Statistical DOE-ILP based power-performance-process (P3) optimization of nano-CMOS SRAM

Saraju P. Mohanty a,*, Jawar Singh b, Elias Kougianos a, Dhiraj K. Pradhan c

a NanoSystem Design Laboratory (NSDL), University of North Texas, USA
b Department of Electronics and Communication Engineering, Jaypee University of Engineering and Technology, India
c Department of Computer Science, University of Bristol, UK

A R T I C L E   I N F O

Article history:
Received 26 April 2010
Received in revised form 24 April 2011
Accepted 11 July 2011
Available online 22 July 2011

Keywords:
Nanoscale CMOS
Process-variation aware design
Low-power design
Static random access memory
Design of Experiments
Integer Linear Programming

A B S T R A C T

As technology continues to scale, maintaining important figures of merit of Static Random Access Memory (SRAMs), such as power dissipation and an acceptable Static Noise Margin (SNM), becomes increasingly challenging. In this paper, we address SRAM instability and power (leakage) dissipation in scaled-down technologies by presenting a novel design flow for simultaneous power minimization, performance maximization and process variation tolerance (P3) optimization of nano-CMOS circuits. The 45 and 32 nm technology node standard 6-Transistor (6T) and 8T SRAM cells are used as example circuits for demonstration of the effectiveness of the flow. Thereafter, the SRAM cell is subjected to a dual threshold voltage (dual-Vth) assignment based on a novel statistical Design of Experiments-Integer Linear Programming (DOE-ILP) approach. Experimental results show 61% leakage power reduction and 13% increase in the read SNM. In addition, process variation analysis of the optimized cell is conducted considering the variability effect in twelve device parameters. To the best of the authors’ knowledge, this is the first study which makes use of statistical DOE-ILP for optimization of conflicting targets of stability and power in the presence of process variations in SRAMs.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction and contributions

SRAM is a volatile memory that retains data bits as long as power is being supplied. It provides fast access to data and is very reliable. Degraded bitcell currents and leakages, and poor SRAM bitcell noise margins, when a large number of devices are integrated into a single die, result in process and design variability which in turn leads to a great loss of parametric yield [1]. A sufficiently large Static Noise Margin (SNM), reduced power consumption and a process variation tolerant circuit are needed in order to prevent substantial loss of parametric yield caused by the technology scaling-induced side effects. Thus, the operations of SRAM have become very critical with the advancement of CMOS technology. In this section, we discuss the importance of the factors that have been considered for optimization, and present the motivation behind the research presented in this paper. By reducing the power consumption significantly, and maximizing the static noise margin we can increase the efficiency and reliability of the SRAM cell. However, the SRAM cell becomes susceptible to process variation at lower supply voltages which in turn decrease its noise handling capacity.

SRAM arrays are widely used as cache memory in microprocessors and Application-Specific Integrated Circuits (ASICs) and occupy a large portion of the die area. Large arrays of fast SRAM help improve the performance of the system. Thus, balancing these requirements is driving the effort to minimize the footprint of SRAM cells [1].

Power dissipation: Embedded systems, particularly those targeted toward low duty cycles and portable applications (e.g., mobile phones), require extremely low energy dissipation as they are typically battery powered. In such systems, a significant amount of power is consumed during memory accesses, which affects the battery life. Hence, efficient active and leakage power saving SRAM designs need to be explored for higher reliability and longer operation of battery powered systems. Different design methods have been proposed, such as decrease in supply voltage, which reduces the dynamic power quadratically and reduces the leakage power linearly [2]. However, with technology scaling, leakage current increases exponentially and reliability is affected significantly due to poor stability noise margins and process variation. These technology scaling-induced side effects are further exacerbated by reduced supply voltage introduced in order to achieve energy efficiency. Fig. 1 shows the comparison of normalized read Static Noise Margin (SNM) and leakage current of
a 6T SRAM cell for different technology nodes. The minimum feature sized device with cell ratio (β = 2) is used for simulation using the Predictive Technology Model (PTM) [3]. It can be seen from Fig. 1 that the read SNM of a 6T SRAM cell is gradually decreasing with technology scaling, while the leakage current is exponentially increasing. Moving from the 132 to the 32 nm technology node, there is 55% reduction in the read SNM while there is 86% increase in leakage current. Therefore, alternative cell topologies or optimization methodologies are needed for nano-regime technologies that provide low standby power (leakage) and higher stability margins (SNM). Along this line, several SRAM cell topologies have been proposed in the recent past to address the ultra-low power requirements [4-8]. Hence, in this paper, standard 6T and 8T SRAM [6,7] cells are used as baseline circuits for optimization.

Performance: SNM can serve as a figure of merit in stability evaluation of SRAM cells. The read SNM is defined as the minimum DC noise voltage which is required to flip the state of the SRAM cell [9] during the read operation. It is measured as the length of the side of the largest square that fits inside the lobes of the butterfly curve of the SRAM. Thus, in this paper we treat the SNM as a measure of performance. The SNM of even defect-free cells is gradually declining with technology scaling, as shown in Fig. 1. SRAM cells with compromised stability can limit the reliability of on-chip data storage making it more sensitive to transistor parameter shift with aging, voltage fluctuations and ionizing radiation [1]. Detection and correction/repair of such cells in modern scaled-down SRAMs become a necessity.

Process variation: Millions of minimum-size SRAM cells are tightly packed making SRAM arrays the densest circuitry on a chip. Such areas on the chip can be especially susceptible and sensitive to manufacturing defects and process variations [1]. Variations in the device parameters translate into variations in SRAM attributes, such as power and stability. Under adverse operating conditions, such SRAMs may inadvertently corrupt the stored data. In SRAMs, it is observed that as the supply voltage is reduced, the sensitivity of the circuit parameters to the process variation increases [10]. For system integration, SRAM must be compatible with subthreshold combinational logic operating at ultra-low voltages. However, this leads to increase in sensitivity to parameter variability. This problem will worsen in nanometer technologies with ultra-low voltage operation and makes SRAM design and stability analysis more challenging. The variations in threshold voltage ($V_{th}$) of SRAM cell transistors due to random dopant fluctuations are the principal reason for parametric failures. The threshold voltage variation is related to the device geometry (length, width and oxide thickness) and doping profile. Eq. (1) shows how the threshold voltage standard deviation ($\sigma_{V_{th}}$) varies with gate oxide thickness ($T_{ox}$), channel dopant concentration ($N_{A}$) and channel length ($L$) and width ($W$) [11]:

$$\sigma_{V_{th}} = \left(\frac{\sqrt{4q^2C_0V_{th}}}{}\right) \frac{T_{ox}}{V_{th}} \left(\frac{\sqrt{N_{A}}}{LW}\right)^2.$$  

(1)

where $q = 2k_B T \ln(N_{A}/n_i)$. $N_{A}$ is the channel dopant concentration, $k_B$ Boltzmann’s constant, $T$ the absolute temperature, $n_i$ the intrinsic carrier concentration, $q$ the elementary charge, and $C_0$ and $\varepsilon_0$ are the permittivity of oxide and silicon, respectively. The above expression is consistent with observations that $\sigma_{V_{th}}$ is inversely proportional to the square root of the device area.

