P3 (Power-Performance-Process) Optimization of Nano-CMOS SRAM using Statistical DOE-ILP

Garima Thakral[1], Saraju P. Mohanty[2], Dhruva Ghai[3], and Dhiraj K. Pradhan[4]
Department of Computer Science and Engineering, University of North Texas, USA.[1,2,3]
Department of Computer Science, University of Bristol, UK.[4]
Email-ID: saraju.mohanty@unt.edu[2], pradhan@compsci.bristol.ac.uk[5].

Abstract

In this paper, a novel design flow is presented for simultaneous P3 (power minimization, performance maximization and process variation tolerance) optimization of nano-CMOS circuits. For demonstration of the effectiveness of the flow, a $45nm$ single-ended 7-transistor SRAM is used as example circuit. The SRAM cell is subjected to a dual-V_{Th} assignment based on a novel statistical Design of Experiments-Integer Linear Programming (DOE-ILP) approach. Experimental results show 44.2$\%$ power reduction (including leakage) and 43.9$\%$ increase in the read static noise margin compared to the baseline design. The process variation analysis of the optimized cell is carried out considering the variability effect in 12 device parameters. A 8 × 8 array is constructed to show the feasibility of the proposed SRAM cell. To the best of the authors' knowledge, this is the first study which makes use of statistical Design of Experiments and Integer Linear Programming for optimization of conflicting targets of stability, power in the presence of process variations in an SRAM cell.

Keywords

Process Variation, Power, Static Noise Margin, Static Random Access Memory, Circuit Optimization, Nanoscale CMOS

1 Introduction

A typical state-of-the-art microprocessor die has large portion devoted to on-chip memory [15]. Static random access memory (SRAM) is a volatile memory that retains data as long as power is being supplied. It provides faster access to data and is more reliable. The operations of SRAM have become very critical with the advancement of CMOS technology which is used for its fabrication.

In the case of nanoscale circuit process variation is the most important design challenge to maintain the circuit yield. For SRAM, it is observed that as the supply voltage is reduced, the sensitivity of the circuit parameters to the process variation increases [8]. The variations in threshold voltage (V_{Th}) of SRAM cell transistors due to random dopant fluctuations is the principal reason for parametric failures. The threshold voltage variation is related to the device geometry (length, width and oxide thickness) and doping profile. Eqn. 1 shows how the standard deviation of the threshold voltage (σV_{Th}) is affected by the gate-oxide thickness (T_{ox}), the channel dopant concentration (N_{ch}), the channel length (L) and the width (W) [13]:

$$\sigma_{V_{Th}} = \left(\frac{\sqrt[4]{4 \times q^3 \times \epsilon_{Si} \times \phi_B}}{2} \right) \left(\frac{T_{ox}}{\epsilon_{ox}} \right) \left(\frac{\sqrt[4]{N_{ch}}}{\sqrt{W \times L}} \right), \quad (1)$$

where ϕ_B = 2 $\times \kappa_B \times T \times \ln(N_{ch}/n_i)$ (with κ_B Boltzmann's constant, T the absolute temperature, n_i the intrinsic carrier concentration, q the elementary charge), and ϵ_{ox} and ϵ_{Si} are the permittivity of oxide and silicon, respectively. The above expression is consistent with observations that σV_{Th} is inversely proportional to the square root of the device area.

Power consumption is an important factor to be considered in SRAM design when targeted for embedded systems. Different design methods have been proposed like decrease in supply voltage, which reduces the dynamic power quadratically and reduces the leakage power linearly [9]. However, substantial problems have been noted when the traditional six-transistor SRAM cell is subjected to ultra-low voltage supply as it gives poor stability. Read static noise margin (SNM) is defined as the minimum DC noise voltage which is required to flip the state of the SRAM cell [2] during the read operation. It is measured as the length of the side of the largest square that is fitted inside the lobes of the butterfly curve of the SRAM. In this paper, the "read SNM" is treated as a measure of performance.

The *novel contributions* of this paper are as follows:

1. A novel design flow for P3 (Power-Performance-Process variation) optimization in nanoscale SRAM is proposed.
2. A 7-transistor SRAM designed using $45nm$ CMOS technology is subjected to the proposed methodology.
3. For P3 optimization of the SRAM, a novel statistical Design of Experiments (DOE) - Integer Linear Programming (ILP) based algorithm is proposed which achieved 44.2% power reduction and 43.9% SNM increase in the SRAM.
4. An 8×8 SRAM array is constructed using P3 optimized SRAM cell to study the feasibility of P3-optimal SRAM array construction.

[0] This research is supported in part by NSF awards CCF-0702361 and CNS-0854182.

The notations and definitions for various terminologies used in this paper are given in Table 1. The rest of the paper is organized in the following manner: SRAM related research is presented in Section 2. Section 3 discusses the proposed P3 design flow for SRAM cell optimization. This is followed by the baseline SRAM design, discussed in Section 4. Section 5 highlights the statistical DOE-ILP step of P3 design flow. This is followed by conclusions and future research in Section 6.

Table 1. Notation and Definition

P3	: power, performance and process variation
V_{Th}	: threshold voltage
μ_{PWR}	: mean value of power of SRAM cell
μ_{SNM}	: mean value of SNM of SRAM cell
σ_{PWR}	: standard deviation of power of SRAM cell
σ_{SNM}	: standard deviation of SNM of SRAM cell
τ_{PWR}	: designer defined constraint for power
τ_{SNM}	: designer defined constraint for SNM
$S_{\mu_{PWR}}$	: solution set for mean of power
$S_{\mu_{SNM}}$	: solution set for mean of SNM
$S_{\sigma_{PWR}}$	: solution set for standard deviation of power
$S_{\sigma_{SNM}}$	: solution set for standard deviation of SNM
S_{obj}	: final objective set
$\cap$	: set intersection operator
V_N	: static noise voltage source

