ↑
- Power dissipation ↓

Semiconductor memory

Read/Write (R/W) Memory or
Random Access Memory (RAM)

Dynamic RAM (DRAM)

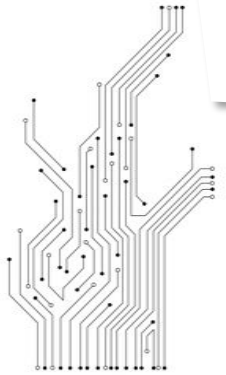
Static RAM (SRAM)

Volatile: Changeable

Read-only Memory (ROM)

(Non-volatile)

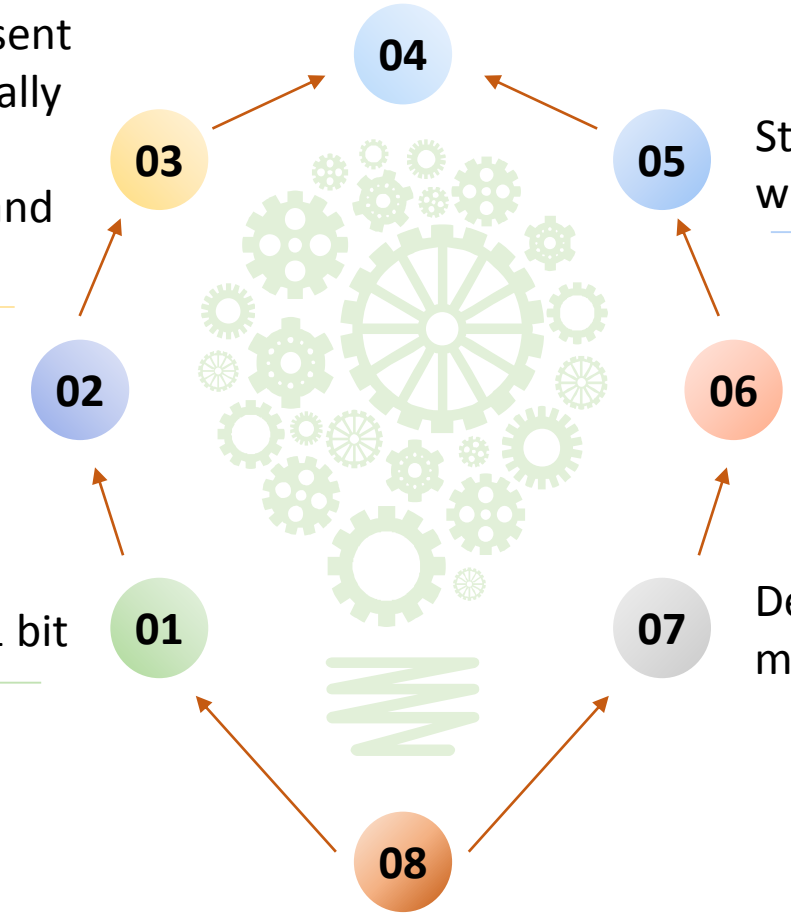
- Mask (Fuse) ROM
- Programmable ROM (PROM)
  - Through UV light ← Erasable PROM (EPROM)
  - Electrically Erasable PROM (E^2PROM)
- Flash Memory
- Ferroelectric RAM (FRAM)



Basic Definitions

Memory Location: A set of cells for storing 1 word.

1 word = Group of Bits to represent 1 entity, the number of bits usually equal to width of data bus.
Typical data bus width = 8, 16, and 32 bits.



Static memory: Data are permanent without refreshing.

1 byte = 8 bits

Dynamic memory: Data needs to be refreshed to be retained.

Cell : Each cell stores 1 bit

Destructive: Once data is read, data in memory is lost.

Volatile: Once power is off, data in memory is lost.

Comments

Each memory cell must be uniquely identified, or addressed.

Memory size is based on power of 2.

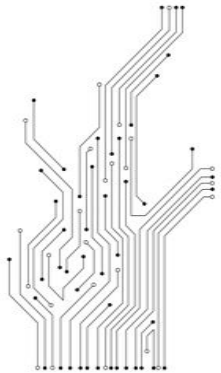
- For example, 1 K = 2^{10}, 4 K = 4 x 2^{10}, 16 K = 16 x 2^{10}, 64 K = 64 x 2^{10}

1m = 2^{20}, 1G = 2^{30}

1T = 2^{40}

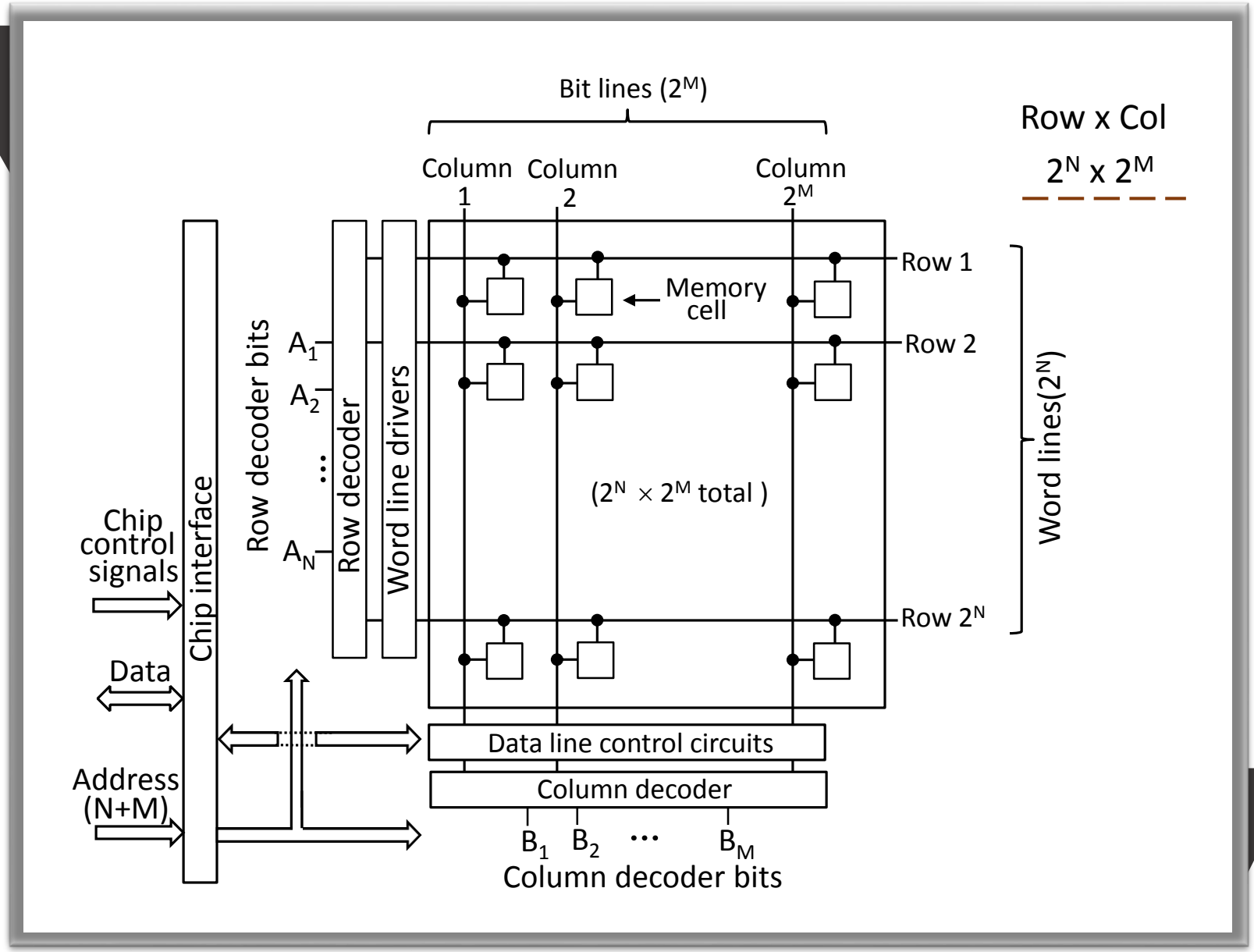
1P = 2^{50}

Memory ICs usually increase in terms of 2^2.



Most current-art fabrication process is used to make memory, especially DRAMs. This is because the basic DRAM memory cell is a minimum dimensioned transistor.

Conceptual Memory Array Organisation



Analysis

The memory is arranged as arrays of single-bit storage cells.

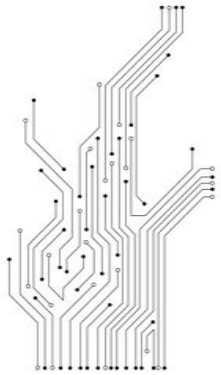
The arrays are encircled by address decoding logic and interface circuitry to external signals.

The memory array is usually square (m=n) to minimise external decoding circuitry. Consider a memory containing 16K 1-bit storage; 16K = 2^{14}.

$2^4 \times 2^{10}$ | $A_{13}, A_{12}, A_{11} \ldots A_0$

If organised as a single row – require 14 – to – 16,384 decoder. This decoder is very complex and probably more complex than memory array.

If organised as a 128 × 128 square – require 2 × 7 – to – 128 decoder, one decoder for row and the other for column. This decoder is substantially simpler and is therefore the practice.

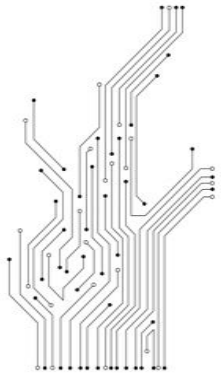


Analysis

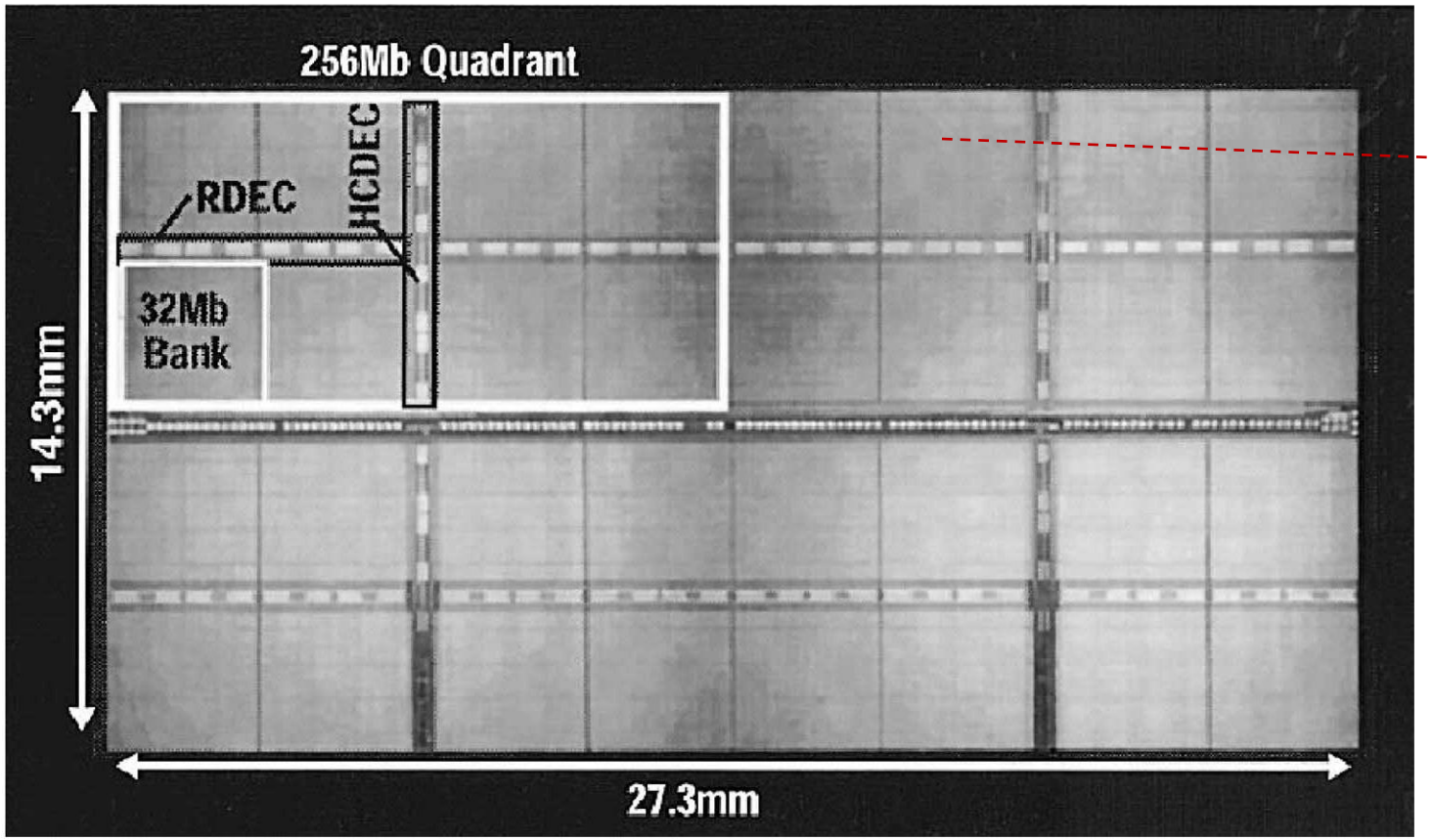
Addressing memory:

- Row Address enables all cells along the selected row, contents of these cells become available along column lines. In circuit terms, the selected row line goes to high (logic “1”).
- Column Address is used to select the particular column containing the desired data bit. This data bit is ultimately routed to drive the output pin. In circuit terms, the selected column becomes an electrical path for the data to flow.