In order to address the above issues, we propose a methodology involving power and performance optimization in the presence of process variations in SRAM cells. However, it is a non-trivial task to simultaneously maintain reduced power dissipation, improved performance (which is SNM in this paper) and process variation tolerance. The distinct contributions of this research are as follows:

1. A novel design flow for simultaneous power–performance–process variation (P3) optimization in nanoscale SRAMs is introduced.
2. The 45 nm standard 6T and 8T SRAM cells are subjected to the proposed methodology.
3. For P3 optimization of the 6T and 8T SRAM cells, we propose a novel statistical Design of Experiments (DOE)–Integer Linear Programming (ILP) based approach. It achieved 61% power reduction and 13% SNM increase.
4. Process variation analysis of the optimal SRAM is conducted considering 12 device parameters and demonstrates the robustness of the design.
5. The proposed methodology for P3 optimization and DOE–ILP approach is also tested on the 32 nm technology node based 6T and 8T SRAM cells.

The notations and definitions used in this paper are given in Table 1. The rest of the paper is organized in the following

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>Cell ratio</td>
</tr>
<tr>
<td>$T$</td>
<td>Temperature</td>
</tr>
<tr>
<td>$n_i$</td>
<td>Intrinsic carrier concentration</td>
</tr>
<tr>
<td>$q$</td>
<td>Electron charge</td>
</tr>
<tr>
<td>$k_B$</td>
<td>Boltzmann constant</td>
</tr>
<tr>
<td>$C_0$</td>
<td>Oxide capacitance</td>
</tr>
<tr>
<td>$\varepsilon_0$</td>
<td>Permittivity of vacuum</td>
</tr>
<tr>
<td>$N_{A}$</td>
<td>Channel dopant concentration</td>
</tr>
<tr>
<td>$T_{ox}$</td>
<td>Gate oxide thickness</td>
</tr>
<tr>
<td>$V_{th}$</td>
<td>Threshold voltage</td>
</tr>
<tr>
<td>$\sigma_{V_{th}}$</td>
<td>Threshold voltage standard deviation</td>
</tr>
<tr>
<td>$\sigma_{SNM}$</td>
<td>SNM standard deviation</td>
</tr>
<tr>
<td>$V_{SNM}$</td>
<td>SNM voltage</td>
</tr>
<tr>
<td>$V_{in}$</td>
<td>Input voltage</td>
</tr>
<tr>
<td>$V_{DD}$</td>
<td>Power supply voltage</td>
</tr>
<tr>
<td>$V_{DDQ}$</td>
<td>Quiescent power supply voltage</td>
</tr>
<tr>
<td>$I_{DQ}$</td>
<td>Output load current</td>
</tr>
<tr>
<td>$I_{DDQ}$</td>
<td>Quiescent output current</td>
</tr>
<tr>
<td>$I_{PVT}$</td>
<td>Process, voltage, and temperature variation</td>
</tr>
<tr>
<td>$V_{DDQ}$</td>
<td>Quiescent power supply voltage</td>
</tr>
<tr>
<td>$I_{DQ}$</td>
<td>Output load current</td>
</tr>
<tr>
<td>$I_{DDQ}$</td>
<td>Quiescent output current</td>
</tr>
<tr>
<td>$I_{PVT}$</td>
<td>Process, voltage, and temperature variation</td>
</tr>
<tr>
<td>$V_{DDQ}$</td>
<td>Quiescent power supply voltage</td>
</tr>
<tr>
<td>$I_{DQ}$</td>
<td>Output load current</td>
</tr>
<tr>
<td>$I_{DDQ}$</td>
<td>Quiescent output current</td>
</tr>
<tr>
<td>$I_{PVT}$</td>
<td>Process, voltage, and temperature variation</td>
</tr>
<tr>
<td>$V_{DDQ}$</td>
<td>Quiescent power supply voltage</td>
</tr>
<tr>
<td>$I_{DQ}$</td>
<td>Output load current</td>
</tr>
<tr>
<td>$I_{DDQ}$</td>
<td>Quiescent output current</td>
</tr>
<tr>
<td>$I_{PVT}$</td>
<td>Process, voltage, and temperature variation</td>
</tr>
</tbody>
</table>

Table 1 Notation and definitions used in this paper.
manner: Related prior research is discussed in Section 2. Section 3 presents the proposed P3 design flow for SRAM cell optimization. The baseline SRAM design and its operation are discussed in Section 4. Section 5 highlights the statistical DOE–ILP step of the P2 design flow. This is followed by conclusions and future research in Section 6.

2. Related prior research in SRAM

Several design and optimization methodologies have been presented in the current literature addressing the nanoscale challenges of SRAM circuits. A high-level overview of a selected subset relevant to this work is presented in Table 2. The stability of the SRAM cell in the presence of random fluctuations is analyzed using a modeling based approach in [12]. Bollapalli et al. [14] quote only the reduced power dissipation. In [10], a Schmitt–Trigger based SRAM is proposed which provides better read stability, write ability and process variation tolerance compared to the standard 6T SRAM cell. A 9-transistor SRAM cell is proposed in [2], which increases the stability and reduces power consumption compared to the traditional 6T SRAM. A method is presented in [9,20], based on dual-Vth and dual-Ith assignments, for low-power design of SRAM while maintaining performance. In [21] a compact model of critical charge of a 6T SRAM cell is presented for estimating the effects of process variations on its soft error susceptibility. Singh et al. [16] have presented a different design methodology of two-port 6T SRAM with multiporput capabilities. Nalam et al. [18] have explored power (only leakage) and SNM parameters using two-phase write and split bit line differential sensing. In [17], a DOE–ILP based methodology is proposed for dual-Vth assignment without accounting for process variations, which is important for nanoscale CMOS. In [15] an SNM enhancement technique is presented that results in undisturbed storage nodes but this achievement comes at the expense of additional transistors. In [22], 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. Singh et al. [19] have presented a 7T SRAM topology, which is suitable for low voltage applications and it is also tolerant to read failures.

This archival journal paper is based on our conference publication [23]. The journal paper includes considerable additional material, such as functional simulation analysis of standard 6T and 8T SRAM cells (different than the previously published one) for different nano-CMOS technology nodes.

Table 2
Comparison of related research in SRAM.

