A Combined DOE-ILP Based Power and Read Stability Optimization in Nano-CMOS SRAM

G. Thakral, S. P. Mohanty and D. Ghai
Dept. of Comp. Science & Engineering
University of North Texas, USA.
Email: saraju.mohanty@unt.edu

Dhiraj K. Pradhan
Dept. of Computer Science
University of Bristol, UK.
Email: pradhan@compsci.bristol.ac.uk

Acknowledgment: This research is supported in part by NSF award numbers CCF-0702361 and CNS-0854182.
Outline of the talk

- Introduction
- Problem Statement: How to decrease power while maintaining performance of SRAM?
- Solutions: Assigning high/low $V_{th}$ to transistors
- Proposed Optimal SRAM Design Flows
- Experimental Results: Nominal an Monte Carlo
- Related Prior Research
- Conclusions and Future Research
Why Efficient SRAM Design?

- Amount of on-die caches increases
- Up to 60% of the die area is devoted for caches in typical processor and embedded application.
- Largely contributes for leakage and power density.

<table>
<thead>
<tr>
<th>Technology</th>
<th>L1 Cache</th>
<th>L2 Cache</th>
<th>L3 Cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>180nm</td>
<td>16KB-I</td>
<td>256KB-I</td>
<td>1.5-25 MB</td>
</tr>
<tr>
<td>130nm</td>
<td>16KB-D</td>
<td>256KB-D</td>
<td></td>
</tr>
<tr>
<td>90nm</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Itanium 2* (L3-9MB) 130nm Technology

SRAM Power Density

SRAM Leakage Power Percentage
Issues in Nano CMOS

- Power
- Leakage
- Performance
- Parasitic
- Delay
- Yield
- Thermal

Nano CMOS

1/6/2010 Mohanty 4
Nano-CMOS SRAM Design Challenges ...

In nano-CMOS regime following are the major issues:

- Data stability and functionality
  - Non-destructive read
  - Successful write
  - Noise sensitivity
- Proper sizing of the transistors
  - To improve the write ability
  - To improve the read stability
  - To improve the data retention
- Minimum size of transistors to maximize the memory density.
- Minimum leakage for low-power design.
- Minimum read access time to improve the performance.
For proper read stability: N1 and N2 are sized wider than N3 and N4.
For successful write: N3 and N4 are sized wider than P1 and P2.
Minimum sized transistors do not provide good stability and functionality.
SRAM cell ratio ($\beta$): ratio of driver transistor’s W/L to access transistor’s W/L.
Single-Ended 7-Transistor SRAM

Highlights of this SRAM:

- Single-ended I/O latch style 7-transistor SRAM.
- Functions in ultra-low voltage regime allowing subthreshold operation.
- Better read stability, better write-ability compared to standard SRAM.
- Improved nanoscale process variation tolerance compared to the standard 6-transistor SRAM.

Load transistors – 2, 4
Driver transistors – 3, 5
Access transistors – 1, 6, 7

Source: Our publication in SOCC 2008
Research Question

How to reduce power dissipation while maintaining/enhancing stability of SRAM.
The Solution Explored in This Paper

- To reduce the power consumption this research investigates the process level technique, called dual-$V_{th}$.
- Important is the selection of appropriate transistors for high-$V_{th}$ assignment so that performance of SRAM is not degraded.
- SRAM is subjected to the dual-$V_{th}$ assignment using a novel combines Design of Experiments-Integer Linear Programming (DOE-ILP) algorithms.
Stability Analysis of SRAM: SNM

Static Noise Margin (SNM): It is the amount of maximum DC voltage (Vn) in this case, that SRAM can tolerate.

Butterfly curve for baseline SRAM.

