Semiconductor Memories

A. Marzuki

Introduction

- Introduction
- DRAM
- SRAM
- MASK ROM

Introduction

- Data storage capacity available on a single integrated circuit grows exponentially being doubled approximately every two years.
- **Capacity** of the dynamic read/write memory (DRAM) chip exceeds now 1 Gigabit.
- **Data transfer speed** of a standard DRAM is at the level of 200Mb/sec/pin.

Introduction

- Non-volatile Memory (NVM) also known as Read-Only Memory (ROM) which retains information when the power supply voltage is off. With respect to the data storage mechanism NVM are divided into the following groups:

– Mask programmed ROM. The required contents of the memory is programmed during fabrication,

– Programmable ROM (PROM). The required contents is written in a permanent way by burning out internal interconnections (fuses). It is a one-off procedure.

– Erasable PROM (EPROM). Data is stored as a charge on an isolated gate capacitor (“floating gate”). Data is removed by exposing the PROM to the ultraviolet light.

– Electrically Erasable PROM (EEPROM) also known as Flash Memory. It is also base on the concept of the floating gate. The contents can be re-programmed by applying a suitable voltages to the EEPROM pins. The Flash Memories are veryimportant data storage devices for mobile applications.

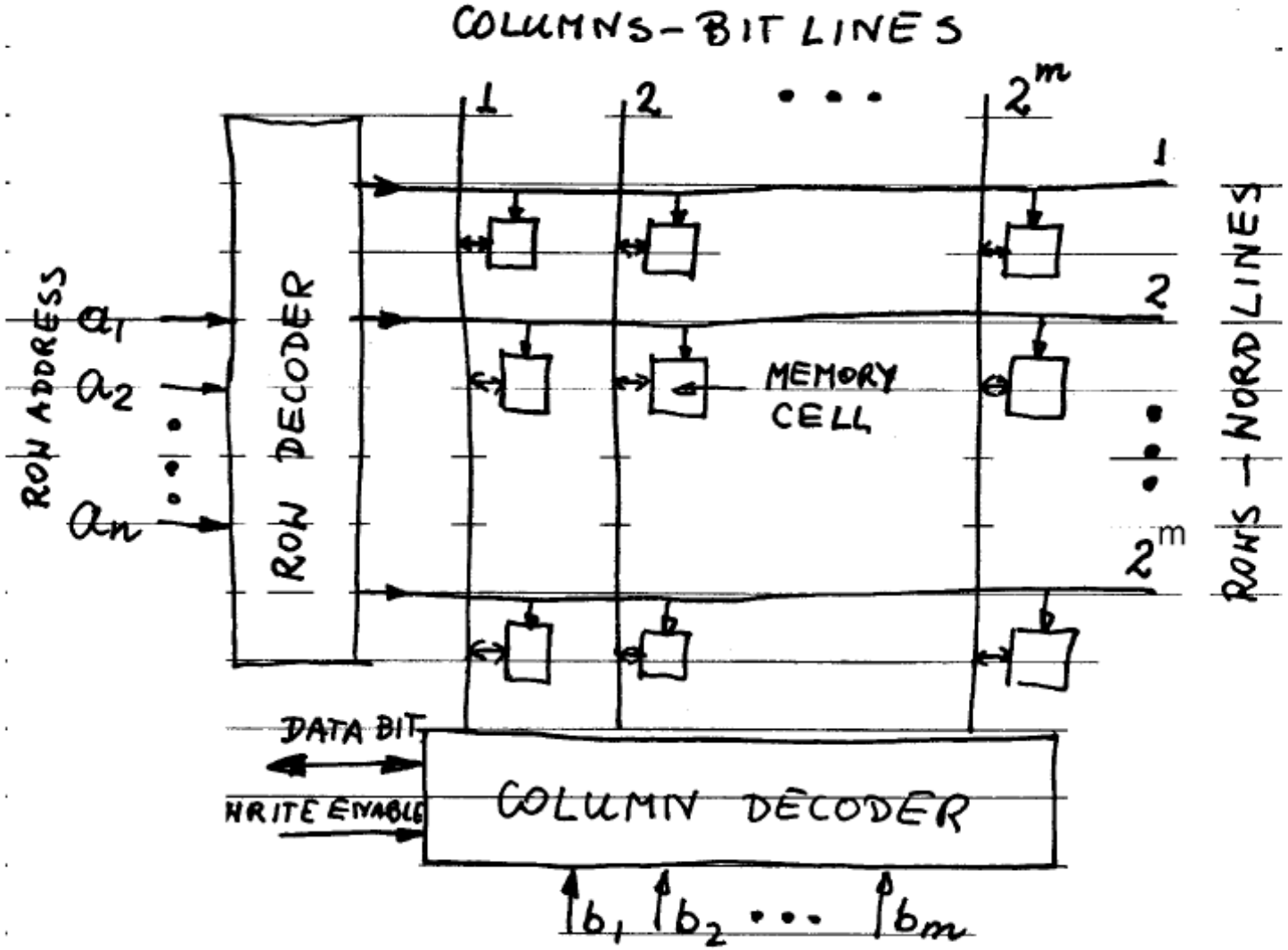
Introduction

- Read/Write (R/W) memory, also known as Random Access Memory (RAM). From the point of view of the data storage mechanism RAM are divided into two main groups:

