UNIVERSITÀ
DEGLI STUDI DI TRIESTE

dipartimento
di ingegneria
e architettura

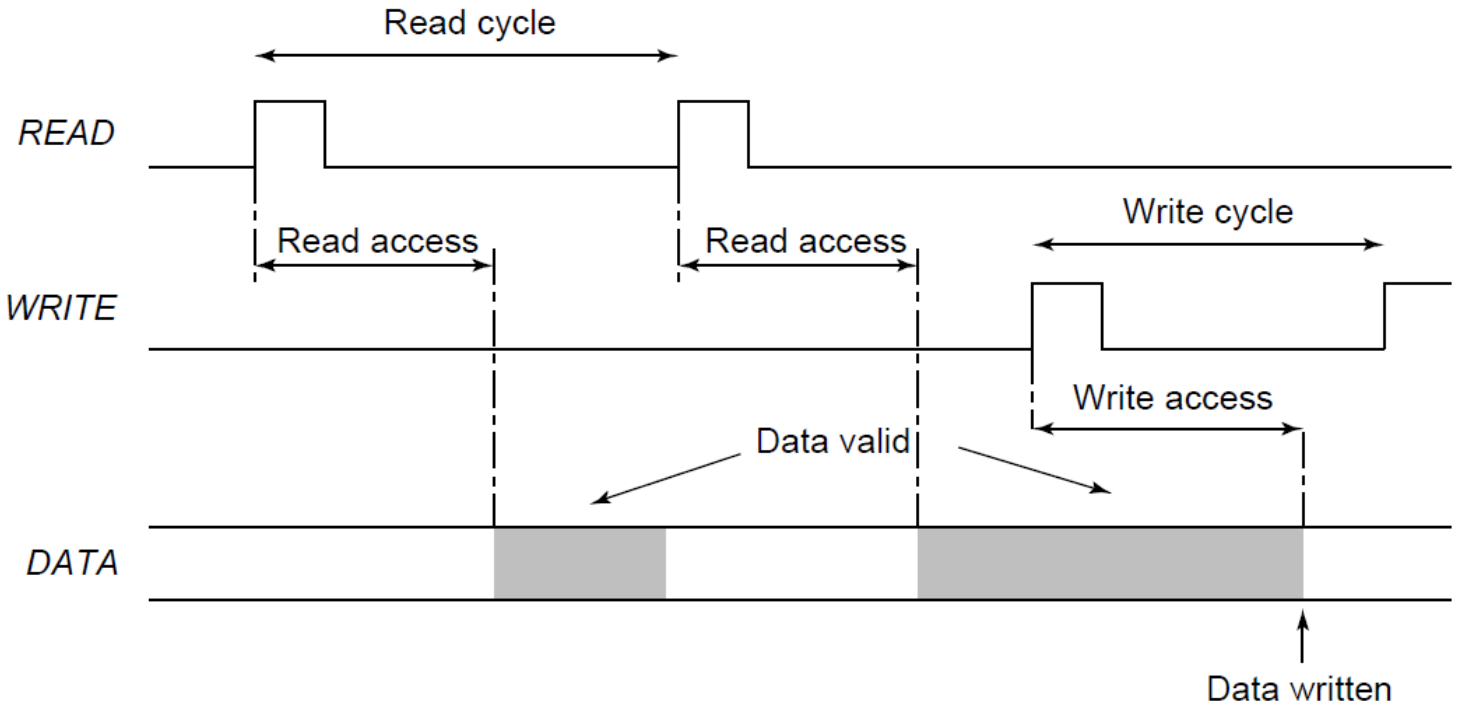
Semiconductor Memories

A. Carini – Digital Integrated Circuits

Memory Classification

- Memories comes in many different format and styles.
- The type of memory unit that is preferable for a given application is a function of many factors: the memory size, the time it takes to access the stored data, the application, the system requirements.
- They can be classified according to
  - Size
  - Timing parameters
  - Function
  - Access Pattern
  - Input/Output architecture
  - Application

Timing parameters



Timing parameters

- *Read-access time:* the time it takes to retrieve (read) from the memory, equal to the delay between the read request and the moment the data is available at the output.
- *Write-access time:* the time elapsed between the write request and the final writing of the input data into the memory.
- *Cycle time* of the memory: the minimum time required between successive reads and writes (normally greater than the access time).

Function

- We will distinguish between ***nonvolatile*** *and* ***volatile memories.***
- *Nonvolatile memories* preserve the information when the power supply is turned off. On the contrary *volatile memories* does not preserve it.
- Another distinction is made between *read-only (ROM)* and *read-write (RWM)* memories.
- ROM encode the information into the circuit topology and are nonvolatile.
- Volatile RWM memories offer both read and write functionality with comparable access times. Data are stored either in flip-flops (*static* memories) or as a charge on a capacitor (*dynamic* memories).
- In nonvolatile RWM (NVRWM), the write operation takes substantially more time than the read. Members of this family are:
  - EPROM (erasable programmable ROM),
  - E^2PROM (electrically erasable programmable ROM),
  - Flash memories.

Access Pattern

- Most memories belong to the **random-access** class, which means memory locations can be read or written in a random order.
- One would expect memories of this class to be called **RAM** modules (random-access memory). For historical reasons, this name has been reserved for the volatile random-access RWM memories.
- Be aware that most ROM or NVRWM units also provide random access, but the acronym RAM should not be used for them.

- Some memory types restrict the order of access, which results in either faster access times, smaller area, or a memory with a special functionality.
- Examples of such are the serial memories: the FIFO (first-in first-out), LIFO (last-in first-out, most often used as a stack), and the shift register.
    - Video memories are an important member of this class.

Access Pattern

- Contents-addressable memories (CAM) represent another important class of nonrandom access memories.
- Instead of using an address to locate the data, a CAM uses a word of data itself as input in a query-style format. When the input data matches a data word stored in the memory array, a MATCH flag is raised.
- Associative memories are an important component of the cache architecture of many microprocessors.

Memory Classification

Read-Write Memory		Non-Volatile Read-Write Memory	Read-Only Memory
Random Access	**Non-Random Access**	EPROM E^2PROM FLASH	Mask-Programmed Programmable (PROM)
SRAM DRAM	FIFO LIFO Shift Register CAM		

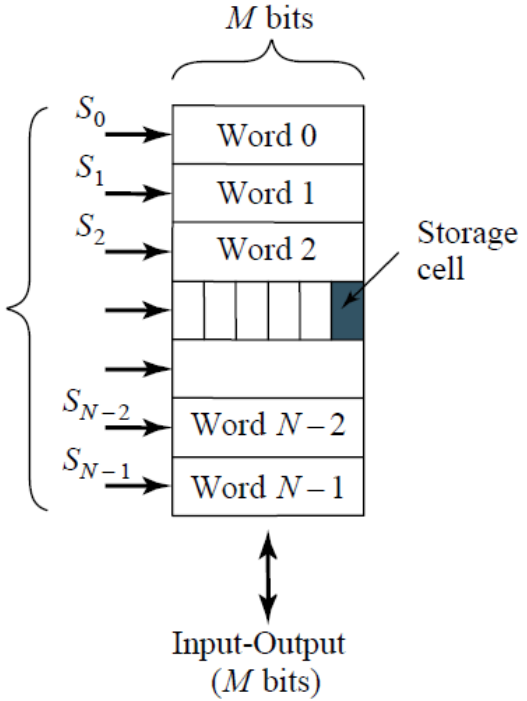
Input/Output Architecture

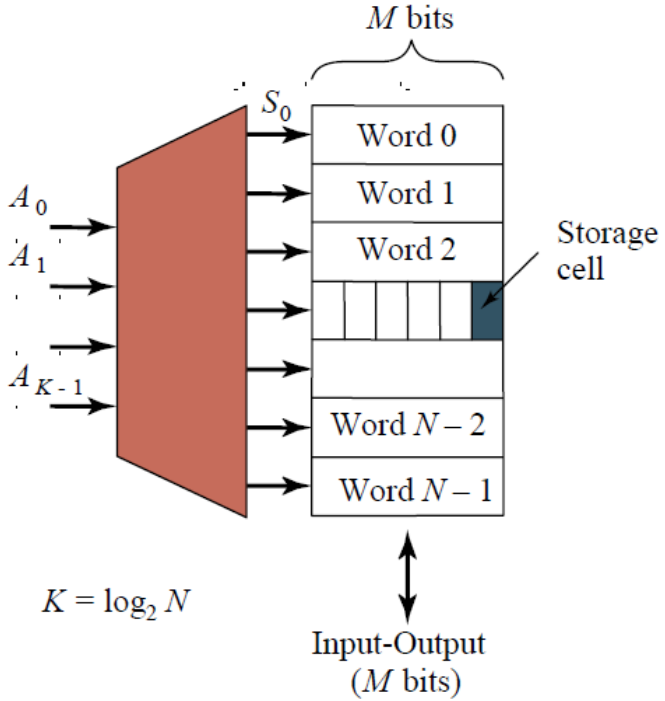
- A final classification of semiconductor memories is based on the number of data input and output ports.
- While a majority of the memory units presents only a *single port* that is shared between input and output, memories with higher bandwidth requirements often have *multiple input and output ports*—and thus are called multiport memories.

Memory Architecture and Building Blocks

- Assume that we would like to implement a memory that holds 1 million ($N = 10^6$) 8-bit ($M = 8$) words.
- When implementing this structure directly selecting the words, we quickly realize that 1 million select signals are needed—one for every word.
- Since these signals are normally provided from off-chip or from another part of the chip, this translates into insurmountable wiring and/or packaging problems.
- A decoder is inserted to reduce the number of select signals.
- A memory word is selected by providing a binary encoded address word.
- The decoder translates this address into $N = 2^k$ select lines, only one of which is active at a time.
- This approach reduces the number of external address lines from 10^6 to 20 ($\log_2 10^6$) in our example, which virtually eliminates the wiring and packaging problems.
- The decoder is typically designed so that its dimensions are matched to the size of the storage cell and the connections between the two.

Memory Architecture and Building Blocks





Intuitive architecture for N x M memory
Too many select signals:
N words == N select signals

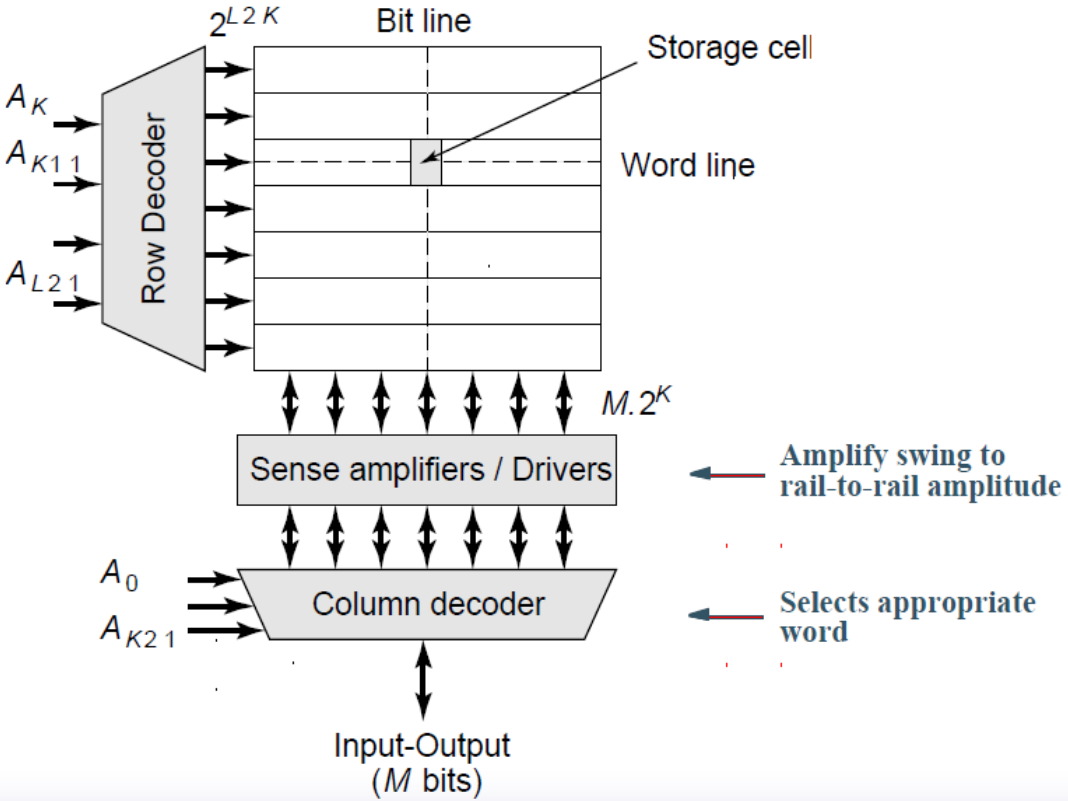
Decoder reduces the number of select signals

$$K = log_2 N$$

Memory Architecture and Building Blocks

- While this resolves the select problem, it does not address the issue of the memory aspect ratio.
- Evaluation of the dimensions of the storage array of our token example shows that its height is approximately 128,000 times larger than its width, assuming the shape of the basic storage cell is approximately square.
- This results in a design that cannot be implemented.
- Besides the bizarre shape factor, the resulting design is also extremely slow.
- The vertical wires connecting the storage cells to the input/outputs become excessively long.
- To address this problem, memory arrays are organized so that the vertical and horizontal dimensions are of the same order of magnitude; thus, the aspect ratio approaches unity.
- Multiple words are stored in a single row and are selected simultaneously.
- To route the correct word to the input/output terminals, a **column decoder** is used.

Memory Architecture and Building Blocks



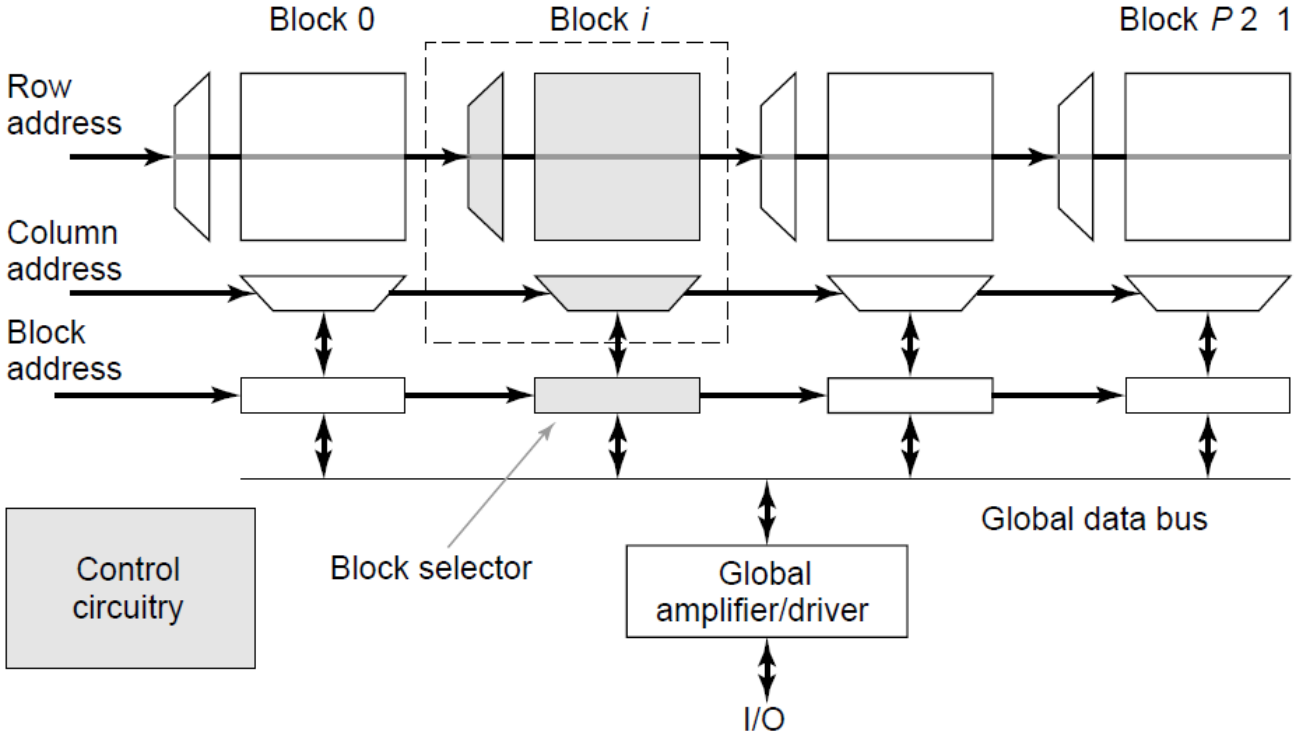
Memory Architecture and Building Blocks

- The area of large memory modules is dominated by the size of the memory core.
- Thus, it is crucial to keep the size of the basic storage cell as small as possible.
- Semiconductor memory cells therefore reduce the cell area by trading off some desired properties of digital circuits, such as noise margin, swing, input/output isolation, fan-out, or speed.
- While a degradation of some of those properties is allowable within the confined domain of the memory core where noise levels be tightly controlled, this is not acceptable when interfacing with the external or surrounding circuitry.
- The desired digital signal properties must be recovered with the aid of peripheral circuitry.

Memory Architecture and Building Blocks

- For example, it is common to reduce the voltage swing on the bit lines to a value substantially below the supply voltage.
- This reduces both the propagation delay and the power consumption.
- Interfacing to the external world, on the other hand, requires an amplification of the internal swing to a full rail-to-rail amplitude.
- This is achieved by the *sense amplifiers.*
- Relaxation of bounds on a number of the coveted digital properties makes it possible to reduce the transistor count of a single memory cell to between **one and six transistors**!

Memory Architecture and Building Blocks



Memory Architecture and Building Blocks

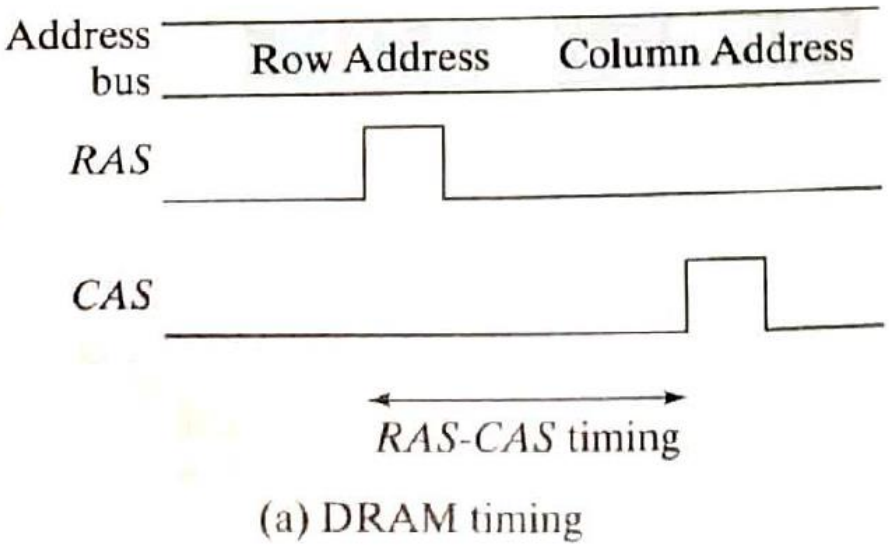
This approach has a dual advantage:

1. length of the local word and bit lines—that is, the length of the lines within the blocks—is kept within bounds, resulting in faster access times,
2. The block address can be used to activate only the addressed block. Nonactive blocks are put in power saving mode with sense amplifiers and row and column decoders disabled. This results in a substantial power saving that is desirable, since power dissipation is a major concern in very large memories.