<table>
<thead>
<tr>
<th>SRAM research</th>
<th>Power</th>
<th>SNM</th>
<th>Tech. node (nm)</th>
<th>Research techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>% Reduction</td>
<td>Value (mV)</td>
<td>% Increase</td>
<td></td>
</tr>
<tr>
<td>Agarwal and Nassif [12]</td>
<td>31.9 nW (leakage)</td>
<td>160 (approx.)</td>
<td>300</td>
<td>65</td>
</tr>
<tr>
<td>Liu and Kursun [13]</td>
<td>0.11 μW (leakage)</td>
<td>78</td>
<td>130</td>
<td>65</td>
</tr>
<tr>
<td>Kulkarni et al. [10]</td>
<td>4.95 nW (standby)</td>
<td>310</td>
<td></td>
<td>65</td>
</tr>
<tr>
<td>Bollapalli et al. [14]</td>
<td>10 mW (total)</td>
<td>-</td>
<td>32</td>
<td>45</td>
</tr>
<tr>
<td>Azam et al. [15]</td>
<td>63.9 μW vs 44.4 μW</td>
<td>299</td>
<td>53–61</td>
<td>65</td>
</tr>
<tr>
<td>Singh et al. [16]</td>
<td>28 (total)</td>
<td>-</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Thakral et al. [17]</td>
<td>105.5 nW</td>
<td>303.3</td>
<td>43.9</td>
<td>45</td>
</tr>
<tr>
<td>Nalam et al. [18]</td>
<td>10–15 (leakage)</td>
<td>10–15</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Anselaert et al. [9]</td>
<td>53.5</td>
<td>43.8</td>
<td>65</td>
<td>45</td>
</tr>
<tr>
<td>Singh et al. [19]</td>
<td>305</td>
<td>65.9</td>
<td>65</td>
<td>Subthreshold 7T-SRAM</td>
</tr>
<tr>
<td>This paper</td>
<td>1.64 nW (leakage)</td>
<td>143.9</td>
<td>4</td>
<td>6T, 45</td>
</tr>
<tr>
<td></td>
<td>2.85 nW</td>
<td>318.2</td>
<td>13</td>
<td>8T, 45</td>
</tr>
<tr>
<td></td>
<td>1.81 nW</td>
<td>81.4</td>
<td>13</td>
<td>6T, 45</td>
</tr>
<tr>
<td></td>
<td>2.34 nW</td>
<td>222.4</td>
<td>12.7</td>
<td>8T, 45</td>
</tr>
</tbody>
</table>

3. The proposed methodology for P3-optimal nano-CMOS SRAM

The proposed design flow to achieve P3-optimal design of both 6T and 8T SRAM circuits is shown in Algorithm 1 in pseudo-code form.

Algorithm 1. P3-optimal design methodology for nano-CMOS SRAM.

1: Input: SRAM topologies (6T and 8T cells) and technology nodes (45 and 32 nm).
2: Output: P3 optimized (power minimization, performance maximization and process variation tolerant) SRAM cell.
3: Perform the baseline design of the SRAM cells.
4: Measure power and performance of baseline SRAM cells.
5: Go to Algorithm 2 for optimizing baseline SRAMs.
6: Re-simulate SRAM cells to obtain P2 (power minimization and performance maximization) SRAM cells.
7: Perform process variation characterization of SRAM cell using device parameters (in this case 12 device parameters).
8: Obtain P3 optimal SRAM cells.
9: Construct SRAM array to observe the feasibility of the SRAM cells.

The input to the proposed design flow is baseline SRAM cells which refer to the 6T and 8T SRAM circuits with nominal sized transistors for a specified technology. Maintaining an acceptable SNM as well as reduced power consumption embedded SRAMs, while scaling the minimum feature size and supply voltages of system-on-a-chip (SoC) is a very challenging task. There are various ongoing research works which discuss techniques to reduce power consumption such as dual-Vth, dual-Vdd, etc. In this paper, we adopt the process-level technique called dual-Vth. Thus, in order to achieve the optimized nano-CMOS circuit we have measured power and SNM values simultaneously using Design of Experiments (DOE). The idea is that leakage is a major component of the total power for the nano-CMOS. Hence, by reducing power through the dual-Vth technique we achieve reduction of total power along with noticeable improvement in performance.

The research problem here is defined as the selection of transistors for high Vth assignment. Further, the assignment is done in such a way that along with the power reduction, the performance metric (i.e. SNM) should not be compromised. To address this research problem of choosing the correct transistors for high-Vth assignment...
we propose a novel statistical Design of Experiments–Integer Linear Programming (DOE–ILP) methodology (Algorithm 2). Design of Experiments or experimental design is the concept of purposeful changes of the inputs in order to study the corresponding changes in the output. A complete full factorial design matrix with two level settings per parameter (low and high voltage threshold) for n transistors would require $2^n$ total runs ($2^2$ for the 6T cell and $2^4$ for the 8T cell). In order to expand the applicability of this approach to large circuits, we followed a Taguchi screening methodology, instead [24]. Taguchi designs are orthogonal with respect to the main effects (in this case the threshold voltages) but contain aliased second order interactions. Since we are subsequently applying ILP techniques, this is not a serious limitation. The implementation of a 2-Level Taguchi design matrix helps in substantially faster optimization time while maintaining good accuracy of the results. Further, ILP combined with DOE is useful for optimizing the linear objective function subject to constraints and to obtain a bound on the optimal value to solve the predictive equations that are formed using DOE. This combined approach has the potential to handle large circuits for optimization in reasonable time.

Once we obtain the P2 optimized SRAM circuit we perform process variation, where variability is considered in 12 device parameters. Detailed discussion is provided in Section 5. After successfully performing the above steps we achieve the target that is a P3 optimal SRAM cell.

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

4. Design and modeling of baseline SRAM circuits

A typical SRAM cell uses two cross-coupled inverters forming a latch and access transistor. The access transistors enable access to the cell during read and write operations and provide cell isolation during the not-accessed state. An SRAM cell is designed to provide non-destructive read access, successful write capability and data storage (or data retention) for as long as the cell is powered.

4.1. Baseline SRAM design for 45 and 32 nm CMOS

In general, the cell design must strike a balance between cell area, robustness, speed, leakage and yield [1]. Smaller cells result in a smaller array area and hence smaller bit line and word line capacitances, which in turn helps improve the access speed performance. Reducing the transistor dimensions is the most effective means to achieve a smaller cell area. However, transistor dimensions cannot be reduced indefinitely without compromising the other parameters. For instance, smaller transistors can compromise the cell stability. Often, performance and stability objectives restrict arbitrary reduction in cell transistor sizes. Similarly, cell area can be traded off for special features such as improved radiation hardening or multiport cell access.