– Static RAM, where data is retained as long as there is power supply on.

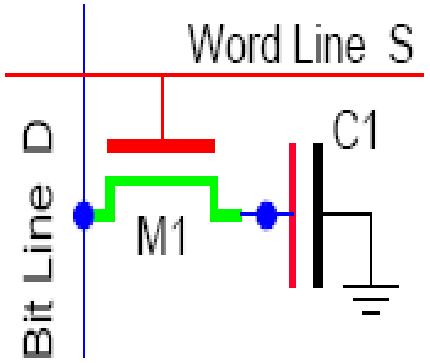
– Dynamic RAM, where data is stored on capacitors and requires a periodic refreshment.

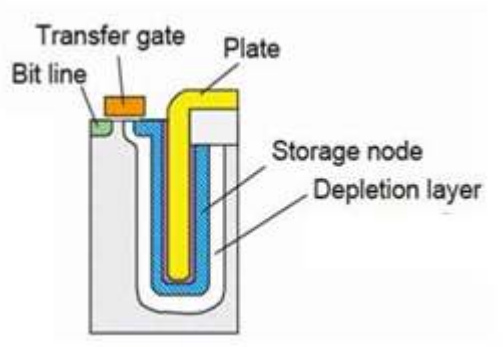
Example of memory organization



DRAM

dynamic RAM binary data is stored as **charge in a capacitor**





The **storage capacitance** C1, which is connected between the drain of the access transistor (the storage node) and the ground, is formed as a **trench** or **stacked** capacitor. The stacked capacitor is created between a second polysilicon layer and a metal plate covering the whole array area. The plate is effectively connected to the ground terminal.

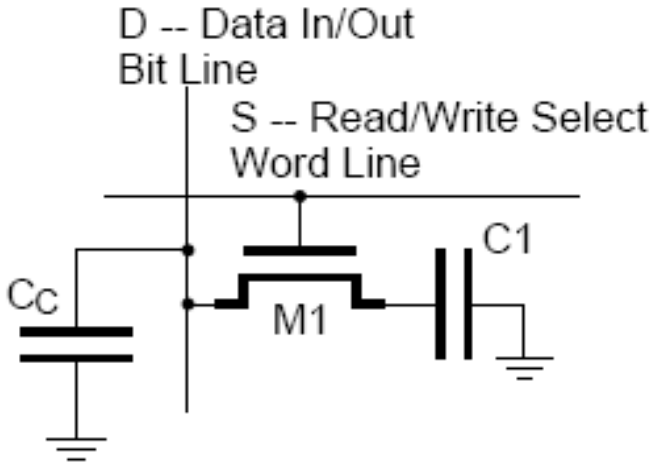
High permittivity dielectric material
Tantalum pentoxide, $\boldsymbol{Ta_2O_5}$

Leakage

Typical storage capacitance has a value of 20 to 50 fF.
Assuming that the voltage on the fully charged storage capacitor is V = 2.5V, and that the leakage current is I = 40pA, then the time to discharge the capacitor C = 20fF to the half of the initial voltage can be estimated as

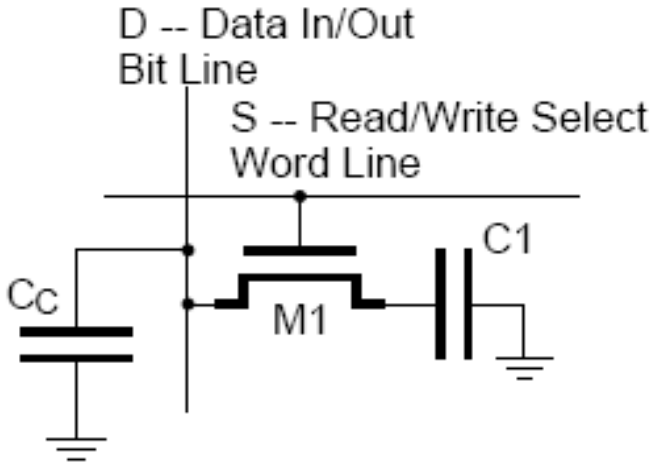
$$t = \frac{1}{2}\frac{C \cdot V}{I} = \frac{20 \cdot 10^{-15} \cdot 2.5}{40 \cdot 10^{-12}} = 0.625\text{ms}$$

