Introduction to SRAM

Jasur Hanbaba
Outline

- Memory Arrays
- SRAM Architecture
  - SRAM Cell
  - Decoders
  - Column Circuitry
- Non-volatile Memory
- Manufacturing Flow
Memory Arrays

Random Access Memory

- Read/Write Memory (RAM) (Volatile)
  - Static RAM (SRAM)
  - Dynamic RAM (DRAM)
- Read Only Memory (ROM) (Nonvolatile)
  - Serial In Parallel Out (SIPO)
  - Parallel In Serial Out (PISO)

Serial Access Memory

- Shift Registers
- Queues

Content Addressable Memory (CAM)

- First In First Out (FIFO)
- Last In First Out (LIFO)

- Electrically Erasable Programmable ROM (EEPROM)
- Flash ROM

- Mask ROM
- Programmable ROM (PROM)
- Erasable Programmable ROM (EPROM)
Memory Types:

- **Volatile:**
  - Random Access Memory (RAM):
    - SRAM "static"
    - DRAM "dynamic"

- **Non-volatile:**
  - Read Only Memory (ROM):
    - Mask ROM "mask programmable"
    - EPROM "electrically programmable"
    - EEPROM "erasable electrically programmable"
    - FLASH memory - similar to EEPROM with programmer integrated on chip
Why is memory design important

- Memory arrays often account for the majority of transistors in modern microprocessor designs. So it is critical to have a memory design that is efficient in terms of area and fast.

- SRAM (static random access memory) is the most widely used in processor design.
  - Simplicity: internal feedback loop that retain its value as long as power is applied.
  - Compatible with standard CMOS processes.
  - Faster than DRAM.
6T SRAM Cell

- Cell size accounts for most of array size
  - Reduce cell size at expense of complexity

- 6T SRAM Cell
  - Used in most commercial chips
  - Data stored in cross-coupled inverters

- Read:
  - Precharge bit, bit_b
  - Raise wordline

- Write:
  - Drive data onto bit, bit_b
  - Raise wordline
1-bit cell of the SRAM
SRAM Read

- Precharge both bitlines high
- Then turn on wordline
- One of the two bitlines will be pulled down by the cell
- Ex: $A = 0$, $A_b = 1$
  - bit discharges, bit_b stays high
  - But $A$ bumps up slightly
- **Read stability**
  - $A$ must not flip
SRAM Write

- Drive one bitline high, the other low
- Then turn on wordline
- Bitlines overpower cell with new value
- Ex: \( A = 0, \ A_b = 1, \) bit = 1, bit_b = 0
  - Force \( A_b \) low, then \( A \) rises high

**Writability**
- Must overpower feedback inverter
  - \( N2 >> P1 \)
SRAM Column Example

Read

Write

Bitline Conditioning

More Cells

SRAM Cell

More Cells

SRAM Cell

data_s1
SRAM Layout

- Tile cells sharing $V_{DD}$, GND, bitline contacts
Decoders
- $n:2^n$ decoder consists of $2^n$ n-input AND gates
  - One needed for each row of memory
  - Build AND from NAND or NOR gates

Static CMOS

Pseudo-nMOS
Large Decoders

- For $n > 4$, NAND gates become slow
  - Break large gates into multiple smaller gates

\[ \text{word0 word1 word2 word3 word15} \]

\[ \text{word0 word1 word2 word3 word15} \]
Many of these gates are redundant
- Factor out common gates into predecoder
- Saves area
- Same path effort
Column Circuitry

- Some circuitry is required for each column
  - Bitline conditioning
  - Sense amplifiers
  - Column multiplexing
Bitline Conditioning

- Precharge bitlines high before reads

- Equalize bitlines to minimize voltage difference when using sense amplifiers
Sense Amplifiers

- Bitlines have many cells attached
  - Ex: 32-kbit SRAM has 256 rows x 128 cols
  - 128 cells on each bitline

- \( t_{pd} \propto (C/I) \Delta V \)
  - Even with shared diffusion contacts, 64C of diffusion capacitance (big C)
  - Discharged slowly through small transistors (small I)

- Sense amplifiers are triggered on small voltage swing (reduce \( \Delta V \))
Differential Pair Amp

- Differential pair requires no clock
- But always dissipates static power
Clocked Sense Amp

- Clocked sense amp saves power
- Requires sense_clk after enough bitline swing
- Isolation transistors cut off large bitline capacitance

![Diagram of Clocked Sense Amp](image)
Column Multiplexing

- \(2^n\) words of \(2^m\) bits each. If \(n >> m\), fold by \(2^k\) into fewer rows of more columns.

- Ex: 2 kword x 16 folded into 256 rows x 128 columns
  - Must select 16 output bits from the 128 columns
  - Requires 16 8:1 column multiplexers

- Good regularity – easy to design

- Very high density if good cells are used
Tree Decoder Mux

- Column mux can use pass transistors
  - Use nMOS only, precharge outputs

- One design is to use k series transistors for $2^k:1$ mux
  - No external decoder logic needed

*Diagram showing a tree decoder mux circuit with inputs A0, A1, A2 and outputs B0 to B7, Y to Y.*

2/21/2011
Single Pass-Gate Mux

- Or eliminate series transistors with separate decoder
Non-volatile Memory

*Used to hold fixed code (ex. BIOS), tables of data (ex. FSM next state/output logic), slowly changing values that persist over power off (date/time)*

- **Mask ROM**
  - Used with logic circuits for tables etc.
  - Contents fixed at IC fab time (truly write once!)

- **EPROM (erasable programmable)** & **FLASH**
  - Requires special IC process (floating gate technology)
  - Writing is slower than RAM. EPROM uses special programming system to provide special voltages and timing.
  - Reading can be made fairly fast.
  - Rewriting is very slow.
  - Erasure is first required, EPROM - UV light exposure, EEPROM – electrically erasable

![Figure 2: Cell bias conditions during programming](image)
The “front end” = wafer fab, E-test, and sort. E-test and sort are also called “wafer test”.

In assembly, wafers are cut into dice and the dice are assembled into packages.

The “back end” = assembly, burn-in, class test and PPV.