Noise model for stability analysis
Currents in 7-Transistor SRAM: Write

Current Path for Write ‘1’

Current Path for Write ‘0’
Currents in 7-Transistor SRAM: Read

Current Path for Read ‘1’

Current Path for Read ‘0’

Parameters | Value  
|-----------|--------
| $P_{sram}$ | 203.6 nW  
| $SNM_{sram}$ | 170 mV  

1/6/2010
Combined DOE-ILP Approach

- **Design of Experiments (DOE)** consists of purposeful changes of inputs (factors) to a process in order to observe the corresponding changes in the outputs (responses).

- **Integer linear programming (ILP)** is a technique for optimization of a linear objective function, subject to linear equality and linear inequality constraints. ILP determines the way to achieve the best outcome (such as maximum profit or lowest cost) in a given mathematical model and given some list of requirements represented as linear equations.
Combined DOE-ILP Approach: Solution 1

1: Input: Baseline circuit, Nominal/High $V_{th}$ models.
2: Output: Objectives set $S_{OBJ} = [f_{PWR}, f_{SNM}]$ with transistors identified for high $V_{th}$ assignment.
3: Setup experiment for transistors of SRAM cell using 2-Level Taguchi L-8 array, where the factors are the transistors and the responses are average $P_{sram}$ and read $SNM_{ram}$.
4: for Each 1:8 experiment of 2-Level Taguchi L-8 array do
5: Perform simulation and record $P_{sram}$ and $SNM_{ram}$.
6: end for
7: Form predictive equations $\hat{f}_{PWR}$ for power, $\hat{f}_{SNM}$ for SNM.
8: Solve $\hat{f}_{PWR}$ using ILP. Solution set: $S_{PWR}$.
9: Solve $\hat{f}_{SNM}$ using ILP. Solution set: $S_{SNM}$.
10: Form $S_{OBJ} = S_{PWR} \cap S_{SNM}$.
11: Assign high $V_{th}$ to transistors based on $S_{OBJ}$.

Algorithm -1
DOE Predictive Equations

\[
\hat{f} = \bar{f} + \sum_{n=1}^{7} \left( \frac{\Delta(n)}{2} \times x_n \right),
\]

Where:

- \( x_n \) is the \( V_{Th} \) state of transistor of nth transistor;
- \( \hat{f} \) is the response of the transistor; (e.g. Power, SNM)
- \( \left( \frac{\Delta(n)}{2} \right) \) is the half-effect of the nth transistor;
Combined DOE-ILP Approach: Solution 2

1: Input: Baseline circuit, Nominal/Hgh - $V_{Th}$ models.
2: Output: Objective set $S_{OBJ} = [f_{PWR}, f_{SNM}]$ with transistors identified for high $V_{Th}$ assignment.
3: Setup experiment for transistors of SRAM cell using 2-Level Taguchi L-8 array, where the factors are the transistors and the responses are average $P_{sram}$ and read SNM$_{sram}$.
4: for Each $1:8$ experiments of 2-Level Taguchi L-8 array do
5: Perform simulations and record $P_{sram}$ and SNM$_{sram}$.
6: end for
7: Form normalized predictive equations: $\hat{f}_{PWR}$* and $\hat{f}_{SNM}$*.
8: Form $\hat{f}_{OBJ} = \left( \frac{\hat{f}_{PWR}^*}{\hat{f}_{SNM}^*} \right)$.
9: Solve $\hat{f}_{OBJ}^*$ using ILP. Solution set: $S_{OBJ}^*$.
10: Assign high $V_{Th}$ to transistors based on $S_{OBJ}^*$.

