# Semiconductor Memory Classification

<table>
<thead>
<tr>
<th>Read-Write Memory</th>
<th>Non-Volatile Read-Write Memory</th>
<th>Read-Only Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Access</td>
<td>Non-Random Access</td>
<td>EPROM</td>
</tr>
<tr>
<td>SRAM</td>
<td>FIFO</td>
<td>$^{E_2}$PROM</td>
</tr>
<tr>
<td>DRAM</td>
<td>LIFO</td>
<td>FLASH</td>
</tr>
<tr>
<td>Shift Register</td>
<td>CAM</td>
<td>Mask-Programmed</td>
</tr>
<tr>
<td>CAM</td>
<td></td>
<td>Programmable (PROM)</td>
</tr>
</tbody>
</table>
Timing Parameters

Read cycle

Read access

Write cycle

Write access

Data valid

Data written

DATA

WRITE

READ
Memory Architecture: Decoders

Intuitive architecture for $N \times M$ memory

Too many select signals:

$N$ words $\Rightarrow N$ select signals

Decoder reduces the number of select signals

$K = \log_2 N$
Array-Structured Memory Architecture

Problem: ASPECT RATIO or HEIGHT >> WIDTH

- Rows are controlled by Row Decoder
- Columns are controlled by Column decoder
- Sense amplifiers convert input swing to rail-to-rail amplitude
- Selects appropriate word
- Bit line and Word line for data transfer
Hierarchical Memory Architecture

Advantages:
1. Shorter wires within blocks
2. Block address activates only 1 block => power savings
Block Diagram of 4 Mbit SRAM

- Clock generator
- Z-address buffer
- X-address buffer
- Predecoder and block selector
- Bit line load
- Transfer gate
- Column decoder
- Sense amplifier and write driver
- CS, WE buffer
- I/O buffer
- x1/x4 controller
- Y-address buffer
- X-address buffer
Contents-Addressable Memory
Memory Timing: Approaches

Address bus: Row Address, Column Address

RAS

CAS

RAS-CAS timing

Address Bus: Address transition initiates memory operation

DRAM Timing: Multiplexed Addressing

SRAM Timing: Self-timed
Read-Only Memory Cells

1

Diode ROM

0

MOS ROM 1

MOS ROM 2
MOS OR ROM

![Diagram of MOS or ROM circuit]
MOS NOR ROM

\[ \text{Pull-up devices} \]

\[ V_{DD} \]

\[ WL[0] \]

\[ WL[1] \]

\[ WL[2] \]

\[ WL[3] \]

\[ BL[0] \]

\[ BL[1] \]

\[ BL[2] \]

\[ BL[3] \]

\[ GND \]
MOS NOR ROM Layout

Cell (9.5λ x 7λ)

Programming using the Active Layer Only

- Red: Polysilicon
- Blue: Metal1
- Green: Diffusion
- Gray: Metal1 on Diffusion
MOS NOR ROM Layout

Cell \((11\lambda \times 7\lambda)\)

Programming using the Contact Layer Only

- Polysilicon
- Metal1
- Diffusion
- Metal1 on Diffusion
MOS NAND ROM

All word lines high by default with exception of selected row
MOS NAND ROM Layout

Cell \((8\lambda \times 7\lambda)\)

Programming 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
NAND ROM Layout

Cell (5\(\lambda\) x 6\(\lambda\))

Programming using Implants Only

- Polysilicon
- Threshold-altering implant
- Metal1 on Diffusion
Equivalent Transient Model for MOS NOR ROM

Model for NOR ROM

- **Word line parasitics**
  - Wire capacitance and gate capacitance
  - Wire resistance (polysilicon)
- **Bit line parasitics**
  - Resistance not dominant (metal)
  - Drain and Gate-Drain capacitance
Equivalent Transient Model for MOS NAND ROM