DRAM-Read & Write



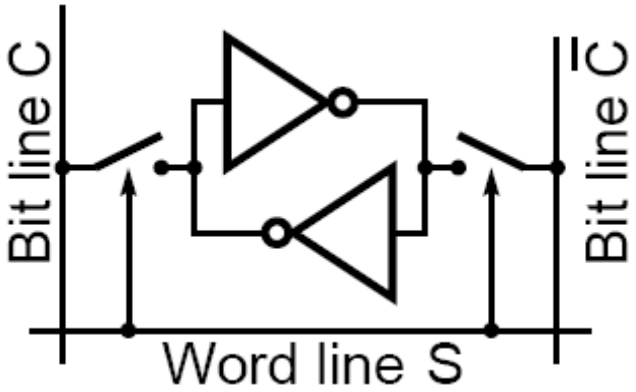
During **read operation** there is a flow of charges between the storage capacitance C1 and the column capacitance, C_C. As a result the column voltage either increases (read ‘1’) or decreases (read ‘0’)
slightly. This difference can then be amplified by the sense amplifier.

DRAM-Read & Write



The **write operation** is performed by applying either high or low voltage to the bit line thus charging (write ‘1’) or discharging (write ‘0’) the storage capacitance through the access transistor.

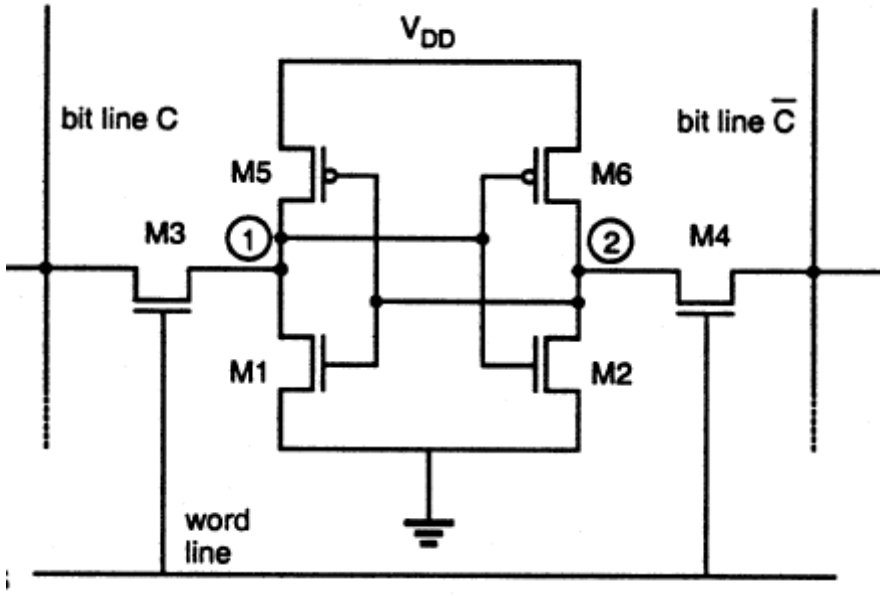
SRAM



The two-inverter latch is able to store one bit data. In order to access the cell the word line is activated with high-level signal S, which closes access switches on both sides of the cell.
The state of the cell (and its complement) is now available on two complemented bit lines and the **read operation** can be performed. In order to perform write operation the data and its complement is supplied through the bit line.

