Semiconductor Memories

- Semiconductor memories array capable of storing large quantities of digital information are essential to all digital systems.
- Maximum realizable data storage capacity of a single chip doubles every two years.
- On-chip memory have become a widely used subsystem in many VLSI circuit.

Semiconductor memory

- read-write: RWM
- read-only: ROM
- nonvolatile read-write: NRWM
- non random access
- random access
- EPROM
- E'PROM
- FLASH
- FIFO, LIFO
- static: SRAM
- dynamic: DRAM
Memory characteristic

- Capability to store data, Memory Size (bit, byte or word)
- Time needed to retrieve data, read-access (delay between the read request and the time the data is available)
- Time elapsed between a write request and the final writing of data, write-access
- Minimum time required between successive reads or writes, cycle time (it is greater than the access times. It can be different for write or read operation, but it is assumed equal for simplicity)
Memory architectures

- A decoder is inserted to reduce the number of select signals (A₀…A_{K-1} address word)

- In order to optimize the area of the memory the shape of the core should be almost square
- It is common to reduce voltage swing of bit line to reduce propagation delay, power consumption, cross talk.

interfacing with external word is achieved via sense amplifiers which give full rail-to-rail swing.

- The architecture works well up to range of 64 Kbits to 256 Kbits (larger memories suffer of speed degradation due to excessively long bit and word lines).

- Higher memories are partitioned into P smaller block (it is added the block address).
  - the local word and bit lines are reduced
  - non active block can be powered off.
- I/O interface has an enormous impact on the global memory performance (control and timing)

- DRAM use multiplexed addressing scheme (the lower and upper halves of the address are presented sequentially)
  - reduced number of package pins
  - it is needed strobe signals for each part of the address

- SRAM uses a self-timed approach (the complete address is presented at once)

---

**Read-Only Memories**

**NOR ROM**

---
- A bit has a 0 value if the a MOS is present

- Each column works has a pseudo-NMOS NOR
  - the design follows the same guidelines of the ratioed logic (pseudo-NMOS)
  - $V_{OL}$ equal to 1 to 3 volt for a 5 V power supply

- The array is constructed by repeating the same cell (in both the horizontal and vertical directions), mirroring the odd cells around the horizontal axis to share the ground line

- The memory is programmed by a selective addition of metal-to-diffusion contact. (the presence of a contact means a 0-bit)
A denser solution uses an extra implant step to be able to set the MOS threshold to a voltage higher than the power supply (in that case transistor is off).

A denser solution is NAND ROM since does not require ground lines.

A transistor represent a 1 bit value.

The word line operates in reverse logic mode (all word line are high for default).
➢ The core area can be almost two times lower than other solutions

➢ Worst performance since a wider pull-up load is needed (higher parasitic capacitances)

➢ To reduce ROM power dissipation a precharged approach can be adopted
Nonvolatile Read-Write Memories

- The architecture of the NVRW memories is virtually identical to the ROM structure.
- They make use of transistor threshold which can be electrically altered.
- To reprogram the memory, it needs to be erased.
- The erasing methods differ for the various classes of NVRW.
- The programming takes typically more time than the reading time (an order of magnitude).

Floating Gate Transistor

- The core of every NVRW is made by a floating-gate transistor.
- *Floating-gate transistor* has the threshold voltage programmable.

![Floating gate diagram](image)
To program the memory an high voltage in the range 15-20 V between source and gate-drain must be used.

When the programmable voltage is applied the high electric field causes avalanche injection (Electrons acquire sufficient energy to become hot and traverse the first oxide insulation).

The phenomenon can occur with oxide as thick as 100 nm (simply to realize) and is self-limiting.

For the programming mechanism it is also called FAMOS (Floating-gate Avalanche-injection MOS).

Avalanche injection
the floating gate change its voltage which is ideally half of the gate

Removing programming voltages leaves charge trapped
It is equivalent to a voltage offset

Although the gate is at the power supply, the floating gate voltages is much lower than threshold voltage.
**EPROM**
Erasable-Programmable Read-Only Memory

- is erasable using ultraviolet light via a transparent window in the package
- are simple and dense
- erasure is slow (from seconds to minutes)
- programming takes several (5-10) \( \mu s/\text{word} \)
- erase/program cycle is limited to one thousand

---

**EEPROM (E²PROM)**
Electrically-Erasable Programmable Read-Only Memory

- provides electrical-erasure procedure
- are based on a modified floating gate transistor called FLOTOX (floating-gate tunneling oxide), which uses the Fowler-Nordheim tunneling effect
- The effect arises using dielectric thickness lower than 10 nm
The main advantage of the effect is that it is reversible.

The bidirectionality of the effect poses the problem of threshold control (removing too much charge can determine a depletion device always on), and to remedy to it a transistor in series is introduced (it acts as access device).

EEPROM are larger than EPROM (two transistors in each cell and FLOTOX is larger than FAMOS due to extra area for tunneling oxide).

