Chapter 7
Optimizing Power @ Design Time – Memory

Slide 7.1
Role of Memory in ICs

- Memory is very important
- Focus in this chapter is embedded memory
- Percentage of area going to memory is increasing

[Ref: V. De, Intel 2006]
Processor Area Becoming Memory Dominated

- On-chip SRAM contains 50–90% of total transistor count
  - Xeon: 48M/110M
  - Itanium 2: 144M/220M
- SRAM is a major source of chip static power dissipation
  - Dominant in ultra low-power applications
  - Substantial fraction in others

*Intel Penryn™* (Picture courtesy of Intel)
Chapter Outline

- Introduction to Memory Architectures
- Power in the Cell Array
- Power for Read Access
- **Power** for Write Access
- New Memory Technologies

Slide 7.4
SRAM Metrics

- Functionality
  - Data retention
  - Readability
  - Writability
  - Soft Errors
- Area
- Power

Why is functionality a "metric"?

- Process variations increase with scaling
- Large number of cells requires analysis of tails (out to 6σ or 7σ)
- Within-die $V_{TH}$ variation due to Random Dopant Fluctuations (RDFs)
Where Does SRAM Power Go?

- Numerous analytical SRAM power models
- Great variety in power breakdowns
- Different applications cause different components of power to dominate
- Hence: Depends on applications: e.g., high speed versus low power, portable
SRAM cell

Three tasks of a cell

- Hold data
  - WL = 0; BLs = X

- Write
  - WL = 1; BLs driven with new data

- Read
  - WL = 1; BLs precharged and left floating

Traditional 6-Transistor (6T) SRAM cell

Slide 7.8
Key SRAM cell metrics

Key functionality metrics
- Hold
  - Static Noise Margin (SNM)
  - Data retention voltage (DRV)
- Read
  - Static Noise Margin (SNM)
- Write
  - Write Margin

Metrics:
Area is primary constraint
Next, Power, Delay

Traditional 6-Transistor (6T) SRAM cell
Static Noise Margin (SNM)

SNM gives a measure of the cell's stability by quantifying the DC noise required to flip the cell.

SNM is the length of the side of the largest embedded square on the butterfly curve.

[Ref: E. Seevinck, JSSC'87]

Slide 7.10
Static Noise Margin with Scaling

- Typical cell SNM deteriorates with scaling
- Variations lead to failure from insufficient SNM

Variations worsen tail of SNM distribution

(Results obtained from simulations with Predictive Technology Models – Ref: PTM, Y. Cao ‘00)
Variability: Write Margin

- Positive SNM
- Negative "SNM"

Cell stability prior to write:

Successful write:

Write failure:

Normalized Q vs. Normalized GB

Slide 7.12
Variability: Cell Writability

Write margin limits $V_{DD}$ scaling for 6T cells to 600 mV, best case.

- 65 nm process, $V_{DD} = 0.6$ V
- Variability and large number of cells makes this worse
Leakage Power dominates while the memory holds data.

Importance of Gate tunneling and GIDL depends on technology and voltages applied.

Sub-threshold leakage
Using Threshold Voltage to Reduce Leakage

- High-$V_{TH}$ cells necessary if all else is kept the same
- To keep leakage in 1 MB memory within bounds, $V_{TH}$ must be kept in 0.4–0.6 V range

[Ref: K. Itoh, ISCAS'06]
Multiple Threshold Voltages

Dual $V_{TH}$ cells with low-$V_{TH}$ access transistors provide good tradeoffs in power and delay

[Ref: Hamzaoglu, et al., TVLSI'02]

Use high-$V_{TH}$ devices to lower leakage for stored '0', which is much more common than a stored '1'.

[Ref: N. Azizi, TVLSI'03]

Slide 7.16
Multiple Voltages

- Selective usage of multiple voltages in cell array
  - e.g., 16 fA/cell at 25°C in 0.13 μm technology

- High $V_{TH}$ to lower sub-$V_{TH}$ leakage
- Raised source, raised $V_{DD}$, and lower BL reduce gate stress while maintaining SNM

[Ref: K. Osada, JSSC’03]
Power Breakdown During Read

- Accessing correct cell
  Decoders, WL drivers
  - For Lower Power:
    • hierarchical WLs
    • pulsed decoders

- Performing read
  - Charge and discharge large BL capacitance
  - For Lower Power:
    • SAs and low BL swing
    • Hierarchical BLs
    • Lower BL precharge
  • Lower $V_{DD}$
    - May require read assist

Slide 7.18
Hierarchical Wordline Architecture

- Reduces amount of switched capacitance
- Saves power and lowers delay

[Ref's: Rabaey, Prentice'03; T. Hirose, JSSC’90]
Hierarchical Bitlines

- Divide up bitlines hierarchically
  - Many variants possible
- Reduces $RC$ delay, also decreases $CV^2$ power
- Lower BL leakage seen by accessed cell