2 Related Prior Research in SRAM

Extensive literature is obtained on designing SRAM for low-power operation using nanoscale technology ranges. In [8], a Schmitt-trigger based SRAM is proposed which provides better read-stability, write-ability, and process variation tolerance compared to the standard 6-transistor SRAM cell. A 9-transistor SRAM cell is proposed in [9], which increases the stability and reduces power consumption compared to traditional 6-transistor SRAM. The stability of SRAM cell is analyzed in the presence of random fluctuations using a modeling based approach in [1]. In [2], the combined dual-V_{Th} and dual-T_{ox} assignment is presented for SRAM cell which improves power (*only leakage is considered*) by 53.5% and SNM by 43.8%. The desired results are obtained by using both dual-V_{Th} and dual-T_{ox} assignment which will need more number of masks during fabrication of the SRAM chip. In this paper, dynamic power along with the leakage power is accounted which results in *reduction in total power* by 44.2% and SNM by 43.9% as compared with the baseline design. Also by considering only dual-V_{Th} the manufacturing cost is reduced, as compared to [2]. In [5], the authors present a compact model for critical charge of a 6T SRAM cell for estimating the effects of process variations on its soft error susceptibility. In [14], a DOE-ILP based methodology is proposed for dual-V_{Th} assignment, but the process variation analysis is done after optimization and has not been considered explicitly as a part of the optimization methodology. In [17], the effect on performance and yield of the SRAM cell has been presented from BEOL (Back-end-of-line design) lithography effects, which is important in terms of manufacturing of the SRAM chip. In [12], a 7-transistor read-failure tolerant SRAM topology is introduced, which is suitable for low voltage applications. This 7-transistor SRAM is is used for demonstration of the methodology. However, the proposed methodology is also applicable to other variants present in literature. A comparison of the proposed research with the existing literature in Table 2 shows that a low power and high stability SRAM design is obtained.

3 Proposed Methodology for P3 Optimality

In this section, the proposed design flow is discussed for P3-optimal SRAM with reduced power dissipation, increased performance (i.e. SNM), and process-variation awareness. Fig. 1 shows the proposed design flow.

A well-established process-level technique, called dual-V_{Th} (threshold voltage) is used for reduction of power consumption. It is a very important to choose appropriate transistors for high-V_{Th} assignment, thus, the statistical DOE-ILP methodology is proposed. The DOE, approach helps in reducing the search space and convergence solutions faster. Further, ILP is useful for optimizing the linear objective function subjected to constraints and obtain a bound on the optimal value to solve the predictive equations formed using DOE. Minimum sized transistors are taken for the baseline design. The input to the flow is a baseline SRAM cell.

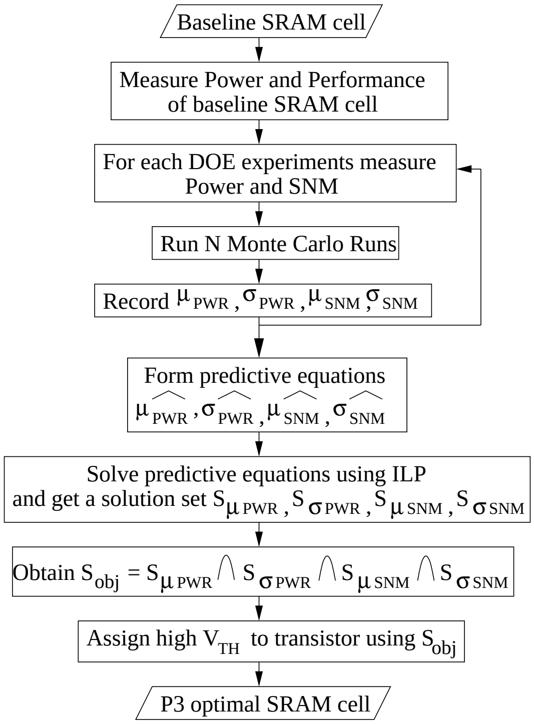

Baseline SRAM cell
Measure Power and Performance of baseline SRAM cell
For each DOE experiments measure Power and SNM
Run N Monte Carlo Runs
Record μ_{PWR}, σ_{PWR}, μ_{SNM}, σ_{SNM}
Form predictive equations $\widehat{\mu_{PWR}}$, $\widehat{\sigma_{PWR}}$, $\widehat{\mu_{SNM}}$, $\widehat{\sigma_{SNM}}$
Solve predictive equations using ILP and get a solution set $S_{\mu PWR}$, $S_{\sigma PWR}$, $S_{\mu SNM}$, $S_{\sigma SNM}$
Obtain $S_{obj} = S_{\mu PWR} \cap S_{\sigma PWR} \cap S_{\mu SNM} \cap S_{\sigma SNM}$
Assign high V_{TH} to transistor using S_{obj}
P3 optimal SRAM cell

Figure 1. Proposed flow for P3-Optimal SRAM.

Fig. 2 shows the theory behind the ILP formulations presented in this paper. The idea is that $\mu_{baseline}$ of the quantity (power or SNM) under consideration needs to be shifted left or right depending on whether it needs to be minimized ($\mu_{minimized}$) or maximized ($\mu_{maximized}$). Also, the $\sigma_{baseline}$ of the quantity (which is a measure of the spread) needs to be minimized to $\sigma_{minimized}$.