The baseline standard 6T and 8T cells are shown in Fig. 3(a) and (b), respectively. The standard 6T cell topology has been most commonly used in the industry, while 8T has received great attention in the recent past, as low-power substitute with significant improvement in the read the SNM as compared to the 6T cell [6,7]. In a standard 6T cell, both read and write operations are performed via the same pass gate access transistors (i.e. $M_5$ and $M_6$) as shown in Fig. 3(a). As a result, there is always a conflicting read and write requirement, since we cannot simultaneously optimize both devices for read and write operations. Hence, the standard 6T cell has low read SNM which further diminishes with voltage scaling. In order to address this conflicting requirement and poor read noise margin problem, isolated read and write operation based SRAM cells are proposed. In the 8T cell, both read and write operations are isolated. The write operation is performed via pass gate access transistors (i.e. $M_5$ and $M_6$), while the read operation is performed via a separate read port which is comprised of transistor $M_7$ and $M_8$, as shown in Fig. 3(b). The isolated read port provides significant improvement in read SNM, since we can optimize the SRAM cell independently for both operations. The SRAM cells have been designed at the 45 nm technology node with supply voltage $V_{DS} = 0.9$ V. The sizes of all the transistors are estimated with pull-up ratio $\alpha = 1$ and cell ratio $\beta = 2$.

The power consumption and SNM of the baseline cells are measured from functional simulations and are tabulated as shown in Table 3. $T_{WRITE}$ and $T_{SNM}$ are designer defined constraints in the optimization methodology. In this paper, we have taken the parameters $T_{WRITE}$ and $T_{SNM}$ as baseline values which are shown in Table 3. We discuss each of the modes of operation of the 6T and 8T cells in detail in the following section.
4.2. Modes of operation for the 6T and 8T cells

4.2.1. Read operation
Prior to initiating a read operation, the bit lines (BL and BLB) are precharged to $V_{DD}$. The read operation is initiated by enabling the word line (WL) and connecting the precharged bit lines to the internal nodes of the cell via access transistors (i.e. $M_3$ and $M_6$), as shown in Fig. 3(a). During read access, BLB starts discharging via node QB, and as a result there will be a potential difference between BL and BLB. This potential difference is sensed by the sense amplifier and information is read out. In order to ensure a non-destructive read operation the sizes of the transistors must be chosen carefully. For example, $M_3$ and $M_4$ must be stronger than $M_2$ and $M_6$ to keep the node voltage lower than the trip voltage of the inverters. Similarly, for a successful write operation $M_5$ and $M_6$ must be stronger than $M_1$ and $M_2$.

However, the read operation of the 8T cell is entirely different from the standard 6T cell, as shown in Fig. 3(b). In the 8T cell, the read bit line (RBL) is precharged to $V_{DD}$ before commencing the read operation. During read access, the precharged bit line starts discharging if the node QB holds '0', otherwise RBL remains high. The status of RBL is sensed by the sense amplifier to read out the information. In the 8T cell there are separate read and write ports. Therefore, the sizing requirements are relaxed and each port can be sized according to the read/write requirement.

4.2.2. Write operation
The write operation of standard 6T and 8T cells is identical. In both cells, the write operation begins with precharging the bit lines (BL and BLB). During write access, the word line (WL) is enabled connecting both access transistors to the internal data storage nodes (Q and QB). In order to flip the state of the cell as shown in Fig. 3, the write driver pulls down the bit line (BL), which is connected to node Q, while keeping the BLB high.

4.2.3. Hold operation
The hold operation has its own significance, particularly for data retention. During hold mode, word lines (WL and RWL) are disabled and the cross-coupled inverters are tightly connected to each other for longer data retention. However, hold SNM of the 6T cell is usually higher than the read SNM. In the 8T M cell, the hold SNM is almost equal to the read SNM because of the separate read port.

4.3. Leakage measurement
Leakage power plays a vital role in the nano-regime and in certain SoC applications it dominates the dynamic power. This section deals with different leakage power measurements of standard 6T and 8T cells under the idle state.

4.3.1. Power model
The major sources of power dissipation for a nano-CMOS circuit are due to capacitive switching, subthreshold leakage, and gate leakage. Both dynamic and static powers are significant fractions of total power dissipation. Each one of them has several forms and origins; they flow between different terminals and in different operating conditions of a transistor. It is essential to study the power consumption profile of SRAMs in order to estimate and minimize their power consumption, especially when they are made of nanoscale CMOS transistors. An SRAM consumes dynamic power only when the bit line or word line is switching their level from low-to-high or high-to-low for Write or Read operations. On the other hand, including the hold (idle) state, power dissipation happens continuously in the form of gate oxide leakage and subthreshold leakage. In general, SRAM contributes to the major portion of the total leakage power in a modern processor during idle states.

4.3.2. Leakage model
The leakage model consists of subthreshold leakage current and gate oxide leakage current. We discuss each of them in brief.

The subthreshold leakage is modeled as follows [1]:

$$I_{sub} = I_s \exp \left( \frac{V_{GS} - V_t}{nV_T} \right) \left( 1 - \exp \left( -\frac{V_{GS}}{V_t} \right) \right).$$

(2)

where $n = (1 + C_3/C_4)$, $V_t = (kT/q)$ is the thermal voltage, $V_{TH}$ is the threshold voltage, $I_s$ is the current when $V_{GS}$ equals $V_{DD}$, $V_{DD}$ is the gate-to-source voltage, and $V_{gs}$ is the drain-to-source voltage.

The gate oxide leakage current is modeled using the following expression [25]:

$$I_g = A W E \left( \frac{T_{ox}}{\lambda_{ox}} \right) ^ n \left( \frac{V_{DD}V_{max}}{V_T} \right) \exp \left( -\frac{V_{GS} - \phi}{\lambda_{ox}} \right) \left[ 1 - \gamma \left( \frac{V_{GS}}{V_T} \right) \right].$$

(3)

where $A = \left( q^2/8\pi E \phi_0 \right)$, $B = (8\pi \sqrt{2\pi m_e \phi_0 T_{ox}} / 3h)$, $m_{ox}$ is the effective carrier mass in the oxide, $\phi_0$ is the tunneling barrier height, $T_{ox}$ is the oxide thickness, $\lambda_{ox}$ is the reference oxide thickness at which all parameters are extracted, $n$ is a fitting parameter, $V_{max}$ is an auxiliary function which approximates the density of tunneling carriers as well as available states, and $\alpha$, $\beta$, and $\gamma$ are the controlling parameters for electron tunneling.

In addition, leakages consist of diode leakage flowing in the transistors of the cell. The diodes are formed between the diffusion region of the transistors and the substrate consumes power in the form of reverse bias current which is drawn from the power supply.

