Semiconductor Memories
Increasing number of transistors in uprocessors are devoted to cache memories….more than 60%, see ITRS for more details…..

High-performance workstations and desktops have several Gbytes of memory

Audio (MP3) and Video players (MPEG4) requires large amount of memory

Can we store Memory using registers? …..yes but the area required will be excessive (need > 10 transistors/bit)

Memory cells are therefore combined into large arrays, which minimizes the overhead caused by the peripheral circuits and increases storage density

Memory design can be classified as high-performance, high density, low-power circuit design
Memory Classification

- Size
- Timing Parameters
- Function
- Access Pattern
Memory Size

- Depends on the level of abstraction
- **Bits**: (used by circuit designers) are equivalent to the number of individual cells (FFs or Registers) to store data
- **Bytes**: (used by chip designers) are groups of 8 or 9 bits or their multiples: Kbyte, Mbyte, Gbyte, Tbyte
- **Words**: (used by system designers) represent a basic computational entity. For example, a group of 32 bits represent a word in a computer that operates on 32 bit data
Timing Parameters

- **READ-Access Time**: time it takes to retrieve (read) from the memory. This is equal to the delay between the read request and the moment the data becomes available at the O/P.

- **WRITE-Access Time**: time elapsed between a write request and the final writing of the input data into the memory.

- **CYCLE Time**: minimum time required between successive reads or writes.
Memory Timing: Definitions

- **Write cycle**
  - **Write access**
  - **Data written**

- **Read cycle**
  - **Read access**
  - **Data valid**

- **DATA**
  - **WRITE**
  - **READ**
Function

- **Read-Only Memory (ROM):**
  - encode the information into the circuit topology-by removing or adding transistors. The topology is hard wired and the data cannot be modified.
  - They belong to the class of **Non-volatile** memories. Disconnection of the supply voltage does not result in a loss of the stored data.

- **Read-Write Memories (RWM):** called as **RAM** (Random-Access Memories).
  - **Static** (retains data if Vdd is retained): example SRAM
  - **Dynamic** (needs periodic refreshing): example DRAM
  - They use active circuitry to store information and belong to the class of **Volatile** memories.
Function....cont’d

- Non-Volatile Read-Write (NVRWM):
  - Recent Non-Volatile Memories can read and write---although write function is substantially slower
  - Fastest growing among semiconductor memories

- Examples:
  - EPROM: Electrically Programmable ROM
  - E²PROM: Electrically Erasable and Programmable ROM
  - Flash memory
Access Pattern

- **Random-Access (RAM):**
  - Memory locations can be read or written in a random manner
  - Most ROMs and NVRWMs allow random access….but “RAM” is used for the RWMs only

- **Serial Access:**
  - Restricts the order of access. Results in faster access times, smaller area, or allows special functionality
  - Examples: (Video Memories)
    - FIFO (first-in first-out)
    - LIFO (last-in first-out)
    - Shift Register

- **Content-Addressable Memory (CAM):**
  - Also known as associative memory
  - Doesn’t use an address to locate the data
  - Uses a word of data itself as input…when input data matches a data word stored in memory array, a MATCH flag is raised
  - Important component of the cache architecture of most microprocessors
# 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></td>
</tr>
<tr>
<td>SRAM</td>
<td>FIFO</td>
<td>EPROM</td>
</tr>
<tr>
<td>DRAM</td>
<td>LIFO</td>
<td>E²PROM</td>
</tr>
<tr>
<td></td>
<td>Shift Register</td>
<td>FLASH</td>
</tr>
<tr>
<td></td>
<td>CAM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mask-Programmed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Programmable (PROM)</td>
</tr>
</tbody>
</table>
More Classification

- **I/O Architecture:**
  - Based on the number of data input and output ports
  - Most memories use a single I/O port
  - *Multiport memories* offer higher bandwidth
    - Example: register files used in RISC processors
    - Adds more complexity to the design

- **Application:**
  - Embedded Memories in SoCs
  - For massive storage (multiples of Gbytes and beyond), more cost effective solutions are to use *magnetic tapes* and *optical disks*—they however, tend to be slower and provide limited access pattern
Semiconductor Memory Trends (up to the 90's)

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

From [Itoh01]
Trends in Memory Cell Area

From [Itoh01]
Semiconductor Memory Trends

Technology feature size for different SRAM generations
Memory Architecture: Decoders

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

K = \log_2 N
Decoder Basic

- Recall that a decoder is a combinational circuit with \( k \) inputs and at most \( 2^k \) outputs.
- Its characteristics property is that for every combination of input values only ONE output =1 at the same time.
- Used to route input data to specific output line.