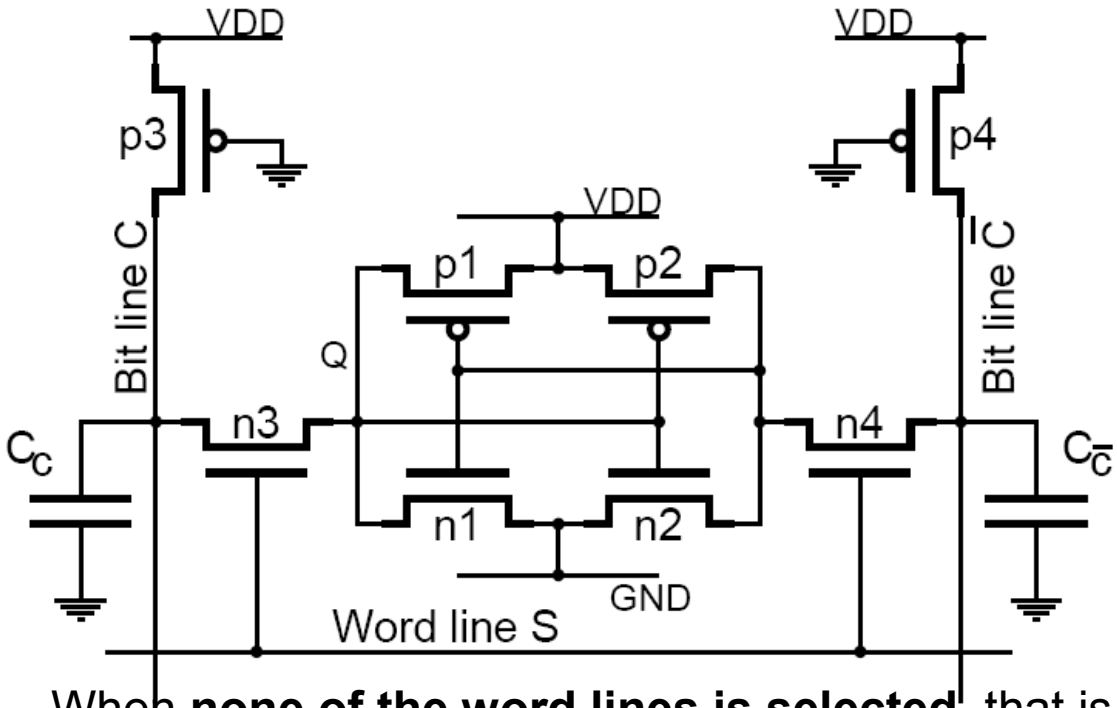
Static Read/Write (or Random Access) memory (SRAM) is able to read and write data into its memory cells and retain the memory contents as long as the power supply voltage is provided.
Currently SRAM are manufactured in the CMOS technology which offers very low static power dissipation, superior noise margin and switching speed.

The schematic of a CMOS SRAM



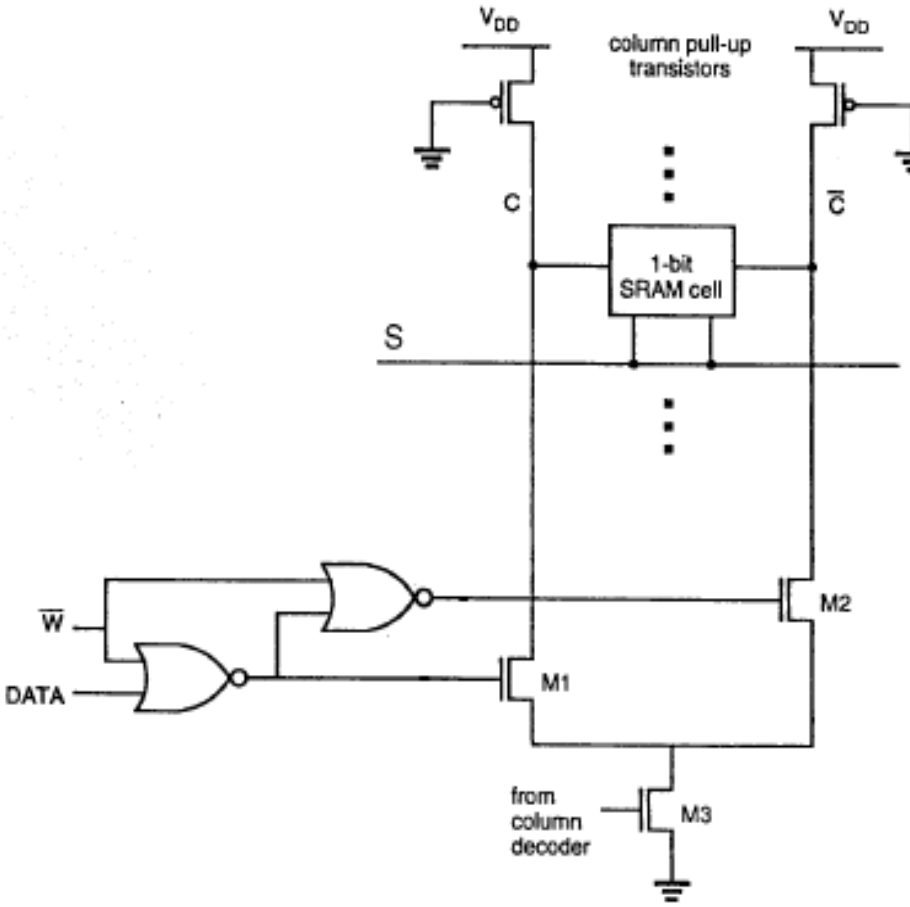
The cell consists of six transistors: four nMOS a two pMOS. Two pairs of transistors form a pair of inverters and two nMOS transistors form the access switches.

