# PROCESS-VOLTAGE-TEMPERATURE AWARE NANOSCALE CIRCUIT OPTIMIZATION

Garima Thakral, B.E., M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

December 2010

APPROVED:
Saraju P. Mohanty, Major Professor
Elias Kougianos, Co-Major Professor
Murali Varanasi, Committee Member
Armin Mikler, Committee Member
Bill Buckles, Coordinator of Graduate Studies
Ian Parberry, Chair of the Department of Computer Science and Engineering
Costas Tsatsoulis, Dean of the College of Engineering
James Meernik, Acting Dean of the Robert B. Toulouse School of Graduate Studies

Thakral, Garima. Process-Voltage-Temperature Aware Nanoscale Circuit Optimization. Doctor of Philosophy (Computer Science and Engineering), December 2010, 157 pp., 20 tables, 82 illustrations, 88 references.

Embedded systems which are targeted towards portable applications are required to have low power consumption because such portable devices are typically powered by batteries. During the memory accesses of such battery operated portable systems, including laptops, cell phones and other devices, a significant amount of power or energy is consumed which significantly affects the battery life. Therefore, efficient and leakage power saving cache designs are needed for longer operation of battery powered applications. Design engineers have limited control over many design parameters of the circuit and hence face many challenges due to inherent process technology variations, particularly on static random access memory (SRAM) circuit design. As CMOS process technologies scale down deeper into the nanometer regime, the push for high performance and reliable systems becomes even more challenging. As a result, developing low-power designs while maintaining better performance of the circuit becomes a very difficult task. Furthermore, a major need for accurate analysis and optimization of various forms of total power dissipation and performance in nanoscale CMOS technologies, particularly in SRAMs, is another critical issue to be considered.

This dissertation proposes power-leakage and static noise margin (SNM) analysis and methodologies to achieve optimized static random access memories (SRAMs). Alternate topologies of SRAMs, mainly a 7-transistor SRAM, are taken as a case study throughout this dissertation. The optimized cache designs are process-voltage-temperature (PVT) tolerant and consider individual cells as well as memory arrays.

ACKNOWLEDGEMENTS

I am most grateful to my Major Professor, Dr. Saraju P. Mohanty for admitting me to his research laboratory and providing me with continuous support throughout my Ph.D. program. He has always motivated me with his dynamic ideas and has been a great source of inspiration.

I wish to sincerely thank my Co-Major Professor, Dr. Elias Kougianos for his well advised approach of research problems. It has immensely helped me in defining my research.

I am sincerely grateful to Dr. Murali Varanasi for being in my doctoral committee. I would also like to extend my gratitude to Dr. Armin Mikler for his insightful suggestions. I have to sincerely thank them all for what I have achieved.

I am much indebted to my grandfather, Shri K.L. Thakral, for believing in me and for his strong support and guidance to achieve my goals. My parents provided me with their unconditional love and encouragement. I would also like to thank my little brother, Gaurav. They may be far physically but they are always close to my heart and soul. It would not have been possible to fulfill my aim without their support.

I would like to thank to my colleagues at Nano System Design Laboratory (NSDL, http: //nsdl.cse.unt.edu), Dhruva, Oleg, Asha and Karo for technically helping me with fruitful discussions. I cannot find proper words to adequately express and record my gratitude to my dedicated teachers right from Kindergarten to M.S. at the University of North Texas (UNT, http://www.unt.edu), who have made me worthy of being a successful research scholar. There are very many noble, generous and good people who have directly and indirectly helped me to achieve what I have done at UNT. I thank them all.

