**Static Random Access Memory (SRAM)**

"Word" lines control operation
"Bit" lines carry data in/out of cell

Figure 13.1 General SRAM cell

- Based on bi-stable circuit
- Holds its value as long as power is applied (static)
- Mosfets MAL, MAR provide I/O access

(a) 6T cell
(b) 4T cell with poly resistors

Figure 13.2 CMOS SRAM circuits

- 6 transistor cell (6T) is most popular, although 4T cell with poly resistors can be smaller depending on resistors available in your process. However, the 4T cell requires extra process & mask steps => more expensive process!
SRAM Design Considerations

Figure 13.3 6T SRAM cell design parameters

(a) Write 1 operation
(b) Resistor model

Figure 13.5 Writing to an SRAM

- $\beta \text{N}/\beta \text{P}$ ratio controls inverter characteristics, but these devices are typically both min size to get the smallest possible SRAM cell.
- $\beta \text{A}$ must be $> \beta \text{N}$, to allow writes ($\beta \text{A} \approx 2$ typical).
  $\Rightarrow V_{DD} \left( \frac{\beta \text{N}}{\beta \text{A} + \beta \text{N}} \right) > V_{midpoint}$ to change cell value.

"Butterfly Plot" used to characterize SRAM cell. (SNM) describes the cells resistance to noise.

Figure 13.4 Butterfly plot
TYPICAL SRAM CELL LAYOUT:

![SRAM Cell Layout](image)

CAN YOU FIND THE ERROR IN THIS LAYOUT? (HINT: IS $\beta_1 > \beta_2$?)

**Figure 13.6** Example of a basic SRAM cell layout

- **Note perpendicular metal routes for supplies and signals**
- **Need to simultaneously minimize cell size and provide I/O access**

**CAN ALSO HAVE MULTIPLE I/O PORTS**:

![2-port CMOS SRAM Cell](image)

**Figure 13.7** A 2-port CMOS SRAM cell
SRAM ARRAYS

- **WE** = ACTIVELOW "WRITE ENABLE"
- **EN** = ACTIVELOW "ENABLE", ALSO CALLED "CHIP SELECT"

Figure 13.9 High-level view of an SRAM

- **SRAM CHIPS ARE ORGANIZED IN N x N ARRAYS** E.G., 128k x 8 ⇒ 128k 8-BIT WORDS
- M ADDRESS BITS CAN ADDRESS 2^M WORDS

Figure 13.10 Central SRAM block architecture

- MANY DIFFERENT DESIGNS EXIST (THIS IS JUST ONE!)
- CENTRAL WORD LINE DECODER/DRIVERS SAVES AREA AND ALLOWS LESS CAPACITANCE ON WORD LINES ⇒ FASTER!
- EACH CORE CONTAINS MULTIPLE WORDS
- THE MUX AT THE BOTTOM SELECT WHICH WORD IS OUTPUT/INPUT
**Figure 13.11** Cell arrangement in a core region

- **HIGH DEGREE OF SYMMETRY PROVIDES FOR A DENSE LAYOUT IN THE CORE MEMORY REGIONS**
- **WHEN A WORD LINE GOES ACTIVE, ALL CELLS IN A GIVEN ROW ARE TURNED ON**
  - **MUX AT BOTTOM OF EACH COLUMN USED TO SELECT THE DESIRED BITS**
- **OFTEN INTERDIGITATE THE WORDS TO EASE MUX ROUTING**

**Figure 13.13** Column MUX/DeMUX network for 8-bit words
MEMORY ADDRESSING

Figure 13.16 Basic addressing scheme

- MEMORY ADDRESSES ARE BROKEN DOWN INTO ROW AND COLUMN SELECT LINES TO ADDRESS ANY WORD IN THE MATRIX
- ADDRESS LINES ARE TYPICALLY LATCHED BEFORE GOING INTO THE ROW & COLUMN DECODERS

Example 13.1
A 128K x 8 SRAM chip holds 128K 8-bit words for a total of 1 MB of total storage. The address word must have a width of

$$ m = \log_2(128K) $$

$$ = 17 $$ (13.4)

to select every 8-bit word location.