Slide 7.20
BL Leakage During Read Access

- Leakage into non-accessed cells
  - Raises power and delay
  - Affects BL differential

Bit-line

Slide 7.21
Bitline Leakage Solutions

- Hierarchical BLs
- Raise $V_{SS}$ in cell
- Negative WL voltage
- Longer access FETs
- Alternative bit-cells
- Active compensation
- Lower BL precharge voltage

[Ref: A. Agarwal, JSSC'03]
Lower Precharge Voltage

Lower BL precharge voltage decreases power and improves Read SNM
- Internal bit-cell node rises less
- Sharp limit due to accidental cell writing if access FET pulls internal ‘1’ low

Slide 7.23
- Lower $V_{DD}$ (and other voltages) via classic voltage scaling
  - Saves power
  - Increases delay
  - Limited by lost margin (read and write)
- Recover Read SNM with read assist
  - Lower BL precharge
  - Boosted cell $V_{DD}$ [Ref: Bhavnagarwala’04, Zhang’06]
  - Pulsed WL and/or Write-after-Read [Ref: Khellah’06]
  - Lower WL [Ref: Ohbayashi’06]
Power Breakdown During Write

- Accessing cell
  - Similar to Read
  - For Lower Power:
    - Hierarchical WLs

- Performing write
  - Traditionally drive BLs full swing
  - For Lower Power:
    - Charge sharing
    - Data dependencies
    - Low swing BLs with amplification

Slide 7.25
Charge recycling to reduce write power

- Share charge between BLs or pairs of BLs
- Saves for consecutive write operations
- Need to assess overhead

Basic charge recycling – saves 50% power in theory

[Ref's: K. Mai, JSSC'98; G. Ming, ASIC'05]
Memory Statistics

- 0’s more common
  - SPEC2000: 90% 0s in data
  - SPEC2000: 85% 0s in instructions
- Assumed write value using inverted data as necessary [Ref: Y. Chang, ISLPED’99]
- New Bitcell:

  1R, 1W port
  W0: WZ = 0, WWL = 1, WS = 1
  W1: WZ = 1, WWL = 1, WS = 0

[Ref: Y. Chang, TVLSI’04]

Slide 7.27
Low-Swing Write

- Drive the BLs with low swing
- Use amplification in cell to restore values

\[ V_{\text{DD}, \text{Prech}} = V_{\text{DD}} - V_{\text{TH}} - \Delta V_{\text{BL}} \]

[Ref: K. Kanda, JSSC'04]
Write Margin

- Fundamental limit to most power-reducing techniques
- Recover write margin with write assist, e.g.,
  - Boosted WL
  - Collapsed cell $V_{DD}$ [Itoh’96, Bhavnagarwala’04]
  - Raised cell $V_{SS}$ [Yamaoka’04, Kanda’04]
  - Cell with amplification [Kanda’04]
Non-traditional cells

- Key tradeoff is with functional robustness
- Use alternative cell to improve robustness, then trade off for power savings
- e.g. Remove read SNM

8T SRAM cell

- Register file cell
- 1R, 1W port
- Read SNM eliminated
- Allows lower $V_{DD}$
- 30% area overhead
- Robust layout

[Ref: L. Chang, VLSI'05]
Cells with Pseudo-Static SNM Removal

- Isolate stored data during read
- Dynamic storage for duration of read

Differential read
[Ref: S. Kosonocky, ISCICT'06]

Single-ended read
[Ref: K. Takeda, JSSC'06]
Emerging Devices: Double-gate MOSFET

- Emerging devices allow new SRAM structures
- Back-gate biasing of thin-body MOSFET provides improved control of short-channel effects, and re-instates effective dynamic control of $V_{TH}$.

**Double-gated (DG) MOSFET**
- Gate length = $L_g$
- Width = $T_{Si}$
- Fin Height $H_{TH} = W/2$

**Back-gated (BG) MOSFET**
- Independent front and back gates
- One switching gate and $V_{TH}$ control gate

[Ref: Z. Guo, ISLPED'05]
6T SRAM Cell with Feedback

- Double-Gated (DG) NMOS pull-down and PMOS load devices
- Back-Gated (BG) NMOS access devices dynamically increase $\beta$-ratio
  - SNM during read ~300 mV
  - Area penalty ~ 19%

[Ref: Z. Guo, ISLPED’05]
Summary and Perspectives

- Functionality is main constraint in SRAM
  - Variation makes the outlying cells limiters
  - Look at hold, read, write modes
- Use various methods to improve robustness, then trade off for power savings
  - Cell voltages, thresholds
  - Novel bit-cells
  - Emerging devices
- Embedded memory major threat to continued technology scaling – innovative solutions necessary
### References

**Books and Book Chapters**

**Articles**
References (cont.)

References (cont.)


Slide 7.37