Basic Principle



When **none of the word lines is selected**, that is, all S signals are '0',
the pass transistors n3, n4 are turned off and the data is retained in all memory cells. The column capacitances are charged by the drain currents of the pull-up pMOS transistors, p3, p4

SRAM Write Circuit

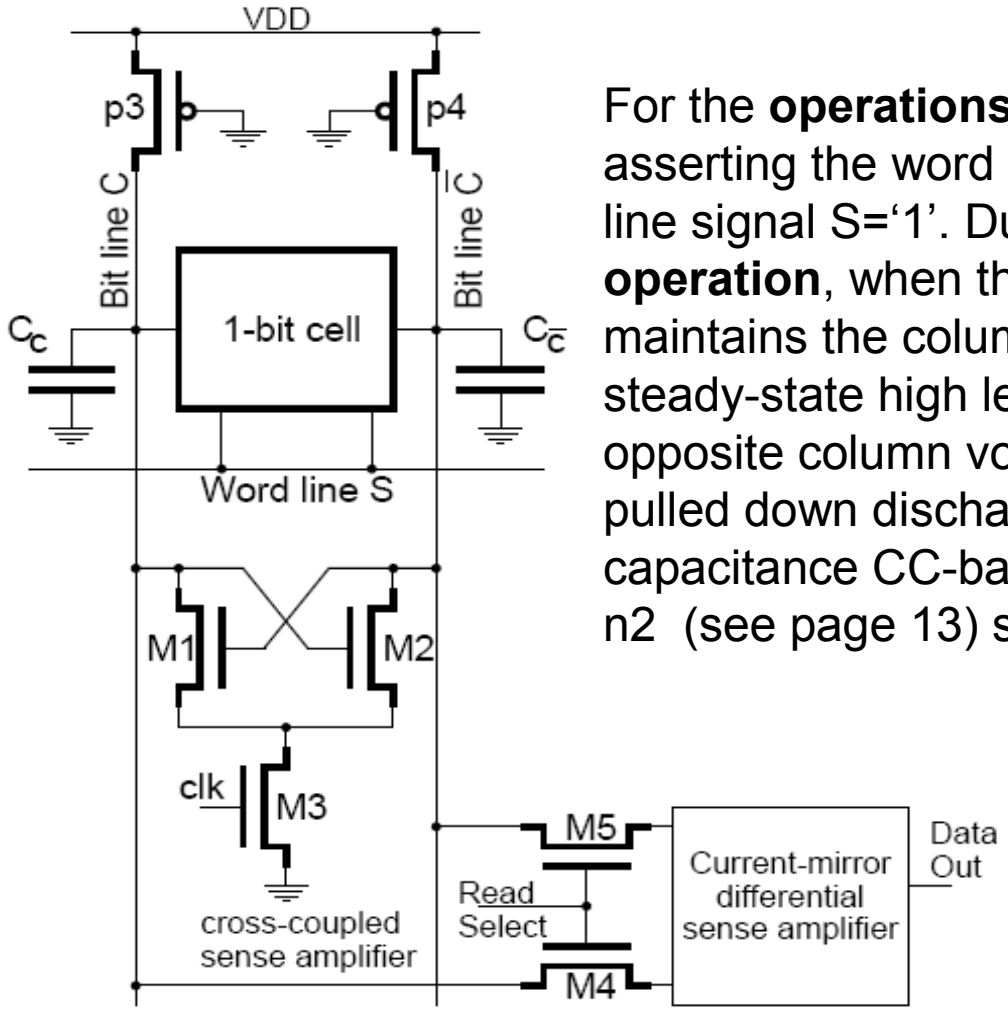


To **write '0'** in the cell, the column voltage VC is forced to low (C = 0).

To **write '1'** in the cell, the opposite column voltage VC-bar is forced to low (C-bar = 0)

This is achieved by connecting either C or C-bar to the ground through the transistor M3 and either M1 or M2.