In most cases, *n* = 8 bits are output simultaneously – data from *n* columns are selected and gated to *n* output pins.



Typical Configuration of DRAM Chip (1 GB SDRAM Chip)



2^2

256 mb x 4 = 1 GB

(*Source:* Figure 2, K. Toshiaki et al., "A 390-mm^2, 16-Bank, 1-Gb DDR SDRAM with Hybrid Bitline Architecture," *IEEE J. Solid-State Circuits,* vol. 34, no. 11, pp. 1580–1588, November 1999.)

Dynamic Random Access Memory (DRAM)

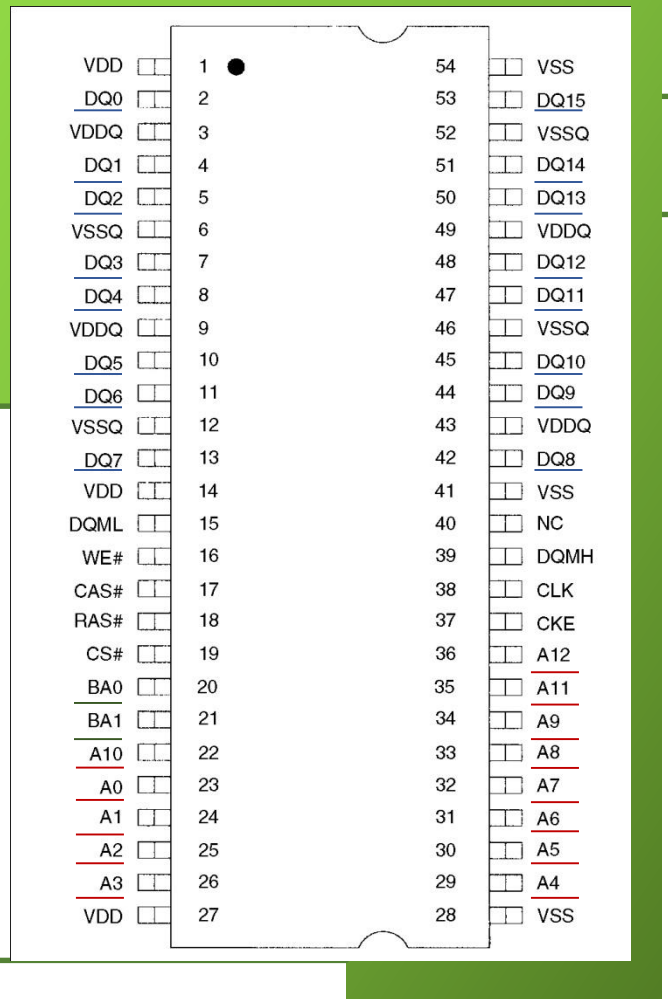
2^{13} = 8K ($A_{12} - A_0$)

8k x 8k = 64M

Row Column

64M x 4 = 256M x 16

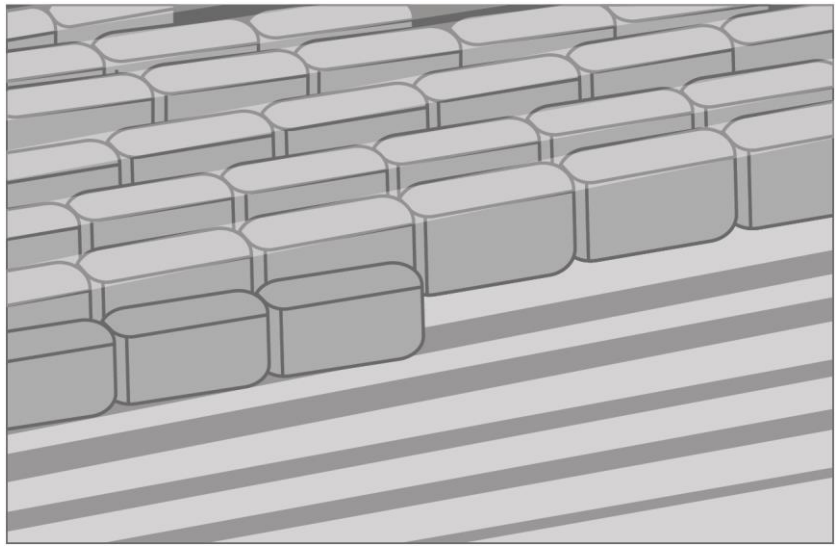
4 Banks



Capacity x Word size

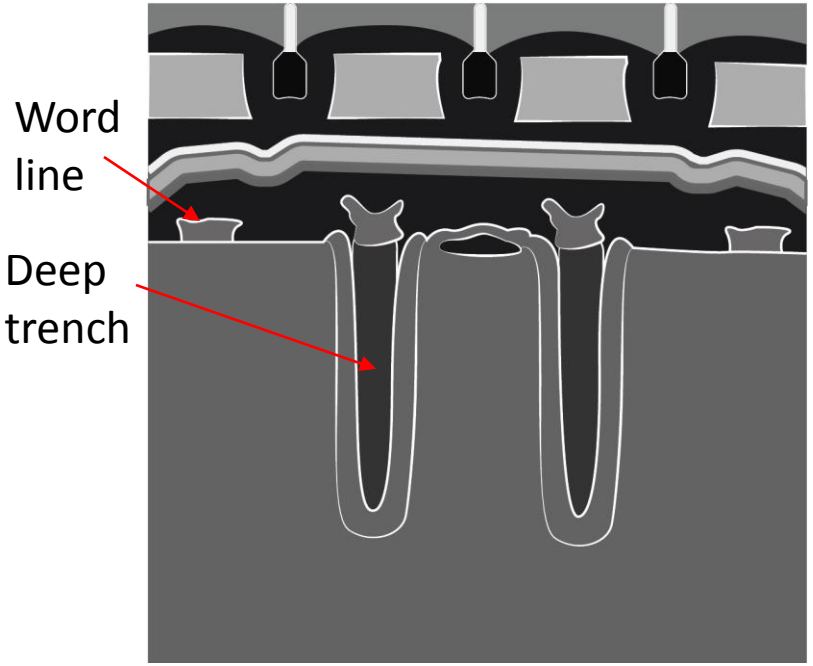
A 256-Mb synchronous DRAM

Various Structures of the One-transistor DRAM Cell



(a)

(a) DRAM cell with a stacked capacitor.



(b) DRAM cell with a trench capacitor.

(*Source (b):* Figure 1, S. Crowder et al., "Integration of trench DRAM into a high-performance 0.18 μm logic technology with copper BEOL, IEDM (International Electron Devices Meeting) '98, pp. 1017–1020, 1998.)

Concept of DRAMs

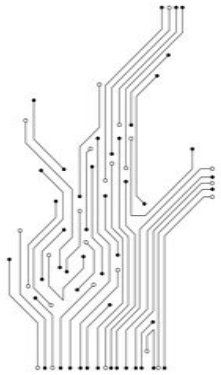
DRAM stores its bit of information as charge on the cell capacitor.

The gate of the transistor is connected to the word line, and its source (drain) is connected to the bit line.

When storing a 1, the capacitor is charged to $V_{DD} - V_{to.}$

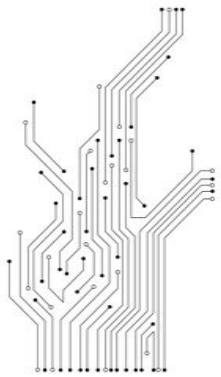
When storing a 0, the capacitor is discharged to 0 V.

nMOS

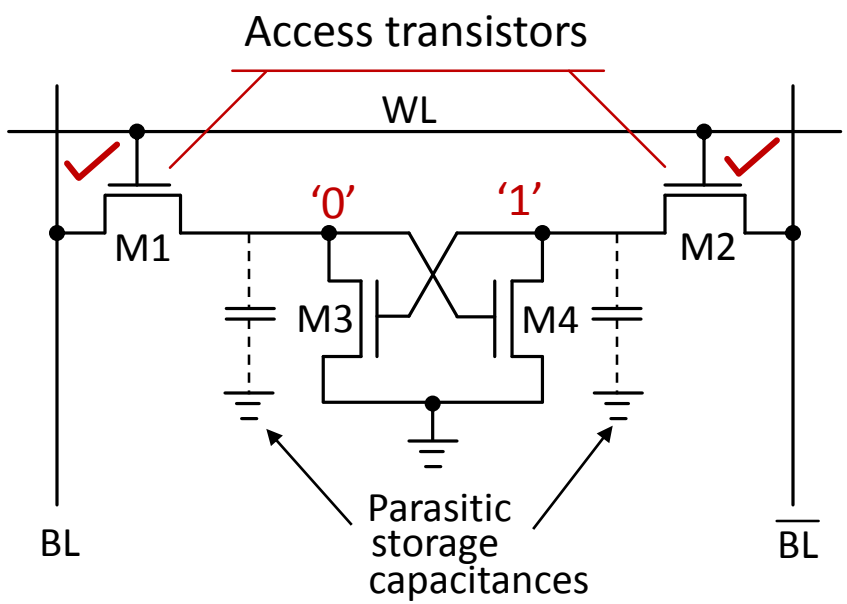


Refresh of DRAM

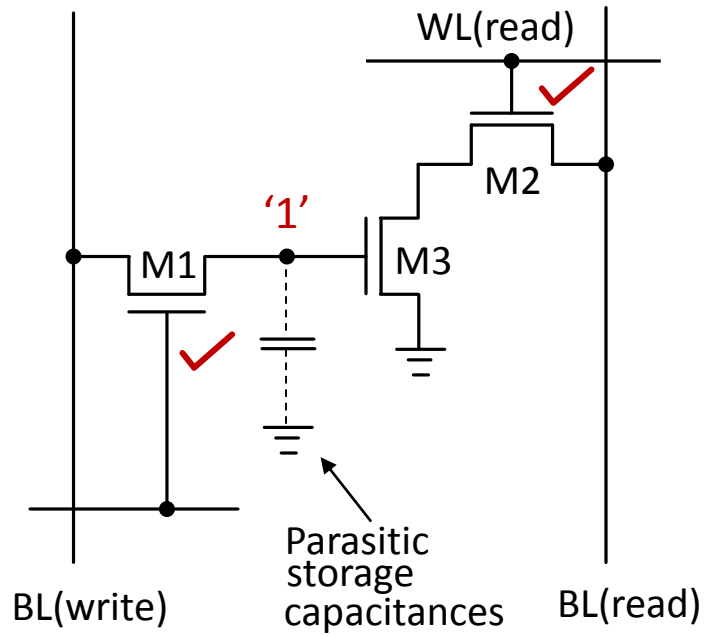
1. Charge leakage, cell must be refreshed periodically.
2. During *refresh*, cell content is ***read*** and data is ***rewritten***, ***restoring*** the capacitor voltage to its proper value.
3. ***Read*** or ***Write*** operation will result in refreshing all cells in selected row.
4. ***Refresh cycle***: time taken to refresh 1 rows of DRAM cells, ~100 ns.
5. ***Refresh period***: time interval between refresh cycles, ~1 ms.
6. Provision must be made for periodic refreshing of the entire memory every 5 ms to 10 ms.
7. During ***Refresh***, the chip will not be available for ***Read*** or ***Write*** operations.
8. ***Refresh*** typically takes up <2% and memory chip is available for normal operation >98% of the time.