For each experiment trial, N Monte Carlo simulations are performed. The mean (μ) and standard deviation (σ) values

Table 2. Comparison of related research

SRAM	Power		SNM		Technology	Research
Research	Value (μW) or (nW)	% Reduction	Value (mV)	% Increase	Node	Highlights
Agrawal [1]	–	–	$160mV$ (approx.)	–	$65nm$	Modeling based approach
Amelifard [2]	–	53.5	–	43.8	$65nm$	Dual-V_{Th} and Dual-T_{ox}
Liu [10]	$31.9nW$ (leakage)	22.9	$300mV$	50	$65nm$	Separate data access mechanism
Kulkarni [8]	$0.11\mu W$ (leakage)	–	$78mV$	58	$130nm$	Schmitt Trigger
Lin [9]	$4.95nW$ (standby)	14.8	$310mV$	52.9	$32nm$	Separate read mechanism
Singh [12]	—	–	$305mV$	65.9	$65nm$	Subthreshold 7T SRAM
Bollapalli [3]	$10mW$ (dynamic + leakage)	53.4	–	–	$45nm$	Separate word line groups
Thakral [14]	$100.5nW$ (dynamic + leakage)	50.6	$303.3mV$	43.9	$45nm$	DOE-ILP for dual-V_{Th}
This research	$113.6nW$ (dynamic + leakage)	44.2	$303.3mV$	43.9	$45nm$	Statistical DOE-ILP for dual-V_{Th}

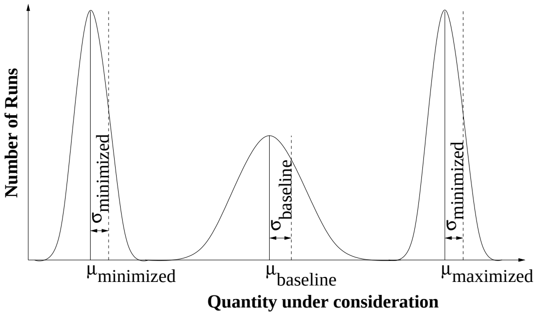


Figure 2. Statistical Optimization of costs.

(Gaussian distribution values) are recorded for average power and performance (SNM) of the SRAM cell. Predictive equations are formed for μ and σ using DOE and are referred as $\widehat{\mu_{PWR}}$, $\widehat{\sigma_{PWR}}$ for power and for SNM as $\widehat{\mu_{SNM}}$, $\widehat{\sigma_{SNM}}$. The predictive equations $\widehat{\mu_{PWR}}$, $\widehat{\sigma_{PWR}}$, $\widehat{\mu_{SNM}}$, $\widehat{\sigma_{SNM}}$ are considered to be linear equations. Each of these linear equations are then solved using integer linear programming (ILP), depending on whether the quantity under consideration is to be maximized or minimized. The solution set for mean and standard deviation of power as $S_{\mu PWR}$, $S_{\sigma PWR}$ and the solution set for mean and standard deviation for SNM as $S_{\mu SNM}$, $S_{\sigma SNM}$ are obtained. For simultaneous power minimization and SNM maximization, the objective S_{obj} is formed as $S_{\mu PWR} \cap S_{\sigma PWR} \cap S_{\mu SNM} \cap S_{\sigma SNM}$ ($\cap$ is defined as the intersection of the sets $S_{\mu PWR}$, $S_{\sigma PWR}$, $S_{\mu SNM}$ and $S_{\sigma SNM}$). Based on S_{obj}, high V_{Th} is assigned to the selected transistors of SRAM cell, and the SRAM cell is re-simulated, to obtain a P3 optimal design. Using this optimized cell, a 8 × 8 array is demonstrated. However, the scope of this paper has been kept at cell-level optimization.

4 Design of Seven Transistor SRAM

The baseline 7-transistor SRAM cell is shown in Fig. 3. This SRAM topology is observed to be suitable for the ultra-low voltage regime. The SRAM cell operates on a single bit line instead of the traditional two bit lines as in case of 6-transistor SRAM cell which performs both read and write operations. It has a read and write access transistor (transistor 1), two inverters (transistors 2, 3, 4 and 5) which are connected back to back in a closed loop fashion in order to store 1 bit information and a transmission gate (transistors 6 and 7). However, the word line is asserted high prior to the read and write operation which is similar to the standard 6-transistor SRAM cell. In the hold mode, the word line (WL) is low and a strong feedback is provided to the cross coupled inverters with the help of transmission gate.

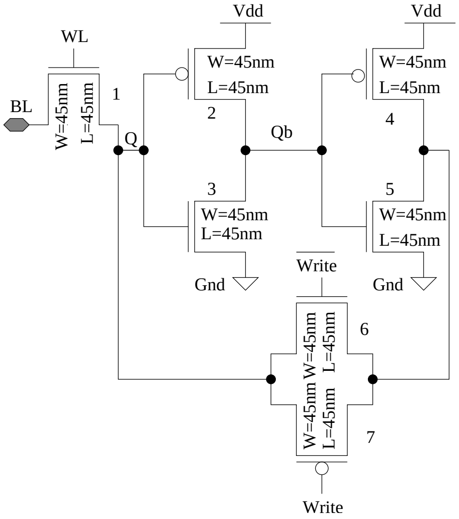


Figure 3. A 7-transistor SRAM cell [12].

4.1 Power and Leakage Measurement

The total power in the nano-CMOS circuit of SRAM cell is the sum of dynamic current, subthreshold leakage current and gate-oxide leakage current. SRAM cell retains it's data for a certain duration of time before it is shut down. Hence the leakage current becomes an important issue as it affects the total power dissipation. It is calculated as Eqn. (2):

$$P_{total} = P_{dynamic} + P_{subthreshold} + P_{gate-oxide}, \qquad (2)$$

where $P_{dynamic}$ is the dynamic power consumption, $P_{subthreshold}$ is the subthreshold leakage in transistors in the "OFF" state and $P_{gate-oxide}$ is referred as the gate-oxide leakage flowing through the transistors [6].

For power dissipation, the current flow in each transistor of SRAM depends on its location in the circuit and operations (read, write or hold) being performed. The current paths for read and write operation have been shown in Fig. 4 for the 7-transistor SRAM cell. The solid arrows shown are for the dynamic current. The dashed arrow represents gate-oxide leakage current and the subthreshold leakage current is shown by dotted arrows which is present in the transistor when it is in the “OFF” state. Basically, when the transistor is in the “ON” state it carries dynamic current alongwith the gate-oxide leakage current and when the transistor is in “OFF” state it will have gate-oxide leakage current as well as subthreshold leakage current.