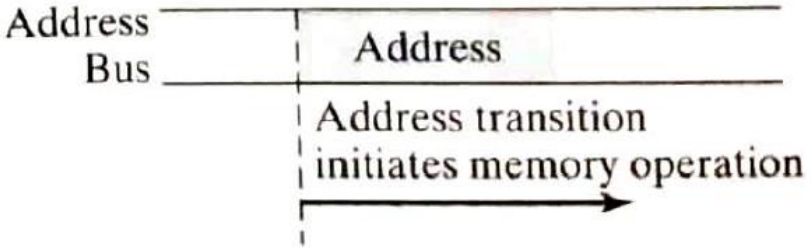
Memory Architecture and Building Blocks

- A component of memory design that is often overlooked is the input/output interface and control circuitry.
- The nature of the I/O interface has an enormous impact on the global memory control and timing.
- This statement is illustrated by comparing the input-output behavior of typical DRAM and SRAM components with the associated timing structure.

Memory Architecture and Building Blocks



(a) DRAM timing



(b) SRAM timing

Memory Architecture and Building Blocks

- Since the early days, DRAM designers have opted for a multiplexed addressing scheme.
- In this model, the lower and upper halves of the address words are presented sequentially on a single address bus.
- This approach reduces the number of package pins and has survived through the subsequent memory generations.
- DRAMs are generally produced in higher volumes.
- Lowering the pin count reduces cost and size at the expense of performance.
- To ensure correct memory operation, a careful timing of the RAS-CAS interval is necessary.
- In fact, the RAS and CAS signals act as clock inputs to the memory module, and are used to synchronize memory events, such as decoding, memory core access, and sensing.

Memory Architecture and Building Blocks

- The SRAM designers, on the other hand, have chosen a self-timed approach.
- The complete address word is presented at once, and circuitry is provided to automatically detect any transitions on that bus.
- No external timing signals are needed.
- All internal timing events, such as the enabling of the decoders and sense amplifiers, are derived from the internally generated transition signal.
- This approach has the advantage that the cycle time of the SRAM is close or equal to its access time.

- Several new approaches to improve the performance of the DRAM for read operations have been introduced.
- Examples are Synchronous DRAM (SDRAM) and Rambus DRAM (RDRAM).
- The main novelty in these new architectures is not in the memory core, but in how they communicate with the outside world.

The Memory Core: Read-Only Memories

- The memory cell:

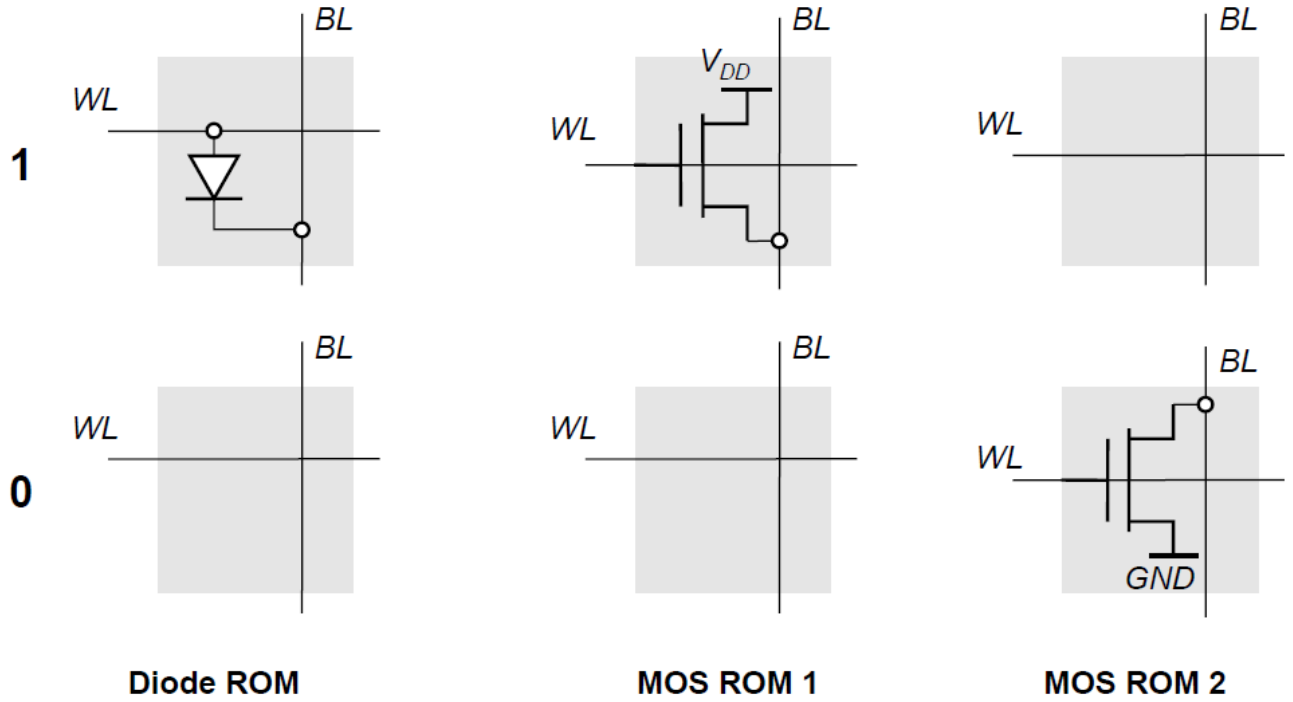


Fig. 12-9

Read-Only Memories: MOS OR ROM

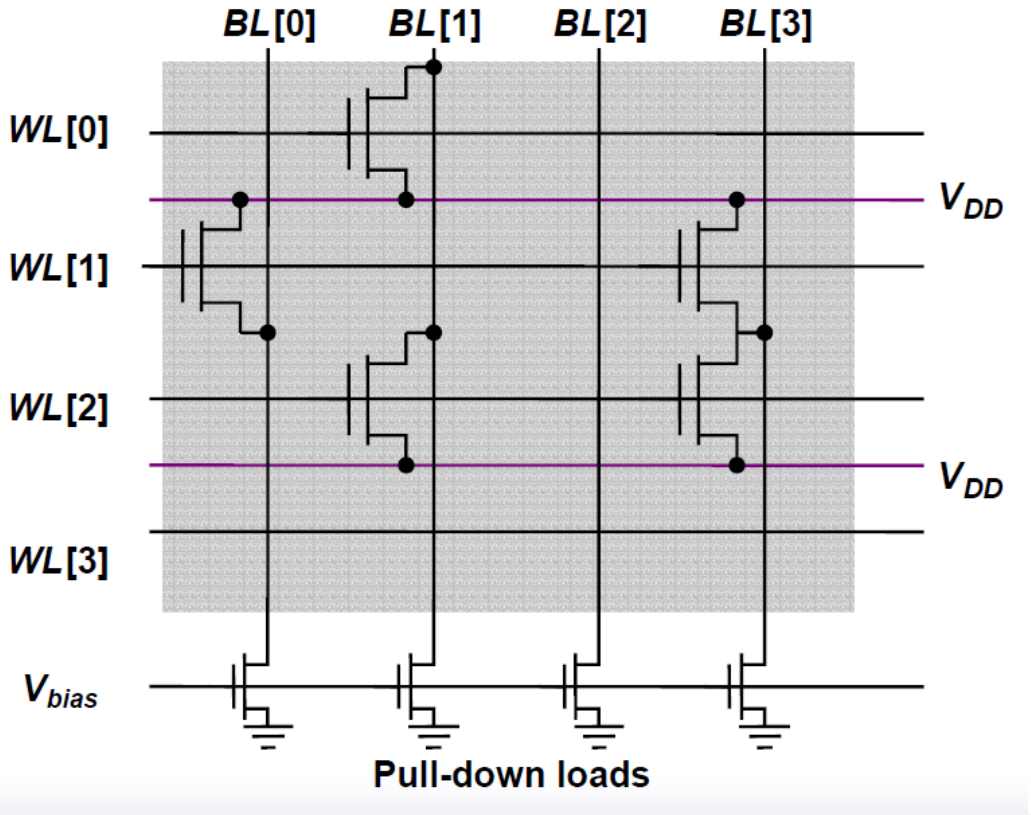


Fig. 12-10

Read-Only Memories: MOS NOR ROM

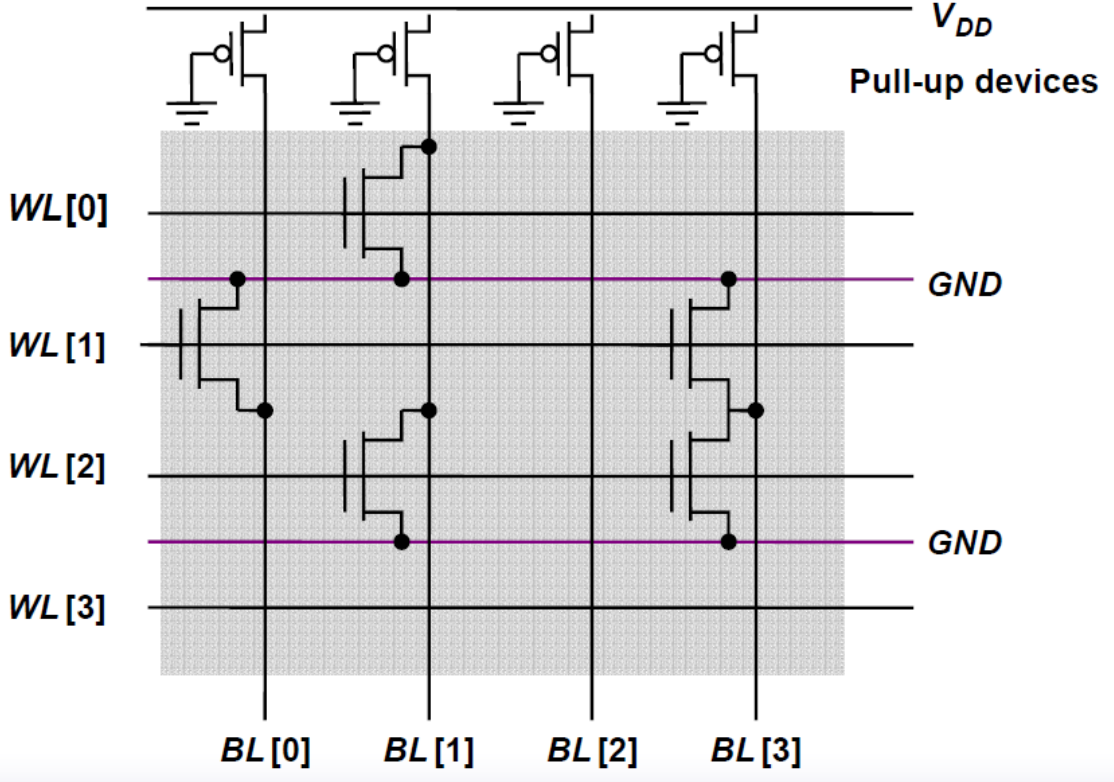
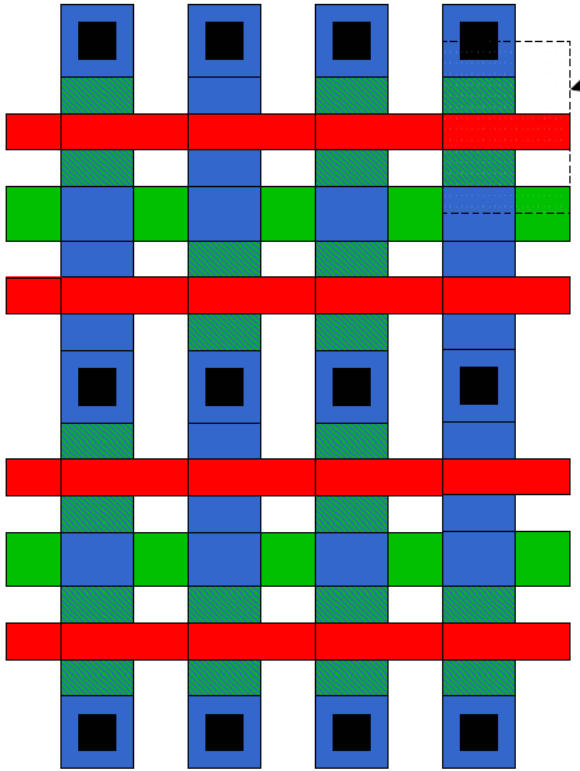


Fig. 12-11

Read-Only Memories: MOS NOR ROM Layout



Cell (9.5λ x 7λ)

Programmming using the Active Layer Only

Transistors in opposite position w.r.t. the previous scheme

- Polysilicon
- Metal1
- Diffusion
- Metal1 on Diffusion

Fig. 12-12a

Read-Only Memories: MOS NOR ROM Layout

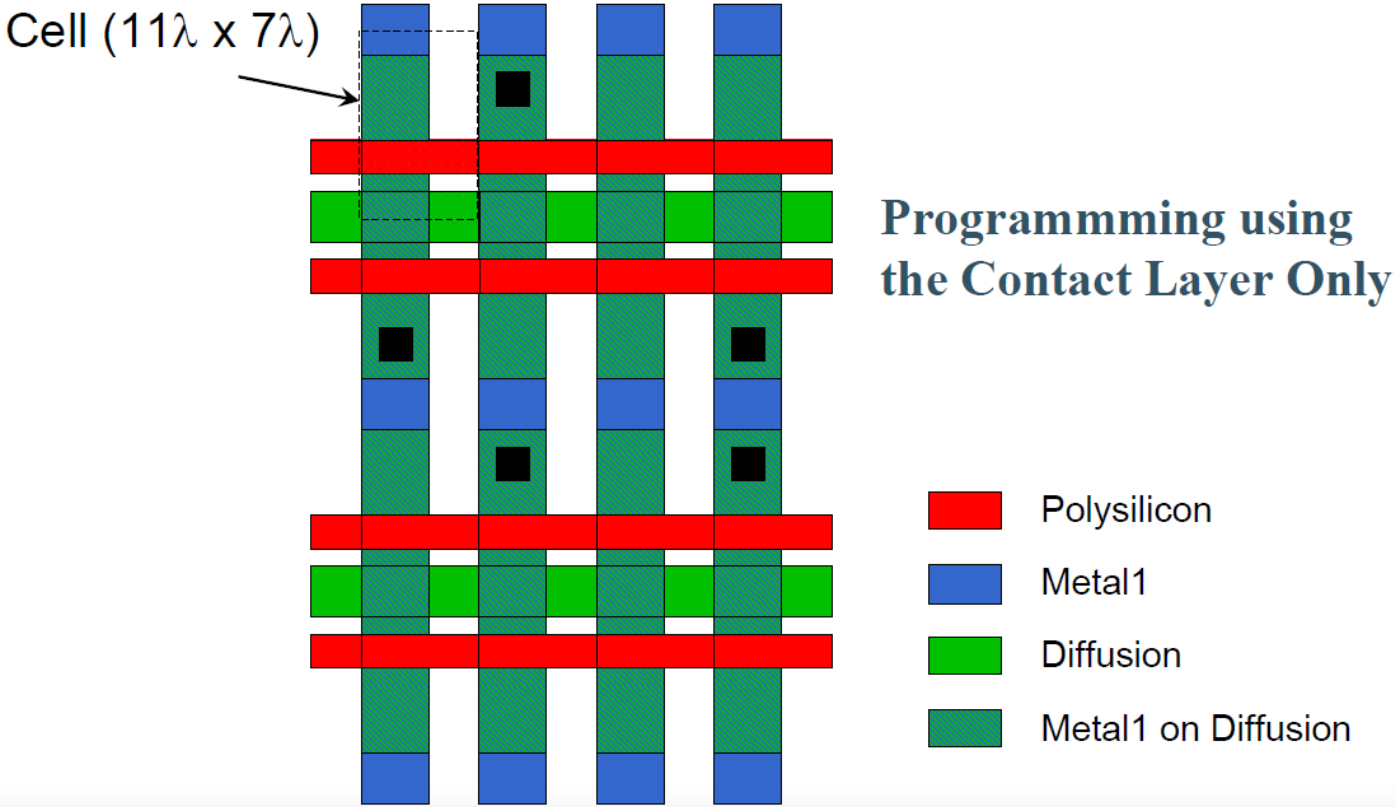


Fig. 12-12b

Read-Only Memories: MOS NAND ROM

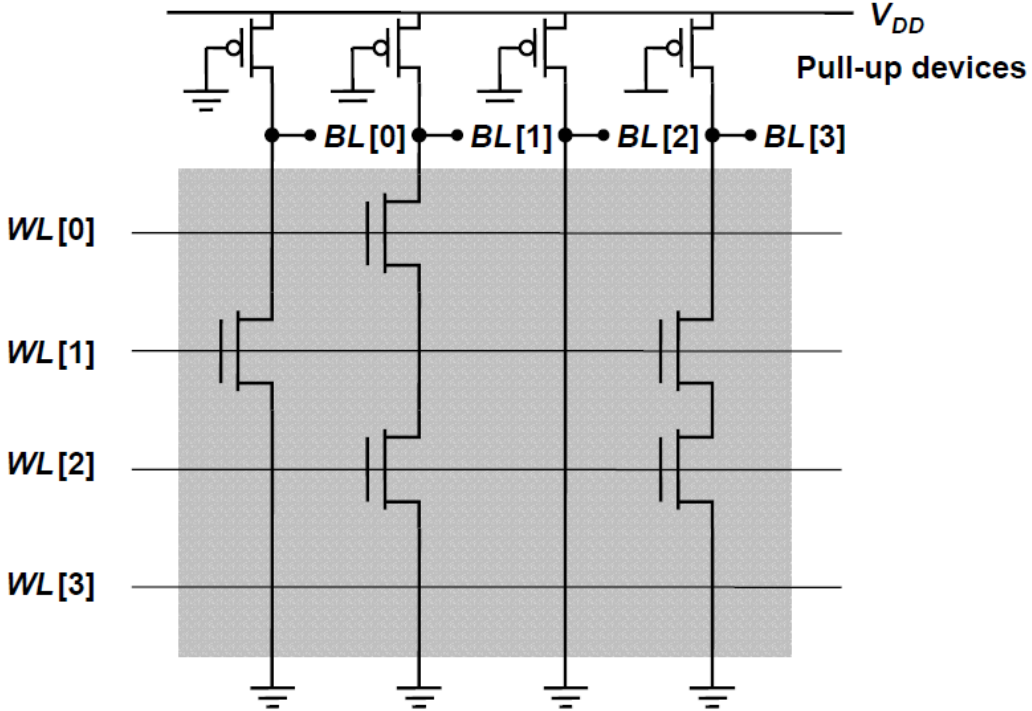
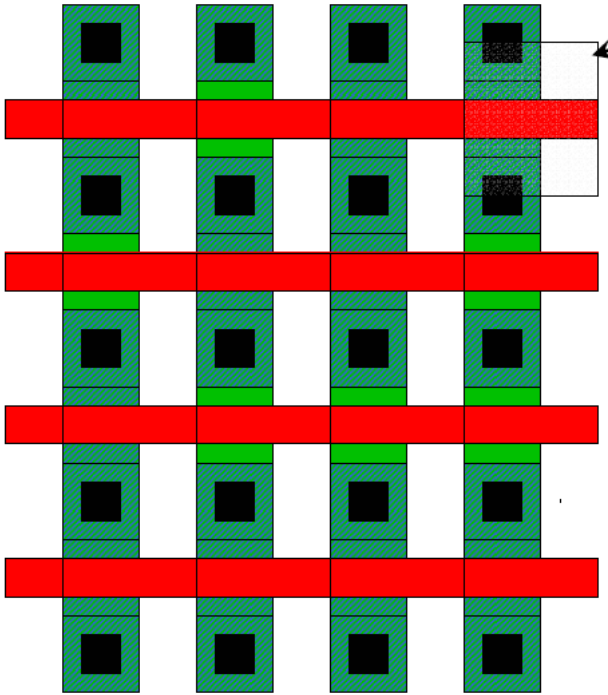