Fabrication of very thin oxide is a challenging and costly manufacturing step.

Advantages are their versatility.

Can support up to $10^5$ erase/write cycle.

Programming can take from almost 10 ms/word (few s/chip).
FLASH
FLASH Electrically-Erasable Programmable Read-Only Memory

➢ Combines EPROM density with E²PROM versatility

➢ Uses avalanche hot-electrons-injection to program the device and the Fowler-Nordheim tunneling to erase it

➢ Erasure is performed in bulk for a complete core or for a subsection. This reduces flexibility, but allows to avoid access transistor (the threshold can be electronically controlled during erasure)

➢ Unlike E²PROM erasure occurs with gate grounded and applying a high voltage (12 V) to the source

---

SRAM

6T cell

➢ Access is enabled by M5 and M6

➢ Two bit line are used (although they are not necessary, this improves noise margin during read and write operations)
6T write operation

- M4, M6 is a pseudo NMOS and to toggle the memory must be its $V_{OL} < V_{DD}/2$

- M5 is diode connected and to toggle its source should be $> V_{DD}/2$ (M1 is in triode region)

\[
\beta_p 4\left[[V_{DD} - V_p] - V_{OL}\right]V_{OL} = \beta_n 6\left[[V_{DD} - V_t] - V_{OL}\right]V_{OL}
\]

\[
\left(\frac{W}{L}\right)_6 \geq \frac{\mu_p}{\mu_n} \left(\frac{W}{L}\right)_4
\]

\[
\beta_{n5}\left[[2(V_{DD} - V_{t0}) - \frac{V_{DD}}{2}]V_{DD}\right] = \beta_{n5}\left[[\frac{V_{DD}}{2} - V_{t5}\right]^2
\]

\[
\left(\frac{W}{L}\right)_5 \geq \left(\frac{3}{2}V_{DD} - 2V_{t0}\right)V_{DD}\left(\frac{W}{L}\right)_5 \geq \left(V_{DD} - V_{t5}\right)^2
\]
6T write operation

- The bit lines are precharged to $V_{DD}$
- To do not destroy information we have to guarantee that the source of the diode connected M5 stay lower than $V_{DD}/2$ (M1 is in triode region)

$$\beta_{ni} \left[ 2(V_{DD} - V_{to}) - \frac{V_{DD}}{2} \right] \frac{V_{DD}}{2} = \frac{\beta_{n5}}{2} \left( \frac{V_{DD} - V_{t5}}{2} \right)^2$$

- This condition is opposite to that found for writing operation, but it is stronger (to write the memory we can only satisfy the first condition)
- The second bit line clamp to $V_{DD}$ making extremely difficult to toggle the cell during the read operation (advantage of the dual bit-line architecture)

- To further prevent the toggling during the read operation the bitline are precharged $V_{DD}/2$ (makes impossible to reach the threshold voltage of the inverter)

- The transient behavior is limited by the read operation since the large bit-line capacitance must be discharged

- The write time is dominated by the propagation delay of the cross-coupled inverter pair

- The 6T cell while simple and reliable it is area-hungry (its dimension are dominated by wiring and interlayer contacts)
4T cell

- 4T cell is 1/3 lower than the 6T cell
- A very high and compact resistance is used which is implemented with undoped polysilicon (sheet resistance $T\Omega$)

DRAM

- To reduce RAM dimension we can eliminate the load which has the only function to replenish the charge lost by leakage
- RAM cells need a periodical refresh, which typically occurs every 1 to 4 ms
- Common DRAM are based on the 3T and 1T cells
- In DRAM the information is represented by charge stored on capacitor
- In contrast to SRAM, no constraint exist on the device ratio
3T cell

- The cell is written placing the appropriate data value on BL1 and asserting WWL (write word line).
- To read the cell the RWL (read word line) is asserted and the data is on BL2 which is precharged (or clamped to a voltage via a load).

Write and read cycle
The cell complexity is substantially reduced compared to the SRAM (576 $\lambda^2$ compared to 1092 $\lambda^2$).

---

**1T cell**

- The cell is written and red asserting WL
- BL is precharged to $V_{DD}/2$
Write and read cycle

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

charge transfer ratio

(1% - 10%)

- \( C_S \) is typically one or two order of magnitude lower than \( C_{BL} \) (\( \Delta V \) around 250 mV).
- 1T DRAM requires sense amplifier for each bit line (in the other memory it is introduce to speed up the read-out)
- Unlike 3T, read-out of the 1T DRAM cell is destructive (after a successful read operation the value must be restored)
- The 1T cell requires the presence of extra capacitance (its minimum value is around 30 fF)
- To avoid the threshold voltage lost when writing a 1, bootstrapping can be used
1T DRAM using a polysilicon-diffusion capacitance

Cross section

Layout

Inversion layer
(induced by plate bias)