Algorithm - 2
Selection of Appropriate Transistors

Configuration for flow 1

Configuration for flow 2
# Experimental Results: 4 Alternatives

<table>
<thead>
<tr>
<th>Design Alternative</th>
<th>Parameter</th>
<th>Value</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>$P_{sram}$</td>
<td>203.6 nW</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$SNM_{sram}$</td>
<td>170mV</td>
<td>-</td>
</tr>
<tr>
<td>$SPWR$</td>
<td>$P_{sram}$</td>
<td>26.34 nW</td>
<td>87.1% decrease</td>
</tr>
<tr>
<td></td>
<td>$SNM_{sram}$</td>
<td>231.9 mV</td>
<td>26.7% increase</td>
</tr>
<tr>
<td>$SSNM$</td>
<td>$P_{sram}$</td>
<td>113.6 nW</td>
<td>44.2% decrease</td>
</tr>
<tr>
<td></td>
<td>$SNM_{sram}$</td>
<td>303.3 mV</td>
<td>43.9% increase</td>
</tr>
<tr>
<td>$SOBJ$ Approach 1</td>
<td>$P_{sram}$</td>
<td>113.6 nW</td>
<td>44.2% decrease</td>
</tr>
<tr>
<td></td>
<td>$SNM_{sram}$</td>
<td>303.3 mV</td>
<td>43.9% increase</td>
</tr>
<tr>
<td>$SOBJ^*$ Approach 2</td>
<td>$P_{sram}$</td>
<td>100.5 nW</td>
<td>50.6% decrease</td>
</tr>
<tr>
<td></td>
<td>$SNM_{sram}$</td>
<td>303.3 mV</td>
<td>43.9% increase</td>
</tr>
</tbody>
</table>
Experimental Results: SNM

Butterfly curve for reduced power SRAM.

Butterfly curve for the optimal SRAM.
Experimental Results: Power/SNM

Supply Voltage ($V_{DD}$)

SNM (mV)

- SNM Baseline
- SNM Optimized

Increase in SNM

Avg. Power (nW)

- Power Baseline
- Power Optimized

Decrease in Power
Monte-Carlo Distribution Results ...

- Butterfly curve for Flow 1
- SNM Distribution for Flow 1
- Power Distribution for Flow 1
- Butterfly curve for Flow 2
- SNM Distribution for Flow 2
- Power Distribution for Flow 2
## Monte Carlo Simulation Results

<table>
<thead>
<tr>
<th>Optimization</th>
<th>Parameter</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{PWR}$</td>
<td>$P_{sram}$</td>
<td>28.91 nW</td>
<td>8.26 nW</td>
</tr>
<tr>
<td></td>
<td>$SNM_{sram}$</td>
<td>180 mV</td>
<td>30 mV</td>
</tr>
<tr>
<td>$S_{SNM}$</td>
<td>$P_{sram}$</td>
<td>147.73 nW</td>
<td>101.4 nW</td>
</tr>
<tr>
<td></td>
<td>$SNM_{sram}$</td>
<td>295 mV</td>
<td>28 mV</td>
</tr>
<tr>
<td>$S_{OBJ}$ : Approach 1</td>
<td>$P_{sram}$</td>
<td>147.73 nW</td>
<td>101.4 nW</td>
</tr>
<tr>
<td></td>
<td>$SNM_{sram}$</td>
<td>295 mV</td>
<td>28 mV</td>
</tr>
<tr>
<td>$S_{OBJ}$ : Approach 2</td>
<td>$P_{sram}$</td>
<td>135.24 nW</td>
<td>101.85 nW</td>
</tr>
<tr>
<td></td>
<td>$SNM_{sram}$</td>
<td>295 mV</td>
<td>28 mV</td>
</tr>
</tbody>
</table>
Conclusions

- A methodology for simultaneous optimization of SRAM power and read stability is presented.
- A 45nm single ended seven transistor SRAM was subjected to the proposed methodology (novel DOE-ILP algorithms) leading to 50.6% power reduction and 43.9% increase in read stability (read SNM).
- The effect of process variation of twelve process parameters on the SRAM is evaluated, and it is found to be process variation tolerant.
- A $8 \times 8$ array has been constructed using the optimized cells whose average power consumption is $4.5\mu W$. 
Future Research

- Future research will involve SRAM-array optimization where variability will be accounted for.
- Along with the states of transistors, the sizes will also be considered which will increase the solution space of the algorithms.
- In addition to the power, performance and process variation, thermal effects will also be taken into account.
Thank you !!!