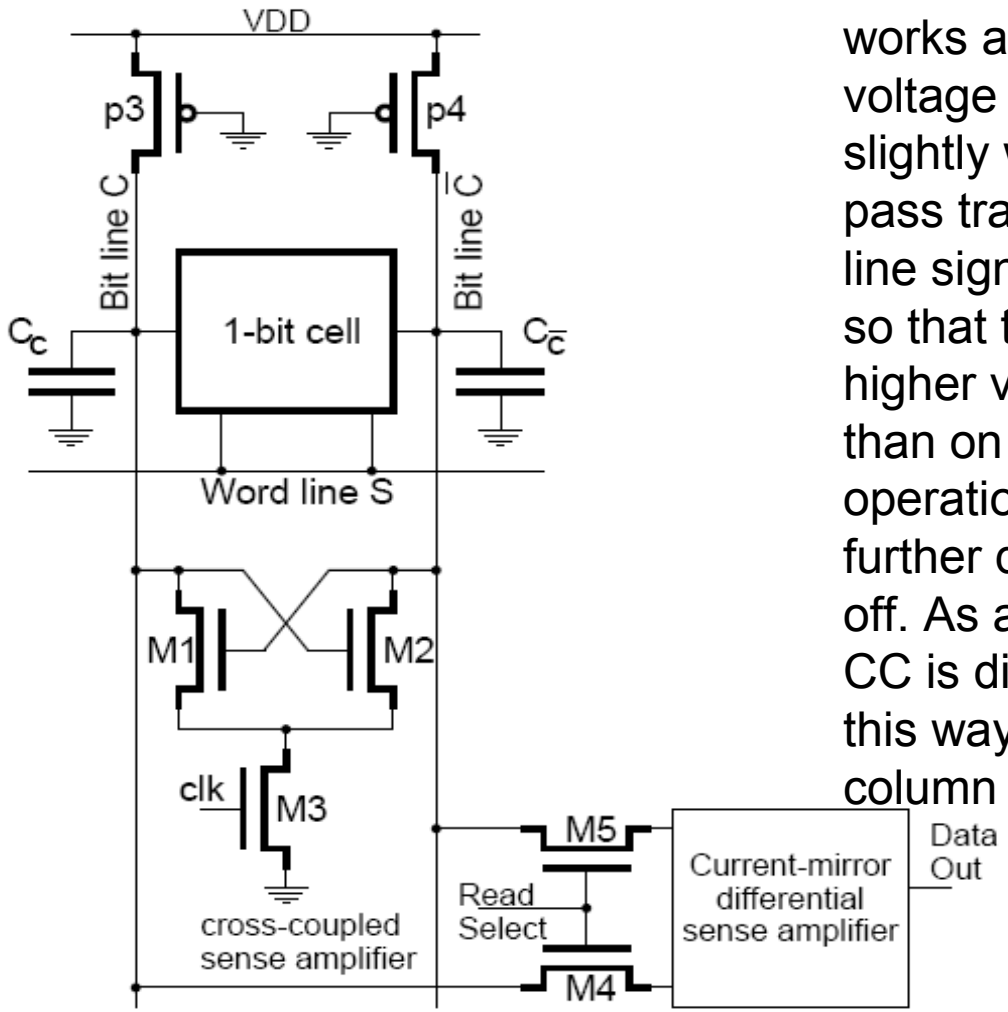
SRAM Read circuit



For the **operations** we select the cell asserting the word line signal S='1'. During the **read '1' operation**, when the stored bit is 1.This maintains the column voltage VC at its steady-state high level (say 3.5V) while the opposite column voltage VC-bar is being pulled down discharging the column capacitance CC-bar through transistors n4, n2 (see page 13) so that VC > VC-bar.

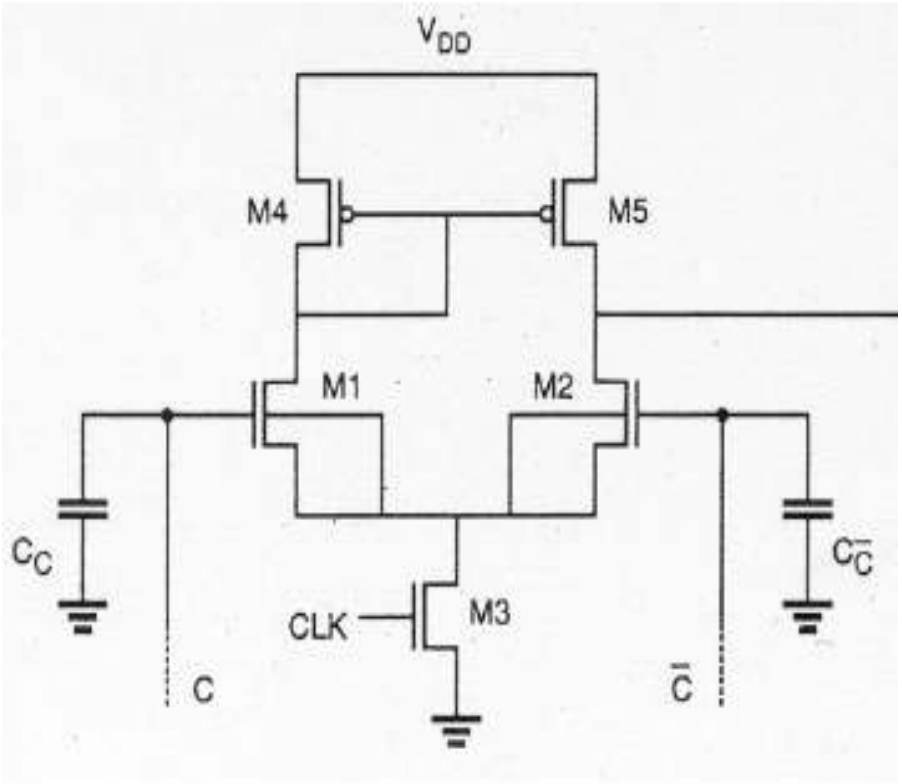
$$\begin{array}{ll} `0' & \text{if} \quad V_C < V_{\bar{C}} \\ `1' & \text{if} \quad V_C > V_{\bar{C}} \end{array}$$