Various Configuration of DRAM Cell



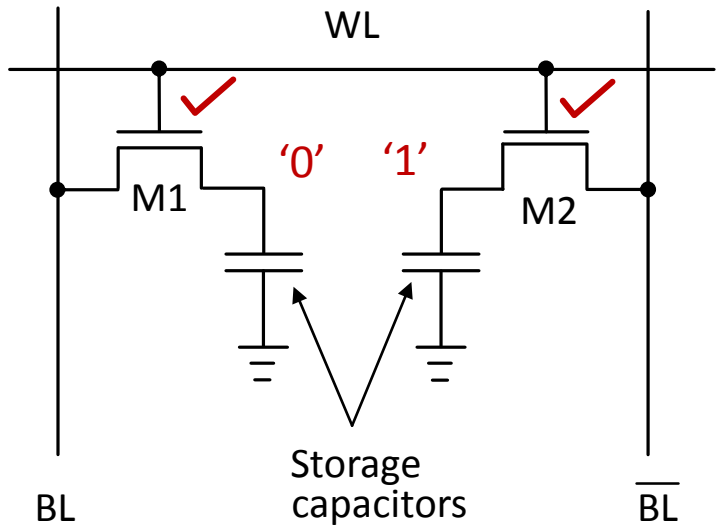
(a) 4-T DRAM cell with 2 storage nodes



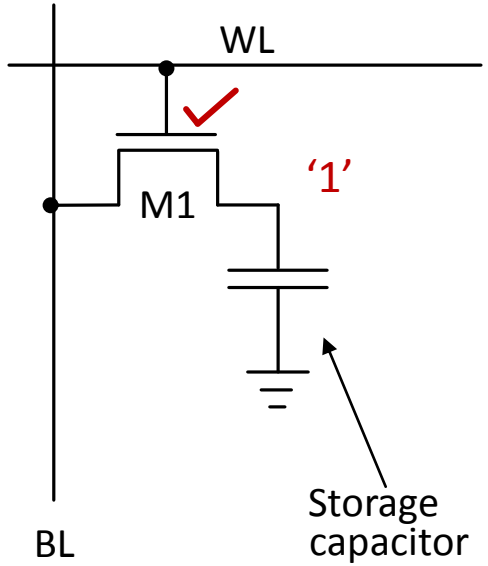
(b) 3-T DRAM cell with 2 bit lines and 2 lines

Contd.

Various Configuration of DRAM Cell



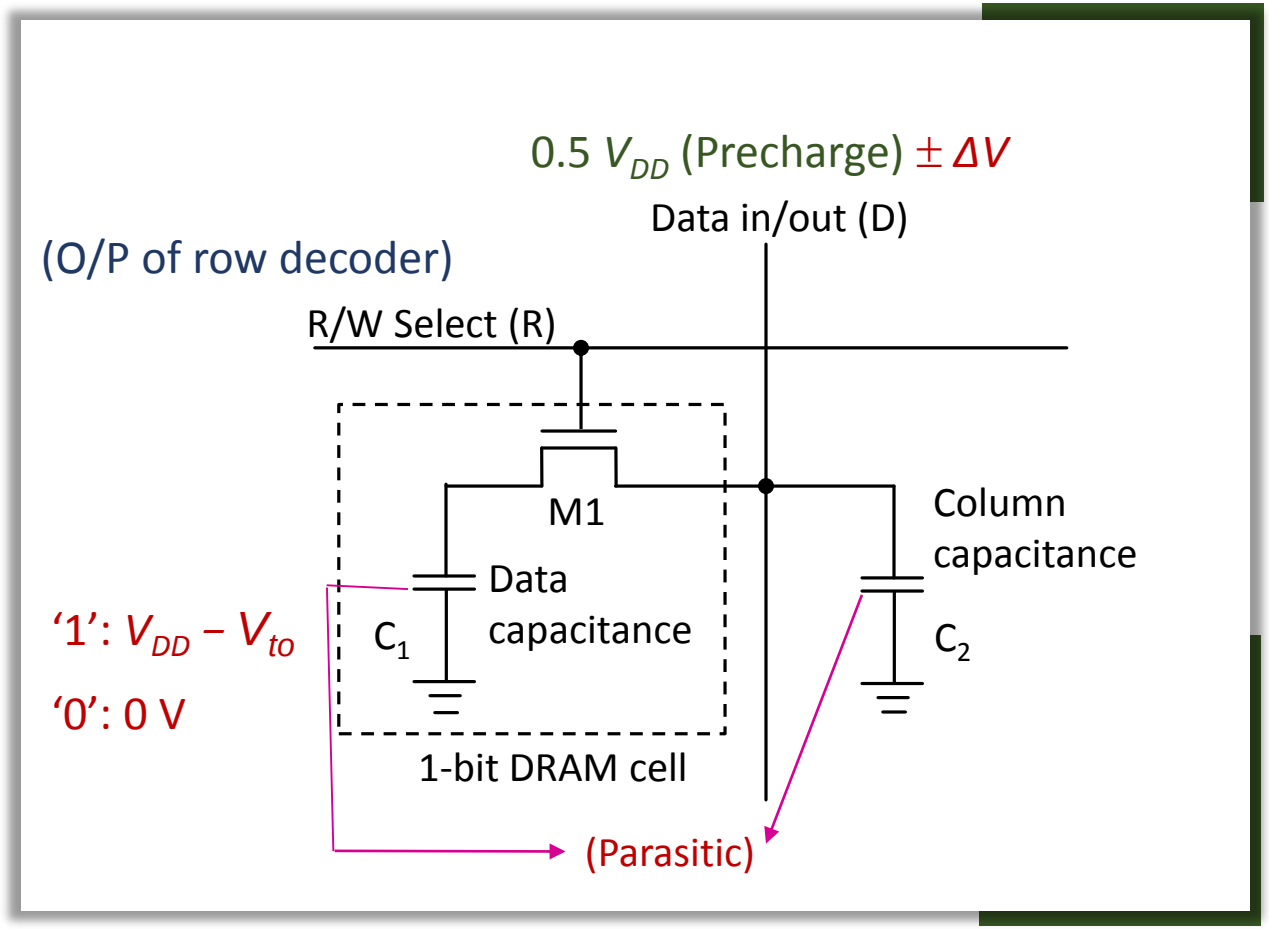
(c) 2-T DRAM cell with 2 bit lines and 1 word line



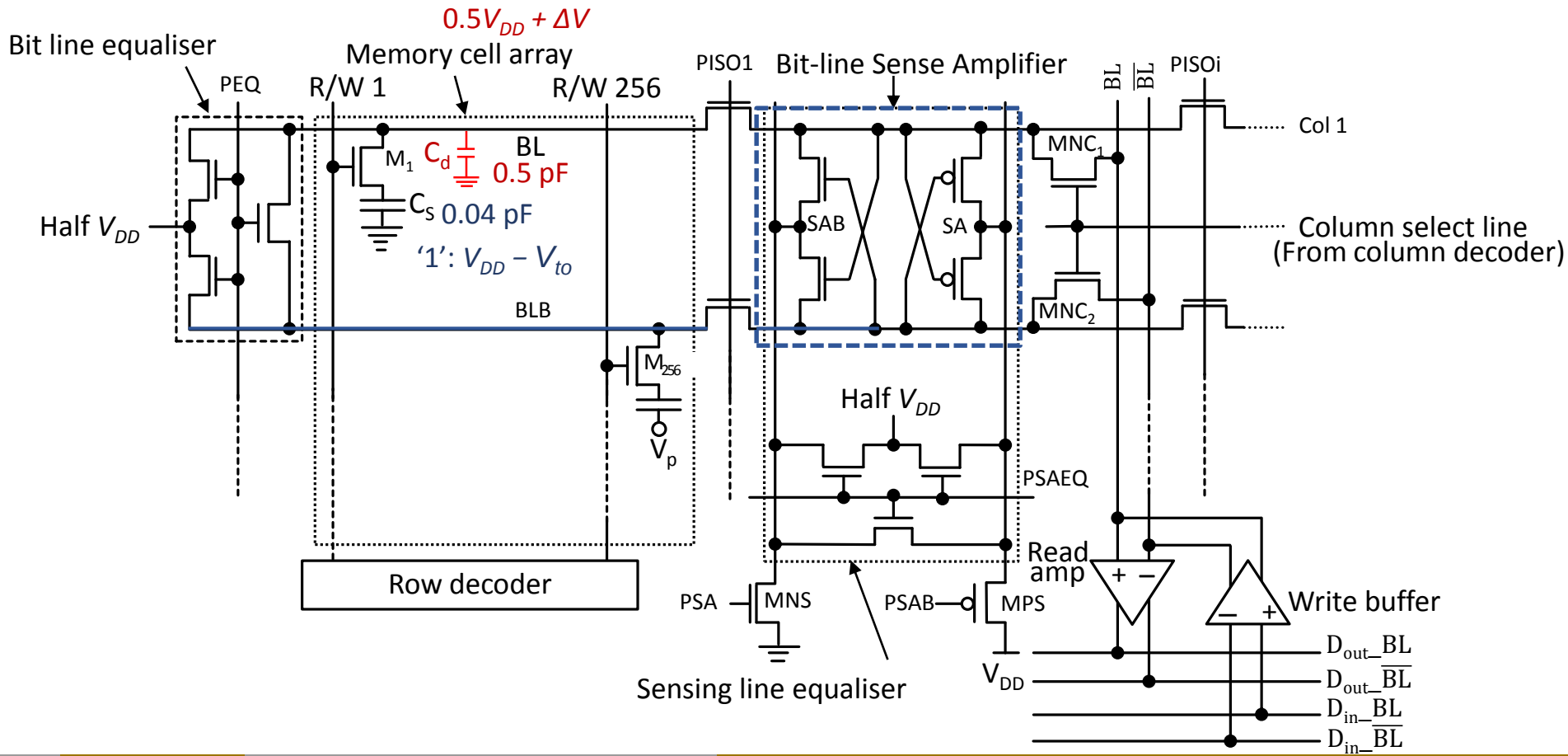
(d) 1-T DRAM cell with 1 bit line and 1 word line

1-T DRAM Cell

Typical one-transistor (1-T) DRAM cell with its access lines.



1-T DRAM Cell – Sense Amplifier



Memory structure of one-transistor DRAM cell array with control circuits, where half-V_{DD} sensing, folded bit line, and shared sense amplifier architectures are incorporated.

1-T DRAM Cell

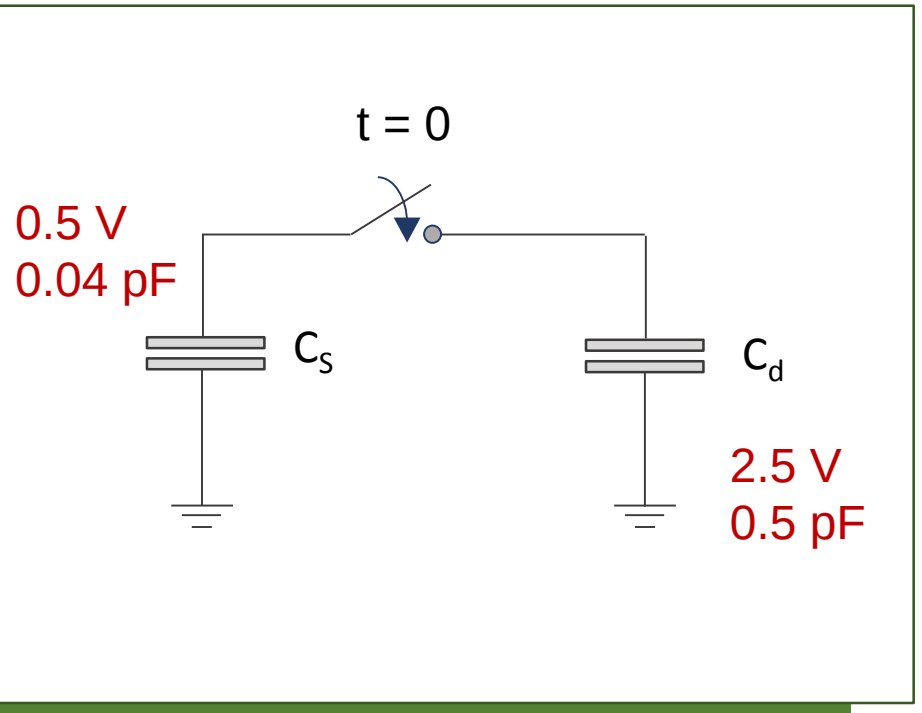
Assume that: C_d is the data line capacitance = 0.5 pF,
C_s is the memory cell capacitance = 0.04 pF, and

$$V_{C_d}|_{(t=0)} = 0.5V_{DD} = 2.5 \text{ V}$$
$$V_{C_S}|_{(t=0)} = V_{DD} = 5 \text{ V}$$

$V_{DD} + \Delta V$
2.5 V + 0.185 V

$$V_{C_d} = V_{C_S} = \frac{2.5 \times 0.5 + 5 \times 0.04}{0.54} \text{V} = 2.685 \text{ V}$$

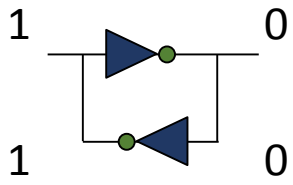
Determine the voltage across C_d and C_s after the switch is closed. This is the final voltage of the memory cell that is output. The content of the memory cell is logic '1'.



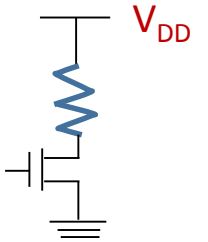
Example

Static RAM

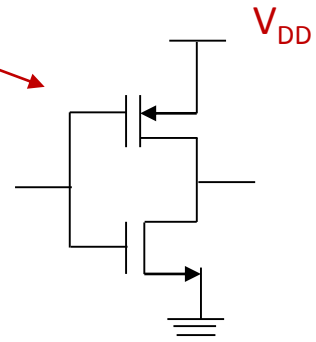
- SRAM is a volatile memory (i.e. it needs continuous power). Changeable
- It is the fastest memory, e.g. used as cache in PCs.
- Each cell contains two inverters. Each cell design is basically two cross coupled inverters. (Latch)



- Old SRAM cell design was based on NMOS (with depletion load device). Some design uses polysilicon resistor load devices (to reduce the cell size).

