Memory Subsystem Design

- DRAM Design
- Static RAM Design
- Row Decoders
- Sense Amplifiers
- Summary
Material from: *CMOS VLSI Design* By Weste and Harris &

*Principles of CMOS VLSI Design*

By Neil E. Weste and Kamran Eshraghian
Memory Arrays

- Random Access Memory
  - Read/Write Memory (RAM) (Volatile)
    - Static RAM (SRAM)
    - Mask ROM
  - Read Only Memory (ROM) (Nonvolatile)
    - Dynamic RAM (DRAM)
    - Programmable ROM (PROM)
    - Erasable Programmable ROM (EPROM)
    - Electrically Erasable Programmable ROM (EEPROM)
- Serial Access Memory
  - Shift Registers
    - Serial In Parallel Out (SIPO)
  - Queues
    - Parallel In Serial Out (PISO)
- Content Addressable Memory (CAM)
  - First In First Out (FIFO)
  - Last In First Out (LIFO)
- Mask ROM
- Programmable ROM (PROM)
- Flash ROM

Read/Write Memory
- (RAM)
- (Volatile)

Read Only Memory
- (ROM)
- (Nonvolatile)
Array Architecture

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

Good regularity – easy to design
Very high density if good cells are used
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*
**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*
  - $N1 >> N2$
**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 Sizing

- High bitlines must not overpower inverters during reads
- But low bitlines must write new value into cell
SRAM Layout

- Cell size is critical: $26 \times 45 \lambda$ (even smaller in industry)
- 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
Read-Only Memories

- Read-Only Memories are nonvolatile
  - *Retain their contents when power is removed*
- Mask-programmed ROMs use one transistor per bit
  - *Presence or absence determines 1 or 0*
**ROM Example**

- **4-word x 6-bit ROM**
  - *Represented with dot diagram*
  - *Dots indicate 1’s in ROM*

Word 0: 010101
Word 1: 011001
Word 2: 100101
Word 3: 101010

Looks like 6 4-input pseudo-nMOS NORs
ROM Array Layout

- Unit cell is 12 x 8 \( \lambda \) (about 1/10 size of SRAM)
PROMs and EPROMs

- Programmable ROMs
  - **Build array with transistors at every site**
  - **Burn out fuses to disable unwanted transistors**
- Electrically Programmable ROMs
  - **Use floating gate to turn off unwanted transistors**
  - **EPROM, EEPROM, Flash**
Building Logic with ROMs

- Use ROM as lookup table containing truth table
  - \( n \) inputs, \( k \) outputs requires \( 2^n \) words \( \times k \) bits
  - Changing function is easy – reprogram ROM

- Finite State Machine
  - \( n \) inputs, \( k \) outputs, \( s \) bits of state
  - Build with \( 2^{n+s} \) \( \times (k+s) \) bit ROM and \( (k+s) \) bit reg
**PLAs**

- A *Programmable Logic Array* performs any function in sum-of-products form.
- *Literals*: inputs & complements
- *Products / Minterms*: AND of literals
- *Outputs*: OR of Minterms

**Example: Full Adder**

\[
\begin{align*}
  s &= a\overline{b}\overline{c} + \overline{a}bc + \overline{a}\overline{b}c + abc \\
  c_{\text{out}} &= ab + bc + ac
\end{align*}
\]
**NOR-NOR PLAs**

- ANDs and ORs are not very efficient in CMOS
- Dynamic or Pseudo-nMOS NORs are very efficient
- Use DeMorgan’s Law to convert to all NORs
**PLA Schematic & Layout**

The image shows a schematic and layout diagram of a PLA (Programmable Logic Array). The diagram includes two planes: the AND plane and the OR plane. The inputs are labeled as $a$, $b$, and $c$, and the outputs are $s$ and $c_{out}$. The diagram also includes several lines of the form $bc$, $ac$, $ab$, $abc$, $\overline{abc}$, $\overline{ab}c$, and $\overline{abc}$.
**PLAs vs. ROMs**

- The OR plane of the PLA is like the ROM array
- The AND plane of the PLA is like the ROM decoder
- PLAs are more flexible than ROMs
  - *No need to have* $2^n$ *rows for* $n$ *inputs*
  - *Only generate the minterms that are needed*
  - *Take advantage of logic simplification*
Memory System Summary

- **DRAM Design**
- Static RAM Design
- Row Decoders
- Sense Amplifiers
Scaling

- **Constant Field Scaling by** \( \alpha \)
  - *All horizontal & vertical dimensions*
  - *Device voltages*
  - *Concentration densities*
- **Constant Voltage Scaling**
  - *Keep* \( V_{DD} \) *constant, scale the process*
- **Lateral Scaling**
  - *Scale gate length, all else is kept unchanged*
Scaling Problems