Model for NAND ROM

- **Word line parasitics**
  - Similar to NOR ROM
- **Bit line parasitics**
  - Resistance of cascaded transistors dominates
  - Drain/Source and complete gate capacitance
Precharged MOS NOR ROM

PMOS precharge device can be made as large as necessary, but clock driver becomes harder to design.
Non-Volatile Memories
The Floating-gate transistor (FAMOS)
Floating-Gate Transistor Programming

Avalanche injection

Removing programming voltage leaves charge trapped

Programming results in higher $V_T$. 
A “Programmable-Threshold” Transistor

![Diagram of a Programmable-Threshold Transistor](image)

- "0"-state
- "1"-state
- "ON"
- "OFF"
- $V_{WL}$
- $V_{GS}$
- $D V_T$
FLOTOX EEPROM

FLOTOX transistor

Fowler-Nordheim $I$-$V$ characteristic
EEPROM Cell

Absolute threshold control is hard
Unprogrammed transistor might be depletion
⇒ 2 transistor cell
Flash EEPROM

Many other options …
Cross-sections of NVM cells

Flash

EPROM
Basic Operations in a NOR Flash Memory—Erase

![Diagram of a NOR Flash Memory cell and array with voltages applied to gate (G), source (S), drain (D), bit lines (BL0, BL1), and word lines (WL0, WL1).]
Basic Operations in a NOR Flash Memory—Write
Basic Operations in a NOR Flash Memory—Read
NAND Flash Memory
NAND Flash Memory
### Characteristics of State-of-the-art NVM

**Table 12-1** Comparison between nonvolatile memories ([Itoh01]).

\[ V_{DD} = 3.3 \text{ or } 5 \text{ V}; V_{PP} = 12 \text{ or } 12.5 \text{ V}. \]

<p>| Cell—Nr. of | Cell Area (ratio wrt EPROM) | Mechanism | External Power Supply | Program/Erase Cycles |</p>
<table>
<thead>
<tr>
<th>Transistors</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MASK ROM</td>
<td>1 T (NAND)</td>
<td>0.35–5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>EPROM</td>
<td>1 T</td>
<td>1</td>
<td>UV Exposure</td>
<td>Hot electrons</td>
</tr>
<tr>
<td>EEPROM</td>
<td>2 T</td>
<td>3–5</td>
<td>FN Tunneling</td>
<td>FN Tunneling</td>
</tr>
<tr>
<td>Flash Memory</td>
<td>1 T</td>
<td>1–2</td>
<td>FN Tunneling</td>
<td>Hot electrons</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FN Tunneling</td>
<td>FN Tunneling</td>
</tr>
</tbody>
</table>
Read-Write Memories (RAM)

- STATIC (SRAM)
  - Data stored as long as supply is applied
  - Large (6 transistors/cell)
  - Fast
  - Differential

- DYNAMIC (DRAM)
  - Periodic refresh required
  - Small (1-3 transistors/cell)
  - Slower
  - Single Ended
6-transistor CMOS SRAM Cell
CMOS SRAM Analysis (Read)

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

\[ \Delta V = \frac{V_{DSATn} + CR(V_{DD} - V_{Tn}) - \sqrt{V_{DSATn}^2(1 + CR) + CR^2(V_{DD} - V_{Tn})^2}}{CR} \]
CMOS SRAM Analysis (Read)

\[ CR = \frac{W_1/L_1}{W_5/L_5} \]

Voltage Rise (V)

Cell Ratio (CR)
CMOS SRAM Analysis (Write)

\[
k_{n,M6} \left( (V_{DD} - V_{tn})V_Q - \frac{V_Q^2}{2} \right) = k_{p,M4} \left( (V_{DD} - |V_{tp}|)V_{DSATp} - \frac{V_{DSATp}^2}{2} \right)
\]

\[
V_Q = V_{DD} - V_{tn} - \frac{1}{2} \left( V_{DD} - V_{tn} \right)^2 - 2 \frac{\mu_p}{\mu_n} PR \left( (V_{DD} - |V_{tp}|)V_{DSATp} - \frac{V_{DSATp}^2}{2} \right),
\]
CMOS SRAM Analysis (Write)
6T-SRAM — Layout
Resistance-load SRAM Cell