Fig. 12-13

All word lines high by default with exception of selected row

Read-Only Memories: MOS NAND ROM



Programmming using the Metal-1 Layer Only

No contact to VDD or GND necessary;
drastically reduced cell size
Loss in performance compared to NOR ROM

Polysilicon

Diffusion

Metal1 on Diffusion

Fig. 12-14a

Read-Only Memories: MOS NAND ROM

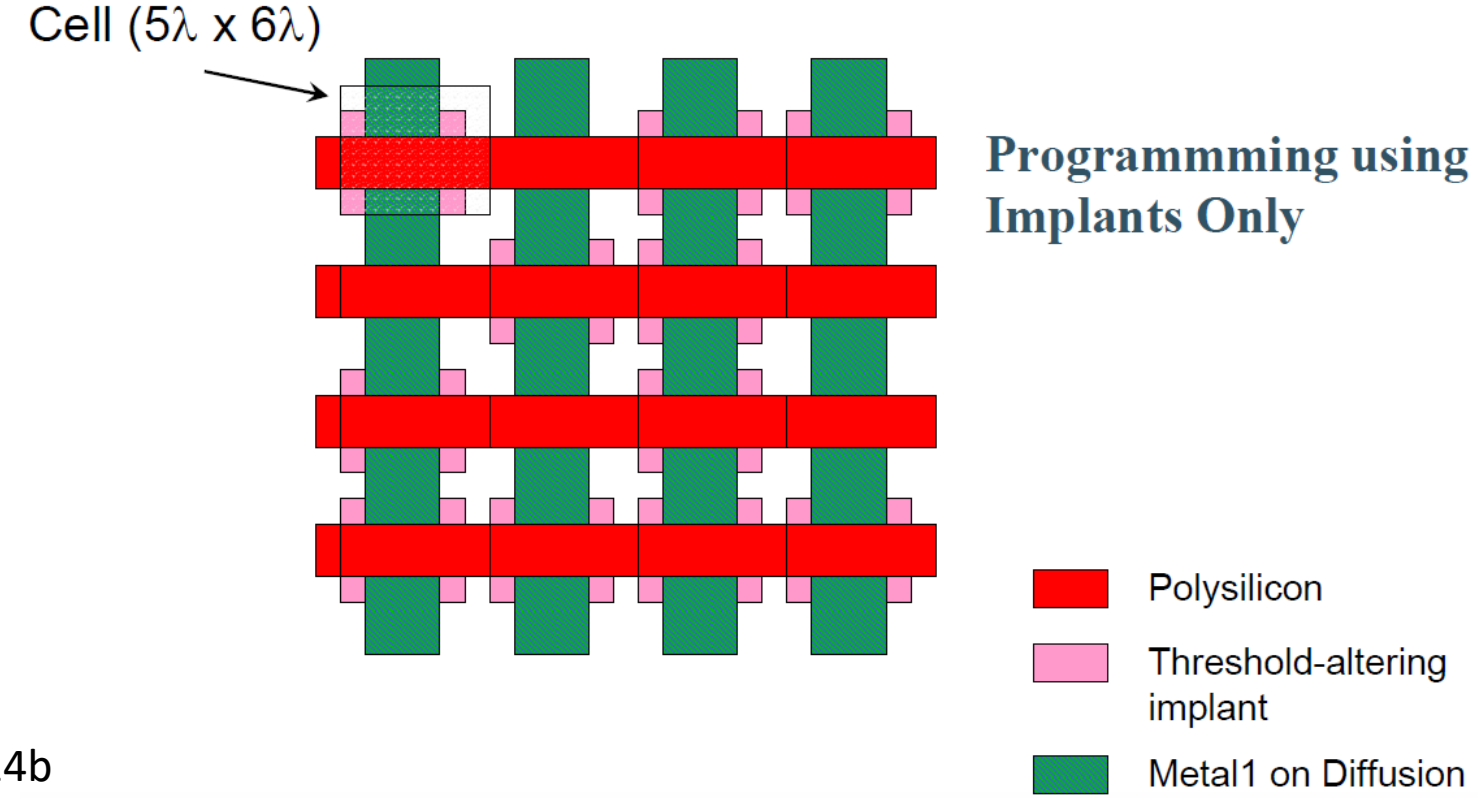


Fig. 12-14b

Read-Only Memories: Precharged Memory Arrays

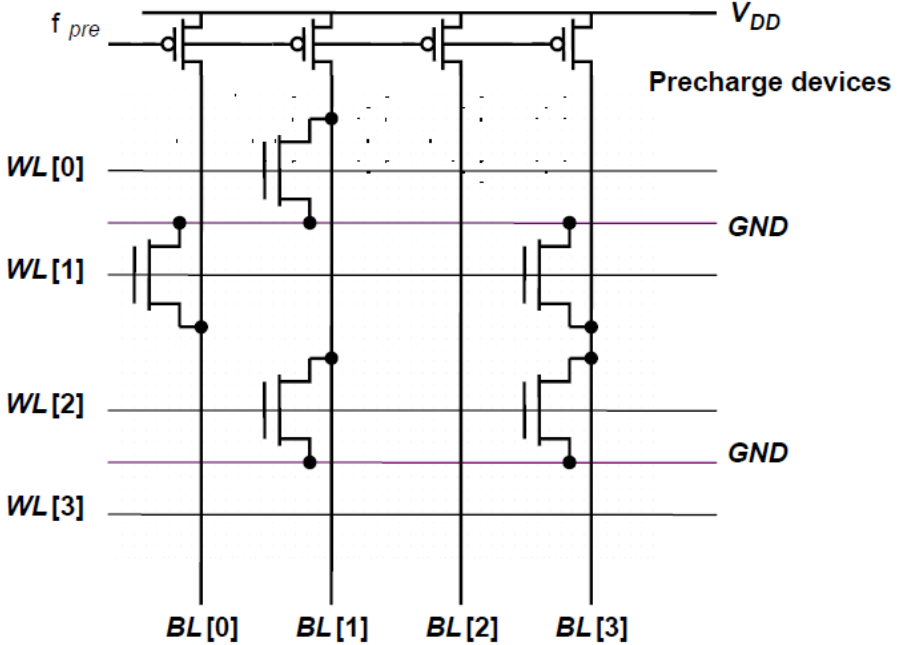


Fig. 12-17

PMOS precharge device can be made as large as necessary, but clock driver becomes harder to design.

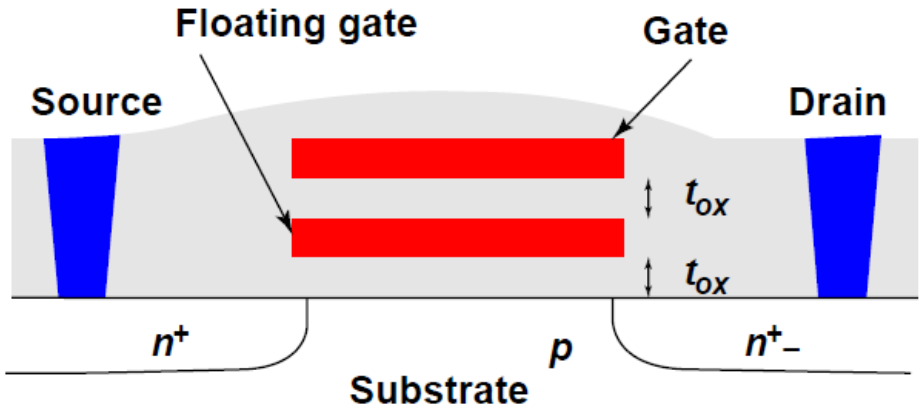
Read-Only Memories: A user perspective

- Classification of ROM memories:
  - Application specific ROMs.
  - Mask-programmable ROMs.
  - Programmable ROMs (PROMs).

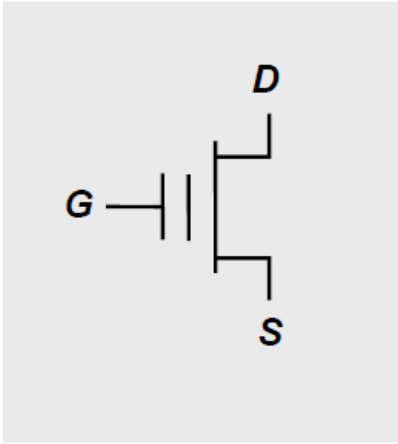
Nonvolatile Read-Write Memories

- The architecture of the NVRW memories is virtually identical to the ROM structure.
- The memory core consists of an array of transistors placed on a word-line/bit line grid. The memory is programmed by selectively disabling or enabling some of those devices. NVRW memories use a modified transistor that permits its threshold to be altered electrically.
- We will see:
  - EPROM (Electrically Programmable ROM)
    - Electrically programmable by the user.
    - Erasable by ultraviolet radiation.
  - EEPROM (Electrically Erasable and Programmable ROM)
    - Every bit can be electrically programmed or erased by the user.
  - FLASH
    - The user can electrically program the single bit, but can only erase the entire memory or large memory banks.