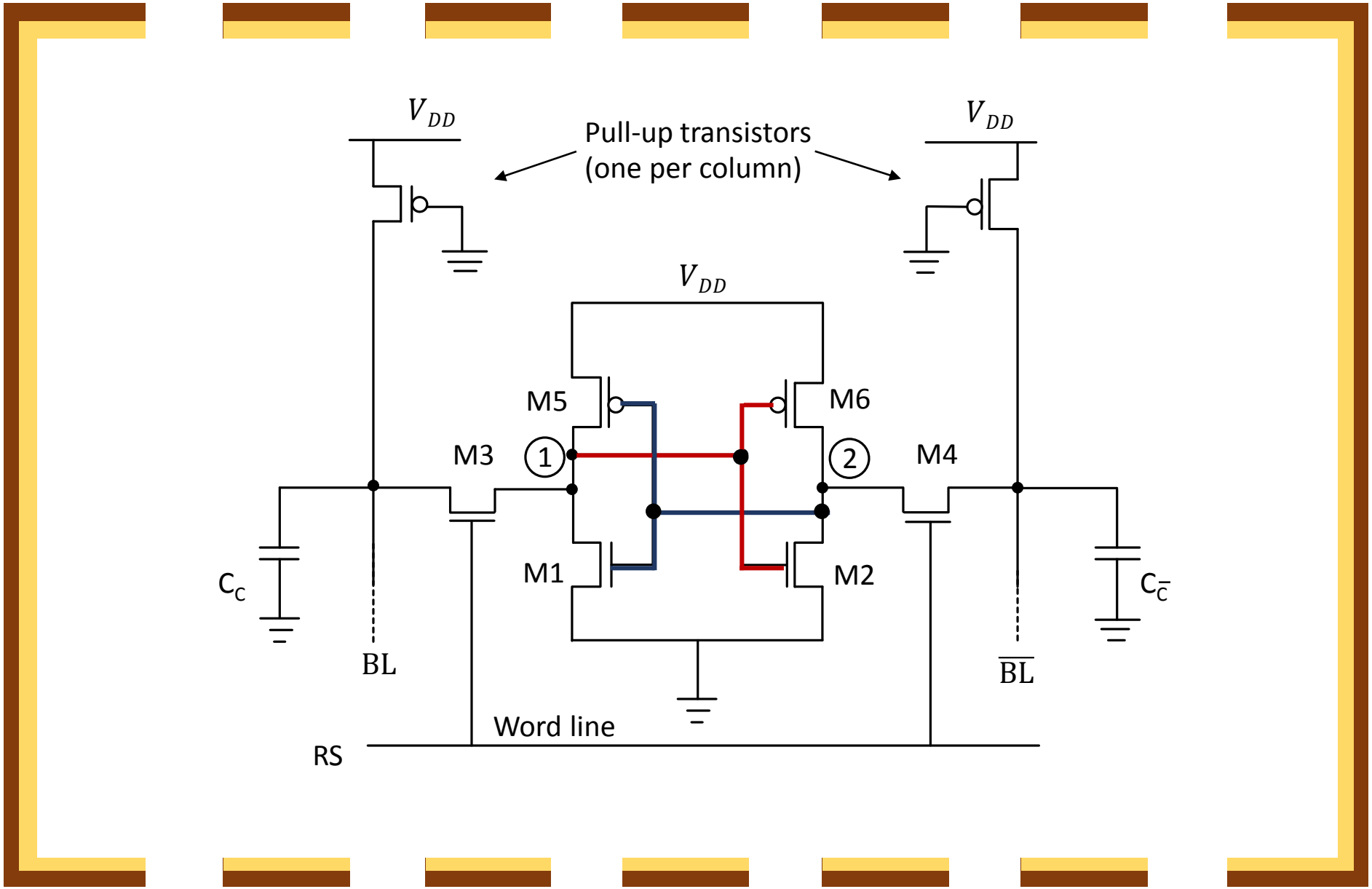


- Most SRAM are now CMOS with PMOS load transistor. Since CMOS inverters have low static power dissipation.

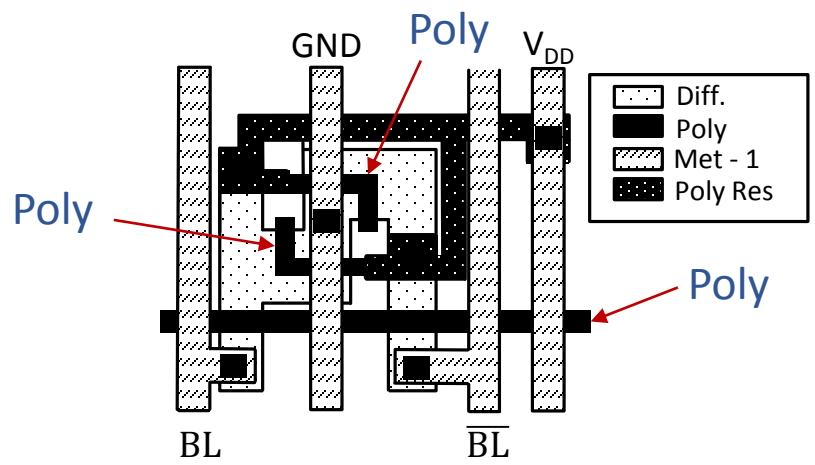


- Each CMOS SRAM cell comprises six transistors (6T – SRAM)