4.3.3. Leakage current paths in the hold state
The current flow in each transistor of the cell depends on its location and the operation being performed. The current paths for hold (idle) state are shown in Fig. 4 for the 6T cell. The solid arrows shown in the figure are for the subthreshold current. The dashed arrows represent gate oxide leakage current which is present in the transistors when they are in the "OFF" state.

Fig. 4. Leakage current paths during the hold state for the 6T (baseline) cell.
Essentially, when the transistor is in the "ON" state it carries dynamic current along with the gate oxide leakage current and when the transistor is in the "OFF" state it will have gate oxide leakage current as well as subthreshold leakage current.

We discuss the hold state current paths in detail, as shown in Figs. 4 and 5, for 6T and 8T cells. In the hold state, the word line is disabled (WL = 0') and the bit lines (BL and BLB) are tied to '1'. Under this state, transistors M₁ and M₅ are in cut-off, carrying gate oxide leakage current. On the other hand, transistors M₂ and M₃ carry subthreshold leakage current and preserve the cell state (i.e. node Q = V_DD and node QB = '0'). However, in the 8T cell the read port (comprised transistors M₇ and M₈) adds two more leakage current components and increases overall leakage power, as shown in Fig. 5. Leakage power in both cells is measured as the power supplied by V_DD, when all word line and bit lines are connected appropriately and data storage nodes (Q and QB) are maintained appropriately for sufficient time to complete the operation under study.

4.4. SNM model and measurement

SNM can serve as a figure of merit in stability evaluation of SRAM cells. The SNM measurement model is described in this section. The SNM of even defect-free cells is declining with technology scaling, as discussed in the previous sections. SRAM cells with compromised stability can limit the reliability of on-chip data storage making them more sensitive to transistor parameter shift with aging, voltage fluctuations and ionizing radiation. Detection and correction/repair of such cells in modern scaled-down SRAMs become a necessity. Fig. 6(a) shows the simulation setup for the 6T cell SNM measurement, consisting of the two inverters (INV-1 and INV-2) in feedback and voltage sources V_T. The same SNM simulation setup can easily be extended for the 8T cell. In other words, the hold SNM setup is equivalent to the hold and read SNM setup of the 8T cell. The two voltage sources are static noise sources. A static noise source can be defined as DC disturbance and mismatch due to variations and processing in the operating conditions of the cell [26]. The two DC voltage sources V_T are placed in adverse direction to the input of the inverters of the SRAM circuit in order to obtain the worst case SNM. The SNM is the maximum amount of noise that can be tolerated at the cell nodes just before flipping the states. In order to obtain the butterfly curve shown in Fig. 6(b), the voltages are varied to and from nodes Q and QB alternatively. The SRAM cell is simulated for 45 nm CMOS technology using the PTM [3] with supply voltage V_DD of 0.9 V and with minimum sized transistors. The worst case SNM obtained from the butterfly curves is also shown in dotted lines in Fig. 6(b) and marked with a small circle.

Table 3 shows leakage power and SNM results for the baseline design (6T and 8T cells). The PVT condition is nominal process voltage variation and temperature is taken as room temperature or 27°C.

It may be noted that SRAM circuits have many other figures of merit, including read delay and hold SNM which can be considered for optimization. However, this particular paper is inspired by our earlier publication which demonstrates that read SNM is a very important figure of merit [27]. The current paper emphasizes mainly two figures of merit, power consumption and read SNM.

5. Statistical DOE–ILP algorithm for P2 optimization

This section discusses in detail the implementation of the statistical Design of Experiments (DOE)-Integer Linear Programming (ILP) algorithm, which is at the heart of the P2 optimization design flow.

5.1. The optimization algorithm

As shown in Algorithm 2, the baseline SRAM cells are taken as the input along with the baseline model file and high-threshold model file. The PVT condition is nominal process values for all devices, nominal power supply and the temperature is taken as room temperature or 27°C. We subject the baseline 6T and 8T cells to a DOE [28] based approach using a 2-Level Taguchi L₉ array. The factors are the V_T states of the different transistors of
the SRAM cells (Fig. 3). Each factor can take a high $V_{th}$ state (1) or a nominal $V_{th}$ state (0). The $I_{th}$ array provides different experimental runs for 6T and 8T cells. Monte Carlo simulations for $N$ runs are performed for each experiment trial. The mean ($\mu$) and standard deviation ($\sigma$) values of the resulting probability density function (approximated by a histogram) are recorded for average power and performance (SNM) of the SRAM cell. Thereafter, using DOE, predictive equations are formed for $\mu$ and $\sigma$ and are denoted by $\mu_{PWR}$, $\sigma_{PWR}$ for power and for SNM as $\mu_{SNM}$, $\sigma_{SNM}$. These predictive equations $\mu_{PWR}$, $\sigma_{PWR}$, $\mu_{SNM}$, $\sigma_{SNM}$ are considered to be linear equations with the constraints being high $V_{th}$ (or state 1) and low $V_{th}$ (or state 0). Each of these linear equations is then solved using Integer Linear Programming (ILP), depending on whether the quantity under consideration is to be maximized or minimized. The complexity of the algorithm otherwise would be $O(2^n)$ where $n$ is the transistor number.

We obtain the solution sets for mean and standard deviation of power as $S_{PWR}$ and the solution sets for mean and standard deviation for SNM as $S_{SNM}$. Since we are interested in power minimization and SNM maximization, we form our final objective $S_{SB}=S_{PWR} \cap S_{SNM}$ (since $S_{SB}$ is the intersection of the sets $S_{PWR}$ and $S_{SNM}$). This is the strength of the proposed algorithm: it allows simultaneous optimization of diverse and conflicting objectives. In the case of different objectives the optimization results in a set of transistors, not a specific value in terms of power or SNM. The sets are then combined depending on the multiple objectives targeted for optimization.

Based on $S_{SB}$, we assign high $V_{th}$ to the transistors of the cell, and re-simulate to obtain a P3 optimal design. The design flow achieves power reduction and read stability increase. Using this optimized cell, the design flow constructs the SRAM array. However, the scope of this paper has been kept at cell-level optimization.