NRWM programmable device

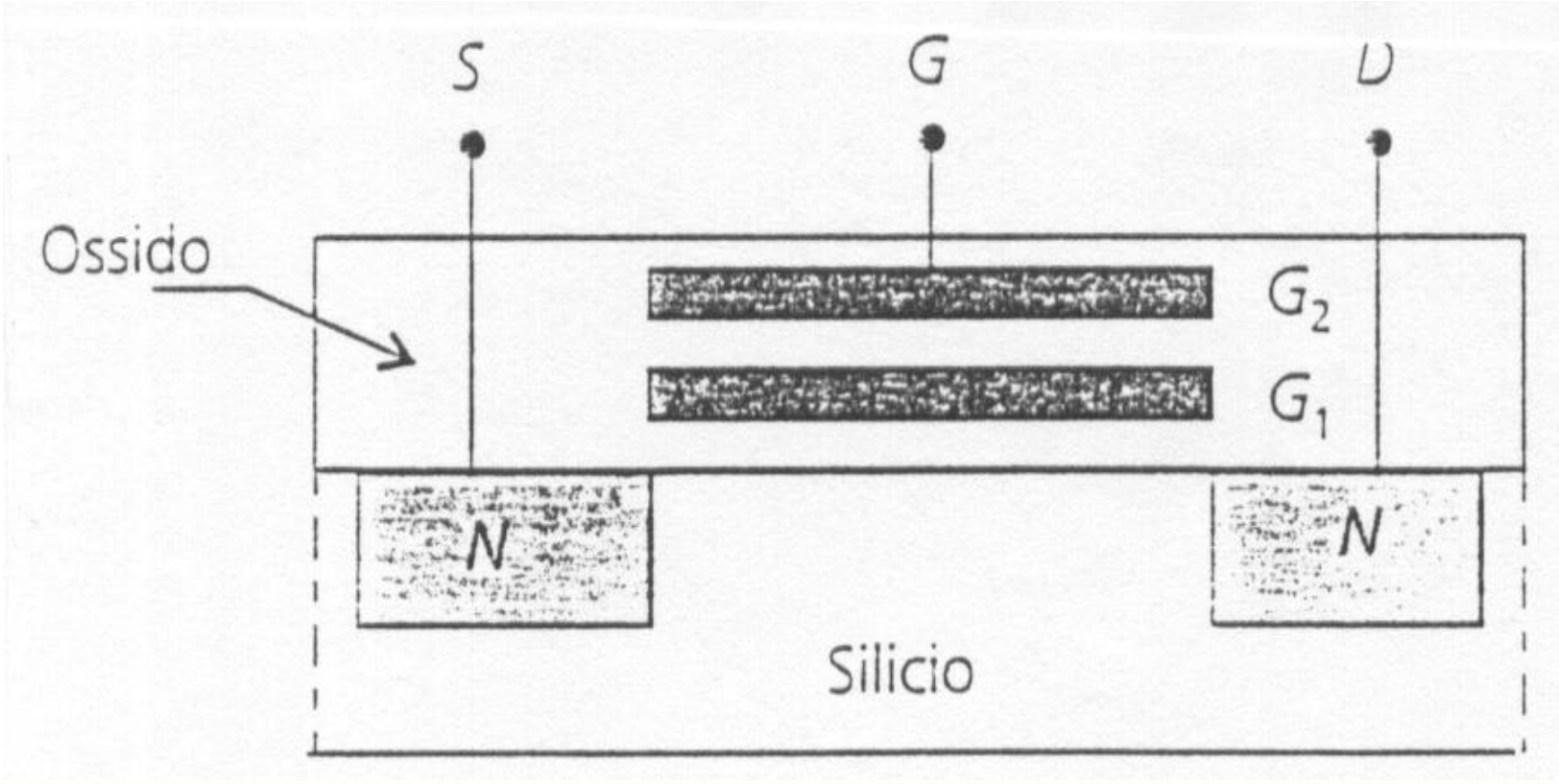


Device cross-section

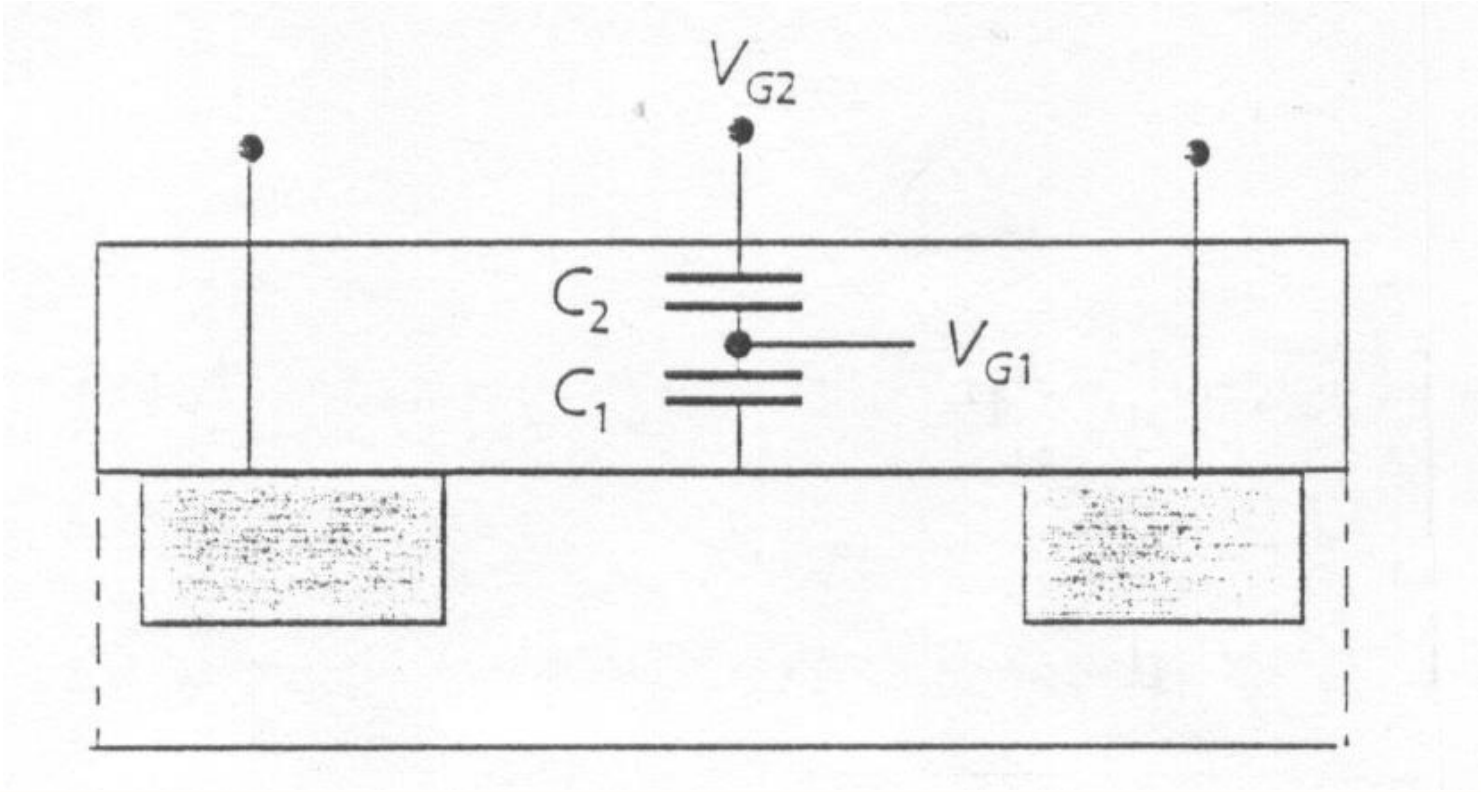


Schematic symbol

NRWM programmable device



NRWM programmable device



NRWM threshold voltage

$$Q_{G1} = 0$$

$$C_1 \, V_{G1} + C_2 \, (V_{G1} - V_{G2}) = 0$$

$$V_{G2} = \left(1 + \frac{C_1}{C_2}\right) V_{G1}$$

$$V_T{}' = \left(1 + \frac{C_1}{C_2}\right) V_T$$

NRWM threshold voltage

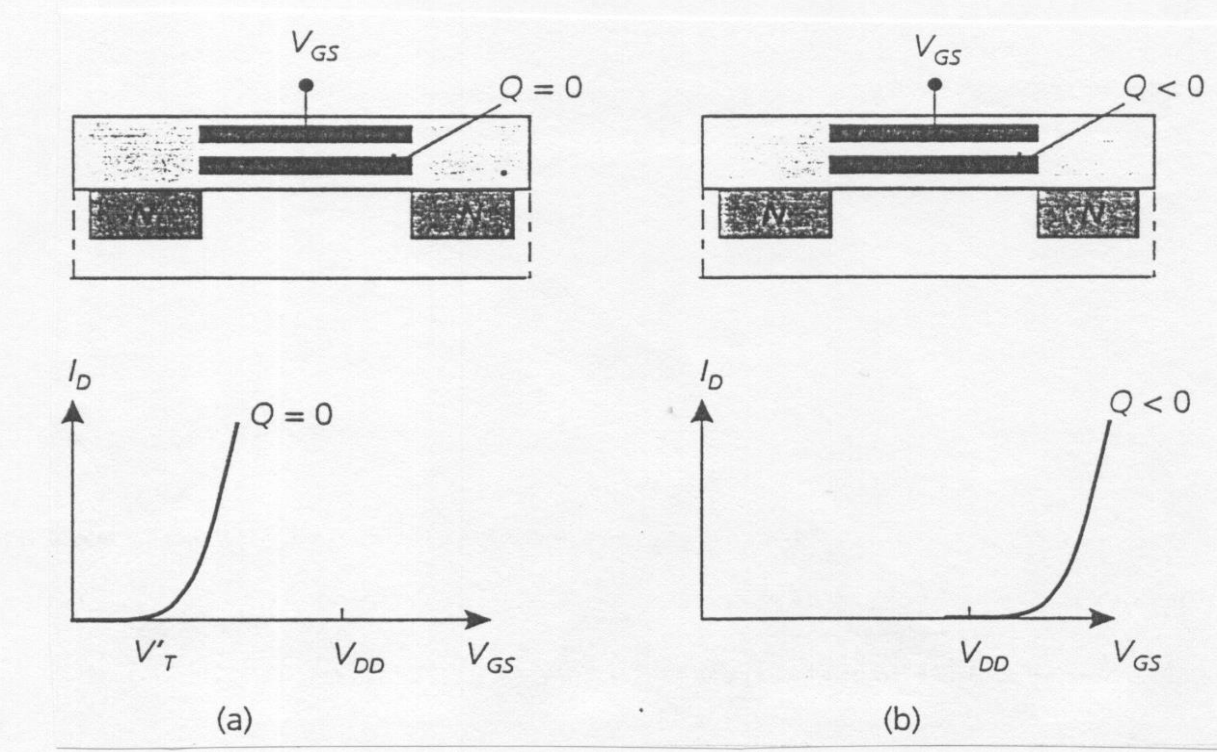
$$Q_{G1} \neq 0$$

$$C_1 V_{G1} + C_2 \left(V_{G1} - V_{G2}\right) = -Q$$

$$V_{G2} = \left(1 + \frac{C_1}{C_2}\right) V_{G1} + \frac{Q}{C_2}$$

$$V_T'' = \left(1 + \frac{C_1}{C_2}\right) V_T + \frac{Q}{C_2}$$

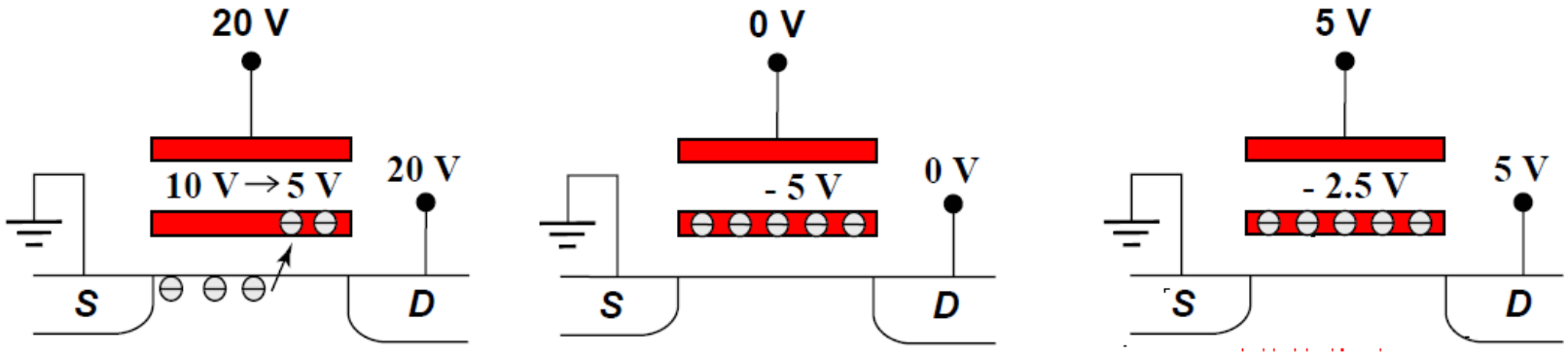
Transfer characteristics



Negative charge injection of the

- The physical mechanisms used to program the dual gate device and take a negative charge on the floating gate are
  - Hot electrons injections
  - Thin oxide tunneling

Hot electron injection



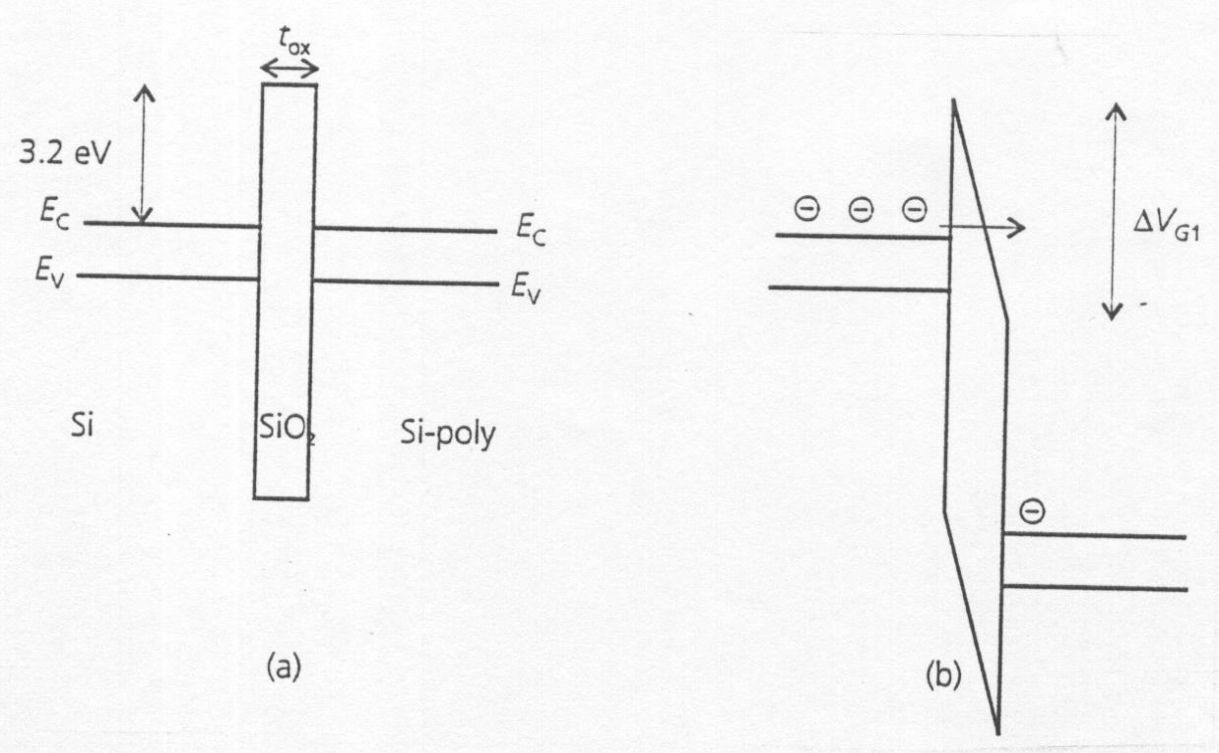
Hot-carrier injection

Removing programming voltage leaves charge trapped

Programming results in higher V_T.

Thin oxide tunnelling

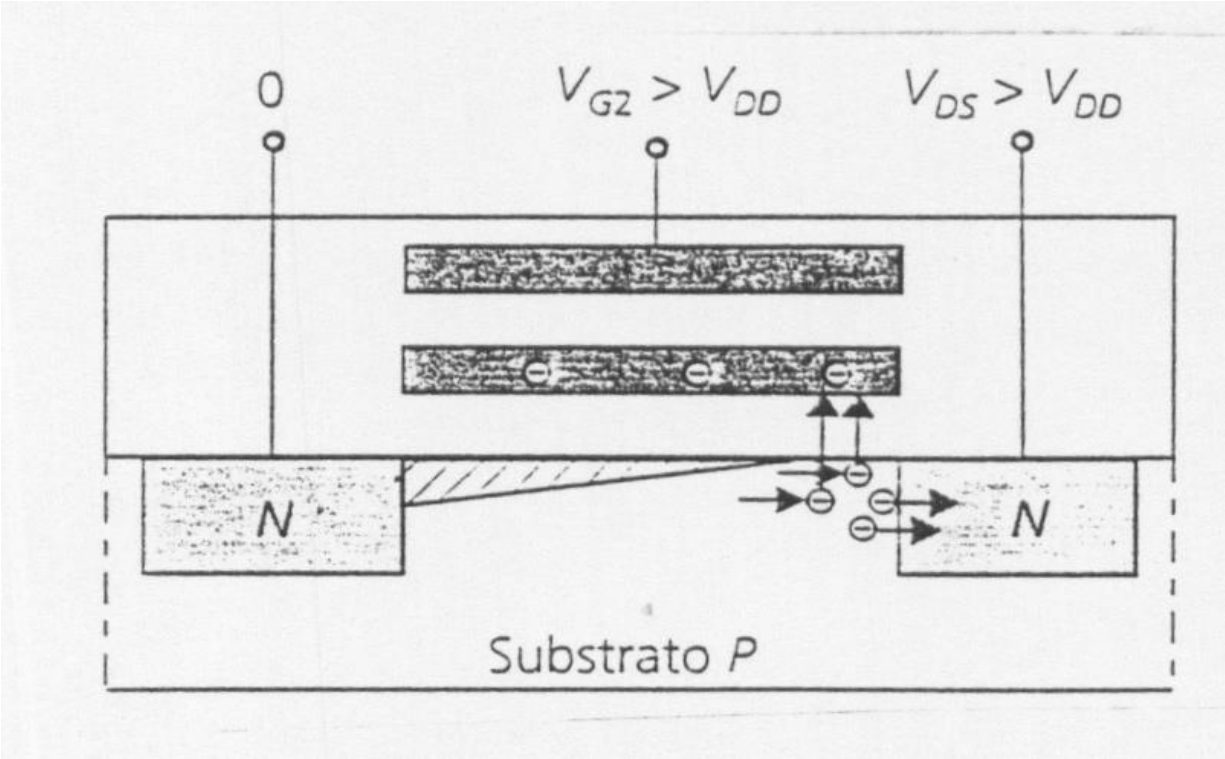
Tunneling Fowler-Nordheim



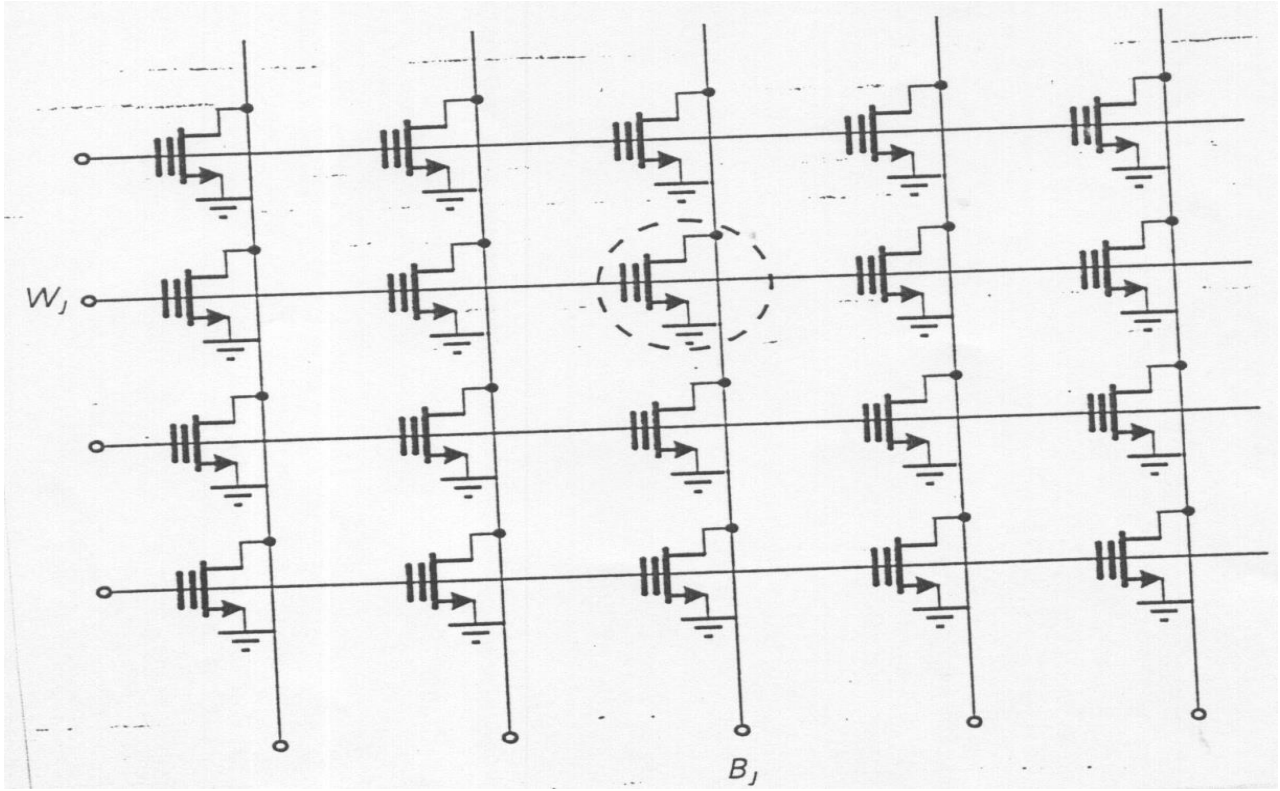
EPROMs

- Programmed using hot electrons injections
- Make use of a Floating Gate Avalanche-injection MOS (FAMOS).
- Erased by ultraviolet radiation exposure for about 20 minutes
- Allow to program the single bits.

EPROMs



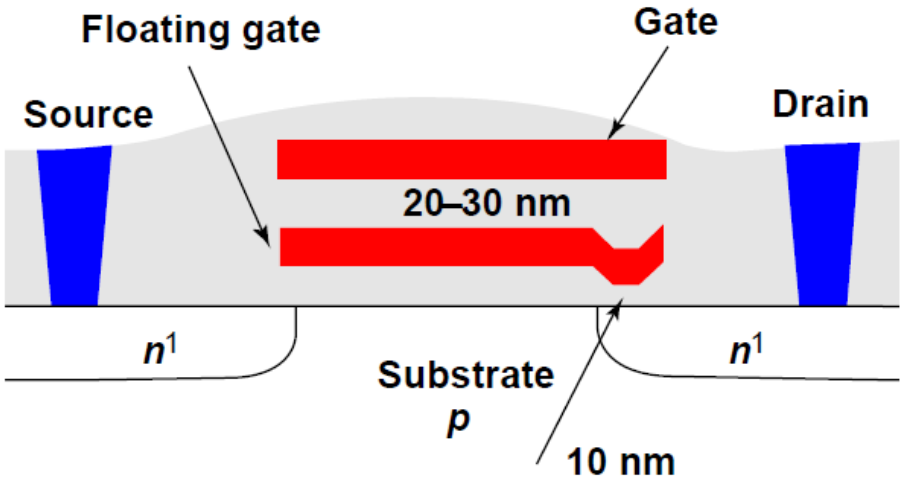
EPROMs



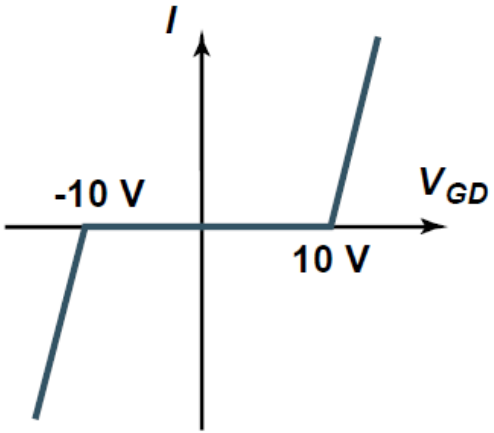
EEPROMs

- Allow to electrically program or erase the single memory bits.
- Use thin oxide tunneling both for programming and erasing.
- Make use of a FLOTOX (floating gate tunneling oxide) transistor.
- Necessitate of an access transistor that double the area of the memory cell.

