SEMICONDUCTOR MEMORIES
Chapter Overview

• Memory Classification
• Memory Architectures
• The Memory Core
• Periphery
• Reliability
# Semiconductor Memory Classification

<table>
<thead>
<tr>
<th>RWM</th>
<th>NVRWM</th>
<th>ROM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random Access</strong></td>
<td><strong>Non-Random Access</strong></td>
<td><strong>EPROM</strong> &lt;br&gt; <strong>E²PROM</strong> &lt;br&gt; <strong>FLASH</strong></td>
</tr>
<tr>
<td>SRAM</td>
<td>FIFO &lt;br&gt; LIFO &lt;br&gt; Shift Register &lt;br&gt; CAM</td>
<td>Mask-Programmed Programmable (PROM)</td>
</tr>
</tbody>
</table>
Memory Architecture: Decoders

N words => N select signals
Too many select signals

Decoder reduces # of select signals
\[ K = \log_2 N \]
Array-Structured Memory Architecture

Problem: ASPECT RATIO or HEIGHT >> WIDTH

Amplify swing to rail-to-rail amplitude
Selects appropriate word
Hierarchical Memory Architecture

Advantages:
1. Shorter wires within blocks
2. Block address activates only 1 block => power savings
Memory Timing: Definitions
Memory Timing: Approaches

DRAM Timing
Multiplexed Addressing

SRAM Timing
Self-timed
MOS NOR ROM

Pull-up devices

| WL[0] | | | |
| WL[1] | | | |
| WL[2] | | | |
| WL[3] | | | |

V_{DD}

GND

GND

Digital Integrated Circuits

Memory

© Prentice Hall 1995
MOS NOR ROM Layout

Only 1 layer (contact mask) is used to program memory array. Programming of the memory can be delayed to one of last process steps.
Threshold raising implants disable transistors
MOS NAND ROM

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

No contact to VDD or GND necessary; drastically reduced cell size
Loss in performance compared to NOR ROM
Equivalent Transient Model for MOS NOR ROM

Word line parasitics
- Resistance/cell: \((7/2) \times 10 \, \Omega/q = 35 \, \Omega\)
- Wire capacitance/cell: \((7\lambda \times 2\lambda) \times (0.6)^2 \times 0.058 + 2 \times (7\lambda \times 0.6) \times 0.043 = 0.65 \, \text{fF}\)
- Gate capacitance/cell: \((4\lambda \times 2\lambda) \times (0.6)^2 = 1.76 = 5.1 \, \text{fF}\).

Bit line parasitics:
- Resistance/cell: \((8.5/4) \times 0.07 \, \Omega/q = 0.15 \, \Omega\) (which is negligible)
- Wire capacitance/cell: \((8.5\lambda \times 4\lambda) \times (0.6)^2 \times 0.031 + 2 \times (8.5\lambda \times 0.6) \times 0.044 = 0.83 \, \text{fF}\)
- Drain capacitance/cell: \(((3\lambda \times 4\lambda) \times (0.6)^2 \times 0.3 + 2 \times 3\lambda \times 0.6 \times 0.8) \times 0.375 + 4\lambda \times 0.6 \times 0.43 = 2.6 \, \text{fF}\)
Equivalent Transient Model for MOS NAND ROM

Model for NAND ROM

Word line parasitics:
- Resistance/cell: \((6/2) \times 10 \ \Omega/q = 30 \ \Omega\)
- Wire capacitance/cell: \((6\lambda \times 2\lambda) (0.6)^2 \times 0.058 + 2 \times (6\lambda \times 0.6) \times 0.043 = 0.56 \ \text{fF}\)
- Gate Capacitance/cell: \((3\lambda \times 2\lambda) (0.6)^2 \times 1.76 = 3.8 \ \text{fF}\).