Example 13.2
The 128K x 8 SRAM chip requires a 17-bit address word. If we use a dual core arrangement with one word per word line, then we need 64K word lines. If we expand each word line to 64 bits (8 words) then the number of word lines is reduced to 8K. The 17-bit address $A_{16}$ ... $A_0$ can then be divided into a 4-bit column address group of $A_6A_5A_4A_3$ and a 13-bit row address group of $A_{12}$ ... $A_0$. Other array sizes divide the address word proportionately.
Figure 13.18 Expanded view of column circuitry

When "Write Enable" (WE) goes active, the bit drivers are connected to the I/O busses and drive the SRAM cell inputs selected by the column MUX.

Figure 13.19 Write circuitry example
**Sense Amplifiers for Read**

- Often > 1 stage of amplification is used.
- Dynamic operation to save power (only on during reads).
- Outputs of 1st stage reset between reads to speed up operation.

**Figure 13.20** Example of a sensing scheme for the read operation.

- **Simple Differential Amplifier with Active Loads (MOS).**
- \( V_0 = A(V_1 - V_2) \)
- \( A > 1 \) required.

**Figure 13.21** Single-ended differential amplifier.

**Figure 13.22** Dual-amplifier scheme for the sense amplifier network.
**Dynamic Random Access Memory (DRAM)**

Figure 13.23  IT DRAM cell

- "IT" DRAM cell used for maximum density
- Very simple! Voltage is stored on capacitor:
  
  \[ V_s = \frac{Q_s}{C_s} \]

  to store a "1", charge the cap
to store a "0", discharge the cap

  But, charge leaks off over time (more on this later...)

- Many novel cap structures used to increase density!

Figure 13.24  A DRAM cell using a trench capacitor

- Caps built out of deep trenches
  
  \[ N^+ - SiO_2 - POLY \]

  forms cap

Figure 13.29  Visualization of stacked capacitor structure

- Caps can be built out of stacked poly layers above the silicon surface

  - Use of 3-D structures greatly increases memory density!
Figure 13.24 Write and hold operations in a DRAM cell

- To write a bit into the cell, the word line is turned on and the switch charges the cap.
- When the switch turns off, the value is held.

Figure 13.27 Read operation in a DRAM cell

- To read the contents of a cell, the switch is turned on and its charge is dumped onto the bit line for the sense amp to amplify. But, since \( C_{bit} \approx C_s \), the change in voltage is typically small.

Example 13.3

Suppose that we have a DRAM cell with \( C_p = 50 \text{ fF} \) and a bit line capacitance of \( C_{bit} = 8 \times C_p \). Assuming a maximum voltage of \( V_j = V_{max} = 2.5 \text{ V} \) on the storage capacitor, the final voltage during a logic 1 read operation is

\[
V_j = \left( \frac{1}{2} \right) (2.5) = 278 \text{ mV}
\]  

(13.20)

A stored logic 0 would result in \( V_j = 0 \text{ V} \), so that the sense amplifier must be able to distinguish between 0.7 V and 0.28 V to determine the value of the stored bit.
**Refreshing DRAMs**

![DRAM cell diagram](image)

**Figure 13.25** Charge leakage in a DRAM cell

- The memory is "dynamic" because the stored charge leaks off over time:
  \[ I_L = C \frac{dV_S}{dt} \quad I_L = \text{leakage current due to } \text{PN junctions and MOS sub-threshold slope} \]

To prevent data loss, the stored value must be "refreshed" before \( V_{stored} \) falls below the minimum value the sense amp can detect.

**Example!**

If: \( C = 50 \mu F \)
\[ I_L = 1mA \]
\[ V_{MAX} - V_{MIN} = 1V \]

Then:
\[ \tau_H = \left( \frac{C}{I_L} \right) \left( \frac{1V}{1V} \right) \]
\[ \tau_H = 0.5 \mu s \]

**Figure 13.26** Refresh operation summary

- To refresh a cell, first read the stored value and then re-write it into the cell.