EEPROMs

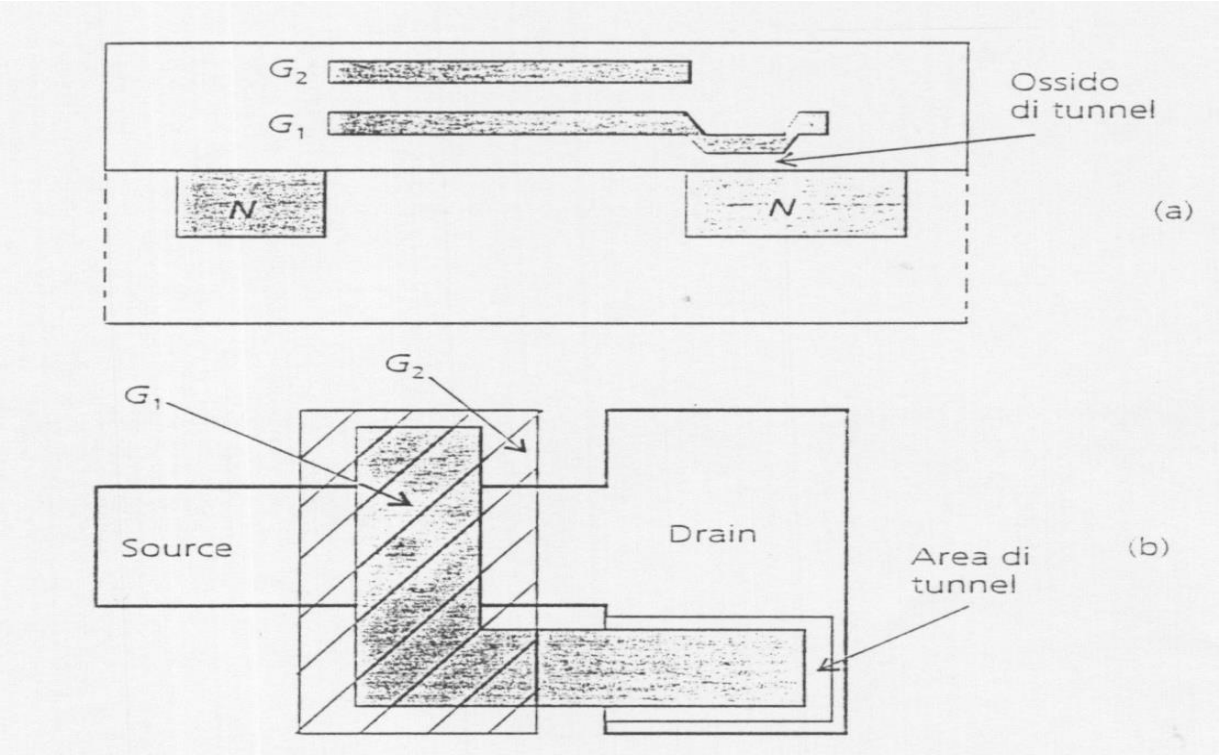


FLOTOX transistor

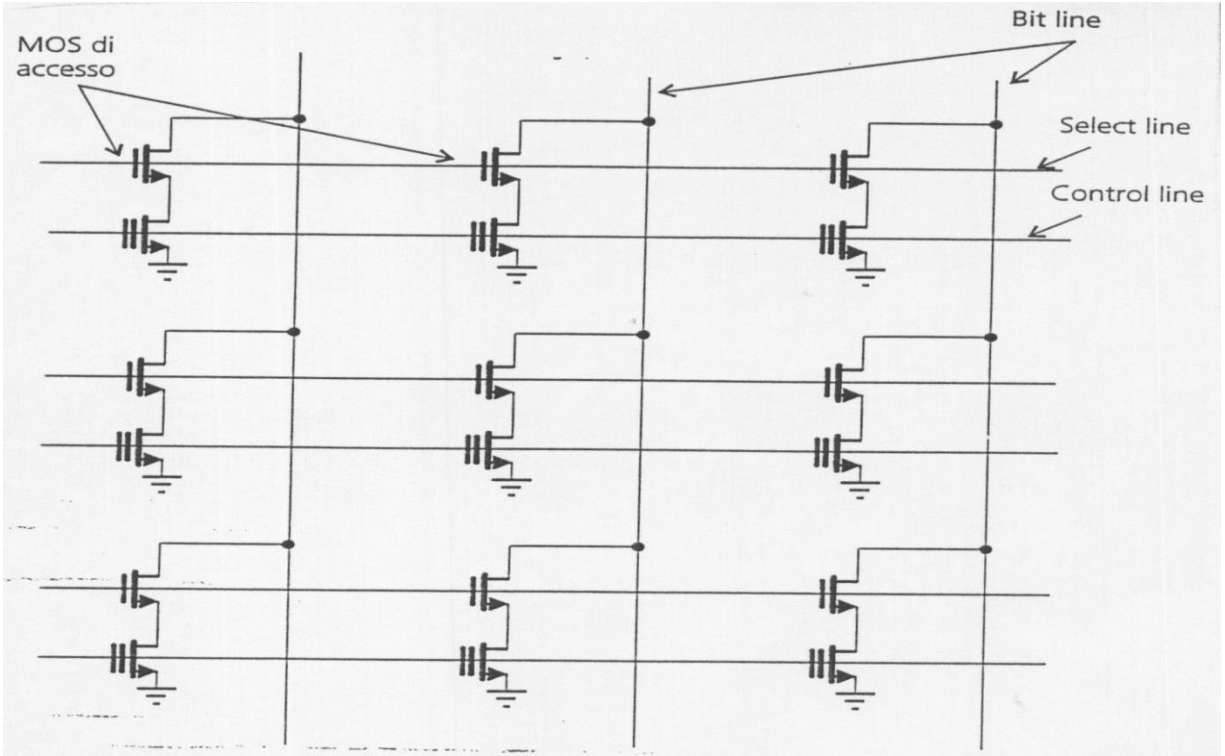


Fowler-Nordheim
I-V characteristic

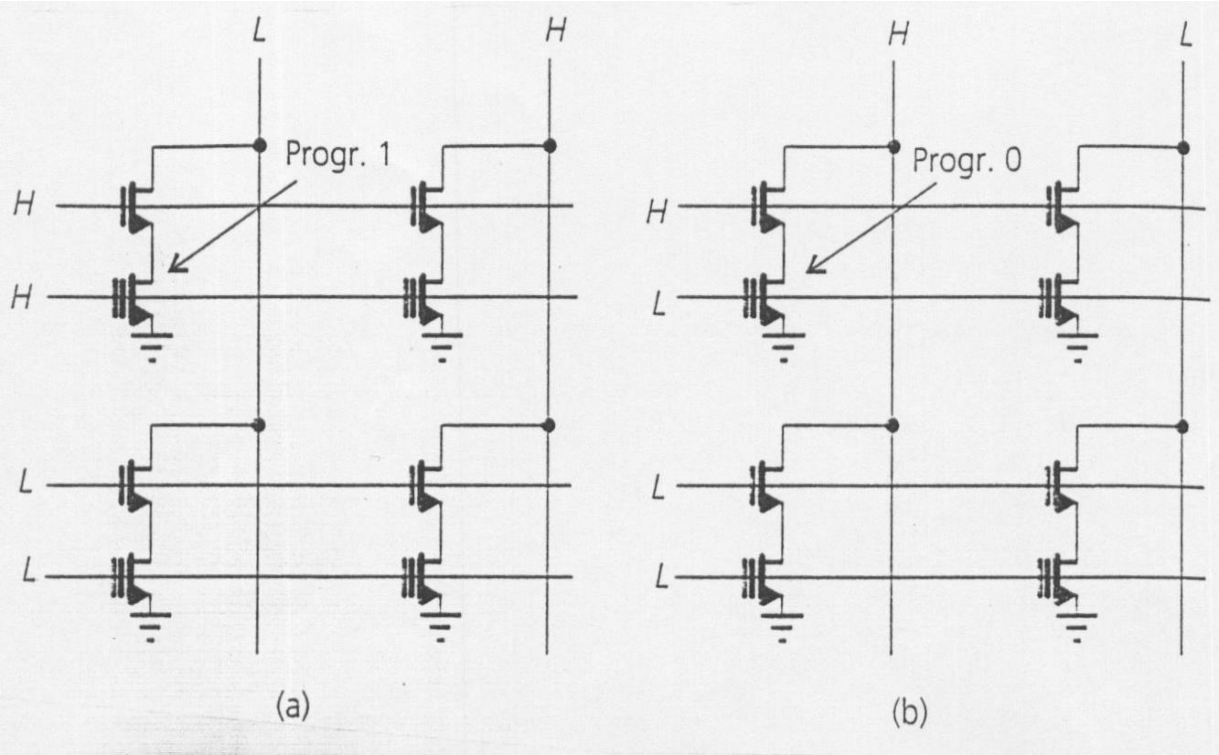
EEPROMs



EEPROMs



EEPROMs

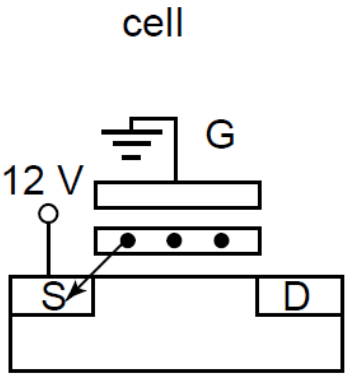


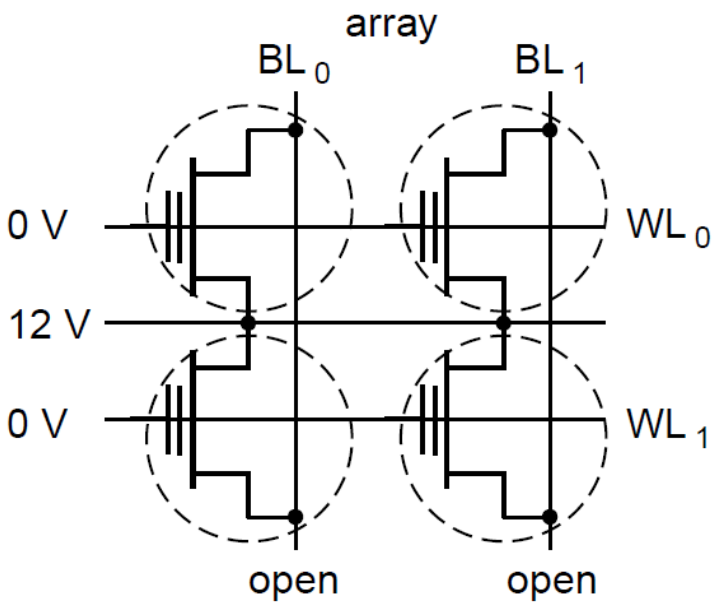
FLASH memories

- Allow to electrically program the single memory bits, while the entire memory or large memory blocks are erased at the same time.
- Many Flash memories use hot electrons injection for programming and tunneling for erasing.

FLASH memories

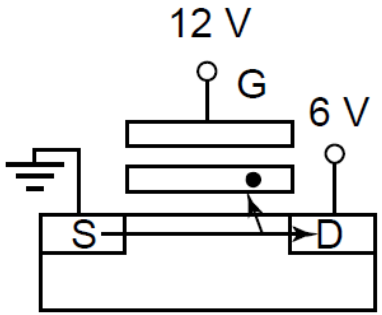
- NOR Erase:

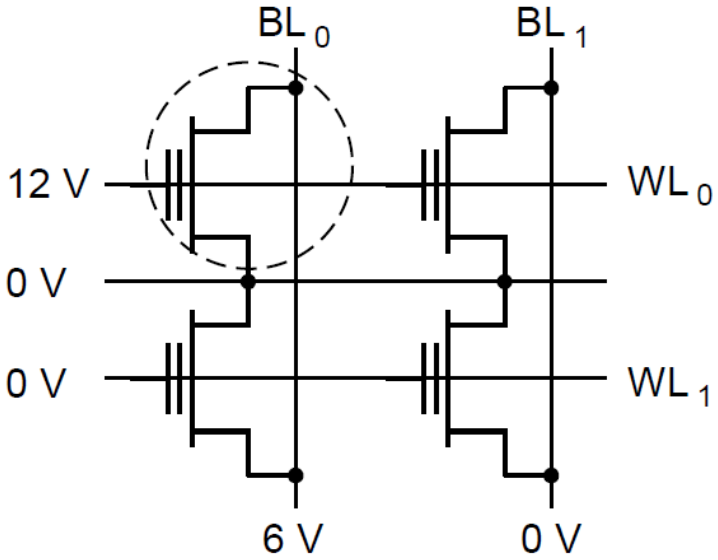




FLASH memories

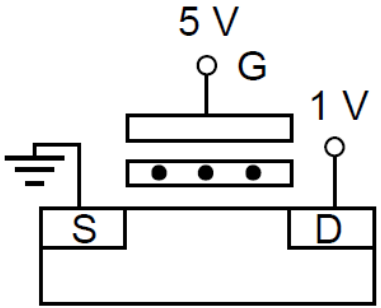
- NOR Write:

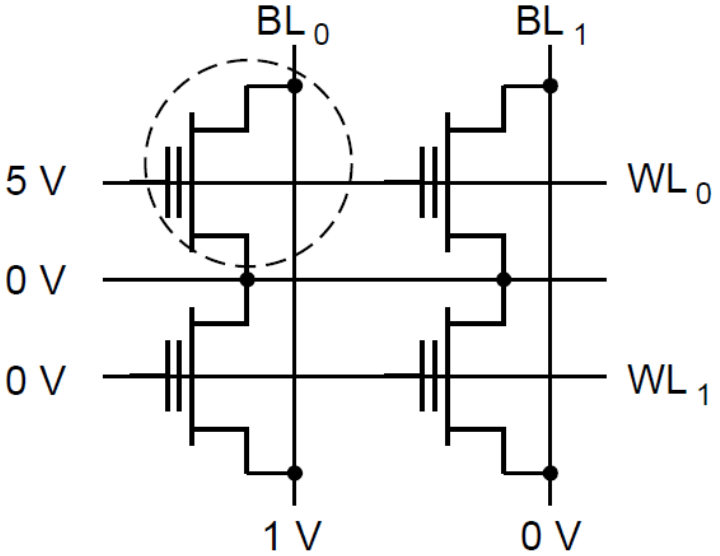




FLASH memories

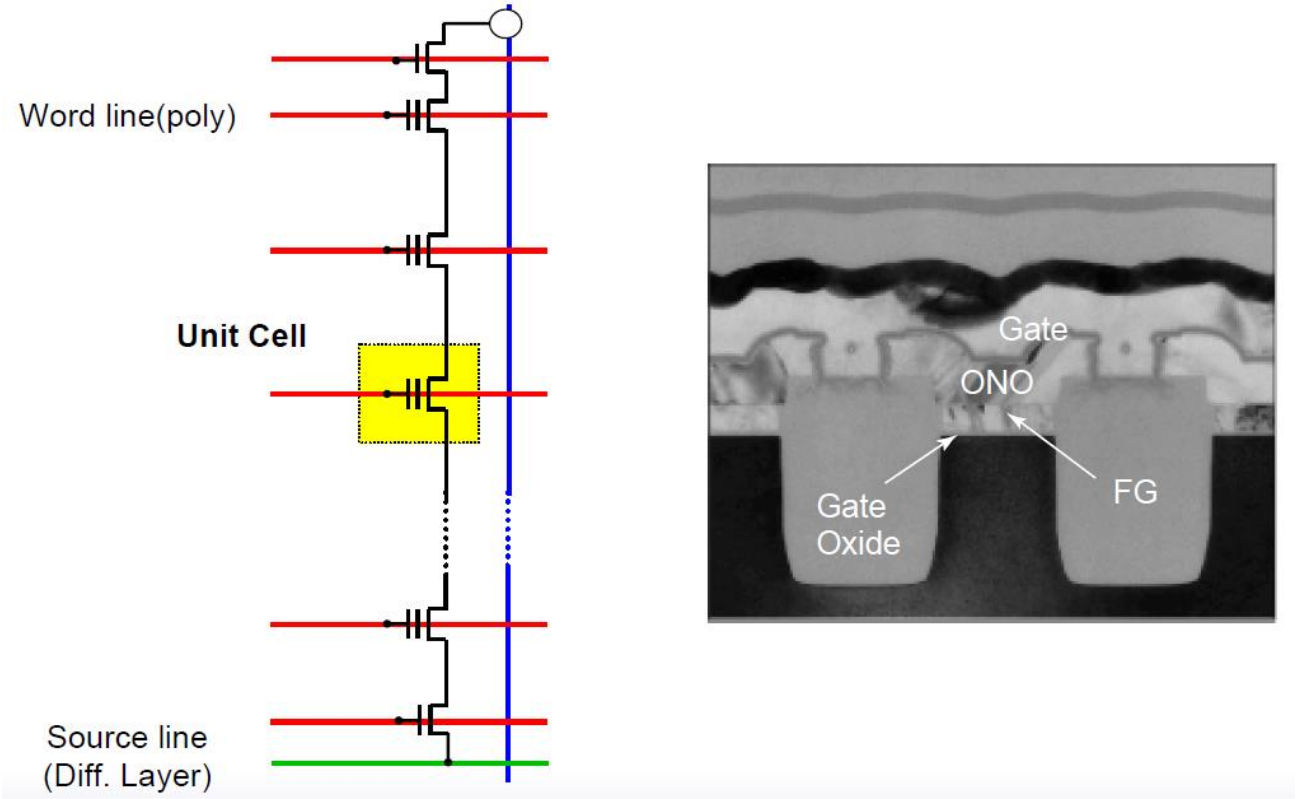
- NOR Read:





FLASH memories

- NAND Flash:



NVRM memories comparison

Table 12-1 Comparison between nonvolatile memories ([Itoh01]).
V_{DD} = 3.3 or 5 V; V_{PP} = 12 or 12.5 V.

	Cell—Nr. of Transistors	Cell Area (ratio wrt EPROM)	Mechanism		External Power Supply		Program/Erase Cycles
			Erase	Write	Write	Read	
MASK ROM	1 T (NAND)	0.35–5	—	—	—	V_{DD}	0
EPROM	1 T	1	UV Exposure	Hot electrons	V_{PP}	V_{DD}	~100
EEPROM	2 T	3–5	FN Tunneling	FN Tunneling	V_{PP} (int)	V_{DD}	10^4–10^5
Flash Memory	1 T	1–2	FN Tunneling	Hot electrons	V_{PP}	V_{DD}	10^4–10^5
			FN Tunneling	FN Tunneling	V_{PP} (int)	V_{DD}	10^4–10^5

Read—Write Memories (RAM)

- Providing a memory cell with roughly equal read and write performance requires a more complex cell structure.
- While the contents of the ROM and NVRWM memories are ingrained in the cell topology or programmed into the device characteristics, storage in RAM memories is based on either *positive feedback* or *capacitive charge.*
- They are labeled as either SRAMs or DRAMs, depending on the storage concept used.