Bit line parasitics:
- Resistance/cell: \(\sim 10 \ \text{k}\Omega\), the average transistor resistance over the range of interest.
- Wire capacitance/cell: Included in diffusion capacitance
- Source/Drain capacitance/cell: \(((3\lambda \times 3\lambda) (0.6)^2 \times 0.3 + 2 \times 3\lambda \times 0.6 \times 0.8) \times 0.375 + (3\lambda \times 2\lambda) (0.6)^2 \times 1.76 = 5.2 \ \text{fF}\)
Propagation Delay of NOR ROM

Word line delay
Consider the $512 \times 512$ case. The delay of the distributed $rc$-line containing $M$ cells can be approximated using the expressions derived in Chapter 8.

$$t_{\text{word}} = 0.38 \left( r_{\text{word}} \times c_{\text{word}} \right) M^2 = 0.38 \left( 35 \ \Omega \times (0.65 + 5.1) \ \text{fF} \right) 512^2 = 20 \ \text{nsec}$$

Bit line delay
Assume a (2.4/1.2) pull-down device and a (8/1.2) pull-up transistor. The bit line switches between 5 V and 2.5 V.

$$C_{\text{bit}} = 512 \times (2.6 + 0.8) \ \text{fF} = 1.7 \ \text{pF}$$

$$I_{avHL} = \frac{1}{2} \left( \frac{2.4}{0.9} \right) \left( 19.6 \times 10^{-6} \right) \left( \frac{(4.25)^2}{2} + (4.25 \times 3.75 - (3.75)^2/2) \right) - \frac{1}{2} \left( \frac{8}{0.9} \right) \left( 5.3 \times 10^{-6} \right) \left( 4.25 \times 1.25 - (1.25)^2/2 \right) = 0.36 \ \text{mA}$$

$$t_{HL} = \frac{(1.7 \ \text{pF} \times 1.25 \ \text{V})}{0.36 \ \text{mA}} = 5.9 \ \text{nsec}$$

The low-to-high response time can be computed using a similar approach.

$$t_{LH} = \frac{(1.7 \ \text{pF} \times 1.25 \ \text{V})}{0.36 \ \text{mA}} = 5.9 \ \text{nsec}$$
Decreasing Word Line Delay

(a) Driving the word line from both sides

(b) Using a metal bypass

(c) Use silicides
Precharged MOS NOR ROM

PMOS precharge device can be made as large as necessary, but clock driver becomes harder to design.
Floating-gate transistor (FAMOS)

(a) Device cross-section

(b) Schematic symbol
Floating-Gate Transistor Programming

Avalanche injection. Removing programming voltage leaves charge trapped. Programming results in higher $V_T$. 
FLOTOX EEPROM

(a) Flotox transistor

(b) Fowler-Nordheim $I-V$ characteristic

(c) EEPROM cell during a read operation
Flash EEPROM

Control gate

Floating gate

Thin tunneling oxide

$\text{erasure}$

$\text{programming}$

$p^+$ source

$p$-substrate

$n^+$ drain

$n^+$ source
Cross-sections of NVM cells

Flash

Courtesy Intel

EPROM

Digital Integrated Circuits

Memory

© Prentice Hall 1995
## Characteristics of State-of-the-art NVM

<table>
<thead>
<tr>
<th></th>
<th>EPROM [Tomita91]</th>
<th>EEPROM [Terada89, Pashley89]</th>
<th>Flash EEPROM [Jinbo92]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory size</td>
<td>16 Mbit (0.6 μm)</td>
<td>1 Mbit (0.8 μm)</td>
<td>16 Mbit (0.6 μm)</td>
</tr>
<tr>
<td>Chip size</td>
<td>7.18 x 17.39 mm²</td>
<td>11.8 x 7.7 mm²</td>
<td>6.3 x 18.5 mm²</td>
</tr>
<tr>
<td>Cell size</td>
<td>3.8 μm²</td>
<td>30 μm²</td>
<td>3.4 μm²</td>
</tr>
<tr>
<td>Access time</td>
<td>62 nsec</td>
<td>120 nsec</td>
<td>58 nsec</td>
</tr>
<tr>
<td>Erasure time</td>
<td>minutes</td>
<td>N.A.</td>
<td>4 sec</td>
</tr>
<tr>
<td>Programming time/word</td>
<td>5 μsec</td>
<td>8 msec/word, 4 sec/chip</td>
<td>5 μsec</td>
</tr>
<tr>
<td>Erase/Write cycles</td>
<td>100</td>
<td>$10^5$</td>
<td>$10^3$-$10^5$</td>
</tr>
</tbody>
</table>