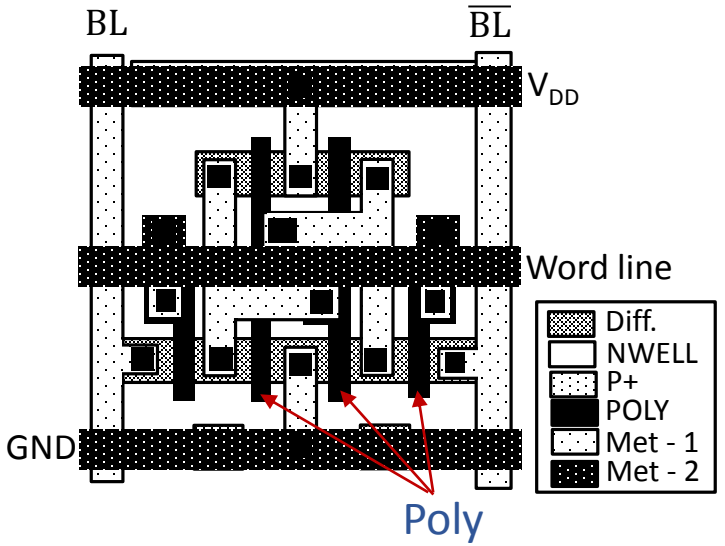
CMOS SRAM Cell



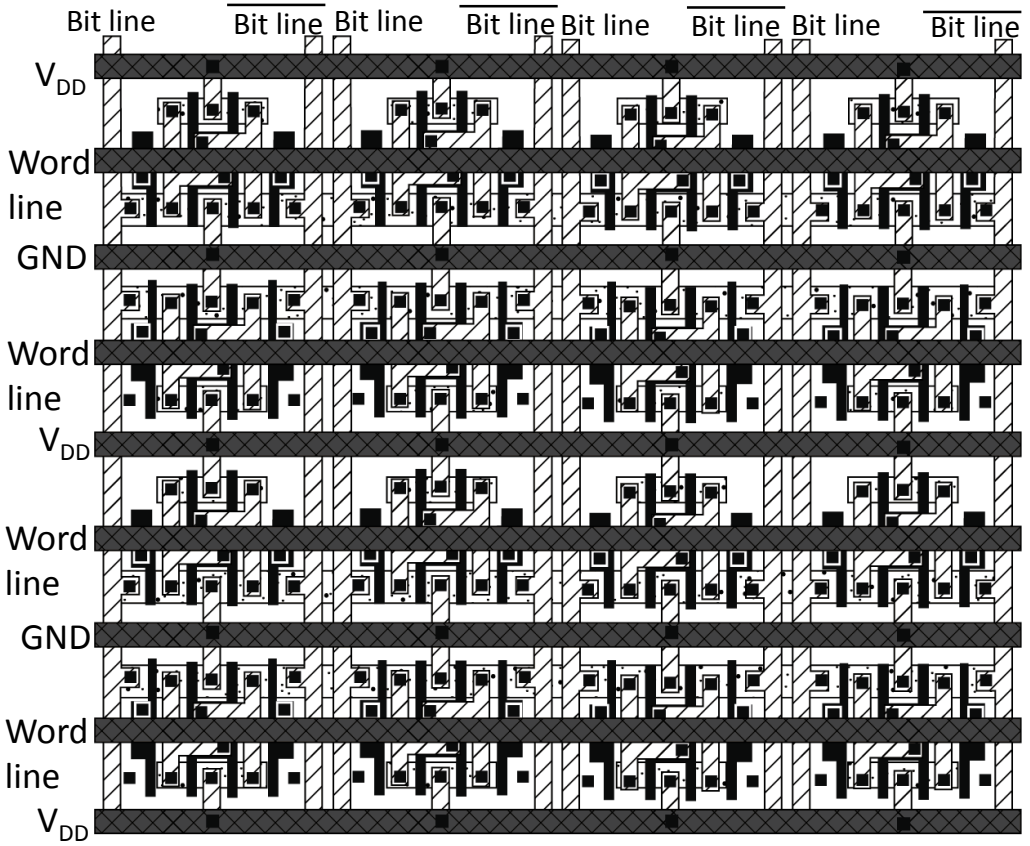
CMOS SRAM Cell Design Strategy



(a) Resistive load SRAM cell

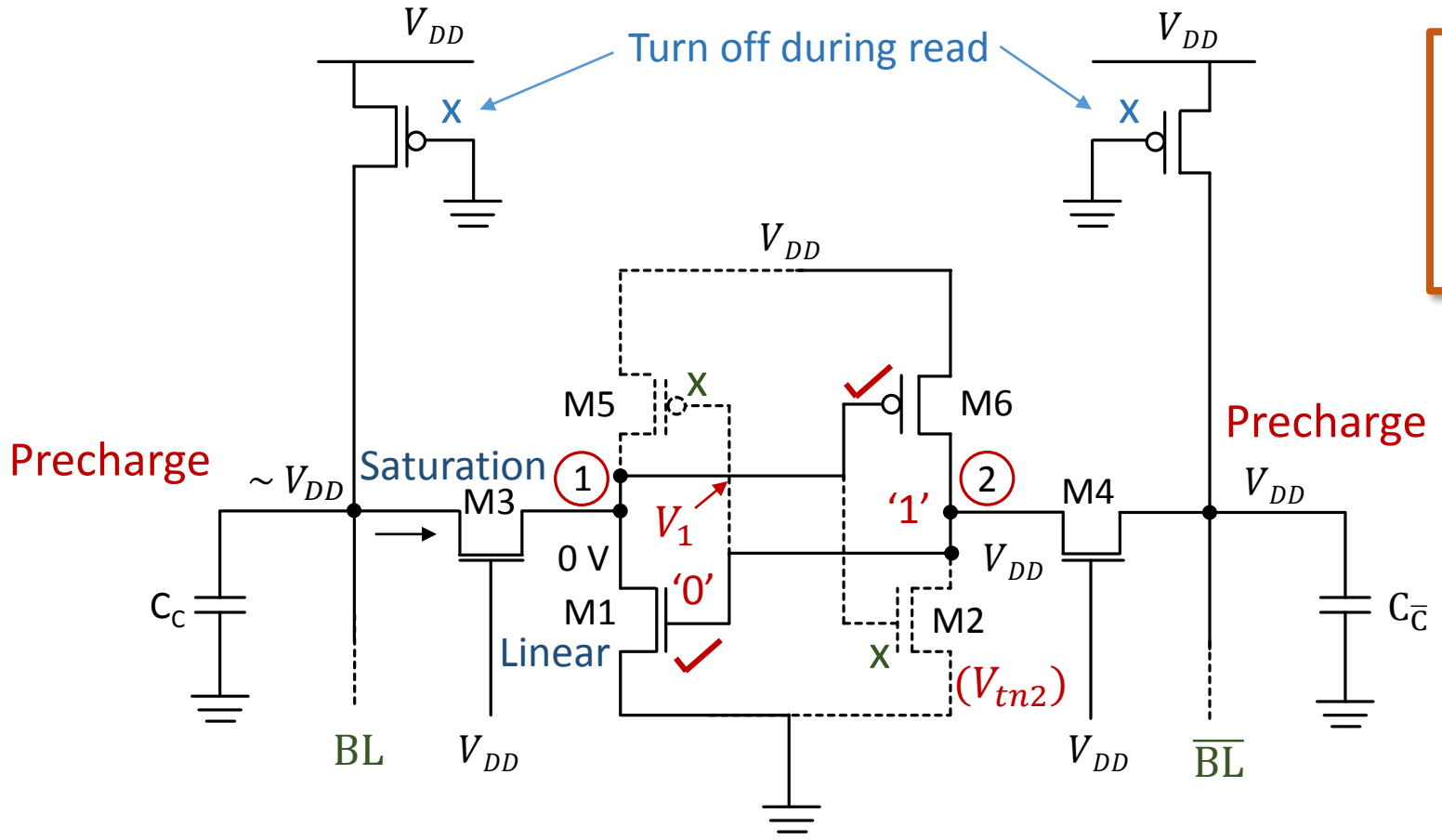


(b) Layout of the CMOS SRAM cell



(c) 4-bit x 4-bit SRAM array

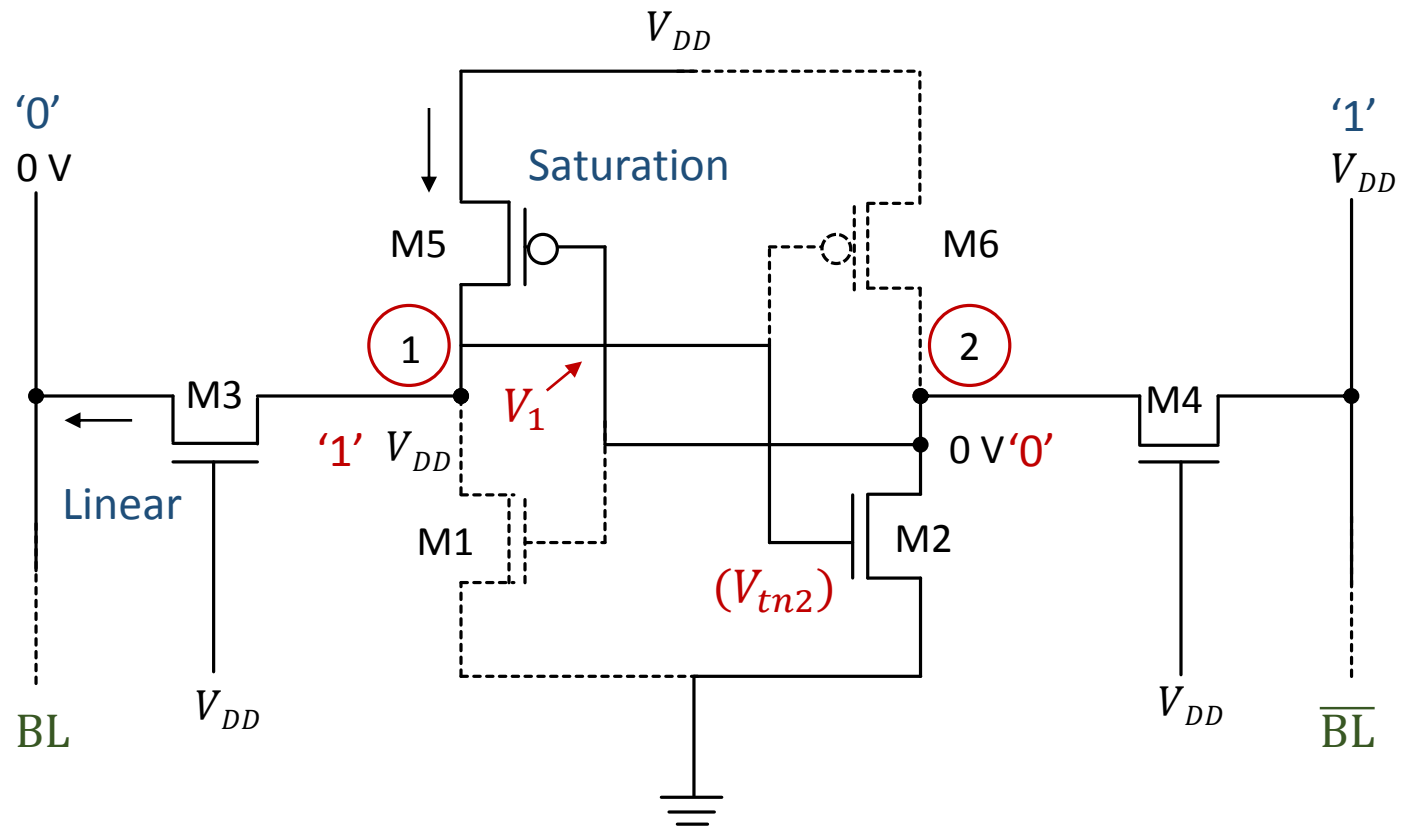
SRAM – READ Operation



$$V_1 < V_{tn2} \Rightarrow \frac{\left(\frac{W}{L}\right)_{M3}}{\left(\frac{W}{L}\right)_{M1}} < \frac{2(V_{DD} - 1.5V_{tn})V_{tn}}{(V_{DD} - 2V_{tn})^2}$$

Referring to points ① and ②, Voltage levels in the SRAM cell at the beginning of the "read" operation.

SRAM – WRITE Operation

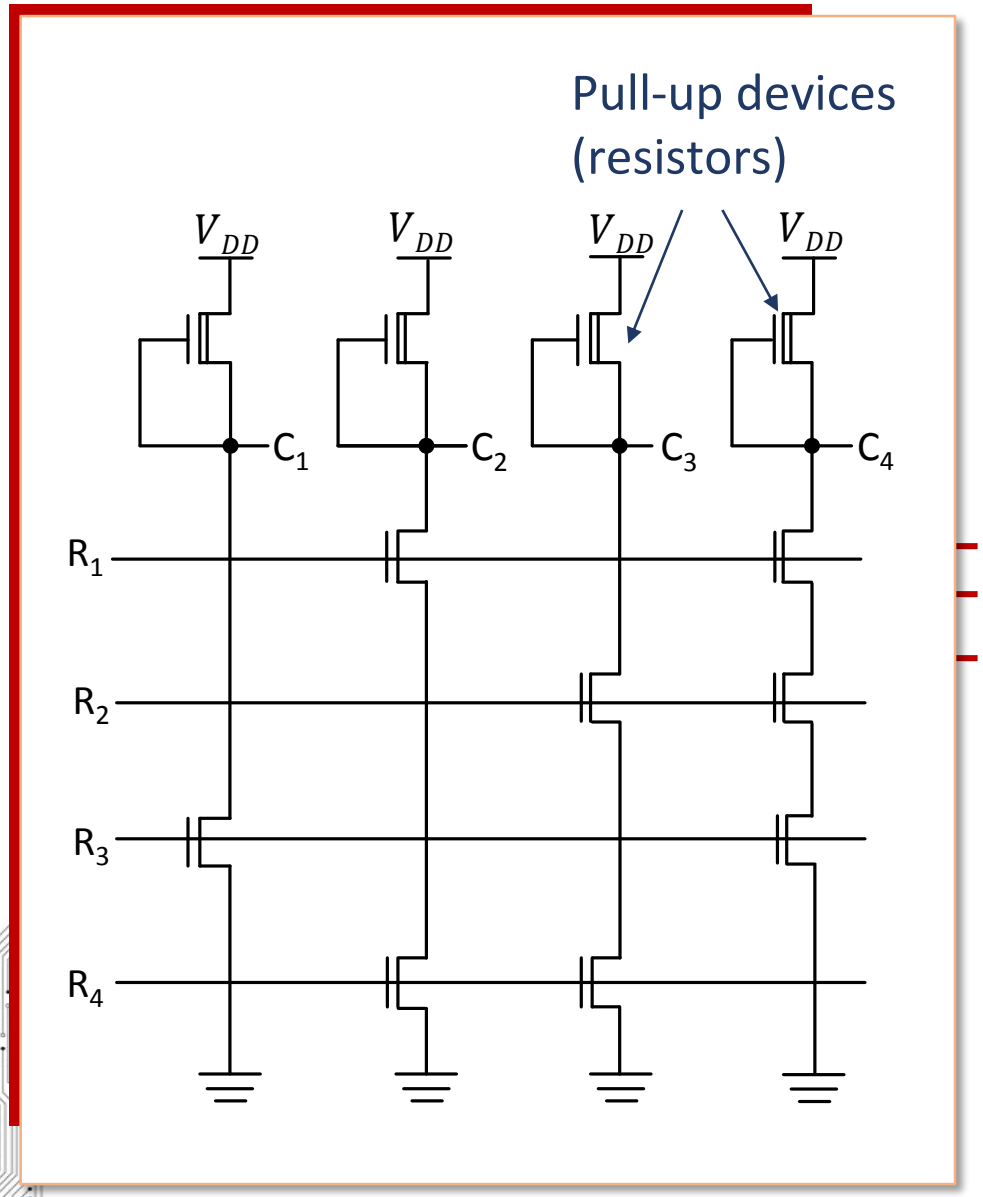


$$V_1 < V_{tn2} \Rightarrow \frac{\left(\frac{W}{L}\right)_{M5}}{\left(\frac{W}{L}\right)_{M3}} < \frac{\mu_n}{\mu_p} \cdot \frac{2(V_{DD} - 1.5V_{tn})V_{tn}}{(V_{DD} + 2V_{tp})^2}$$

Referring to points (1) and (2), Voltage levels in the SRAM cell at the beginning of the "write" operation.

Read Only Memory (ROM)

Addressing a Column



(Transistors in Series)

NAND Implementation

"Active Low"

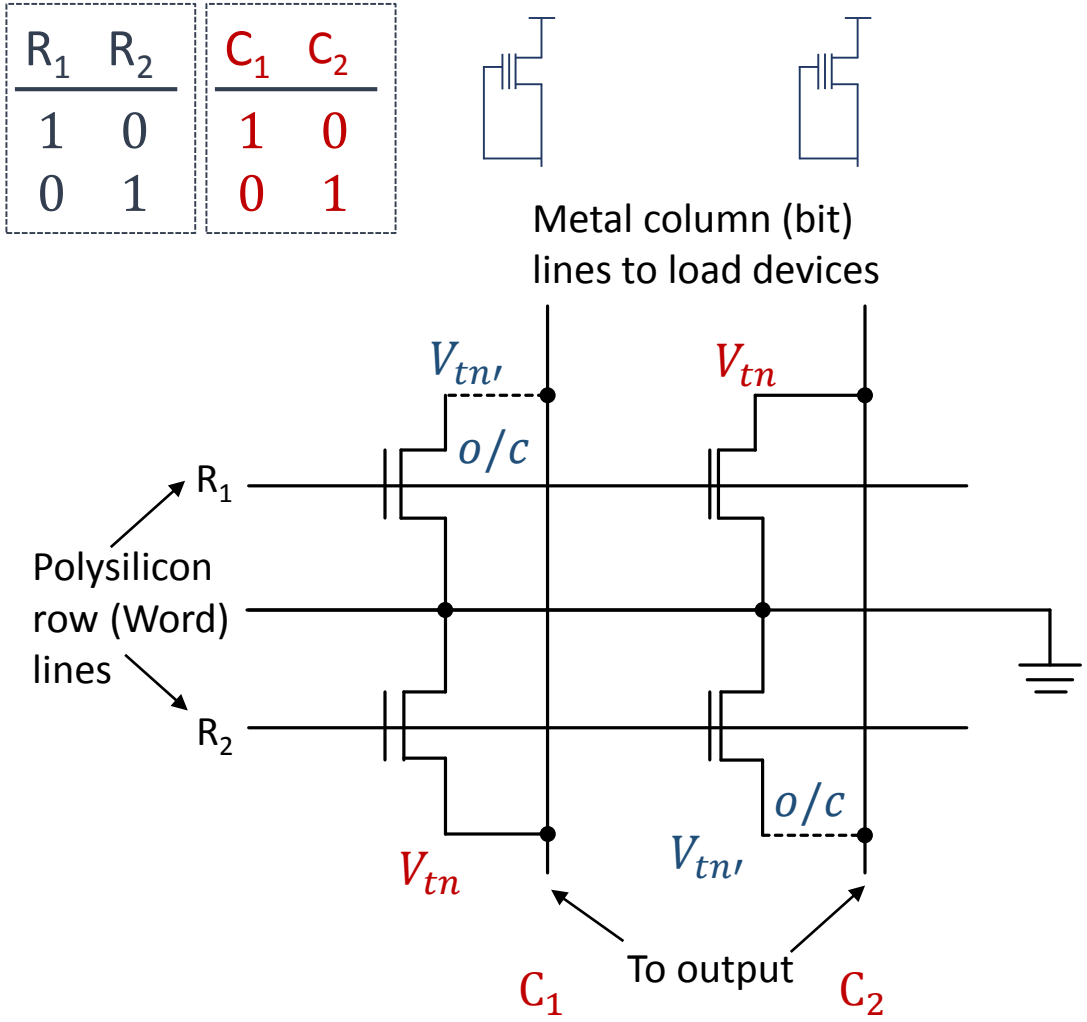
R_1	R_2	R_3	R_4
0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	0