For detailed understanding, the read “1” and write “0” operations are discussed. Fig. 4(d) shows the read “1” operation of the SRAM cell. In this case, WL and BL will be at high level in order to read a value. So, Q node will have “1” and transistor 2 and transistor 5 will be in “OFF” state, carrying gate-oxide leakage current and subthreshold leakage current. Transistor 3 and transistor 4 will have dynamic current along with gate-oxide leakage current, as they are in “ON” state. Qb will be “0”. In the read operation, the transistors 6 and 7 of the transmission gates will be in “ON” state, hence, carrying dynamic current and gate-oxide leakage current. The write “0” operation is shown in Fig. 4(a). In this case bit line will be “0” and WL is precharged to level “1”. In order to write “0” on the SRAM cell, Q will be “0”. Transistors 2 and 5 are “ON” so they will have dynamic current and gate-oxide leakage current. Transistors 3 and 4 will have subthreshold leakage current and gate-oxide leakage current as they are in “OFF” state. The transistor 6 and transistor 7 will be in “OFF” state in case of write operation, hence will have subthreshold leakage current and gate-oxide leakage current. Similarly, current paths during read “1” and write “0” operations can be identified.

4.2 SNM Model and Measurement

The SNM measurement model is described in this section. Fig. 5 shows the set-up for SNM measurement of the SRAM circuit. It consists of the two inverters (inverter I and inverter II) in feedback and voltage sources V_N. The two voltage sources are the static noise sources. Static noise source is defined as DC disturbances and mismatches due to variations and processing in operating conditions of the cell [11]. The two DC voltage sources V_N are placed in adverse direction to the input of the inverters of the SRAM circuit in order to obtain the worst case SNM. In order to obtain the butterfly curve as shown in Fig. 9(a), the voltages are varied to and from node Q and Qb alternatively. The SRAM cell is simulated at $45nm$ CMOS technology using PTM model [16] with supply voltage V_{dd} of $0.7V$ and with minimum sized transistors.

The power consumption and SNM measurement of the baseline SRAM cell are shown in Table 3. The butterfly curve for baseline SRAM is shown in Fig. 9(a). The supply voltage is (V_{dd}) = $0.7V$. The SRAM cell has been designed at the $45nm$ node [16] with minimum sized transistors. As shown in Table 1, τ_{PWR} and τ_{SNM} are designer defined constraints in the optimization methodology. In this paper the parameters τ_{PWR} and τ_{SNM} are considered as the baseline values which are shown in Table 3.

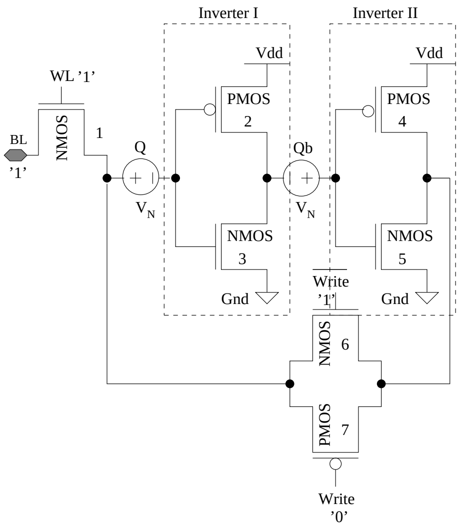


Figure 5. Set-up for SNM measurement.

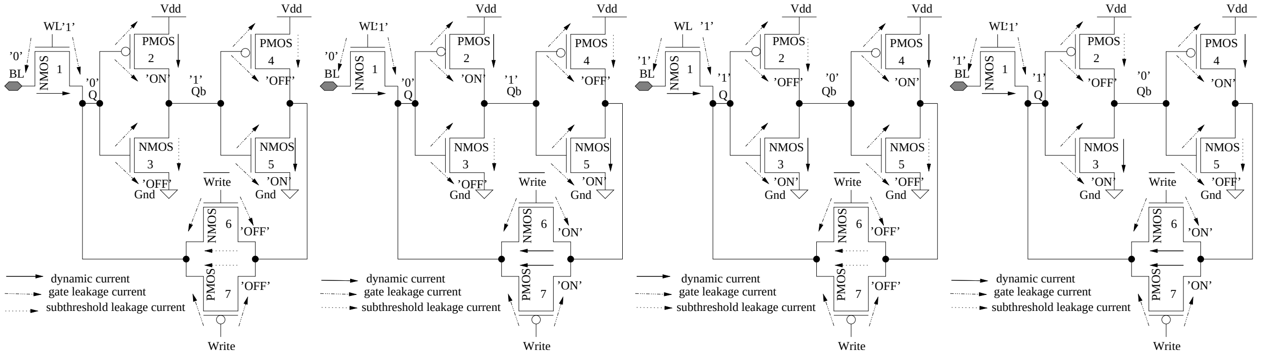
Table 3. Power and SNM for baseline SRAM cell.

Parameter	Value
τ_{PWR}	$203.6\ nW$
τ_{SNM}	$170\ mV$

5 Statistical DOE-ILP Optimization Algorithm

This section discusses the statistical Design of Experiments (DOE)-Integer Linear Programming (ILP) algorithm, which is the heart of the P3 optimization design flow. As shown in Algorithm 1, the baseline SRAM cell is taken as the input alongwith the baseline model file and high-threshold model file. The baseline 7-transistor SRAM is subjected to a DOE [4, 7] based approach using a 2-Level Taguchi L-8 array. The factors are the seven V_{Th} states of the seven transistors of the SRAM cell (Fig. 3). Each factor can take a high V_{Th} state (1) or a nominal V_{Th} state (0). The complexity of the problem is $O(2^n)$ (where n is the transistor number), or in other words, exponential. The L-8 array has a total of 8 experiments. The solution for faster convergence is proposed in the rest of the section.