Required \( = \frac{1}{f_{\text{refresh}}} > \frac{1}{\tau_H} \)

(Note error in text)
**READ-ONLY MEMORIES (ROM)**

- **Used to store data permanently** (e.g., PC BIOS)
- **Built in arrays similar to RAMs**
- **Typically built using selectively placed**
  **vias, contacts, or FETs**

![Logic diagram for a NOR-based ROM](image)

**Figure 13.33** Logic diagram for a NOR-based ROM

- **"Pseudo-NMOS" ROM uses triode PMOS FETs as loads,**
  **with NMOS FETs selectively placed (or connected) as**
  **pull downs**

![ROM array using pseudo-nMOS circuitry](image)

**Figure 13.34** ROM array using pseudo-nMOS circuitry

- **The addressed row is driven high, the rest stay low.**
- **Data is stored based on which rows are tied to each Nor.**
ROM LAYOUTS

- PROGRAMMED BY SELECTIVE PLACEMENT OF NMOS FETS TO ACT AS PULL DOWNS WHEN THE ROW LINE = "1"
- CAN ALSO PLACE AN NMOS FET AT EVERY LOCATION, AND SELECTIVELY CONNECT THEM USING CONTACTS OR VIAS

Figure 13.35 Map for ROM layout

Figure 13.36 ROM layout based on FET map
**PROGRAMMABLE ROMS (PROMS)**

- **Allow the user to store data needed by their application electrically and on the fly.**
- **Early ErPROMs** required UV light to erase them.
- **Today, special MOSFETs are used to make electrically erasable PROMs or "E2PROMs" ("Flash Memory").**

**Figure 13.37** Floating-gate MOSFET

**Figure 13.38** Effect of charge storage on the floating gate

- **Extra "floating gate" is used to trap charge and therefore shift the VT higher.**
- **FETs with low VT turn on when the word line goes high, but FETs with high VT stay off.**

**Figure 13.39** A E2PROM word using floating-gate nFETs
**Figure 13.40** Programming a floating-gate FET

- This structure uses a high drain voltage to program the FET by trapping "hot electrons" on the floating gate (typ ≈ 12-30V)
- Retention times > 10 years

**Figure 13.41** Fowler-Nordheim tunnelling

- An alternate structure modifies the gate near the drain (thinner oxide) to enhance tunnelling of electrons to the floating gate
- Erasure requires reversing the voltages
- "Flash" EPROMs erase all bits at once
Programmable Logic Arrays (PLAs)

- Allows standard "SOP" (Sum-of-Product) logic functions to be easily programmed
- Typically programmed by masks during processing by selectively placing FETs, vias, or contacts

Figure 13.43 Structure of an AND-OR PLA

Figure 13.44 Logic gate diagram of the PLA
GATE ARRAYS

![Gate array base](image)

Figure 13.45 Transistor arrangement in a gate array

- **Gate Arrays are Chips Full of Arrays of FETs Which Are Not Connected**
- **The User Can Wire These FETs in Whatever Connections They Need (NANDs, NORs, FFs) Using Only Metal Masks**
  - Very Flexible, and Takes Much Less Processing Time Since Most of the Process Steps Are Already Done!
  - "Rapid Prototyping"

- **Field-Programmable Gate Arrays** (FPGA) Can Be Programmed by the User Using a PC to "Burn" the Circuit
  - Allows New Chips to Be Created in Real Time
  - Can Have Thousands of Gates (Dense!)
  - Uses "Fuses" or "Antifuses" to Make or Break Connections
ANTIFUSES AND FUSES

Figure 13.49 Programmable antifuse arrangements

• "ANTIFUSE" IS NATURALLY AN OPEN CIRCUIT
  → PROGRAMMED BY FORCING A HIGH CURRENT THROUGH IT, WHICH MELTS THE ANTIFUSE LAYER AND SHORTS METAL LAYERS

• "FUSES" ARE NATURALLY A SHORT CIRCUIT
  → PROGRAMMED BY FORCING A HIGH CURRENT THROUGH IT TO "BLOW" THE CONNECTION
  → TRADITIONALLY MADE FROM THIN ALUMINUM METAL LINES
  → MODERN FUSES USE POLY, WHICH CHANGES IT'S RESISTANCE (REQUIRES A SENSE AMP)

• FUSES/ANTIFUSES OFTEN USED TO ADJUST A CHIP'S PERFORMANCE DURING TESTING ("WAFER SORT")