C_1	C_2	C_3	C_4
0	1	0	1
0	0	1	1
1	1	0	1
1	1	1	0

NOR ROM Array (Active High)

R_1	R_2
1	0
0	1

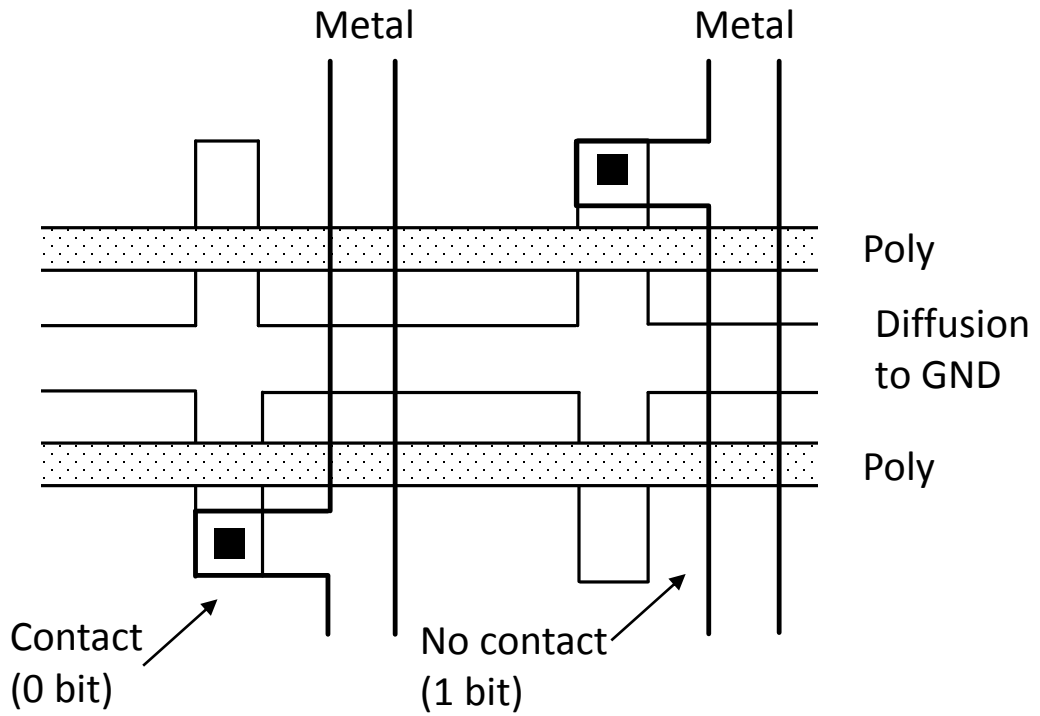
C_1	C_2
1	0
0	1



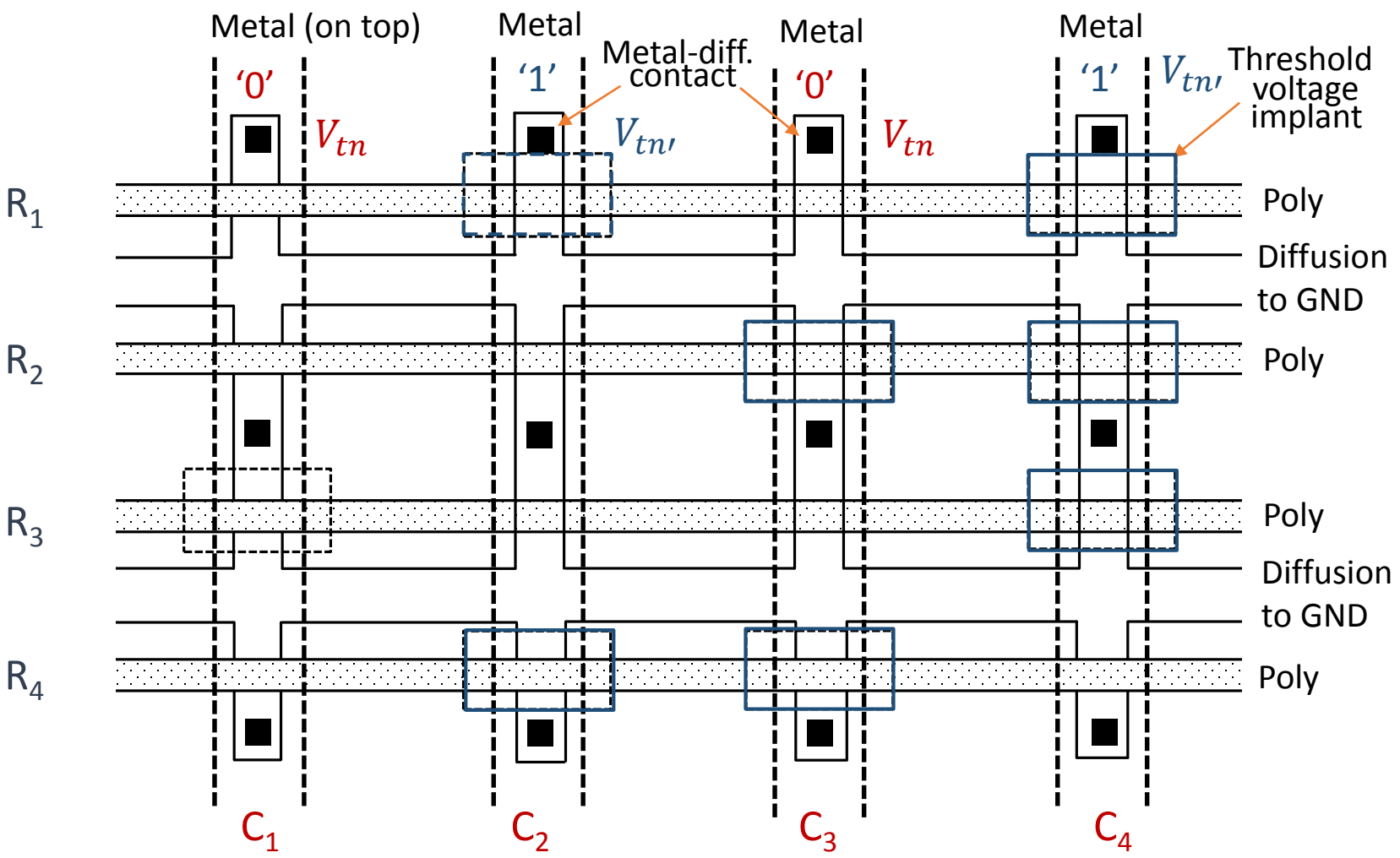
(Transistors in Parallel)

NOR Implementation

"Active High"



4-bit x 4-bit NOR ROM Array



Higher threshold voltage to turn on nMOS

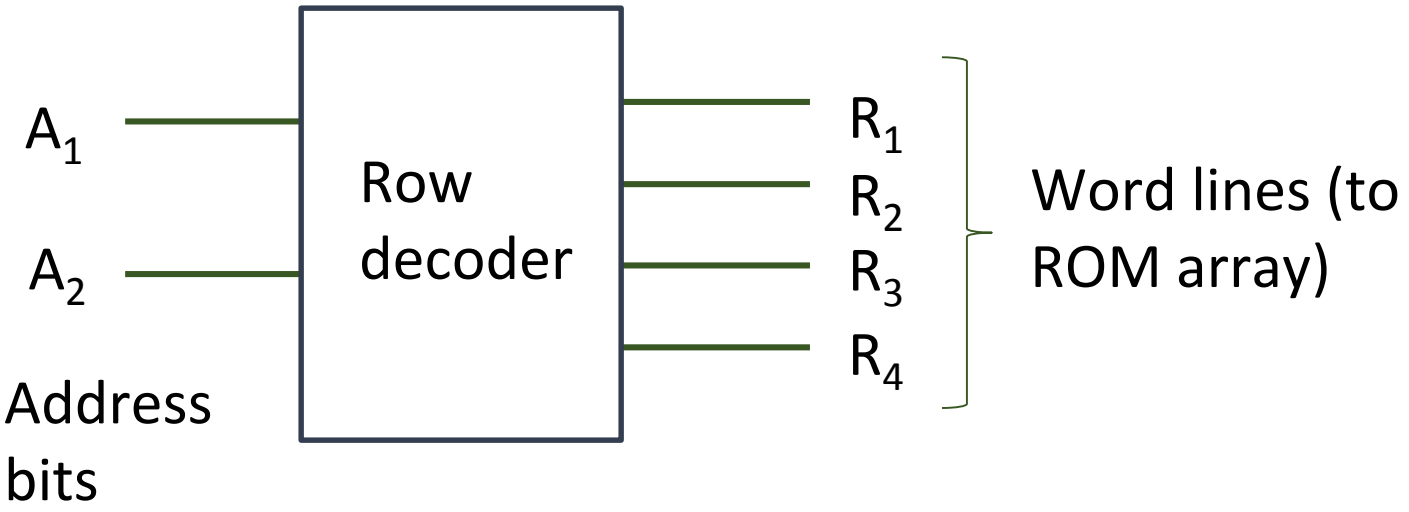
Row Decoder

(NOR based) logic '1' to select a row.