For formation of the linear equations to be subjected to ILP, DOE method is used. The DOE-ILP is a much better approach as compare to the other techniques because is more efficient and faster. The proposed algorithm converges to solution faster using less resources. 100 Monte Carlo simulations are run for each the experiment. Thus, a total of 800 Monte Carlo runs taking 12 process parameters in account. The 12 process parameters considered are as follows: (1) T_{oxn}: NMOS gate oxide thickness (nm), (2) T_{oxp}: PMOS gate oxide thickness (nm), (3) L_{na}: NMOS access transistor channel length (nm), (4) L_{pa}: PMOS access transistor channel length (nm), (5) W_{na}: NMOS access transistor channel width (nm), (6) W_{pa}: PMOS access transistor channel width (nm), (7) L_{nd}: NMOS driver transistor channel length (nm), (8) W_{nd}: NMOS driver transistor channel width (nm), (9) L_{pl}: PMOS load transistor



(a) Current for write "0" (b) Current for read "0" (c) Current for write "1" (d) Current for read "1"

Figure 4. Current paths for the seven transistor SRAM cell during different read and write operations.

```
Algorithm 1 P3 optimization in nano-CMOS SRAM
Input: Baseline PWR and SNM of the SRAM cell, Baseline model file, High-threshold model file.
Output: Optimized objective set $f_{obj}$ = [$f_{PWR}$, $f_{SNM}$] optimal SRAM cell with transistors identified for high $V_{Th}$ assignment.
Setup experiment for transistors of SRAM cell using 2-Level Taguchi L-8 array, where the factors are the $V_{Th}$ states of transistors of SRAM cell, the response for average power consumption is $\widehat{\mu_{PWR}}$, $\widehat{\sigma_{PWR}}$ and the response for read SNM is $\widehat{\mu_{SNM}}$, $\widehat{\sigma_{SNM}}$.
for Each 1:8 experiments of 2-Level Taguchi L-8 array do
    Run 100 Monte Carlo runs
    Record $\mu_{PWR}$, $\sigma_{PWR}$ and $\mu_{SNM}$, $\sigma_{SNM}$
end for
Form linear predictive equations
$\widehat{\mu_{PWR}}$, $\widehat{\sigma_{PWR}}$ for power
$\widehat{\mu_{SNM}}$, $\widehat{\sigma_{SNM}}$ for SNM.
Solve $\widehat{\mu_{PWR}}$ using ILP: Solution set $S_{\mu PWR}$.
Solve $\widehat{\sigma_{PWR}}$ using ILP: Solution set $S_{\sigma PWR}$.
Solve $\widehat{\mu_{SNM}}$ using ILP: Solution set $S_{\mu SNM}$.
Solve $\widehat{\sigma_{SNM}}$ using ILP: Solution set $S_{\sigma SNM}$.
Form $S_{obj} = S_{\mu PWR} \cap S_{\sigma PWR} \cap S_{\mu SNM} \cap S_{\sigma SNM}$.
Assign high $V_{Th}$ to transistors based on $S_{obj}$.
Re-simulate SRAM cell to obtain optimized objective set.
```

channel length (nm), (10) W_{pl}: PMOS load transistor channel width (nm), (11) N_{chn}: NMOS channel doping concentration (cm^{-3}), (12) N_{chp}: PMOS channel doping concentration (cm^{-3}). Amongst these parameters some are independent and others are correlated which is to be considered during the simulation. Each of these process parameters is considered to have a Gaussian distribution with mean (μ) taken as the nominal values specified in the PTM [16] and 3 × standard deviation (3-σ) as 10% of the mean. A correlation coefficient of 0.9 between T_{oxn} and T_{oxp} is assumed. The responses under consideration are mean μ_{PWR} and standard deviation σ_{PWR} of the average power consumption and also the mean μ_{SNM} and standard deviation σ_{SNM} of the read SNM of the cell.

After performing the experiments, and the half-effects are recorded using the following expression:

$$\frac{\Delta(n)}{2} = \frac{avg(1) - avg(0)}{2}, \tag{3}$$

where $\left[\frac{\Delta(n)}{2}\right]$ is the half-effect of nth transistor, avg(1) is the average value of power (or SNM) when transistor n is in high-V_{Th} state, and avg(0) is the average value of power (or SNM) when transistor n is in nominal V_{Th} state.

The normalized predictive equations are used in order to eliminate the effect of two different units that is nW for power and mV for SNM. Normalized predictive equations are formed as follows:

$$\hat{f} = \bar{f} + \sum_{n=1}^{7} \frac{\Delta(n)}{2} \times x_n, \tag{4}$$

where $\hat{f}$ is the response (power, SNM), $\bar{f}$ is the average of the responses, $\left[\frac{\Delta(n)}{2}\right]$ is the half effect of the nth transistor, and x_n is the V_{Th} state of the nth transistor.

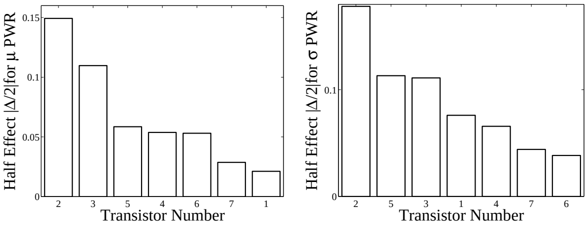
Eqn. 5 shows the predictive equation for mean of the average power consumption of the SRAM cell.

$$\begin{aligned} \widehat{\mu_{PWR}} \quad = \quad & 0.58 - 0.02 \times x_1 - 0.15 \times x_2 \\ & -0.10 \times x_3 - 0.05 \times x_4 - 0.59 \times x_5 \\ & -0.05 \times x_6 + 0.02 \times x_7. \end{aligned} \tag{5}$$

Fig. 6(a) shows the pareto plots of the half-effects of the transistors for μ_{PWR}. In the equation, x_1 represents the V_{Th} state of transistor 1 (Fig. 3), x_2 represents the V_{Th} state of transistor 2, and so on. From this, an ILP problem is formulated as:

$$\begin{aligned} \min \quad & \widehat{\mu_{PWR}} \\ \text{s.t.} \quad & x_n \in \{0,1\} \, \forall n \\ & \mu_{SNM} > \tau_{SNM}. \end{aligned} \tag{6}$$

To minimize power consumption, $\widehat{\mu_{PWR}}$ is minimized. The constraints '1' and '0' represent coded values for high V_{Th} and nominal V_{Th} states, respectively. ILP has been used for smaller circuit, but the methodology is automated, and hence can be used for larger circuits. Solving the ILP problem, the optimal solution is obtained as: $S_{\mu PWR}$ = [x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 1, x_6 = 1, x_7 = 0]. This is interpreted as transistors 1, 2, 3, 4, 5, 6 are high V_{Th} transistors, and transistor 7 is nominal V_{Th} transistor.