**Algorithm 2. P2 optimization in nano-CMOS SRAM**

1. **Input:** Baseline PWR and SNM of the SRAM cell, baseline model file, high-threshold model file.
2. **Output:** Optimized objective set $S_{SB}=\{\mu_{PWR}, \sigma_{SNM}\}$ optimal SRAM cell 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 $V_{th}$ states of transistors of SRAM cell, the response for average power consumption is $\mu_{PWR}$, $\sigma_{PWR}$, and the response for read SNM is $\mu_{SNM}$, $\sigma_{SNM}$.
4. For Each 1:8 experiment of 2-Level Taguchi L-8 array do
   a. Perform N Monte Carlo runs.
   b. Record $\mu_{PWR}$, $\sigma_{PWR}$ and $\mu_{SNM}$, $\sigma_{SNM}$.
5. **end for.**
6. Form linear predictive equations $\mu_{PWR}$, $\sigma_{PWR}$ for power $\mu_{SNM}$, $\sigma_{SNM}$ for SNM.
7. Solve $\mu_{PWR}$ using ILP: Solution set $S_{PWR}$.
8. Same as Solve $\mu_{SNM}$ using ILP: Solution set $S_{SNM}$.
9. Solve $\mu_{PWR}$ using ILP: Solution set $S_{PWR}$.
10. Solve $\mu_{SNM}$ using ILP: Solution set $S_{SNM}$.
11. Solve $\mu_{PWR}$ using ILP: Solution set $S_{SB}$.
12. Form $S_{SB}=S_{PWR} \cap S_{SNM}$.
13. Assign high $V_{th}$ to transistors based on $S_{SB}$.
14. Re-simulate SRAM cell to obtain optimized objective set.

Monte Carlo simulations of 1000 runs are performed for each experiment. Therefore, we have a total of 6K (for 6T SRAM cell) and 8K (for 8T SRAM cell) Monte Carlo runs, taking 12 parameters into account. The 12 process parameters considered are as follows: (1) $T_{mean}$: NMOS gate oxide thickness (nm), (2) $T_{as}p$: PMOS gate oxide thickness (nm), (3) $L_{o}$: NMOS access transistor channel length (nm), (4) $L_{p}$: PMOS access transistor channel length (nm), (5) $W_{p}$: NMOS access transistor channel width (nm), (6) $W_{p}$: PMOS access transistor channel width (nm), (7) $I_{o}$: NMOS driver transistor channel length (nm), (8) $I_{p}$: PMOS load transistor channel length (nm), (9) $I_{n}$: NMOS channel doping concentration (cm$^{-3}$), (10) $I_{n}$: PMOS channel doping concentration (cm$^{-3}$). It may be noted that statistical information about these parameters may not be provided by the foundry. However, they are identified based on various published works [29]. The objective is to make the data characterization as accurate as possible for the current technology. Each of these process parameters is considered to have a Gaussian distribution with mean ($\mu$) taken as the nominal values specified in the PTM [3] and standard deviation ($\sigma$) as 10% of the mean. Amongst these parameters some are independent and others are correlated which is considered during the simulation. A correlation coefficient of 0.9 between $T_{mean}$ and $T_{as}p$ is assumed. The responses under consideration are the mean $\mu_{PWR}$ and standard deviation $\sigma_{PWR}$ of the average power consumption and also the mean $\mu_{SNM}$ and standard deviation $\sigma_{SNM}$ of the read SNM of the cell.

The experiments are performed and the half-effects are recorded using the following expression:

$$\Delta(n) = \frac{(\text{avg}(1)-\text{avg}(0))}{2}.$$  

(4)

where $\Delta(n)/2$ is the half-effect of the nth transistor, $\text{avg}(1)$ is the average value of power when transistor $n$ is in the high-$V_{th}$ state, and $\text{avg}(0)$ is the average value of power when transistor $n$ is in the nominal $V_{th}$ state.

We have taken normalized predictive equations in order to eliminate the effect of two different units, that is, nW for power and nV for SNM. The normalized predictive equations are

$$\bar{\Delta}(n) = \frac{\Delta(n)}{2} \times \frac{x_{n}}{\mu_{SNM}}.$$  

(5)

where $\bar{\Delta}(n)$ is the predicted response, $\bar{\Delta}(n)$ is the average of the responses, $\Delta(n)/2$ is the half-effect of the nth transistor, and $x_{n}$ is the $V_{th}$ state of the nth transistor.

5.2. P3 optimization of the 6T cell

The predictive equation for the mean of the average power consumption of the 6T cell is

$$\mu_{PWR_{6T}} = 0.29 - 0.24x_{1} + 0.05x_{2} - 1.0x_{4} + 0.43x_{5} + 0.03x_{6}.$$  

(6)

Here, $x_{i}$ represents the $V_{th}$ state of transistor $i$ (as in Fig. 3(a)). Fig. 7(a) shows Pareto plots of the half-effects of the 6T transistors for $\mu_{PWR_{6T}}$. From this, we formulate an ILP problem:

$$\min \mu_{PWR_{6T}}$$

s.t. $x_{i} \in \{0, 1\}$ \forall $i$

$$\mu_{SNM} > \mu_{SNM_{min}}$$

As we wish to minimize power consumption, we minimize $\mu_{PWR_{6T}}$. The constraints $1$ and $0$ represent coded values for high $V_{th}$ and nominal $V_{th}$ states, respectively. ILP has been used for small circuits, but the methodology is automated, and hence can be used for larger circuits. Solving the ILP problem, we obtain the optimal solution as $S_{PWR_{6T}} = \{x_{1} = 1, x_{2} = 0, x_{3} = 1, x_{4} = 1, x_{5} = 0, x_{6} = 1\}$. This can also be interpreted as transistors 1, 4, and 6 are high $V_{th}$ transistors, and transistors 2, 3, and 5 are minimal $V_{th}$ transistors.
Fig. 7. Pareto plot of 6T SRAM cell for (a) mean leakage power ($\mu_{\text{PWR}}$) and (b) standard deviation of leakage power ($\sigma_{\text{PWR}}$).

The Pareto plot of the half-effects for $\sigma_{\text{PWR}}$ of 6T SRAM cell is shown in Fig. 7(b). Similarly, Eq. (7) shows the predictive equation for the standard deviation of the leakage power consumption of the SRAM cell:

$$
\sigma_{\text{PWR}} = 0.26 + 0.03x_2 + 1.0x_4 - 0.453x_5 + 0.09x_6.
$$

(7)

From this, we formulate an ILP problem:

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

Since we seek to minimize the standard deviation of leakage power consumption, we minimize $\sigma_{\text{PWR}}$. Solving the ILP problem, we obtain the optimal solution as $S_{\text{PWR},\text{opt}} = \{x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 0, x_6 = 1\}$. This can also be interpreted as transistor 5 is high $V_{Th}$ and transistors 1, 2, 3, 4, and 6 are nominal $V_{Th}$ transistors.

The predictive equation for $\mu_{\text{SNM,6T}}$ for the 6T cell is

$$
\mu_{\text{SNM,6T}} = 0.42 + 0.44x_1 + 0.55x_2 + 0.48x_3 + 1.0x_4 - 0.02x_5 + 0.01x_6.
$$

(8)

Fig. 8(a) shows the Pareto plot of the half-effects of the transistors for $\mu_{\text{SNM,6T}}$ for the 6T cell.