Addressing a row

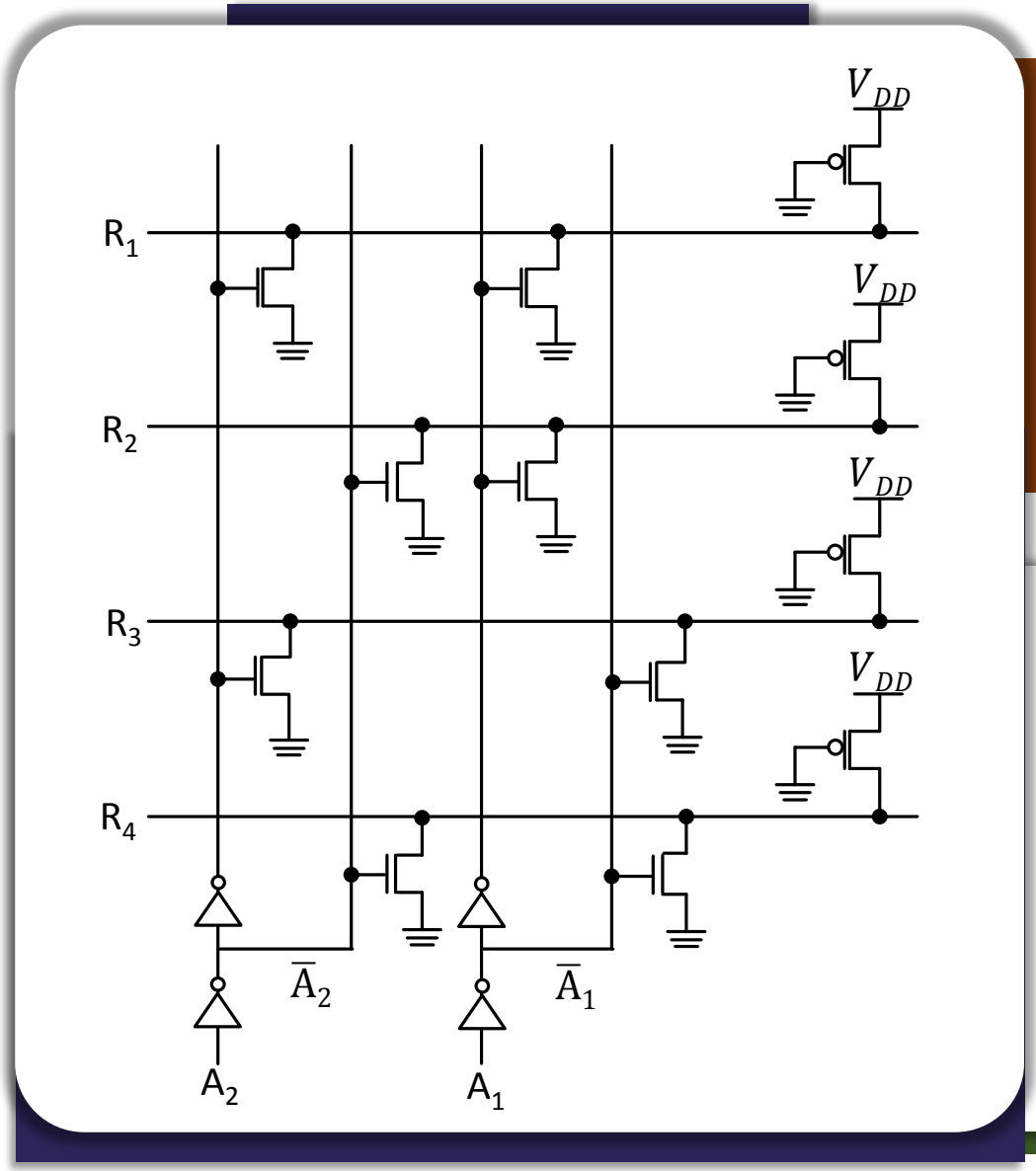
A_1	A_2
0	0
0	1
1	0
1	1

R_1	R_2	R_3	R_4
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1



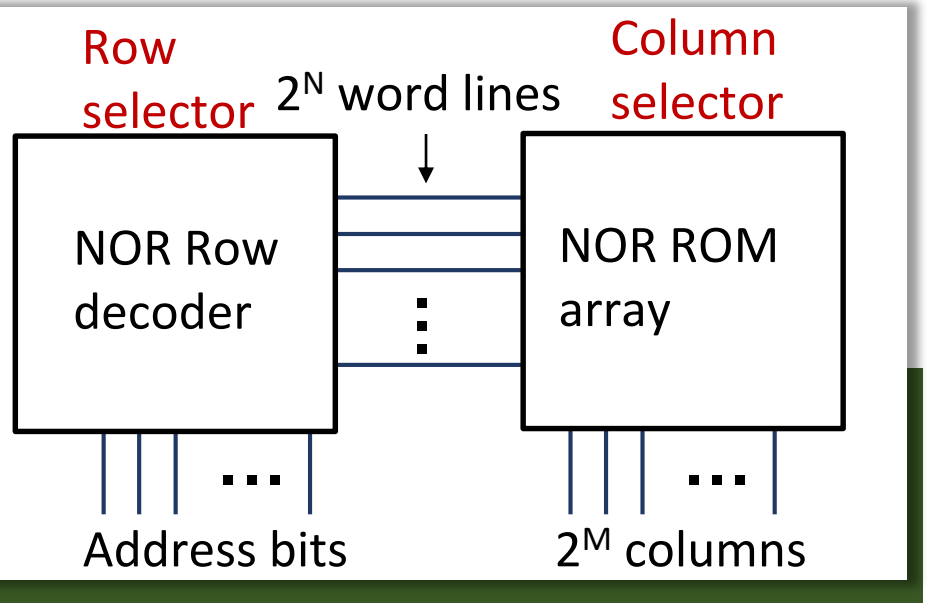
Row Decoder – NOR Array

Active high
Logic ‘1’ to select a row.



A_1	A_2
0	0
0	1
1	0
1	1

R_1	R_2	R_3	R_4
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1



Implementation

Row decoder and the ROM array as:

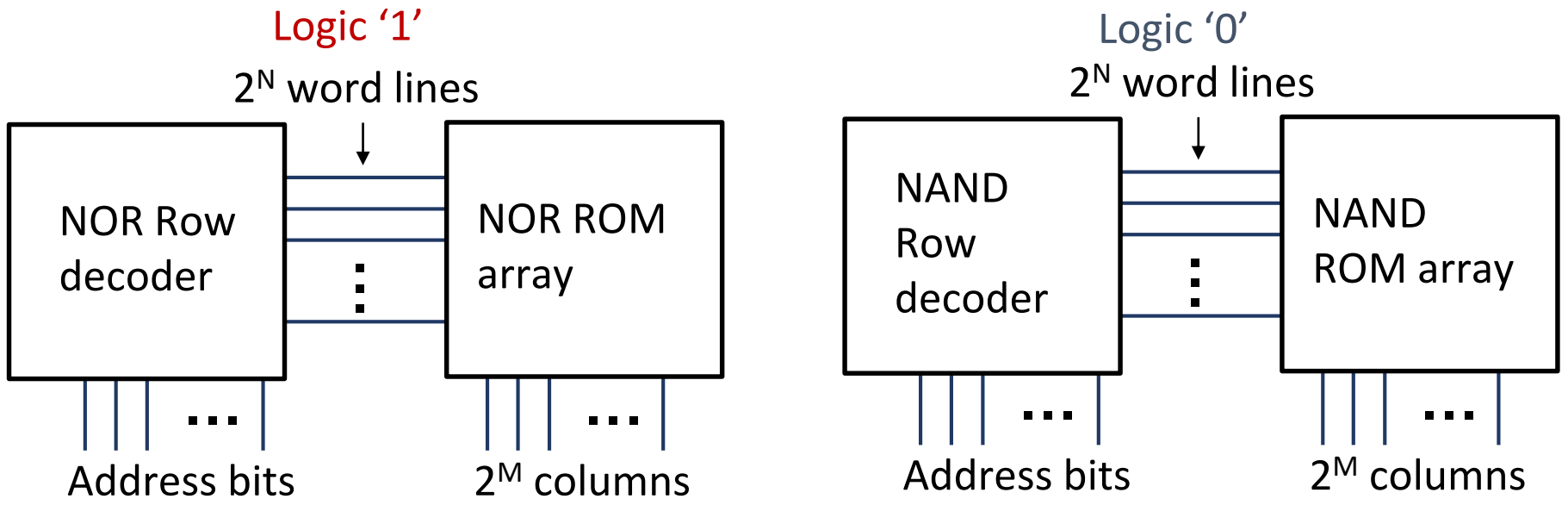
(a) two adjacent NOR planes and

(b) two adjacent NAND planes.