(a) μ_{PWR}

(b) σ_{PWR}

Figure 6. Pareto plot for mean (μ PWR) and standard deviation (σ PWR) of SRAM power.

The pareto plot of the half-effects of the transistor for σ_{PWR} is shown in Fig. 6(b). Similarly, Eqn. 7 shows the predictive equation for the standard deviation of the average power consumption of the SRAM cell.

$$\begin{aligned} \widehat{\sigma_{PWR}} &= 0.61 + 0.07 \times x_1 - 0.18 \times x_2 \\ &\quad -0.11 \times x_3 - 0.06 \times x_4 - 0.11 \times x_5. \end{aligned} \tag{7}$$

From this, an ILP problem is formulated as:

$$\begin{aligned} \min \quad & \widehat{\sigma_{PWR}} \\ \text{s.t.} \quad & x_n \in \{0,1\}\, \forall n \\ & \mu_{SNM} > \tau_{SNM}. \end{aligned} \tag{8}$$

To minimize the standard deviation (which is an indication of the spread) of power, $\widehat{\sigma_{PWR}}$ is minimized. Solving the ILP problem, the optimal solution is obtained as: $S_{\sigma PWR}$ = [x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 1, x_6 = 1, x_7 = 0]. This can also be interpreted as transistors 2, 3, 4, 5, 6 are high V_{Th} transistors, and transistors 1,7 are nominal V_{Th} transistors.

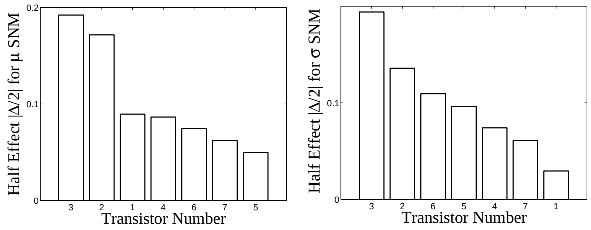
Similarly, the predictive equation for μ_{SNM} is formed as shown in Eqn. 9.

$$\begin{aligned} \widehat{\mu_{SNM}} &= 0.45 - 0.09 \times x_1 + 0.17 \times x_2 \\ &\quad -0.19 \times x_3 - 0.09 \times x_4 + 0.05 \times x_5 \\ &\quad +0.07 \times x_6 - 0.06 \times x_7. \end{aligned} \tag{9}$$

Fig. 7(a) shows the pareto plot of the half-effects of the transistors for μ_{SNM}. From this, an ILP problem is formulated as follows:

$$\begin{aligned} \max \quad & \widehat{\mu_{SNM}} \\ \text{s.t.} \quad & x_n \in \{0,1\}\, \forall n \\ & \mu_{PWR} < \tau_{PWR}. \end{aligned} \tag{10}$$

To maximize SNM, $\widehat{\mu_{SNM}}$ is maximized. Solving the ILP problem, the optimal solution is obtained as: $S_{\mu SNM}$ = [x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 0, x_5 = 1, x_6 = 1, x_7 = 0]. This is interpreted as transistors 2, 5 and 6 are high V_{Th} transistors, and transistors 1, 3, 4 and 7 are nominal V_{Th} transistors.



(a) μ_{SNM}

(b) σ_{SNM}

Figure 7. Pareto plot for mean (μ_{SNM}) and standard deviation (σ_{SNM}) of read SNM.

Fig. 7(b) show the pareto plot of the half-effects of the transistors for σ_{SNM}. The predictive equation for σ_{SNM} is formed as shown in Eqn. 11.

$$\begin{aligned} \widehat{\sigma_{SNM}} &= 0.35 + 0.03 \times x_1 - 0.13 \times x_2 \\ &\quad +0.19 \times x_3 + 0.07 \times x_4 - 0.09 \times x_5 \\ &\quad -0.11 \times x_6 + 0.06 \times x_7. \end{aligned} \tag{11}$$

From this, an ILP problem is formulated as:

$$\begin{aligned} \min \quad & \widehat{\sigma_{SNM}} \\ \text{s.t.} \quad & x_n \in \{0,1\}\, \forall n \\ & \mu_{PWR} < \tau_{PWR}. \end{aligned} \tag{12}$$

To minimize the standard deviation (which is an indication of the spread) of SNM, $\widehat{\sigma_{SNM}}$ is minimized. Solving the ILP problem, the optimal solution is obtained as: $S_{\sigma SNM}$ = [x_1 = 0, x_2 = 1, x_3 =0, x_4 = 0, x_5 = 1, x_6 = 1, x_7 = 0]. This can also be interpreted as transistors 2, 5 and 6 are high V_{Th} transistors, and transistors 1, 3, 4 and 7 are nominal V_{Th} transistors.

The overall objective function S_{obj} for P3 optimality is formulated as follows:

$$S_{obj} = S_{\mu PWR} \cap S_{\sigma PWR} \cap S_{\mu SNM} \cap S_{\sigma SNM}, \tag{13}$$

where $\cap$ is interpreted as the set intersection operator. In other words, the devices which are part of low-power and high-SNM solution sets are picked. The following solution is obtained: S_{obj} = [x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 0, x_5 = 1, x_6 = 1, x_7 = 0], i.e., transistors 2, 5, 6 are high V_{Th} transistors, and transistors 1, 3, 4, 7 are nominal V_{Th} transistors. Fig. 8 shows the SRAM cell with the high V_{Th} transistors circled.