Digital Integrated Circuits  Memory  © Prentice Hall 1995
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 (Write)

\[
\begin{align*}
&k_{n,M6}\left(\frac{V_{DD} - V_{Tn}}{2}\right) - \frac{V_{DD}^2}{8} = k_{p,M4}\left(\frac{V_{DD} - |V_{Tp}|}{2}\right) - \frac{V_{DD}^2}{8} \\
&\frac{k_{n,M5}}{2}\left(\frac{V_{DD}}{2} - V_{Tn}\left(\frac{V_{DD}}{2}\right)^2\right) = k_{n,M1}\left(\frac{V_{DD} - |V_{Tn}|}{2}\right) - \frac{V_{DD}^2}{8} \\
&\frac{(W/L)_{n,M6}}{(W/L)_{p,M4}} \geq 0.33 \\
&(W/L)_{n,M5} \geq 10 (W/L)_{n,M1}
\end{align*}
\]
CMOS SRAM Analysis (Read)

\[
\frac{k_{n,M5}}{2} \left( \frac{V_{DD}}{2} - V_{Tn} \left( \frac{V_{DD}}{2} \right) \right)^2 = k_{n,M1} \left( \left( V_{DD} - |V_{Tn}| \right) \frac{V_{DD}}{2} - \frac{V_{DD}^2}{8} \right)
\]

\[(W/L)_{n,M5} \leq 10 (W/L)_{n,M1} \quad \text{(supercedes read constraint)}\]
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
No constraints on device ratios
Reads are non-destructive
Value stored at node X when writing a “1” = $V_{WWL} - V_{Tn}$
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 $V_{DD}$. 
1-T DRAM Cell

(a) Cross-section

(b) Layout

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

Trench Cell

Stacked-capacitor 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**

$$WL_0 = A_0A_1A_2A_3A_4A_5A_6A_7A_8A_9$$

$$WL_{511} = \overline{A_0A_1A_2A_3A_4A_5A_6A_7A_8A_9}$$

**NOR Decoder**

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

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

Dynamic 2-to-4 NOR decoder

2-to-4 MOS dynamic NAND Decoder

Propagation delay is primary concern
A NAND decoder using 2-input pre-decoders

Splitting decoder into two or more logic layers produces a faster and cheaper implementation.
4 input pass-transistor based column decoder

Advantage: speed (\(t_{pd}\) does not add to overall memory access time)

Disadvantage: large transistor count

Only 1 extra transistor in signal path

advantage: 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

![Diagram of sense amplifier](image)
Differential Sensing - SRAM

(a) SRAM sensing scheme.

(b) Doubled-ended Current Mirror Amplifier

(c) Cross-Coupled Amplifier
Latch-Based Sense Amplifier

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.
Single-to-Differential Conversion

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

(a) reading a zero

(b) reading a one

(c) control signals
Single-Ended Cascode Amplifier
DRAM Timing
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

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

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

Increase Yield using Error Correction and Redundancy
Open Bit-line Architecture — Cross Coupling
Folded-Bitline Architecture
Transposed-Bitline Architecture

(a) Straightforward bitline routing.

(b) Transposed bitline architecture.
Alpha-particles

1 particle ~ 1 million carriers
Yield

Yield curves at different stages of process maturity
(from [Veendrick92])
Redundancy

- Redundant rows
- Redundant columns

Memory Array

Row Address

Row Decoder

Column Decoder

Column Address

Fuse Bank
Redundancy and Error Correction
Programmable Logic Array

Product Terms

AND
PLANE

OR
PLANE

$X_0X_1$

$X_2$

$f_0$

$f_1$
Pseudo-Static PLA
Dynamic PLA
Clock Signal Generation for self-timed dynamic PLA

(a) Clock signals

(b) Timing generation circuitry.
PLA Layout

- **V_{DD}**
- **And-Plane**
- **Or-Plane**
- **GND**

**Pull-up devices**

\[ x_0 \quad x_0 \quad x_1 \quad x_1 \quad x_2 \quad x_2 \]

\[ f_0 \quad f_1 \]
PLA versus ROM

Programmable Logic Array
structured approach to random logic
“two level logic implementation”
NOR-NOR (product of sums)
NAND-NAND (sum of products)

IDENTICAL TO ROM!

Main difference
ROM: fully populated
PLA: one element per minterm

Note: Importance of PLA’s has drastically reduced
1. slow
2. better software techniques (multi-level logic synthesis)
Semiconductor Memory Trends

Memory Size as a function of time: x 4 every three years
Semiconductor Memory Trends

- Increasing die size factor 1.5 per generation
- Combined with reducing cell size factor 2.6 per generation
Semiconductor Memory Trends

Technology feature size for different SRAM generations