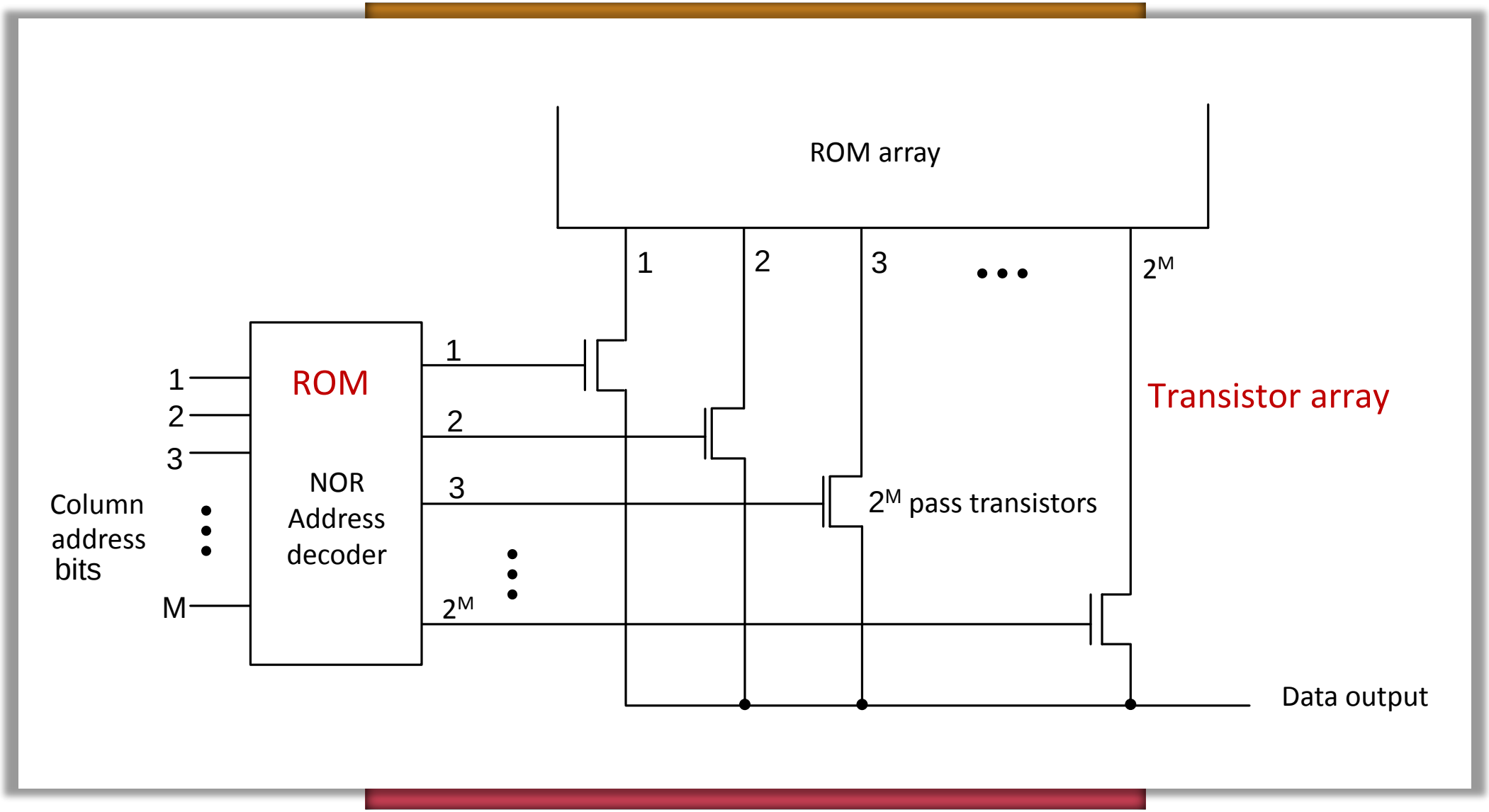
NAND Implementation

A_1	A_2
0	0
0	1
1	0
1	1

R_1	R_2	R_3	R4
0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	0

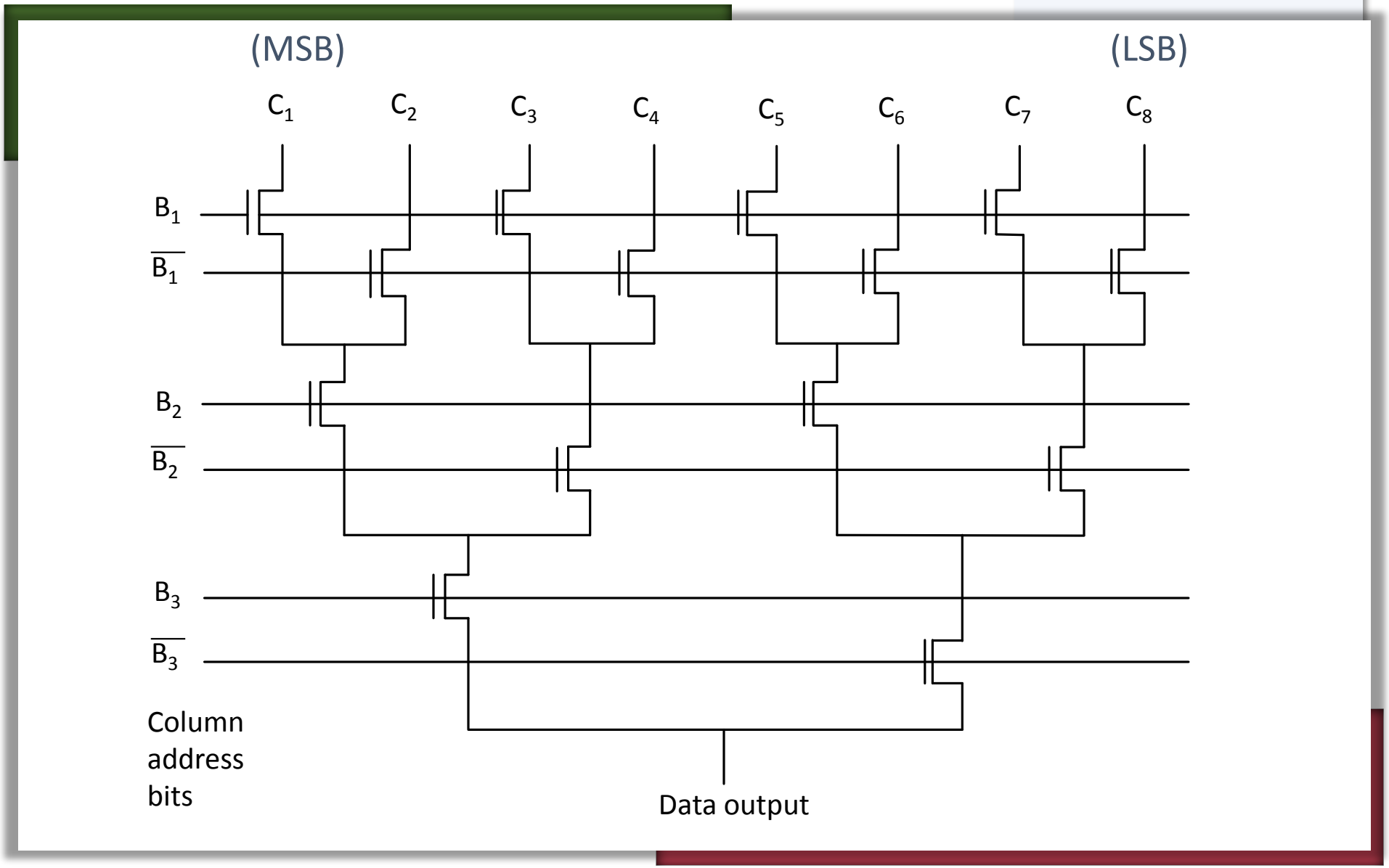


Column Decoder



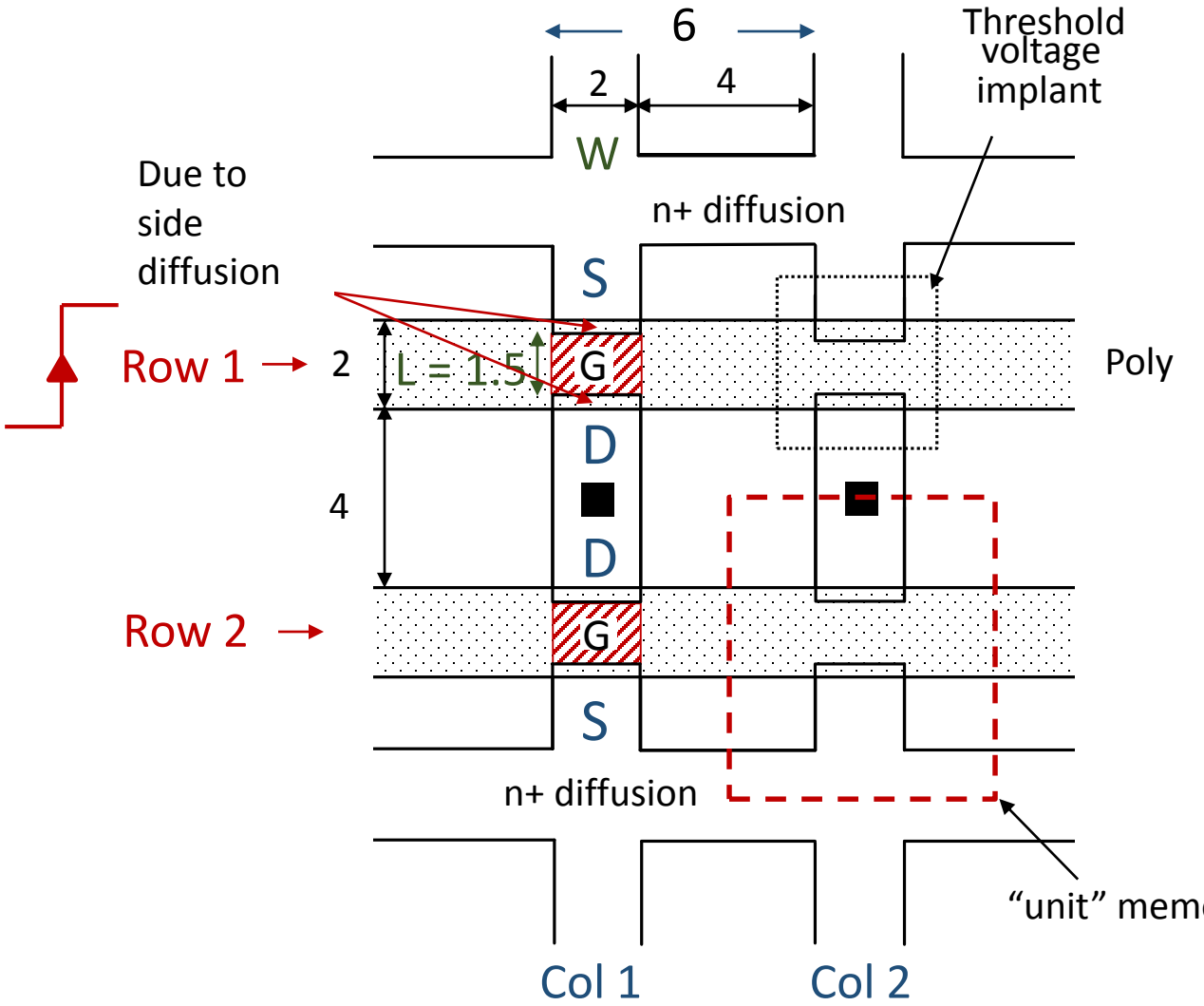
Column Decoder for 8-bit Lines

3 to 8 Tree decoder

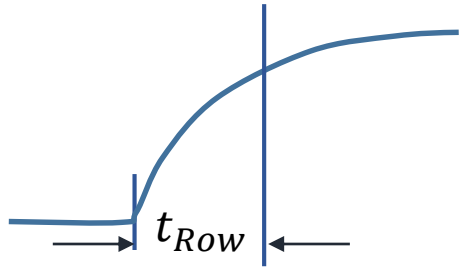


Row Access Time

Example



NOR ROM array with:
W = 2 μm and L = 1.5 μm



Drain contact shared by two adjacent transistors.
(Contact to metal column running on top of each diffusion column)

Row Access Time

$2^8 = 256$

A ROM array has 7 row address bits and 8 column address bits. Find the propagation delay associated with the row decoder circuit. Given that, $C_{ox} = 0.347\ \mu F/cm^2$, $\mu_n C_{ox} = 20\ \mu A/V^2$ and Poly sheet resistance $= 20\ \Omega/\square$.

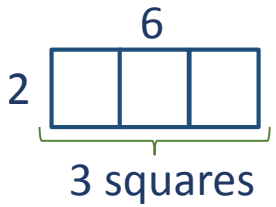
2 µm x 1.5 µm

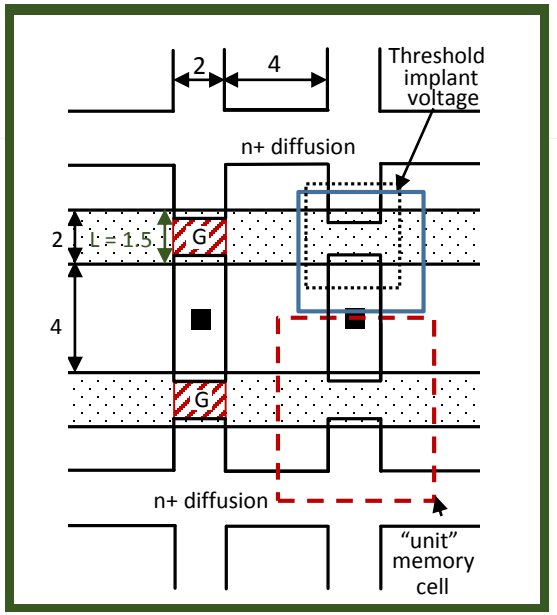
Solution:

$C_{row} = C_{ox}.W.L = 10.4 fF/bit$

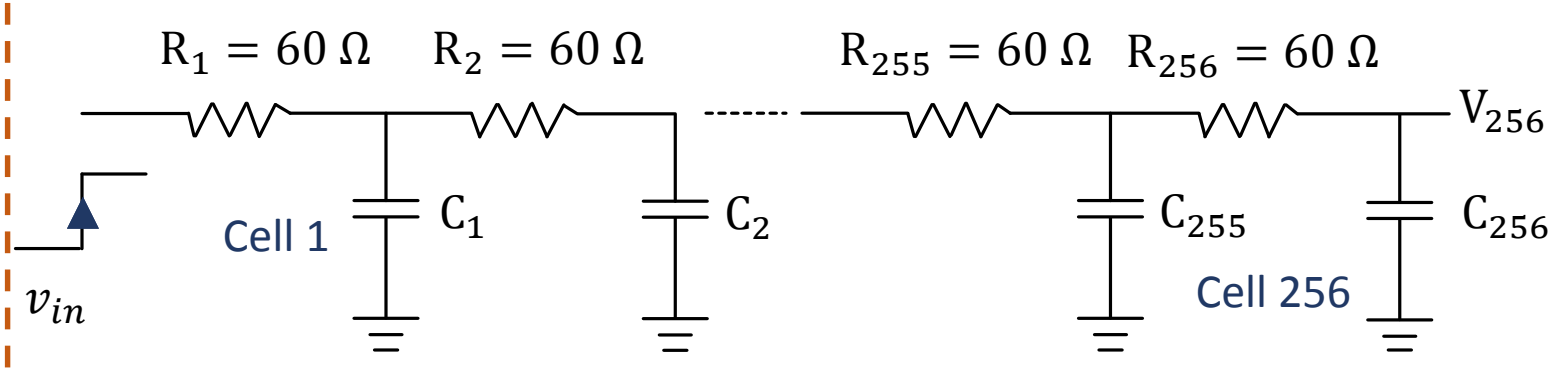
R_{row} =(#of square) × (Polysheet resistance)= $60\ \Omega/bit$

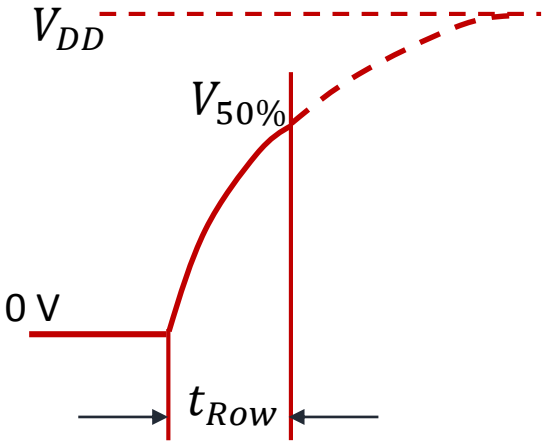
3 20

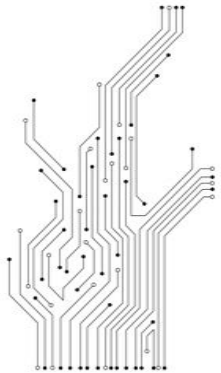




Example Contd.







Row Access Time

By neglecting the signal propagation delay associated with the row address decoder circuit and assuming that the row line is driven by an ideal step voltage waveform, the row access time can be approximated as:

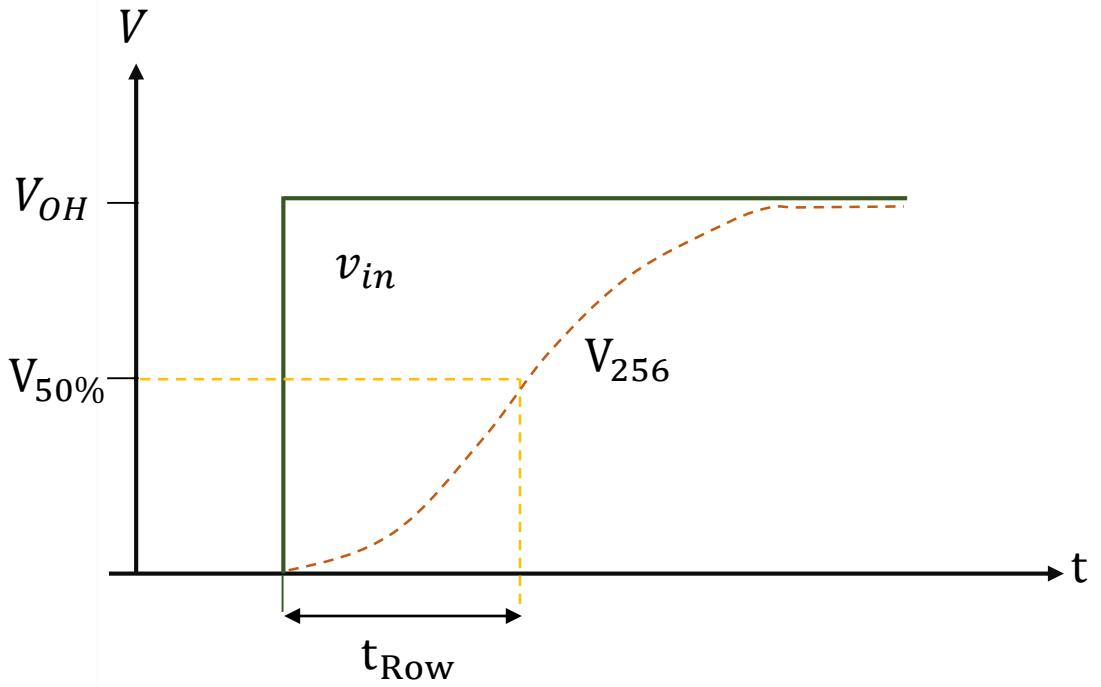
By empirical (experiment)

Where $t_{Row} \approx 0.38 . R_T . C_T = 15.53\,\text{ns}$

Example Contd.

$$R_T = \sum_{\text{all col}}^{256} R_i = 15.36\,\text{k}\Omega$$

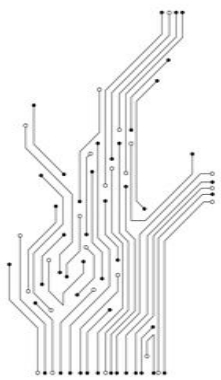
$$C_T = \sum_{\text{all col}}^{256} C_i = 2.66\,\text{pF}$$



Summary

Here are the key takeaways from this lesson.

- There are two different types of semiconductor memory, RAM and ROM. There are two types of RAMs, DRAM and SRAM.
- Read, write and refresh are the various operations of DRAM.
- Read and write are the various operations of SRAM.
- Read is the only operation for ROM.
    - Row and Column decoders are used to decode the address to a selected row and column.
    - There are two types of implementation, NAND and NOR based implementations.
    - Row access time is the propagation delay from the first column to the last column of each row of the memory array.



Thank You

Semiconductor Memories