Table 4 shows that the dual-V_{Th} assignment in SRAM shows 44.2% power reduction and 43.9% increase in read SNM over the baseline design. The optimized butterfly curve is shown in Fig. 9(b). Fig. 10 shows the comparison of baseline and P3 optimized SRAM cell power and SNM for various values of V_{dd}. As per the design flow, an 8×8 array is constructed using the optimized cell, shown in Fig. 11. The average power

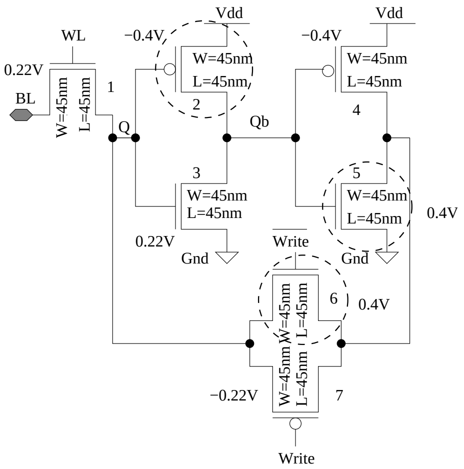
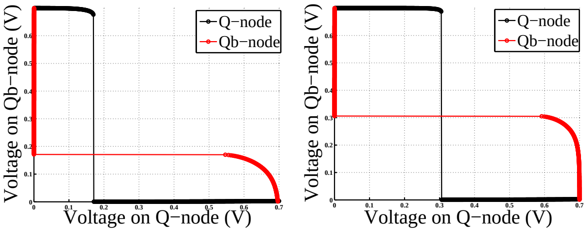


Figure 8. P3 optimized 7T SRAM cell with the circled transistors having high V_{Th}.

consumption of the array is $4.5\mu W$. The results are comparable to [14] where process variation is not considered. Thus, the current paper that accounted process variation could yield similar results, which proves its effectiveness.

Table 4. Results for 7-transistor SRAM cell.

Optimization	Parameter	Value	Change
S_{obj}	Average power P_{SRAM}	$113.6\ nW$	44.2%
S_{obj}	SNM	$303.3\ mV$	43.9%



(a) For baseline

(b) For P3 optimized

Figure 9. Butterfly curves of the SRAM.

Fig. 12(a) shows the effect of process variations on the butterfly curve of the P3 optimized SRAM. Fig. 12(b) shows the distributions for "SNM High" and "SNM Low" extracted from the Monte Carlo simulations. "SNM Low" is treated as the actual SNM. Table 5 shows the corresponding statistical data. Fig. 12(c) shows the distribution of average power of the P3 optimized SRAM cell under process variations. It shows a Lognormal nature. The results are consistent with [14]. However, the distributions are going to change when the optimization will be performed on parasitic-extracted netlist contrary to the transistor level netlist of the current paper. This is being investigated as ongoing research.

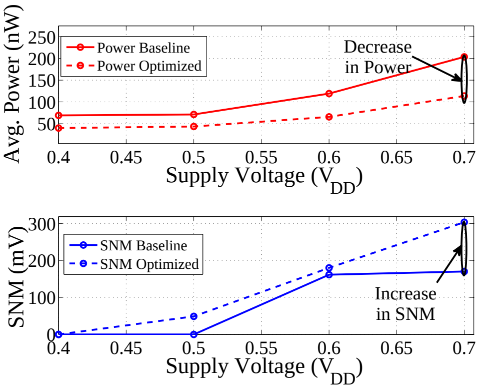


Figure 10. Power and read SNM comparison.

Table 5. Statistical Results for SNM.

Read SNM	μ (mV)	σ (mV)
SNM Low	295	28
SNM High	350.4	71

6 Conclusions and Future Research

A statistical DOE-ILP approach has been presented in this paper for simultaneous P3 (power-performance-process) optimization of SRAM cell. The read SNM has been treated as the performance metric. The optimization has been carried out at cell level. For this, a single ended 7-transistor SRAM cell of 45nm has been subjected to the proposed approach which leads to 44.2% power reduction (including leakage) and 43.9% increase in performance (read SNM). For process variation effect, 12 parameters are considered. Using the P3 optimized cell a 8 × 8 array is constructed and data is presented for power consumption. As part of extension of this research, a P4 optimal methodology is under consideration, where the 4th "P" would be parasitics. Thermal effects will also be incorporated in the future which will lead to what is envisioned as P4VT optimal; V stands for voltage and T stand for temperature. Also, array-level optimization of SRAM with mismatch and process variation will be considered as part of the design flow.

References

[1] K. Agarwal and S. Nassif. Statistical Analysis of SRAM Cell Stability. In *Proceedings of the Design Automation Conference*, pages 57–62, 2006.

[2] B. Amelifard, F. Fallah, and M. Pedram. Reducing the Subthreshold and Gate-tunneling Leakage of SRAM Cells using Dual-Vt and Dual-Tox Assignment. In *Proceedings of the Design Automation and Test in Europe*, pages 1–6, 2006.

[3] K. Bollapalli, R. Garg, K. Gulati, and S. Khatri. Low power and high performance sram design using bank-based selective forward body bias. In *Proceedings of the 19th ACM Great Lakes symposium on VLSI*, pages 441–444, 2009.