Static Random-Access Memory (SRAM)

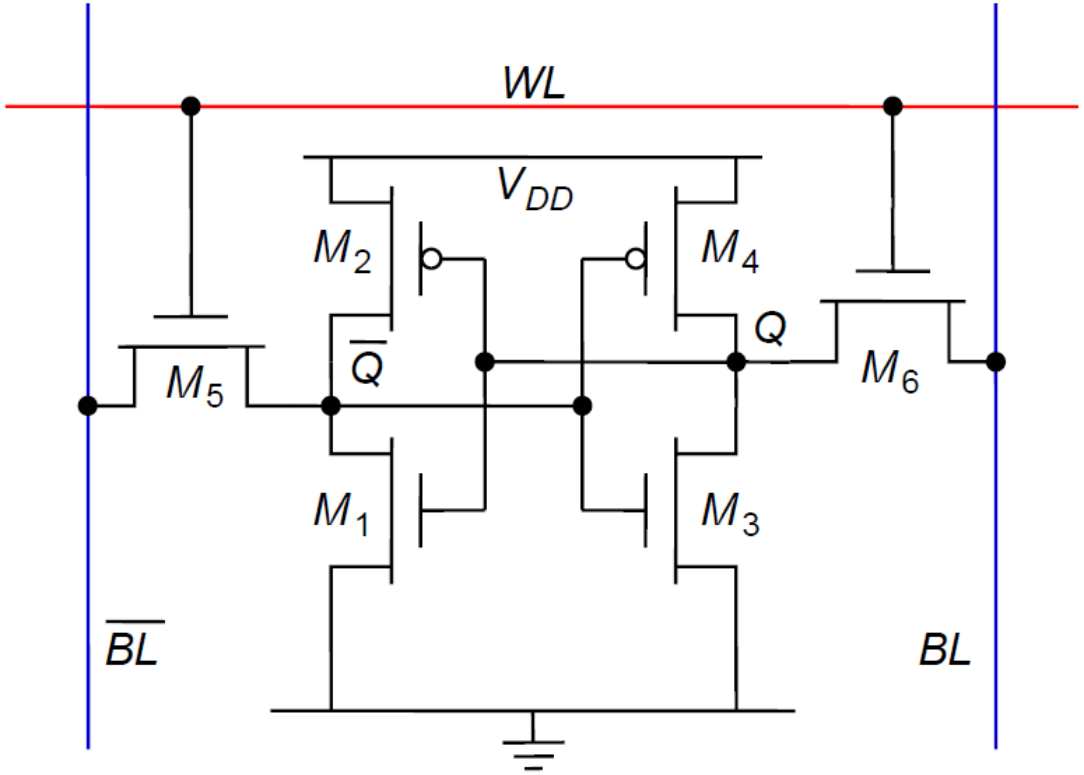
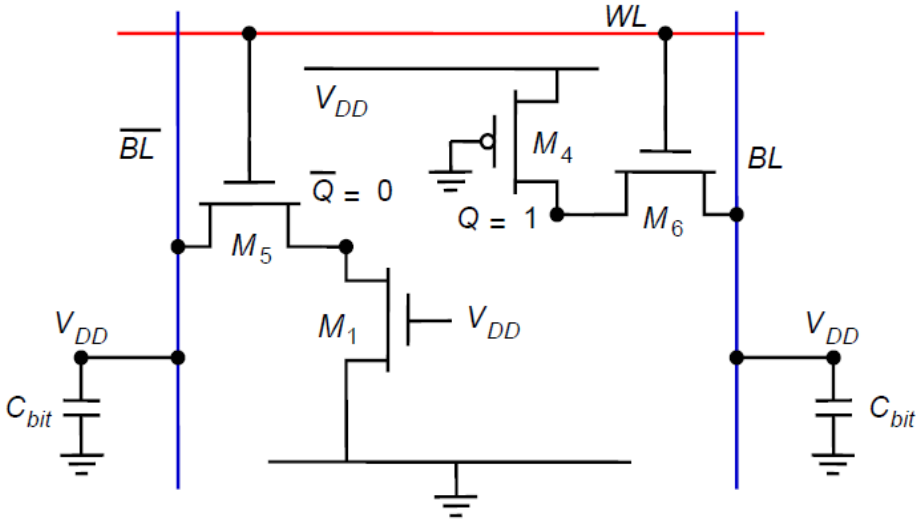


Fig. 12.27

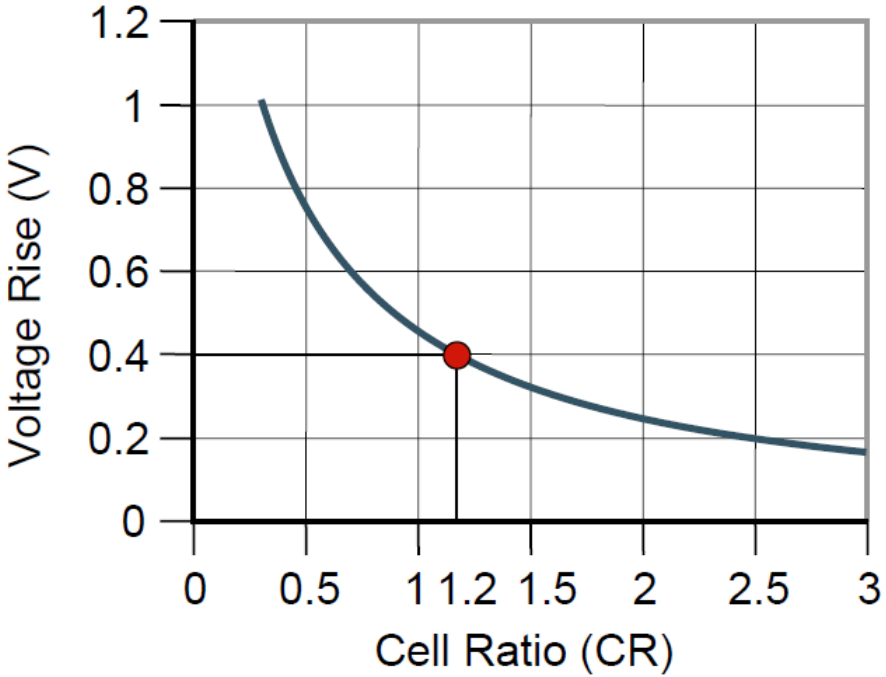
SRAM: Read



$$k_{n,M5}\left((V_{DD}-\Delta V-V_{Tn})V_{DSATn}-\frac{V_{DSATn}^2}{2}\right) = k_{n,M1}\left((V_{DD}-V_{Tn})\Delta V-\frac{\Delta V^2}{2}\right)$$

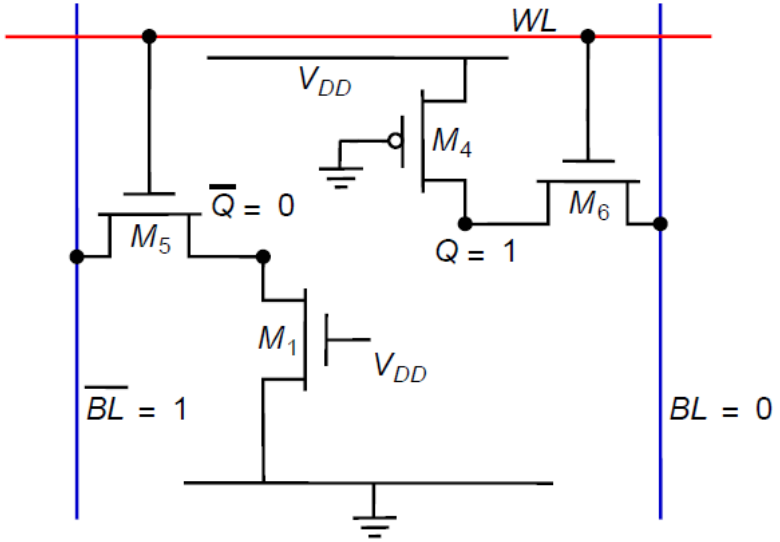
$$\Delta V = \frac{V_{DSATn}+CR(V_{DD}-V_{Tn})-\sqrt{V_{DSATn}^2(1+CR)+CR^2(V_{DD}-V_{Tn})^2}}{CR}$$

SRAM: Read



$$CR = \frac{W_1/L_1}{W_5/L_5}$$

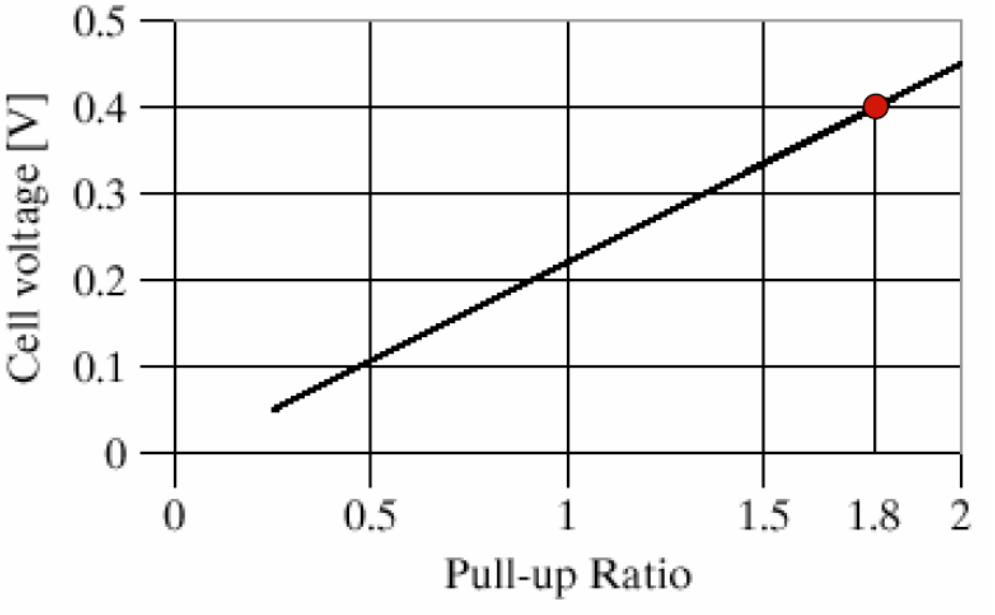
SRAM: Write



$$k_{n,M6}\left((V_{DD}-V_{Tn})V_Q-\frac{V_Q^2}{2}\right)=k_{p,M4}\left((V_{DD}-|V_{Tp}|)V_{DSATp}-\frac{V_{DSATp}^2}{2}\right)$$

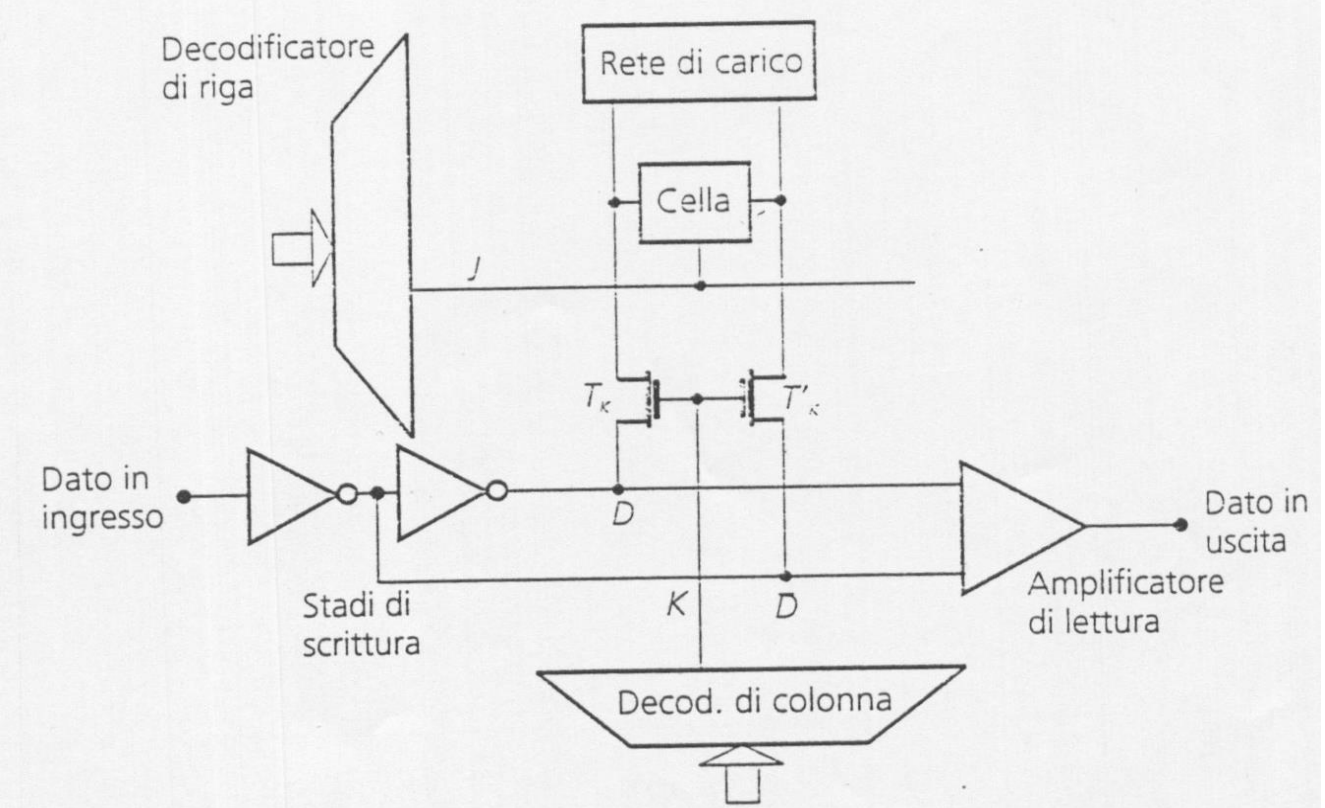
$$V_Q=V_{DD}-V_{Tn}-\sqrt{(V_{DD}-V_{Tn})^2-2\frac{\mu_p}{\mu_n}PR\left((V_{DD}-|V_{Tp}|)V_{DSATp}-\frac{V_{DSATp}^2}{2}\right)},$$