SRAM Read circuit



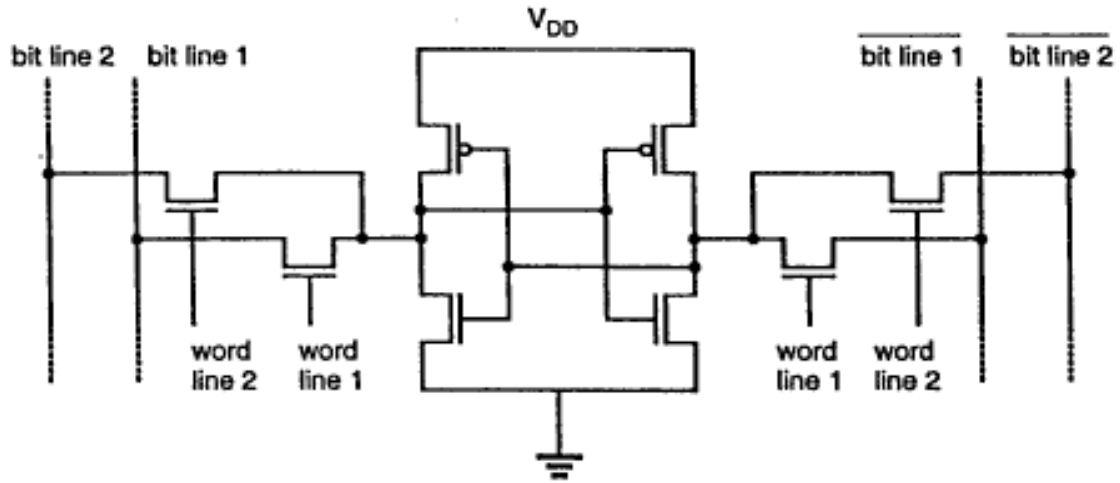
The **cross-coupled sense amplifier** works as a latch. Assume that the voltage on the bit line C start to drop slightly when the memory access pass transistors are activated by the word line signal S, and that the clk signal is high so that the transistor M3 is turned on. Now, higher voltage on the gate of M1 transistor than on the gate of M2 starts the latching operation which pulls the VC voltage further down switching the transistor M2 off. As a result the parasitic capacitance, CC is discharged through M1 and M3. In this way a small difference between column voltages is amplified.

Differential current-mirror sense amplifier



Dual-Port SRAM

- Dual port SRAM allows simultaneously access to the same location in the memory array.