[4] D. Ghai, S. P. Mohanty, and E. Kougianos. Variability-aware optimization of nano-CMOS Active Pixel Sensors using design and analysis of Monte Carlo experiments. In *Proceedings of the*

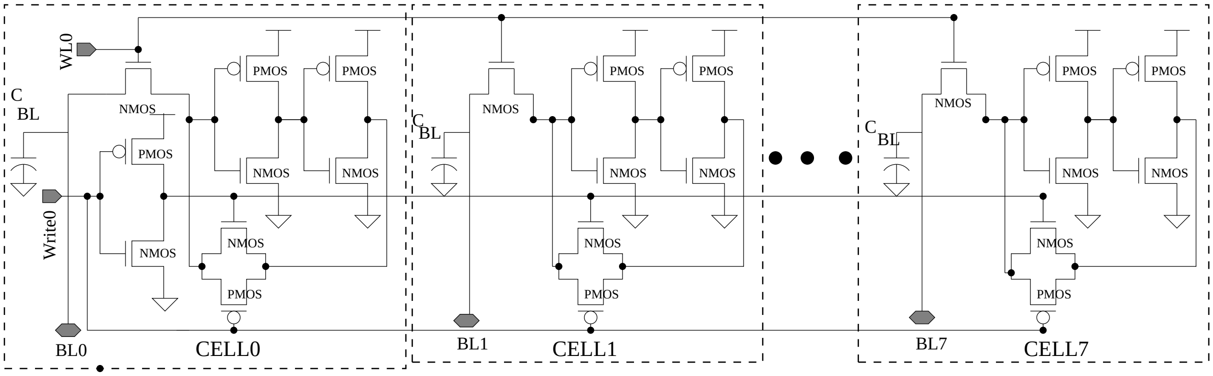
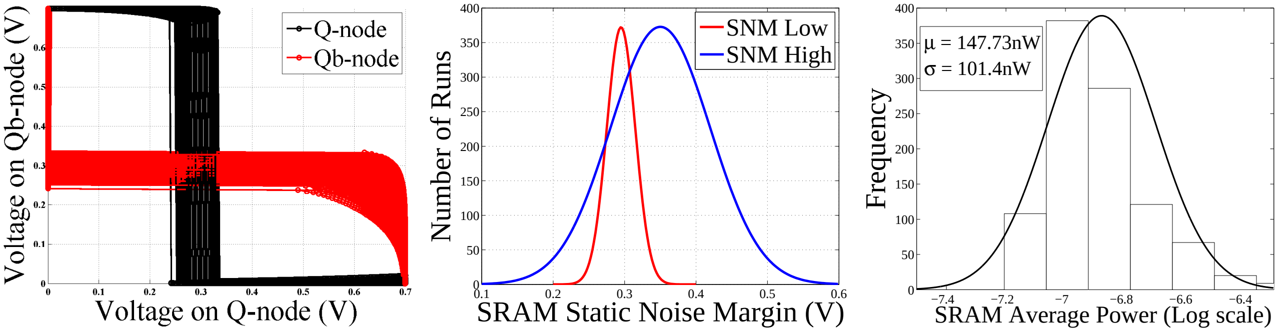


Figure 11. One row of the 8×8 array constructed using P3 optimized 7-transistor SRAM cells.



(a) Butterfly Curve

(b) SNM Distribution

(c) Power Distribution

Figure 12. Process variation study of the SRAM.

International Symposium on Quality Electronic Design, pages 172–178, 2009.

[5] S. Jahinuzzaman, M. Sharifkhani, and M. Sachdev. Investigation of Process Impact on Soft Error Susceptibility of Nanometric SRAMs Using a Compact Critical Charge Model. In *Proceedings of the International Symposium on Quality Electronic Design.*, pages 207–212, 2008.

[6] E. Kougianos and S. P. Mohanty. Metrics to Quantify Steady and Transient Gate Leakage in Nanoscale Transistors: NMOS Vs PMOS Perspective. In *Proceedings of the 20th IEEE International Conference on VLSI Design (VLSID)*, pages 195–200, 2007.

[7] E. Kougianos and S. P. Mohanty. Impact of Gate-Oxide Tunneling on Mixed-Signal Design and Simulation of a Nano-CMOS VCO. *Elsevier Microelectronics Journal*, 40(1):95–103, January 2009.

[8] J. Kulkarni, K. Kim, S. Park, and K. Roy. Process variation tolerant SRAM array for ultra low voltage applications. In *Proceedings of the Design Automation Conference*, pages 108–113, 2008.

[9] S. Lin, Y. B. Kim, and F. Lombardi. A low leakage 9t sram cell for ultra-low power operation. In *Proceedings of the ACM Great Lakes symposium on VLSI*, pages 123–126, 2008.

[10] Z. Liu and V. Kursun. High Read Stability and Low Leakage Cache Memory Cell. In *Proceedings of the International Symposium on Circuits and Systems*, pages 2774–2777, 2007.

[11] E. Seevinck and et. al. Static noise margin analysis of MOS SRAM cells. *IEEE Journal of Solid-State Circuits*, 22(5):748754, October 1987.

[12] J. Singh, J. Mathew, D. K. Pradhan, and S. P. Mohanty. A Subthreshold Single Ended I/O SRAM Cell Design for Nanometer CMOS Technologies. In *Proceedings of the International SOC Conference*, pages 243–246, 2008.

[13] P. A. Stolk, F. P. Widdershoven, and D. B. M. Klaassen. Modeling Statistical Dopant Fluctuations in MOS Transistors. *IEEE Transactions on Electron Devices*, 45(9):1960–1971, September 1998.

[14] G. Thakral, S. P. Mohanty, D. Ghai, and D. K. Pradhan. A Combined DOE-ILP Based Power and Read Stability Optimization in Nano-CMOS SRAM. In *Proceedings of the 23rd IEEE International Conference on VLSI Design (ICVD)*, 2010.

[15] N. Yoshinobu and et. al. Review and future prospects of low voltage RAM circuits. *IBM journal of research and development*, 47(5/6):525–552, 2003.

[16] W. Zhao and Y. Cao. New Generation of Predictive Technology Model for sub-$45nm$ Design Exploration. In *Proceedings of the International Symposium on Quality Electronic Design*, pages 585–590, 2006.

[17] Y. Zhou, R. Kanj, K. Agrawal, Z. Li, R. Joshi, S. Nassif, and W. Shi. The impact of BEOL lithography effects on the SRAM cell performance and yield. In *Proceedings of the International Symposium on Quality Electronic Design*, pages 607–612, 2009.