Eq. (8) shows the predictive equation for mean of the read SNM of the 6T cell. From this, we formulate an ILP problem:

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

Since we want to maximize SNM, we maximize $\mu_{\text{SNM}}$. Solving the ILP problem, we obtain the optimal solution as $S_{\text{SNM,6T},\text{opt}} = \{x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 0, x_6 = 1\}$. This can also be interpreted as transistors 1, 2, 3, 4 and 6 are high $V_{Th}$ transistors, and transistor 5 is nominal $V_{Th}$ transistor.

Fig. 8(b) shows the Pareto plot of the half-effects of the transistors for $\sigma_{\text{SNM}}$. The predictive equation for $\sigma_{\text{SNM}}$ is formed as shown in Eq. (9). Next, we compute the standard deviation of the read SNM for 6T SRAM cell:

$$
\sigma_{\text{SNM,6T}} = 0.64 - 0.35x_1 + 0.57x_2 + 0.34x_3 + 0.56x_4 + 1.0x_5 - 1.0x_6.
$$

(9)

From this, we formulate an ILP problem for the 6T cell as follows:

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

As we want to minimize the standard deviation (which is an indication of the spread) of read SNM, we minimize $\sigma_{\text{SNM}}$. Solving the ILP problem, we obtain the optimal solution as...
\[ S_{SNM_{pr}} = \{x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0, x_6 = 1 \}. \] This can also be interpreted as transistors 1, 4 and 6 are high \( V_{th} \) transistors, and transistors 2, 3, and 5 are nominal \( V_{th} \) transistors.

Our final objective function \( S_{obj_{pr}} \) is formed as follows:

\[ S_{obj_{pr}} = S_{\mu_{PWR_{pr}}} \cap S_{\sigma_{PWR_{pr}}} \cap S_{\mu_{SNM_{pr}}} \cap S_{\sigma_{SNM_{pr}}}, \tag{10} \]

where \( \cap \) is interpreted as the set intersection operator. In other words, we pick devices which are part of low-power and high-SNM solution sets. We form normalized equations for power and SNM so that there is no unit interference. We obtain, \( S_{\mu_{PWR_{pr}}} = \{x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0, x_6 = 1 \} \), i.e. transistors 1, 4, and 6 are high \( V_{th} \) transistors, and transistors 2, 3, and 5 are nominal \( V_{th} \) transistors. Fig. 9(a) shows the P3 optimized standard 6T SRAM cell having high \( V_{th} \) transistors are hatched.

In order to demonstrate the effectiveness of the proposed algorithm (DOE-ILP P3-Optimization), we simulated the 6T and 8T cells for different technology nodes (45 and 32 nm). Figs. 10 and 11 show the DOE-ILP based dual-\( V_{th} \) assignment results of standard 6T SRAM cell. There is a marginal increase in the read SNM of the 45 and 32 nm nodes, while there is a significant reduction (60\%) in the mean leakage power under P3 optimized approach. However, the small increase in read SNM of the 6T cell is mainly due to the very strict optimization space available. These results are comparable to previous approaches which did not account for process variations [17].

5.3. P3 optimization of the 8T cell

The predictive equations for the mean and standard deviation of leakage power consumption of the 8T cell are

\[ \mu_{PWR_{pr}} = 0.3 + 0.24x_1 + 0.06x_2 + 1.0x_4 + 0.43x_5 + 0.01x_7 + 0.02x_8, \tag{11} \]

\[ \sigma_{PWR_{pr}} = 0.10 + 0.01x_1 + 0.03x_2 + 1.0x_4 + 0.44x_5 + 0.01x_7 + 0.01x_8. \tag{12} \]

Fig. 12(a) and (b) shows the Pareto plots of the half-effects of the transistors for \( \mu_{PWR_{pr}} \) and \( \sigma_{PWR_{pr}} \), respectively. From this, we formulate the ILP problem for minimization of \( \mu_{PWR_{pr}} \) and \( \sigma_{PWR_{pr}} \):

\[ \begin{align*}
\text{min} & \quad \mu_{PWR_{pr}} \\
\text{min} & \quad \sigma_{PWR_{pr}} \\
\text{s.t.} & \quad x_i \in \{0,1\} \quad \forall n \\
& \quad \mu_{SNM} > \mu_{SNM_{pr}}.
\end{align*} \]

Since we wish to minimize the leakage power consumption, we minimize \( \mu_{PWR_{pr}} \) and \( \sigma_{PWR_{pr}} \). Solving the above formulated ILP problem, we obtain the optimal solution as \( S_{\mu_{PWR_{pr}}} = \{x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0, x_6 = 1 \} \). This can be interpreted as transistors 1, 4, 6, 7 and 8 are high \( V_{th} \) transistors, and transistors 2, 3, and 5 are nominal \( V_{th} \) transistors. Similarly for \( S_{\sigma_{PWR_{pr}}} = \{x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 1, x_6 = 0, x_7 = 1 \} \). This can also be interpreted as transistors 4, 5, 7 and 8 are high \( V_{th} \) transistors, and transistors 1, 2, 3 and 5 are nominal \( V_{th} \) transistors.

Pareto plots of the half-effects of the transistors for \( \mu_{SNM_{pr}} \) and \( \sigma_{SNM_{pr}} \), respectively, for the 8T cell are shown in Fig. 13(a) and (b). Eqs. (13) and (14) show the derived predictive equation for mean and standard deviation of the read SNM of the 8T cell:

\[ \mu_{SNM_{pr}} = 0.40 + 0.91x_1 + 0.03x_2 + 1.0x_3 + 0.58x_4 - 0.04x_5 + 0.4x_6. \tag{13} \]

\[ \sigma_{SNM_{pr}} = 0.37 + 0.15x_1 + 0.35x_2 + 0.15x_3 - 0.33x_4 + 1.0x_5 + 1.0x_6. \tag{14} \]

In order to maximize the predictive equations formed above for \( \mu_{SNM_{pr}} \) and \( \sigma_{SNM_{pr}} \), we formulate an ILP problem:

\[ \begin{align*}
\text{max} & \quad \mu_{SNM_{pr}} \\
\text{min} & \quad \sigma_{SNM_{pr}} \\
\text{s.t.} & \quad x_i \in \{0,1\} \quad \forall n \\
& \quad \mu_{SNM} < \mu_{SNM_{pr}}.
\end{align*} \]

Fig. 10. Statistical mean and standard deviation of read SNM of a nominal and P3 optimized 6T SRAM cell for 45 and 32 nm technology node.

Fig. 11. Statistical mean and standard deviation of leakage power of a nominal and P3 optimized 6T SRAM cell for 45 and 32 nm technology node.