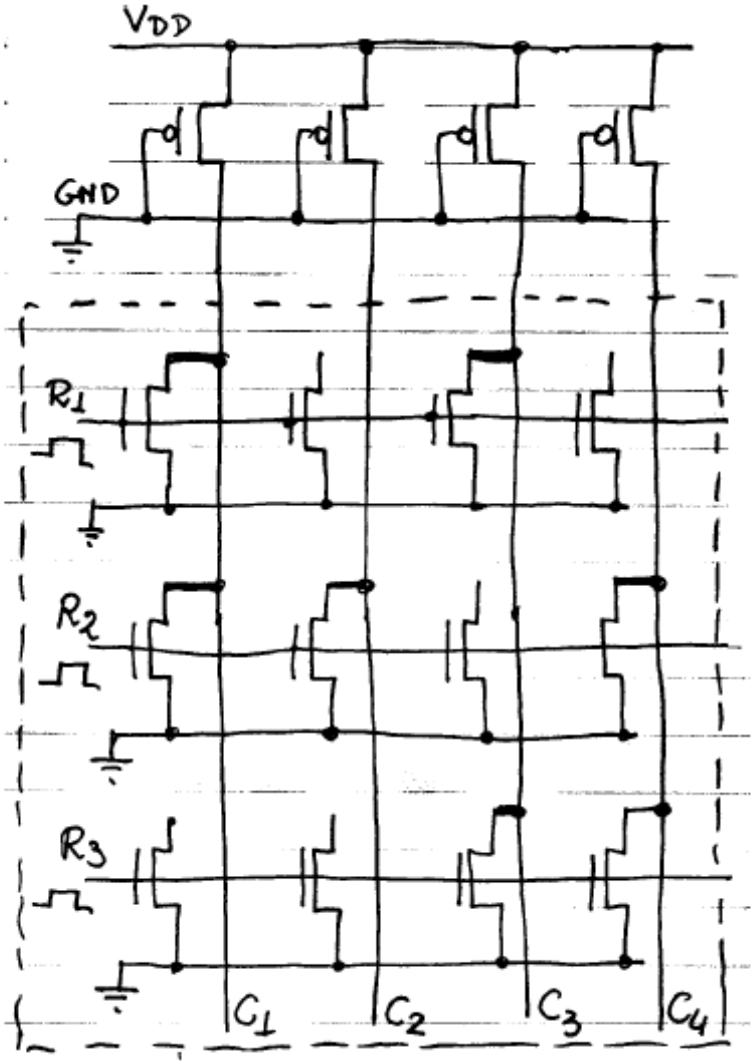


To resolve contention (timing conflict) due to more than one processor.

Mask PROM

- NOR-based ROM
- NAND-based ROM

NOR-based ROM

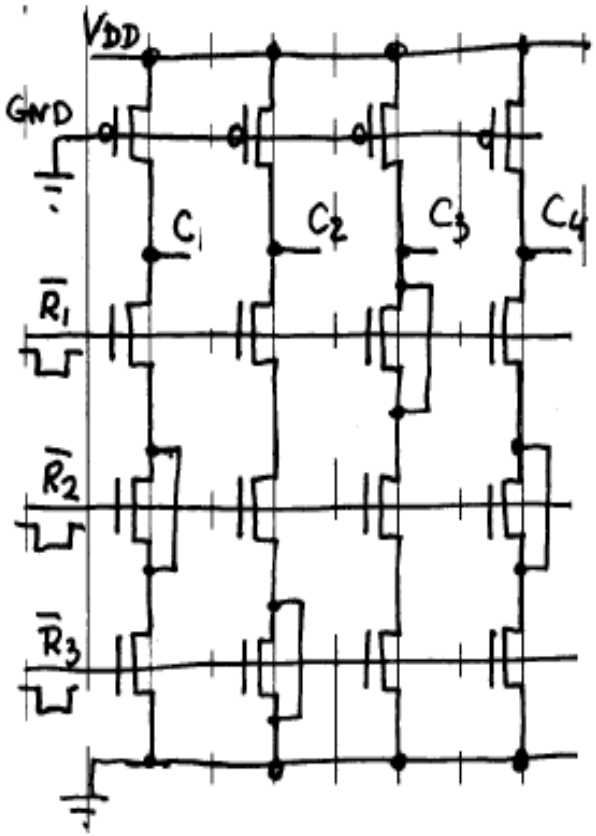


Each memory cell is represented by one nMOS transistor and a binary **information is stored** by **connecting or not** the drain terminal of such a transistor to the **bit line**.

	C_1	C_2	C_3	C_4
R_1	0	1	0	1
R_2	0	0	1	0
R_3	1	1	0	0

For every row address only one **word line** is activated by applying a high signal to the gates of nMOS transistors in a row. If a selected transistor in the i-th column is connected to a bit line then the **logic '0'** is stored in this memory cell. if the transistor is not connected, then the **logic '1'** is stored.

NAND-based ROM



When no word line is activated, all nMOS transistors are on and the line signals, Ci are all low. When a word line is activated all transistors in the row are switched off and the respective Ci signals are high. If a transistor in the selected row is short-circuited, then the respective Ci signal is low.

	C_1	C_2	C_3	C_4
$\overline{R_1}$	1	1	0	1
$\overline{R_2}$	0	1	1	0
$\overline{R_3}$	1	0	1	1

In other words, the **logic '0'** is stored when a transistor is replaced with a wire, whereas the **logic '1'** is stored by an nMOS transistor being present.

References

- A.P.Paplinski's Lecture Material.
- S-M. Kang and Y. Leblebici ,*CMOS Digital Integrated Circuits: Analysis and Design*,, 3rd edition