SRAM: Write

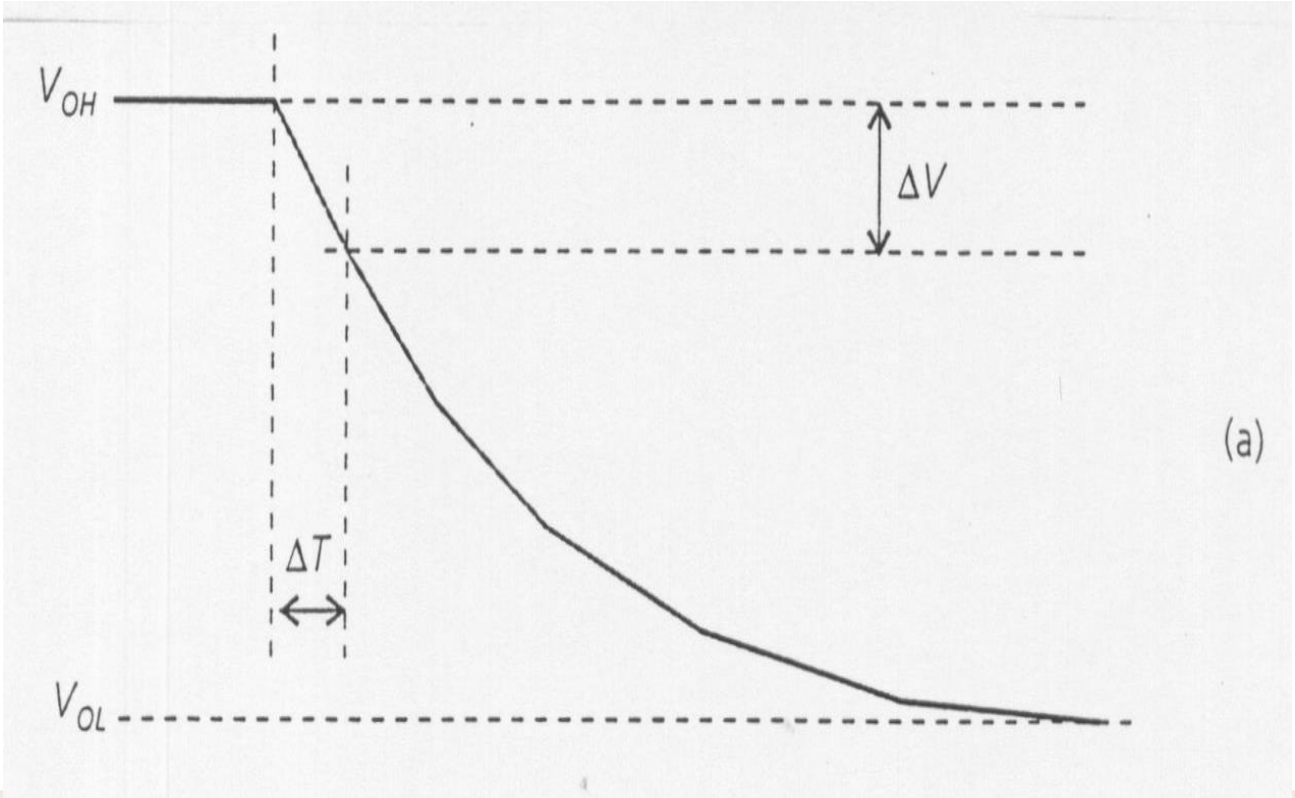


$$PR = \frac{W_4/L_4}{W_6/L_6}.$$

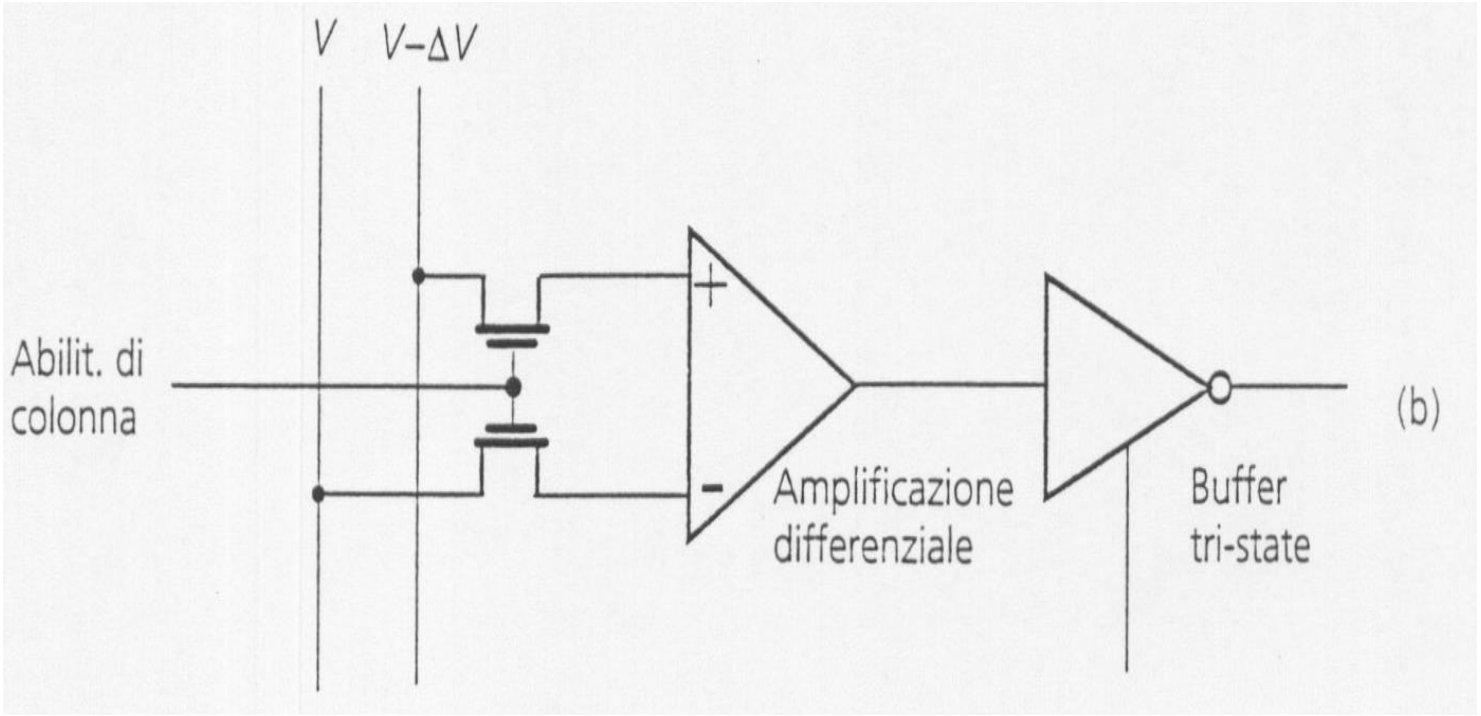
SRAM: Sense Amplifier



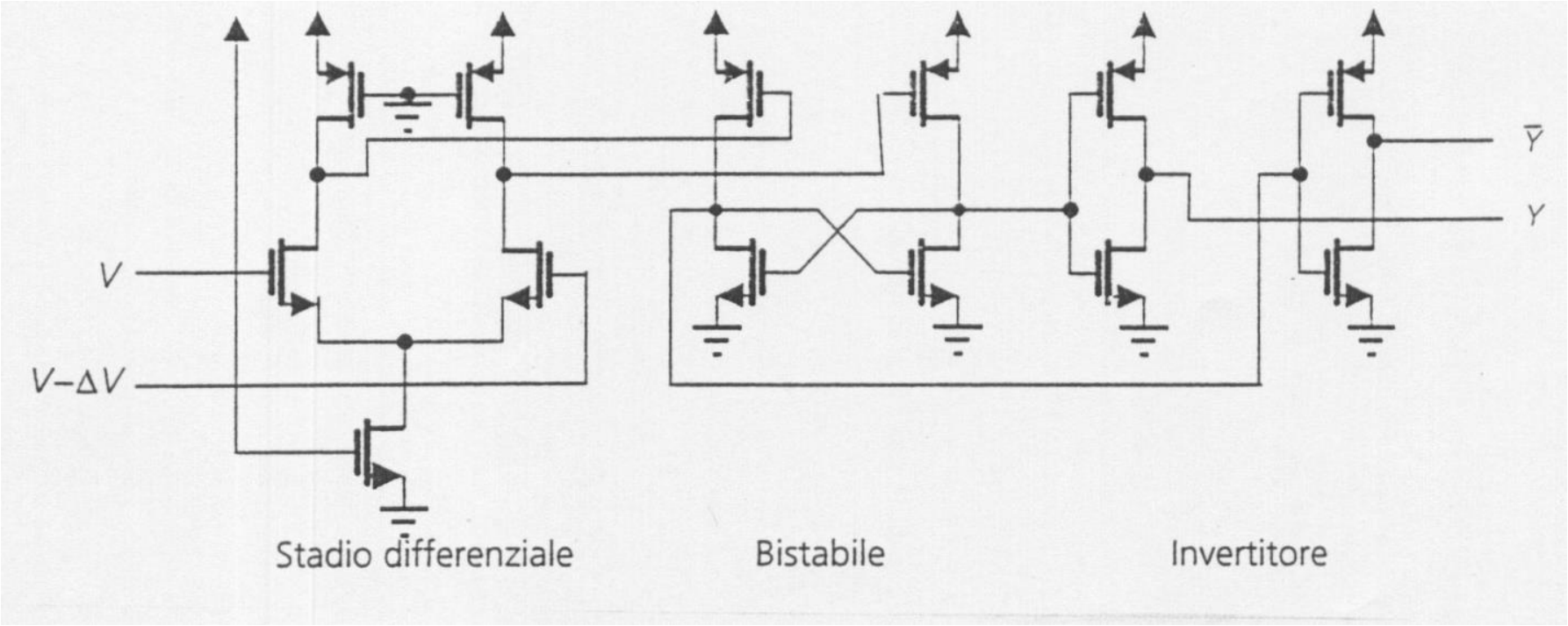
SRAM: Sense Amplifier



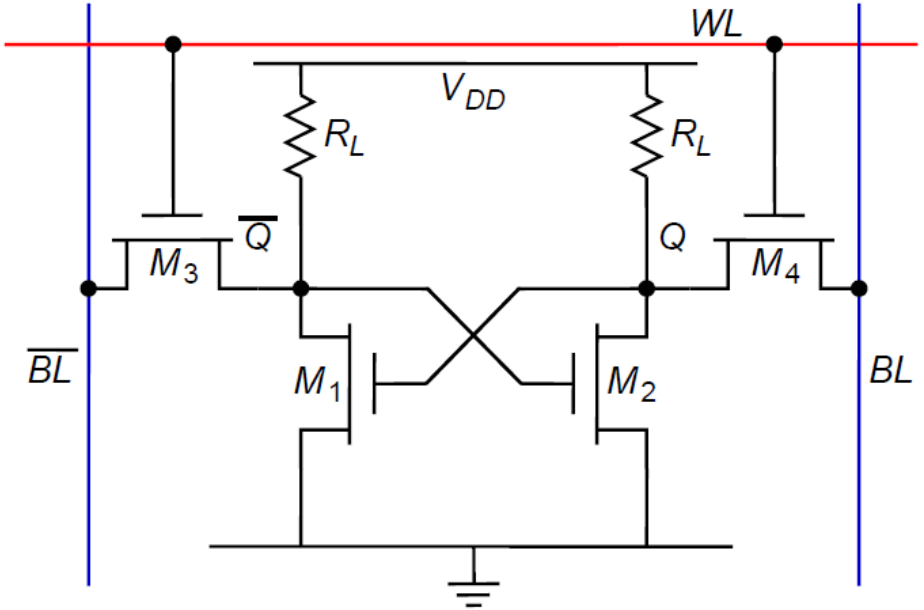
SRAM: Sense Amplifier



SRAM: Sense Amplifier



Resistance-load SRAM cell



Static power dissipation -- Want R_L large
Bit lines precharged to V_{DD} to address t_p problem

Resistance-load SRAM cell

- Keeping the static power dissipation per cell as low as possible is a prime design priority in SRAM cells.
- Consider a 1-Mbit SRAM memory operating at 2.5V and using a 10 kΩ resistor as the inverter load.
- With each cell sinking 0.25 mA in static current, a total standby dissipation of 250 W can be recorded!
- Therefore, the only obvious choice is to make the load resistance as large as possible.
- A very large, yet compact, resistor can be manufactured by using an undoped polysilicon, which has a **sheet resistance of several T Ω** /sq (Tera = 10^{12}).

SRAM cells comparison

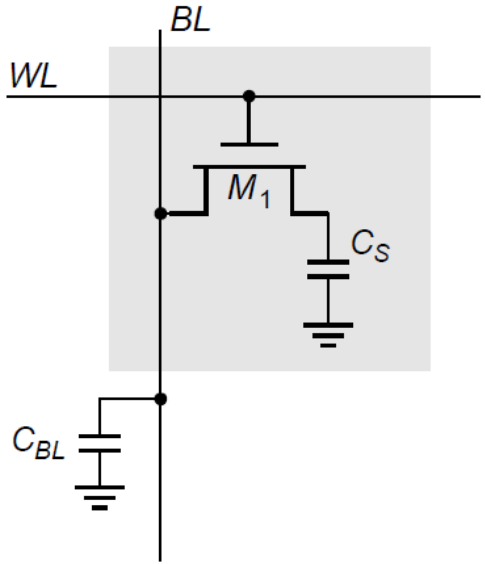
Table 12-2 Comparison of CMOS SRAM cells used in 1-Mbit memory (from [Takada91])

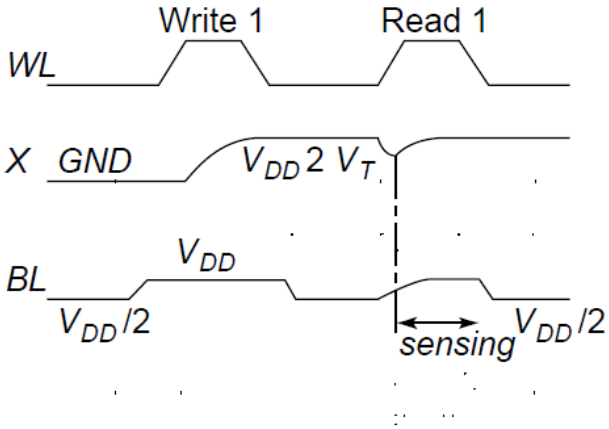
	Complementary CMOS	Resistive Load	TFT Cell
Number of transistors	6	4	4 (+2 TFT)
Cell size	58.2 μm^2 (0.7-μm rule)	40.8 μm^2 (0.7-μm rule)	41.1 μm^2 (0.8-μm rule)
Standby current (per cell)	10^{-15} A	10^{-12} A	10^{-13} A

DRAM

- DRAMs are based on charge storage in a capacitor.
- DRAMs need a *refresh* phase to periodically restore the charge, which cannot be preserved indefinitively.
- They drastically reduce the number of devices needed to implement the memory cell and the number of interconnection lines.

DRAM: 1T cell



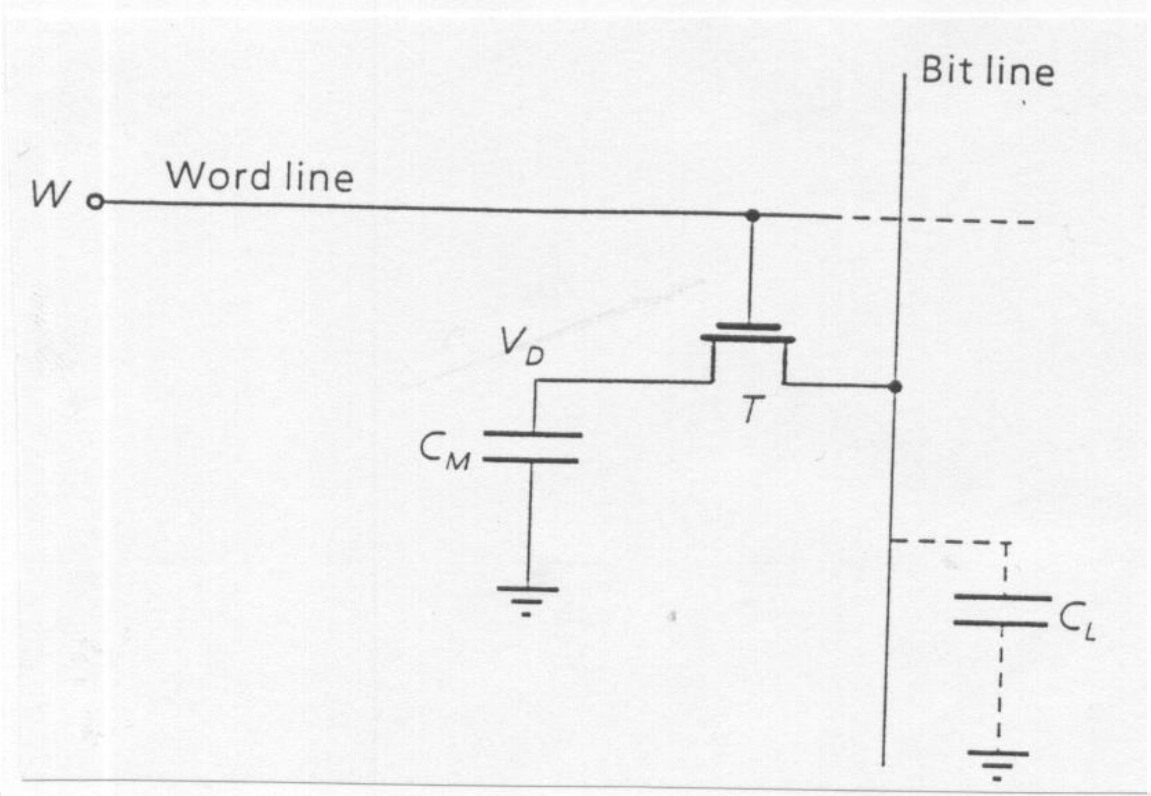


Write: C_S is charged or discharged by asserting WL and BL.
Read: Charge redistribution takes places between bit line and storage capacitance

$$\Delta V = V_{BL} - V_{PRE} = (V_{BIT} - V_{PRE}) \frac{C_S}{C_S + C_{BL}}$$

Voltage swing is small; typically around 250 mV.

DRAM: 1T cell



DRAM: 1T cell

$$Q = C_L V_R + C_M V_D = V_R'(C_L + C_M)$$

$$V_R' = \frac{C_L}{C_L + C_M} V_R + \frac{C_M}{C_L + C_M} V_D$$

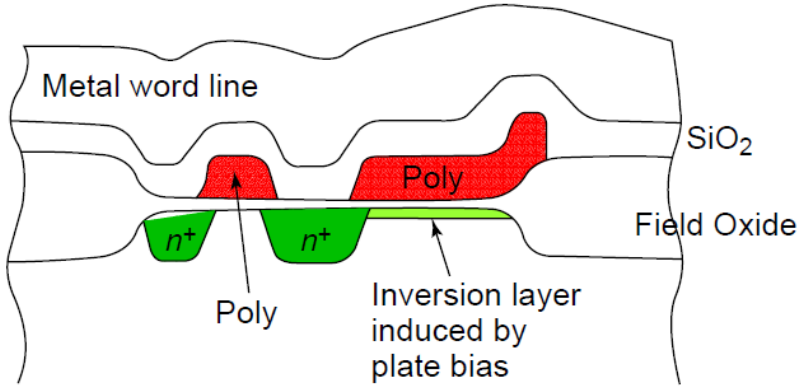
$$\Delta V_R = V_R - V_R' = \frac{C_M}{C_L + C_M}(V_R - V_D)$$

Es:

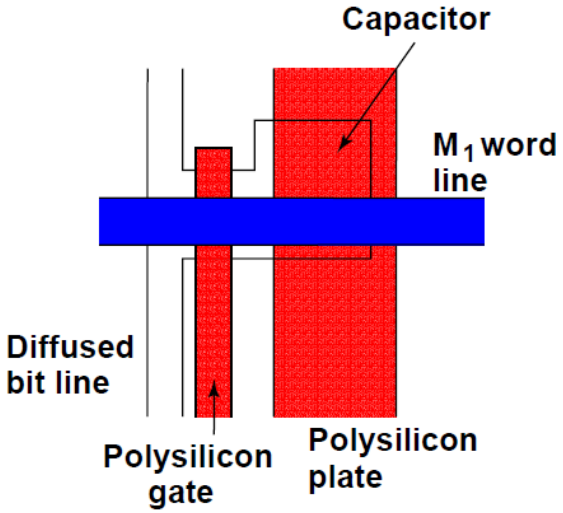
$$C_M \ll C_L \qquad C_M = 25\ fF \qquad C_L = 1\ pF$$