For example: for \( a=b=c=0 \), only \( S0 =1 \).

\[
3 = \log_2 8
\]

\[
S0 = a'b'c' \\
S1 = a'b'c \\
S2 = a'bc' \\
S3 = a'bc \\
S4 = abc' \\
S5 = ab'c \\
S6 = abc' \\
S7 = abc
\]
Array-Structured Memory Architecture

Problem: consider ~1 million \((N=2^{20})\) 8-bit \((M=2^3)\) words, ASPECT RATIO or HEIGHT >> WIDTH

Peripheral circuitry is needed to recover the desired digital signal properties

Make vertical and horizontal dimensions of the same order of magnitude
Store multiple words in one row
Use a column decoder to select the correct word

Amplify swing to rail-to-rail amplitude
For interfacing to the external world
Selects appropriate word

OK for 64 Kbits to 256 Kbits beyond which speed degrades as length, \(C\), and \(R\) of word/bit lines increase excessively

\(K = \log_2 N\)
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
READ Operation:
Assume 1 is stored at Q
Assume both BLs are held high before the read.
Read cycle started by asserting the WL, enables PTs M5 and M6
During a correct read operation values stored in Q and $\overline{Q}$ are transferred to the bit lines leaving BL at its precharge value and by discharging $\overline{BL}$ through M1-M5
A “0” can be read in a similar manner (now BL gets discharged through M6 and M3)

**Major advantage of dual BL:** Q is clamped to Vdd by BL and prevents any inadvertent toggling of the INV pair

SRAM cell should be as small as possible.....but reliable operation requires careful sizing...
CMOS SRAM Analysis (Read “1” operation)

Transistor sizing is needed to avoid writing 1 accidentally, i.e., voltage at $\overline{Q}$ becomes $> V_M$ of Inv M3-M4

$\overline{Q}$ must stay low enough so that there is no substantial current through M3-M4 INV

M1 must be stronger than M5

As difference between BL and BLB builds up, the sense amp. is activated to accelerate the reading process

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

Value of the ripple voltage

$CR = \text{cell ratio} = \frac{M1}{M5}$

$Lecture 14, ECE 225$
CMOS SRAM Analysis (Read)

Node voltage must stay below the Vth of M3: CR must be >1.2

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

Choose M5 to be minimum size and M1 > M5
Assume that $Q=1$

To write a 0 in the cell: set $BL=1$ and $BL=0$

Similar to applying a reset pulse to an SR latch. FF will change state if sized properly.

$\overline{Q}$ cannot be pulled high due to the sizing of $M5$ and $M1$ already done for reading.

New value must be written through $M6$

Reliable writing of the cell is ensured if we can pull node $Q$ low enough—below the $V$'th of $M1$.

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

(M6 in linear)

(M4 in saturation)

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

$PR = \text{pull-up ratio of cell} = M4/M6$
CMOS SRAM Analysis (Write)

PR between the PMOS (M4) pull-up and the NMOS (M6) Pass Transistor must be < 1.8 to keep Vtn < 0.4 V

\[ PR = \frac{W_4}{W_6} \frac{L_4}{L_6} \]

Should be low to keep \( V_Q \) low

0.25um CMOS
Performance of SRAM

- Read operation is more critical. It requires discharging of the large bit line capacitance through the stack of 2 transistors (M1-M5)
- Write time is dominated by the propagation delay of the cross-coupled inverter pair, since the drivers that set BL and BL can be large
- Sense amplifiers used to accelerate Read time….as the difference between BL and BL builds up, sense amplifier is activated, and it discharges one of the bit lines
Sense Amp Operation

![Diagram](image)

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

- Sense amp activated
- Word line activated

$Lecture 14, ECE 225$
6T-SRAM — Layout

6T SRAM
Takes significant area...the two PMOS need n-wells
Resistive-load (4T) SRAM Cell

Reduce area using resistive load inverters

$R_L$ must maintain the state of the cell, that is compensate for the leakage currents ($\sim 10^{-15} \text{A}$)

$I_{\text{load}} > 10^{-13} \text{A}$ to compensate for leakage—puts an upper limit on $R_L$

Replacing the PMOSs by resistors reduces wiring

SRAM cell size reduced by 1/3

Static power dissipation -- Want $R_L$ large (use undoped poly)

Bit lines precharged to $V_{DD}$ to address $t_p$ problem
SRAM Characteristics

Instead of PMOS devices, use parasitic devices on top of cell structure using thin-film transistors (TFTs).