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

<table>
<thead>
<tr>
<th>Table 12-2</th>
<th>Comparison of CMOS SRAM cells used in 1-Mbit memory</th>
<th>(from [Takada91])</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of transistors</strong></td>
<td><strong>Complementary CMOS</strong></td>
<td><strong>Resistive Load</strong></td>
</tr>
<tr>
<td>Number of transistors</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Cell size</strong></th>
<th><strong>Complementary CMOS</strong></th>
<th><strong>Resistive Load</strong></th>
<th><strong>TFT Cell</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>58.2 μm² (0.7-μm rule)</td>
<td>40.8 μm² (0.7-μm rule)</td>
<td>41.1 μm² (0.8-μm rule)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Standby current (per cell)</strong></th>
<th><strong>Complementary CMOS</strong></th>
<th><strong>Resistive Load</strong></th>
<th><strong>TFT Cell</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>10⁻¹⁵ A</td>
<td>10⁻¹² A</td>
<td>10⁻¹³ A</td>
<td></td>
</tr>
</tbody>
</table>
3-Transistor DRAM Cell

No constraints on device ratios
Reads are non-destructive
Value stored at node X when writing a “1” = $V_{WWL} - V_{Th}$
3T-DRAM — Layout
1-Transistor DRAM Cell

Write: $C_S$ is charged or discharged by asserting WL and BL.
Read: Charge redistribution takes place 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 Cell Observations

- 1T DRAM requires a sense amplifier for each bit line, due to charge redistribution read-out.
- DRAM memory cells are single ended in contrast to SRAM cells.
- The read-out of the 1T DRAM cell is destructive; read and refresh operations are necessary for correct operation.
- Unlike 3T cell, 1T cell requires presence of an extra capacitance that must be explicitly included in the design.
- When writing a “1” into a DRAM cell, a threshold voltage is lost. This charge loss can be circumvented by bootstrapping the word lines to a higher value than VDD.
Sense Amp Operation

- $V_{BL}$
- $V_{PRE}$
- $DV(1)$
- $V(1)$
- $V(0)$
- $t$

- Sense amp activated
- Word line activated
1-T DRAM Cell

Cross-section

Uses Polysilicon-Diffusion Capacitance
Expensive in Area
SEM of poly-diffusion capacitor 1T-DRAM
Advanced 1T DRAM Cells

Trench Cell

Stacked-capacitor Cell
Static CAM Memory Cell
Periphery

- Decoders
- Sense Amplifiers
- Input/Output Buffers
- Control / Timing Circuitry
Row Decoders

Collection of $2^M$ complex logic gates
Organized in regular and dense fashion

(N)AND Decoder

$$WL_0 = A_0A_1A_2A_3A_4A_5A_6A_7A_8A_9$$

$$WL_{511} = A_0A_1A_2A_3A_4A_5A_6A_7A_8A_9$$

NOR Decoder

$$WL_0 = \overline{A_0 + A_1 + A_2 + A_3 + A_4 + A_5 + A_6 + A_7 + A_8 + A_9}$$

$$WL_{511} = \overline{A_0 + \overline{A_1} + \overline{A_2} + \overline{A_3} + \overline{A_4} + \overline{A_5} + \overline{A_6} + \overline{A_7} + \overline{A_8} + \overline{A_9}}$$
Hierarchical Decoders

Multi-stage implementation improves performance

NAND decoder using 2-input pre-decoders
Dynamic Decoders

2-input NOR decoder

2-input NAND decoder
4-input pass-transistor based column decoder

Advantages: speed ($t_{pd}$ does not add to overall memory access time)
Only one extra transistor in signal path
Disadvantage: Large transistor count
4-to-1 tree based column decoder