$$V_R = 5V \quad \Rightarrow \quad \boxed{\Delta V_R = 125\,mV}$$

DRAM: 1T cell



Cross-section

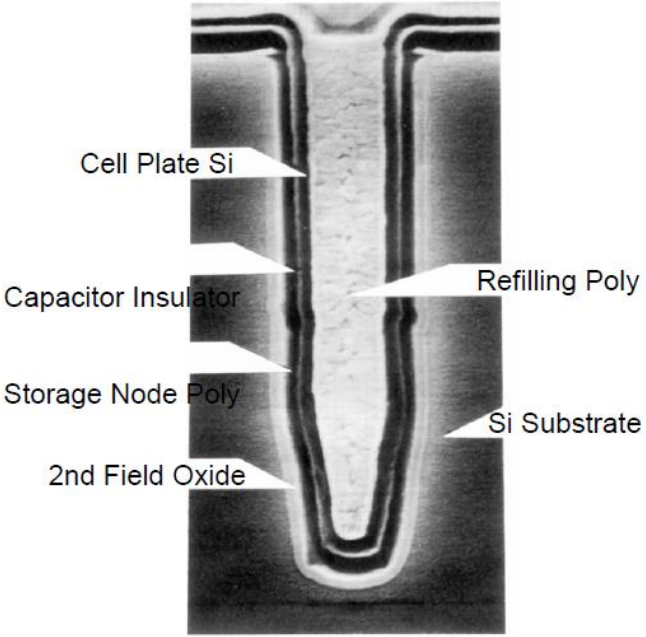


Layout

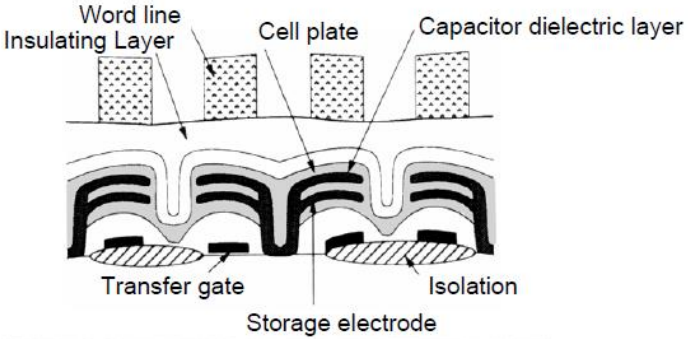
Uses Polysilicon-Diffusion Capacitance

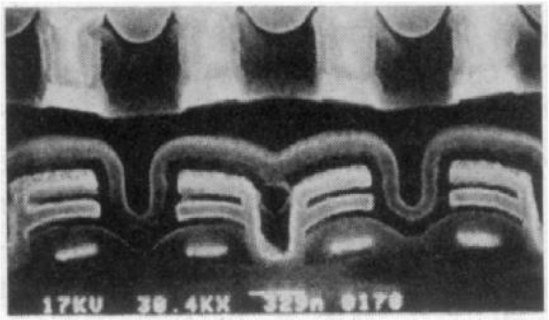
Expensive in Area

DRAM: 1T cell



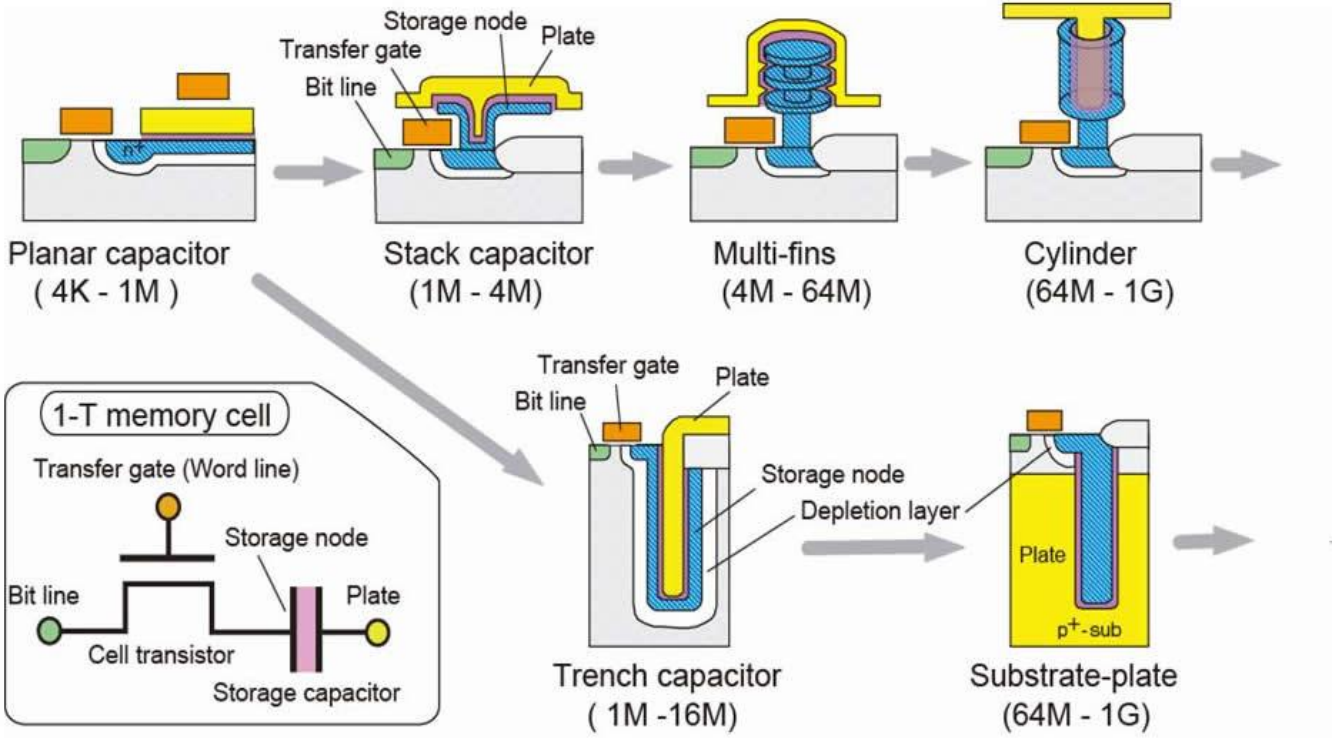
Trench Cell



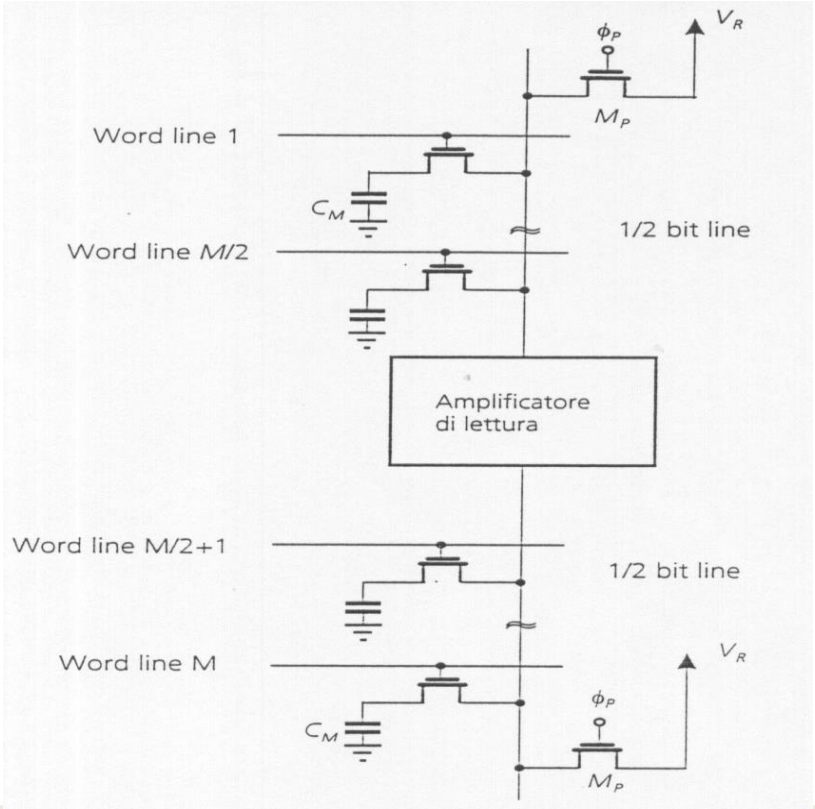


Stacked-capacitor Cell

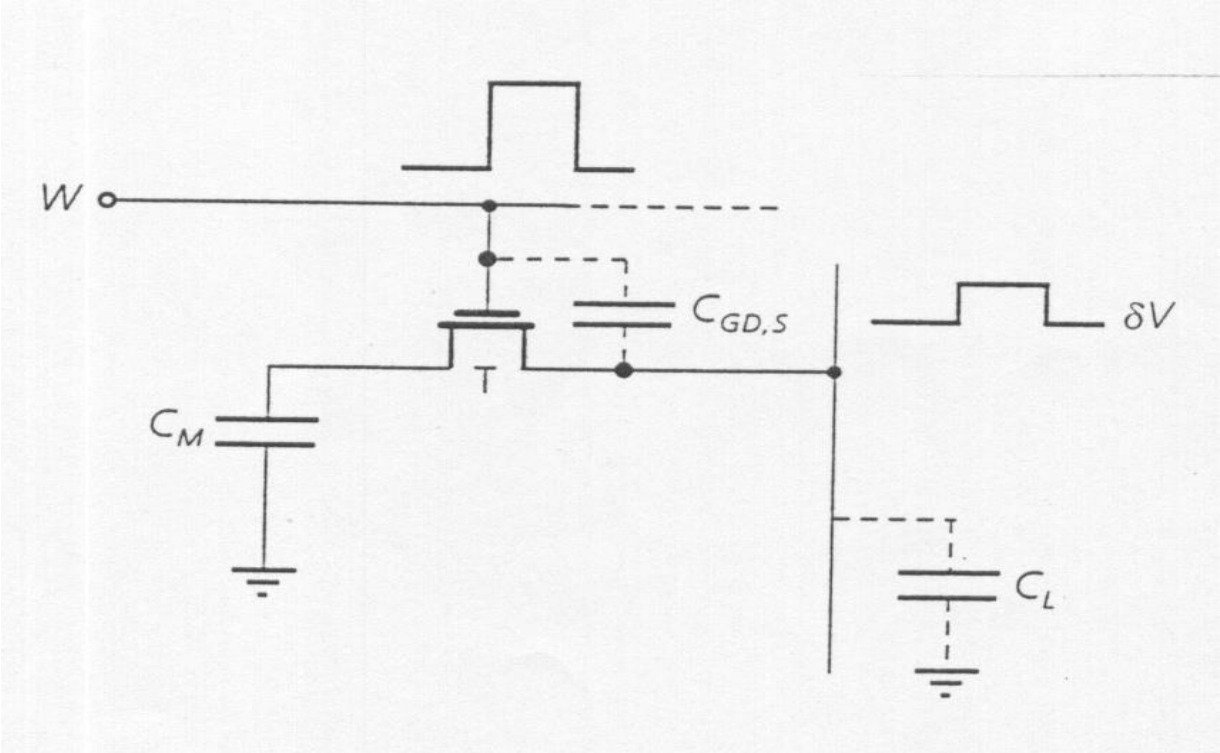
DRAM: 1T cell


Storage node
Plate
Transfer gate
Bit line
Planar capacitor
(4K - 1M)
Stack capacitor
(1M - 4M)
Multi-fins
(4M - 64M)
Cylinder
(64M - 1G)
1-T memory cell
Transfer gate (Word line)
Storage node
Bit line
Plate
Cell transistor
Storage capacitor
Transfer gate
Plate
Bit line
Storage node
Depletion layer
Plate
p+-sub
Trench capacitor
(1M -16M)
Substrate-plate
(64M - 1G)

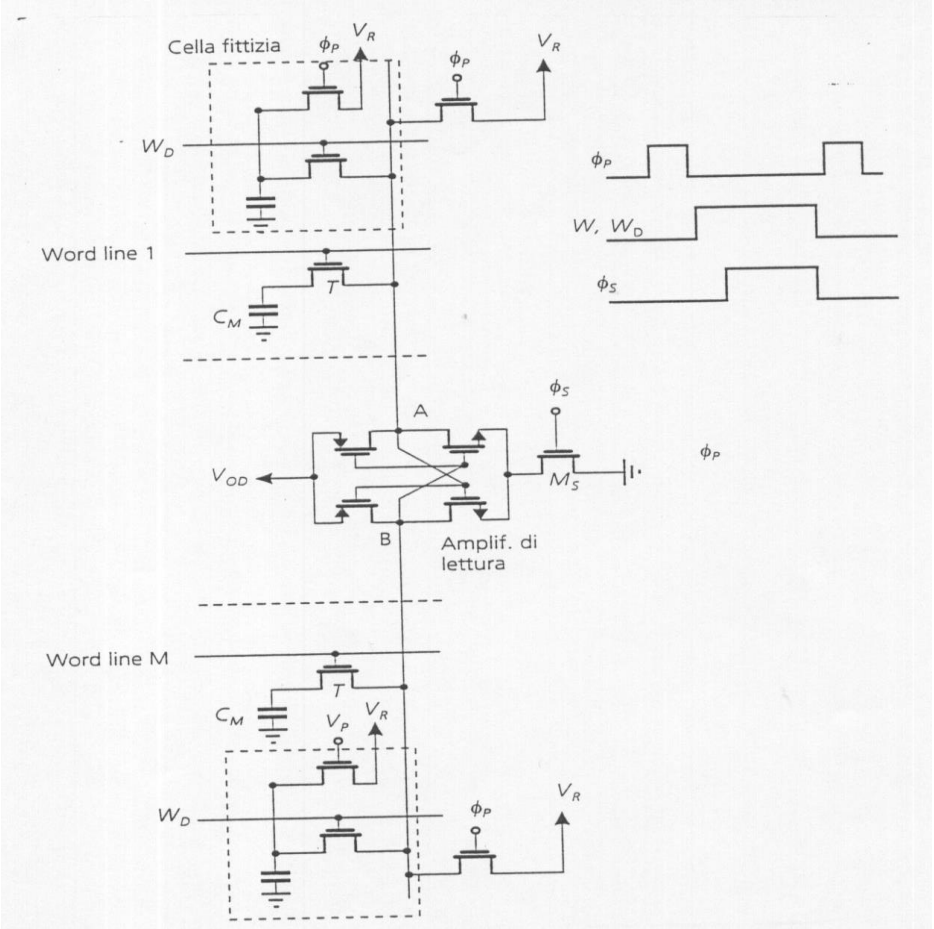
DRAM


ϕ_P
V_R
M_P
Word line 1
C_M
1/2 bit line
Word line $M/2$
Amplificatore
di lettura
Word line M/2+1
1/2 bit line
Word line M
V_R
ϕ_P
C_M
M_P

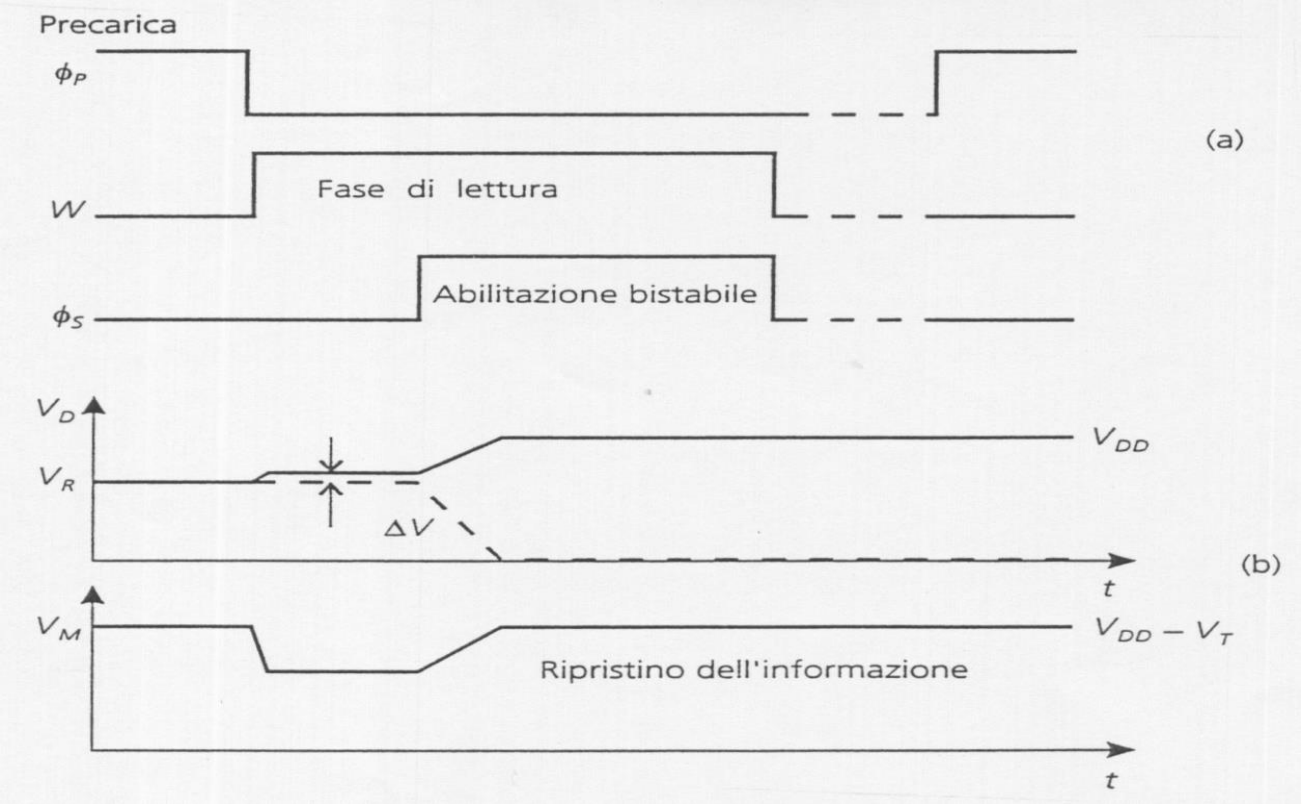
DRAM



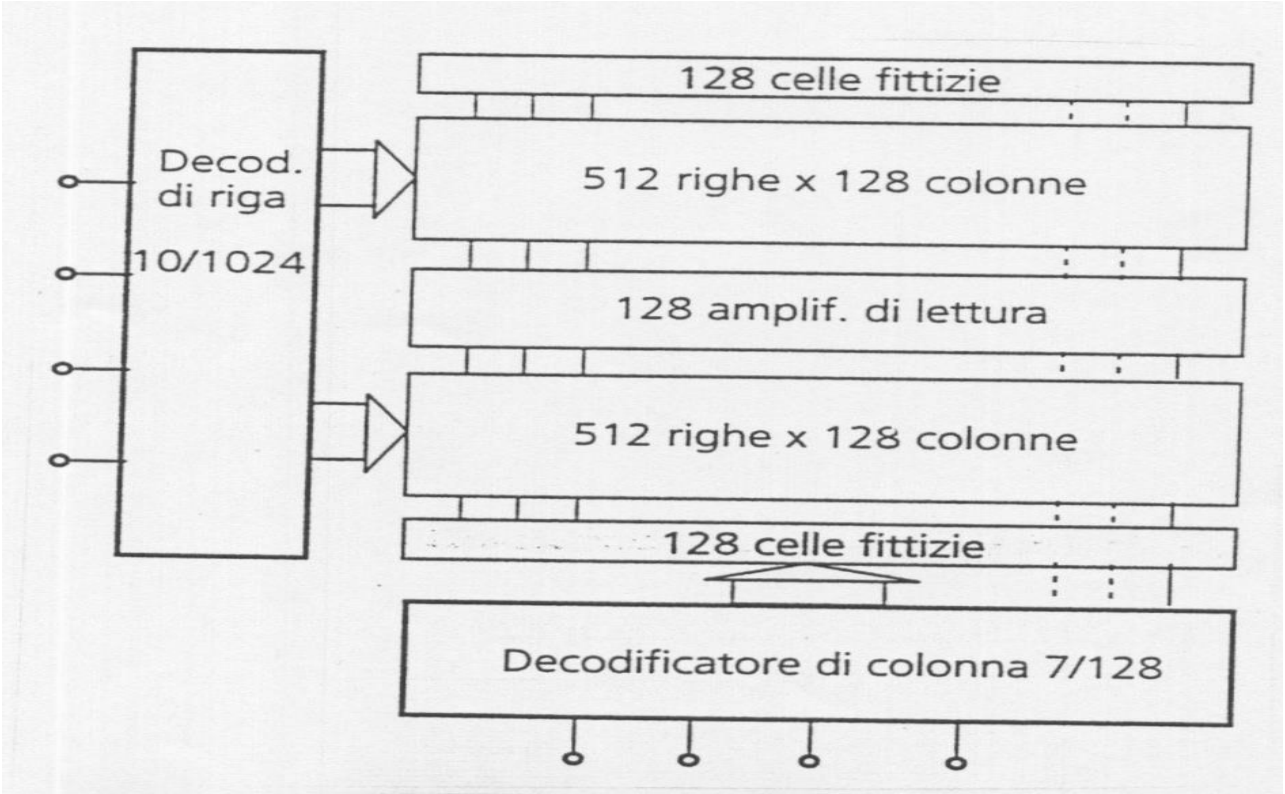
DRAM: reading



DRAM: reading



DRAM: reading



See:

- J. M. Rabaey, A. Chandrakasan, B. Nikolic, «Digital Integrated Circuits: A Design Perspective», Pearson, 2003
  - Cap. 12.1-12.2

- Paolo Spirito, «Elettronica Digitale», McGraw-Hill, 2006
  - Cap. 13.3 e 13.8