However, embedded SRAM cells---used in uprocessor caches, employ 6T cells.
Hierarchical Memory Architecture

For Larger Memories....

Advantages:
1. Shorter wires within blocks: faster access times
2. Block address activates only 1 block => power savings

Control circuitry

Block selector
Enables a single memory block at a time

Global amplifier/driver

I/O

Global data bus

Lecture 14, ECE 225
Kaustav Banerjee
Block Diagram of 4 Mbit SRAM

32 blocks, each containing 128 Kbits
Each block is structured as an array of 1024 rows and 128 columns

[Hirose90]
3-Transistor DRAM Cell (Early Days)

Periodic refresh operation consists of read followed by write.

1 Kbit memory: Intel
Still used in some ASICs

Dynamic: since it involves charge storage on a capacitor

Cell is written by placing value on BL1 and asserting Write Word Line (WWL=1)
Data retained as charge stored on Cs once WWL=0
For reading the cell, RWL=1
M2 can be on or off depending on stored value
BL2 is either clamped to Vdd or is precharged to either Vdd or Vdd-Vt
M2-M3 pulls BL2 low when X=1, otherwise BL2 remains high (cell is inverting: senses the inverse value of the stored signal)
Unlike SRAM, no constraint on device sizes

Read operation is non-destructive

No special process steps needed

Value at node $X = V_{WWL} - V_{tn}$

This reduces the current through $M2$ during read operation and increases read access time: can use a higher value of $V_{WWL}$ to avoid this.
**1-Transistor DRAM Cell**

Write: Place data on BL and assert WL, depending on data value Cs is 1 or 0

Read: before read, precharge BL to $V_{PRE}$

After WL=1, charge redistribution takes places between bit line and storage capacitance resulting in a voltage change on BL

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

$V_{BIT}$ is initial voltage on $C_s$. $V_{BL}$ is final voltage on BL after charge redistribution.

Voltage swing is small since $C_s \ll C_{BL}$; typically around 250 mV.

Lecture 14, ECE 225

Kaustav Banerjee
**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

Cross-section

Layout

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

Trench Cell

Stacked-capacitor Cell

Lecture 14, ECE 225

Kaustav Banerjee
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 $V_{DD}$.
The SNM can be estimated graphically by the length of the side of the square fitted between the VTCs and having the longest diagonal.

As noise increases at the two nodes above, the reverse VTC for INV1 moves upward, while the VTC for INV2 moves to the left.

Once they both move by the SNM value, the curves meet at only two points.....at A and B, and any further noise flips the data.
Static Noise Margin (SNM)

- **Hold Margin**: How strongly the node storing ‘1’ and the node storing ‘0’ are coupled to $V_{DD}$ and $V_{SS}$ respectively.
- **Read Margin**: The difference between $V_{TRIP}$ and $V_{READ}$.
- **Write Margin**: The maximum voltage on a bit-line that allows writing to the cell, while the other bit-line is at $V_{DD}$. (not determined by the butterfly curve)
SNM Dependencies

- Dependence on $V_{DD}$: SNM for a bitcell with ideal VTCs is still limited to $V_{DD}/2$
- Dependence on sizing
SNM Dependencies

- Dependence on random doping variation ($V_{TH}$ mismatch)
- Dependence on global variation

![Vth Mismatch](image1)

![Global Variation](image2)
3-D ICs: Extending Moore’s Law in the Z-Dimension, and more….

K. Banerjee et al., Proceedings of the IEEE, 2001
3-D ICs: Multiple Active Si Layers

- **Advantages**
  - Reduce Interconnect Length by Vertically Stacking Multiple Si Layers
  - Improve Chip Performance
  - Reduce Chip Area
  - Heterogeneous integration possible, e.g., memory, digital, analog, optical, etc.
Integrated SoC to Sense, Process, and Collaborate
3-D Technologies

Epitaxial Lateral Overgrowth

- ELO Islands
- 2nd Layer
- Oxide
- SEG
- SOI layer-1
- SOI layer-2
- Oxide
- Silicon substrate
- Grown Vertical connection of Islands
Solid Phase Crystallization of $\alpha$-Si

- Locally induce nucleation
- Grow laterally, inhibiting additional nucleation
- Build MOSFET in a single grain with SOI like performance
3-D Technologies

Wafer Bonding

Bonding using Glue Layer

Thermocompression Bonding using Cu Pads

Device Layer
Interconnect Layer
Gate
Via
Local Metal
p
n
p
p
ILD/Al
Thinned Si
Polymer
ILD/Al
Thinned Si
Cu
Inter-wafer via
Inter-wafer via