Fig. 9. P3 optimized (a) standard 6T and (b) read SNM free 8T SRAM cells; with hatched transistors having high \( V_{th} \).
As we want to maximize SNM, we maximize \( \bar{\sigma}_{\text{SNM}} \) and \( \bar{\sigma}_{\text{SNM}} \). Solving the ILP problem, we obtain the optimal solution as \( S_{\text{SNM}} = \{ x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0, x_6 = 1, x_7 = 1 \} \). This can also be interpreted as transistors 2, 3 and 5 are nominal \( V_{th} \) transistors, and transistors 1, 4, 6, 7 and 8 are high \( V_{th} \) transistors. Similarly for \( S_{\text{SNM}} = \{ x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 1, x_6 = 1, x_7 = 1 \} \). This can also be interpreted as transistors 2 and 3 are nominal \( V_{th} \) transistors, and transistors 4, 5, 7 and 8 are high \( V_{th} \) transistors.

Our final objective function \( S_{\text{obj}} \) is formed as follows:

\[
S_{\text{obj}} = S_{\mu \text{PWR} \cap S_{\Phi \text{PWR}}} \cap S_{\bar{\sigma}_{\text{SNM}}} \cap S_{\bar{\sigma}_{\text{SNM}}}.
\]

where \( \cap \) is interpreted as the set intersection operator. In other words, we pick devices which are part of low-power and high-SNM solution sets. We form normalized equations for power and SNM so that there is no unit interference because we wish to achieve a low power and high stability in our proposed design.
We obtain $S_{opt} = \{x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0, x_6 = 1, x_7 = 1$ and $x_8 = 1\}$, i.e. transistors 1, 4, 6, 7 and 8 are high $V_{th}$ transistors, and transistors 2, 3, and 5 are nominal $V_{th}$ transistors. Fig. 9(b) shows the P3 optimized 8T SRAM cell with the high $V_{th}$ transistors hatched.

Figs. 14 and 15 show the DOE-ILP based dual-$V_{th}$ assignment results obtained from the P3 optimized 8T cell, shown in Fig. 9(b). The absolute value of the read SNM of the 8T cell is $2 \times$ higher than the 6T cell. However, there is a 13% increase in the read SNM of the 45 and 32 nm nodes with the P3 optimization approach, while the standard deviation of read SNM is almost unchanged.

A significant leakage power reduction (51%) under P3 optimized approach is observed with marginal reduction in the standard deviation of the leakage power. These results are very promising and the proposed approach is more suitable for the read SNM free SRAM cells, such as 8T, 9T and 10T [30-35,5,6,7]. A 13% increase in read SNM of the 8T cell is almost equivalent to 30% of the total read SNM of the standard 6T cell as can be observed from Figs. 10 and 14. Fig. 16(a) and (b) shows the butterfly curves for the P3 optimized 6T and 8T cells simulated for the 32 nm node. The squares embedded inside the butterfly curves are a measure of the read SNM under process variation. It can be observed that the read SNM of the 8T cell is better than that of the 6T cell.

5.4. Comparative analysis of the results

In order to obtain a broad perspective of performance for the current algorithms, we compare with some indirectly related work here. The method presented in [9,20] is based on dual-$V_{th}$ and dual-$T_{ox}$ assignment for low power design while maintaining performance. In [9], a combined dual-$V_{th}$ and dual-$T_{ox}$ assignment is presented 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-$T_{ox}$ assignments, which require a larger number of masks and lithography steps during fabrication. In the current paper, we have taken into account substrate threshold and gate oxide leakage power which results in total improvement in leakage power by 60% for the 6T cell. For the 8T cell total improvement in leakage power by 61% and SNM by 13% is obtained. This is achieved by considering only dual-$V_{th}$, thus significantly reducing manufacturing costs as well.

The 6T and 8T SRAM cells presented in the literature were chosen to experiment with the proposed optimization methodology. It may be noted that the improvement of the power and SNM comes from the identification of the right transistors for proper $V_{th}$ assignment, not from sizing of the transistors. We anticipate that further sizing of the transistors along with $V_{th}$ assignment will further improve the results. However, the proposed optimization methodology is also applicable to other variants present in the literature. Our research is in full swing in SRAM circuit optimization [37,38]. The proposed algorithm and many similar algorithms are being investigated in our research. For example, a high-$x$/metal-gate based 10-transistor SRAM circuit is investigated for 32 nm technology in [38]. From the diverse experiments it is observed that the proposed algorithms are independent of SRAM circuit topology, CMOS technology node, and sizes.

6. Conclusions and future research

A statistical DOE-ILP approach has been presented in this paper for simultaneous P3 (power-performance-process) optimization of 6T and 8T SRAM cells simulated in 45 and 32 nm technology nodes. The read SNM has been treated as the performance metric. The optimization has been performed at cell level. For this, both SRAM cells of 45 and 32 nm have been subjected to the proposed approach which leads to 60% leakage power reduction and 13% increase in performance (read SNM). In order to achieve this objective, the novel statistical DOE-ILP approach is used for power minimization and SNM maximization. For process variation effect, 12 design and technology parameters are considered. As part of extension of this research, we plan to propose a P4VT optimal methodology (where the 4th "P" is parasitics and the "T" is thermal effects) will be incorporated in this study. Further future work of this research involves array-level optimization of SRAM where mismatch and process variation will be considered as part of the design flow.

Acknowledgments

This research is supported in part by NSF awards numbers CNS-0854182 and DUE-0942629. The authors would like to acknowledge the inputs of Dhruv Ghai and Garima Thakral, graduates of the University of North Texas.

References

Dhiraj K. Pradhan is currently a Professor of Computer Science at the University of Bristol, Bristol, UK. Previously, he was a Professor of Electrical and Computer engineering at Oregon State University, Corvallis, and held the COE Endowed Chair Professorship in Computer Science at Texas A&M University, College Station, where he also served as the founder of the Laboratory of Computer Systems, and he held a professorship at the University of Massachusetts, Amherst, where he also served as a coordinator of computer engineering. He also worked at different universities including Stanford University, Stanford, CA. Also he has worked as a Staff Engineer at IBM.

Dr. Pradhan is a fellow of the IEEE, ACM and the Japan Society for the Promotion of Science. He is the recipient of a Humboldt Prize in Germany. In 1997, he was also awarded the Fulbright-Flad Chair in Computer Science. He received Best Paper awards including the 1996 IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems Best Paper Award, with W. Kunz, on “Recursive Learning: A New Implication Technique for Efficient Solutions to CAD Problems Test, Verification and Optimization.” Dr. Pradhan’s research interests include low power designs, soft-error protection and variability issues.