- **Constant Voltage Scaling** increases field across gate oxide – sometimes need *Lightly-Doped Drains*

- Depletion regions determine how small the channel can get
  - *Source-drain length > 2 depletion layer widths*
  - *Can reduce depletion layer width by increasing substrate doping*

- **Constant Field Scaling** – get:
  - \( \frac{I_{ds}}{\alpha} \)
  - *But \( \alpha^2 \) devices / per area*
  - *Therefore, current density goes up by \( \alpha \)*
Packages

• Package functions
  – *Electrical connection of signals and power from chip to board*
  – *Little delay or distortion*
  – *Mechanical connection of chip to board*
  – *Removes heat produced on chip*
  – *Protects chip from mechanical damage*
  – *Compatible with thermal expansion*
  – *Inexpensive to manufacture and test*
Package Types

- Through-hole vs. surface mount
Multichip Modules

- Pentium Pro MCM
  - Fast connection of CPU to cache
  - Expensive, requires known good dice
Chip-to-Package Bonding

- Traditionally, chip is surrounded by *pad frame*
  - *Metal pads on 100 – 200 μm pitch*
  - *Gold bond wires attach pads to package*
  - *Lead frame distributes signals in package*
  - *Metal heat spreader helps with cooling*
Interdigitated (a), Core-limited (b) or Pad-limited (c) Pads
Bump Pad

- Can be put anywhere on chip
  - Plate pads with solder bumps
  - Invert chip & reflow bond to substrate
- Invented by IBM:
  - No pad connection failure EVER in IBM equipment
Advanced Packages

- Bond wires contribute parasitic inductance
- Fancy packages have many signal, power layers
  - Like tiny printed circuit boards
- Flip-chip places connections across surface of die rather than around periphery
  - Top level metal pads covered with solder balls
  - Chip flips upside down
  - Carefully aligned to package (done blind!)
  - Heated to melt balls
  - Also called C4 (Controlled Collapse Chip Connection)
Package Parasitics

- Use many $V_{DD}$, GND in parallel
  - Inductance, $I_{DD}$
Heat Dissipation

- 60 W light bulb has surface area of 120 cm$^2$
- Itanium 2 die dissipates 130 W over 4 cm$^2$
  - Chips have enormous power densities
  - Cooling is a serious challenge
- Package spreads heat to larger surface area
  - Heat sinks may increase surface area further
  - Fans increase airflow rate over surface area
  - Liquid cooling used in extreme cases ($$$)
Input / Output

• Input/Output System functions
  – Communicate between chip and external world
  – Drive large capacitance off chip
  – Operate at compatible voltage levels
  – Provide adequate bandwidth
  – Limit slew rates to control di/dt noise
  – Protect chip against electrostatic discharge
  – Use small number of pins (low cost)
I/O Pad Design

- Pad types
  - $V_{DD}$ / GND
  - Output
  - Input
  - Bidirectional
  - Analog
Output Pads

- Drive large off-chip loads (2 – 50 pF)
  - *With suitable rise/fall times*
  - *Requires chain of successively larger buffers*

- Guard rings to protect against latchup
  - *Noise below GND injects charge into substrate*
  - *Large nMOS output transistor*
  - *p+ inner guard ring*
  - *n+ outer guard ring*
    - In n-well
**Input Pads**

- **Level conversion**
  - *Higher or lower off-chip V*
  - *May need thick oxide gates*

- **Noise filtering**
  - *Schmitt trigger*
  - *Hysteresis changes $V_{IH}$, $V_{IL}$*

- **Protection against electrostatic discharge**
ESD Protection

- Static electricity builds up on your body
  - *Shock delivered to a chip can fry thin gates*
  - *Must dissipate this energy in protection circuits before it reaches the gates*

- ESD protection circuits
  - *Current limiting resistor*
  - *Diode clamps*

- ESD testing
  - *Human body model*
  - *Views human as charged capacitor*
**Bidirectional Pads**

- Combine input and output pad
- Need tristate driver on output
  - *Use enable signal to set direction*
  - *Optimized tristate avoids huge series transistors*
Analog Pads

• Pass analog voltages directly in or out of chip
  – No buffering
  – Protection circuits must not distort voltages
MOSIS I/O Pad

- 1.6 \( \mu \text{m} \) two-metal process
  - Protection resistors
  - Protection diodes
  - Guard rings
  - Field oxide clamps