Number of devices drastically reduced
Delay increases quadratically with # of sections; prohibitive for large decoders
Solutions: buffers
  progressive sizing
  combination of tree and pass transistor approaches
Decoder for circular shift-register
Sense Amplifiers

\[ t_p = \frac{C \cdot \Delta V}{I_{av}} \]

make $\Delta V$ as small as possible

Idea: Use Sense Amplifier

small transition $\rightarrow$ s.a. $\rightarrow$ output

input
Differential Sense Amplifier

![Diagram of Differential Sense Amplifier]

Directly applicable to SRAMs
Differential Sensing — SRAM

(a) SRAM sensing scheme

(b) Two stage differential amplifier
Latch-Based Sense Amplifier (DRAM)

Initialized in its meta-stable point with EQ
Once adequate voltage gap created, sense amp enabled with SE
Positive feedback quickly forces output to a stable operating point.
Charge-Redistribution Amplifier

Concept

Transient Response

$V_{in}$
$V_L$
$V_S$
$V_{ref}$
$V_{ref} = 5 \, 3V$

$t$ (nsec)

0.0
1.00
2.00
3.00
Charge-Redistribution Amplifier—EPROM

\[ V_{DD} \]

\[ SE \rightarrow M_4 \rightarrow \text{Out} \]

\[ V_{casc} \rightarrow M_3 \rightarrow C_{out} \]

\[ WLC \rightarrow M_2 \rightarrow C_{col} \]

\[ WL \rightarrow M_1 \rightarrow C_{BL} \]

Load

Cascode device

Column decoder

EPROM array
Single-to-Differential Conversion

How to make a good $V_{\text{ref}}$?
Open bitline architecture with dummy cells
DRAM Read Process with Dummy Cell

![Graphs showing DRAM read process with dummy cell.](image-url)
Voltage Regulator
Charge Pump
DRAM Timing
RDRAM Architecture
Address Transition Detection
Reliability and Yield

- Semiconductor memories trade off noise-margin for density and performance

Highly Sensitive to Noise (Crosstalk, Supply Noise)

- High Density and Large Die size cause Yield Problems

\[ Y = 100 \frac{\text{Number of Good Chips on Wafer}}{\text{Number of Chips on Wafer}} \]

\[ Y = \left[ \frac{1 - e^{-AD}}{AD} \right]^2 \]

Increase Yield using Error Correction and Redundancy
Sensing Parameters in DRAM
Noise Sources in 1T DRam
Open Bit-line Architecture — Cross Coupling
Open Bit-line Architecture — Cross Coupling
Folded-Bitline Architecture
Transposed-Bitline Architecture

(a) Straightforward bit-line routing

(b) Transposed bit-line architecture
Alpha-particles (or Neutrons)

1 Particle ~ 1 Million Carriers
Yield

Yield curves at different stages of process maturity (from [Veendrick92])
Redundancy
Error-Correcting Codes

Example: Hamming Codes

\[ \begin{align*}
P_1 & \oplus B_3 \oplus B_5 \oplus B_7 = 0 \\
1 & = 3
\end{align*} \]
Redundancy and Error Correction
Sources of Power Dissipation in Memories
Data Retention in SRAM

SRAM leakage increases with technology scaling
Suppressing Leakage in SRAM

Inserting Extra Resistance

Reducing the supply voltage
Data Retention in DRAM

The graph shows the relationship between capacity (in bits) and current (in A) for different operating voltages (V) and extrapolated threshold voltages at 25°C (V). The cycle time is 150 ns. The graph includes the labels $I_{AC}$, $I_{AC^T}$, and $I_{DC}$. The x-axis represents capacity (in bits) ranging from 15M to 64G, and the y-axis represents current (in A) ranging from $10^{-8}$ to $10^3$. The operating voltage (V) is shown on the lower scale ranging from 0.8 to 3.3, and the extrapolated threshold voltage at 25°C (V) is shown on the lower scale ranging from 0.13 to 0.53.