Above all, I am grateful and indebted to the benevolent founders and the enlightened, noble and generous management of UNT for providing me with the opportunity and facilities for achieving my M.S. as well as Ph.D.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

# CHAPTER 1

# INTRODUCTION AND MOTIVATION

It has been observed in the last few decades that memory is one of the driving forces behind the fast growth of nanoscale complementary metal-oxide-semiconductor (nano-CMOS) technology. Today's world has been changed by complex integrated circuits (ICs) consisting of billions of transistors with feature sizes of less than 45 nm, fabricated in plants that cost billions of dollars. The first IC was introduced in the year 1958 and consisted of a flip-flop using two transistors. The Intel Pentium 4 microprocessor, introduced in 2003, consisted of 55 million transistors. Today's high-end graphics processing units (GPUs) contain roughly 2 billion transistors. This demonstrates the exponential growth of very large scale integration (VLSI) technology. Thus, the demand for reliability, high performance, and high functional integration density of digital devices has made the scaling of CMOS devices inevitable. These devices give rise to a variety of leakage components because of junctions and capacitors present in the transistors. These leakage components are further aggravated by process and environmental variations and have a very serious impact on power and power-density budgets available for reliable computation. This leakage power contributes in heating issues, product cost, reliability and portability. Therefore, there is a major need for the careful analysis, characterization and optimization of various forms of total power dissipation (including leakage) in nano-CMOS technologies.

For accurate analysis, characterization, estimation and optimization of nano-CMOS designs, a case study is performed in this dissertation using memory circuits and state-of-the-art CMOS process. The background for this research is discussed in the following sections.

## 1.1. Nanoscale-CMOS Background

### 1.1.1. Technology Scaling

The technology scaling and demand for better performance of nano-CMOS circuits has enabled the embedding of millions of static random access memory (SRAMs) cells into

ICs. Technology scaling of CMOS implies reduction of the geometrical features (device characteristics) as well as process parameters of the transistors. The importance of technology scaling is growing. Technology shrinks by 0.7x per generation, with every new generation able to integrate 2x more functions per chip while the chip cost does not increase significantly. This level of complexity includes a great variety of circuits, such as large scale integration (LSI), very large scale integration (VLSI), ultra large scale integration (ULSI), and wafer scale integration (WSI). This research will emphasize memory circuits of VLSI IC complexity.

The idea behind scaling was initiated with Moore's law, which had a phenomenal impact as a driving force for the entire semiconductor industry. Gordon Moore (Intel Corporation) noted in 1965 that the main advantages of integration and technology scaling are the reduction of cost per function of the system. This implied that the cost of a component is nearly inversely proportional to the number of components. The principle idea here is that as the component integration is growing, the yield starts to deteriorate. He noted that "the number of transistors increased at a rate of approximately a factor of two every 18 months" [66]. Design goals for scaling have thus been well defined, the main goal being to double the transistor density, which will result in more components on the same chip.

1.1.1.1. Nanoscale-CMOS technology

The technology node is characterized by the channel length of the device that a process follows. Current trends in industrial processes have now reached the nanometer level. The industry has already put the 32 nm node process into production. The benefits of CMOS technology scaling are multifold including the following:

- Increase in chip density
- Increase in performance
- Increase in speed
- Decrease of device size

- Decrease of manufacturing costs
- Decrease in supply voltage
- Decrease in power dissipation

However, there are some obvious drawbacks that can be added to scaling, including the following:

- The power-delay-area product does not scale efficiently
- Parametric failures increase since these failures are caused by variations in device parameters
- Multiple effects on performance of device dimension reduction
- Highly expensive
- Issues in reliability and portability
- Increasing mask costs and fabrication related issues
- Increase in complexity
- Decreasing design flexibility
- Random dopant fluctuations
- Increasing process variations at nanometer levels
- Soft failures and hard failures

The International Technology Roadmap for Semiconductors (ITRS) is the main source of information about the current state of semiconductor technology. It is a sort of roadmap which identifies the major technological challenges which are faced by the semiconductor industry in the upcoming years. According to ITRS, the working effective gate length (year 2010) for general high-performance digital nano-CMOS process is approximately 45 nm with eventual reduction to as low as 9 nm by the year 2020 [61]. By this we observe how important is, and will be, the nano-CMOS process for at least a decade.

The ITRS has also made a very important observation regarding the power supply ($V_{DD}$) reduction from approximately 1 V today to as low as 0.5 V by the year 2020. The value

of $V_{DD}$ plays a very important role in the power dissipation profile [61]. This discussion supports the ITRS report for the redistribution of various sources of power-dissipation in nano-CMOS technologies shown in Figure 1.1 [61].

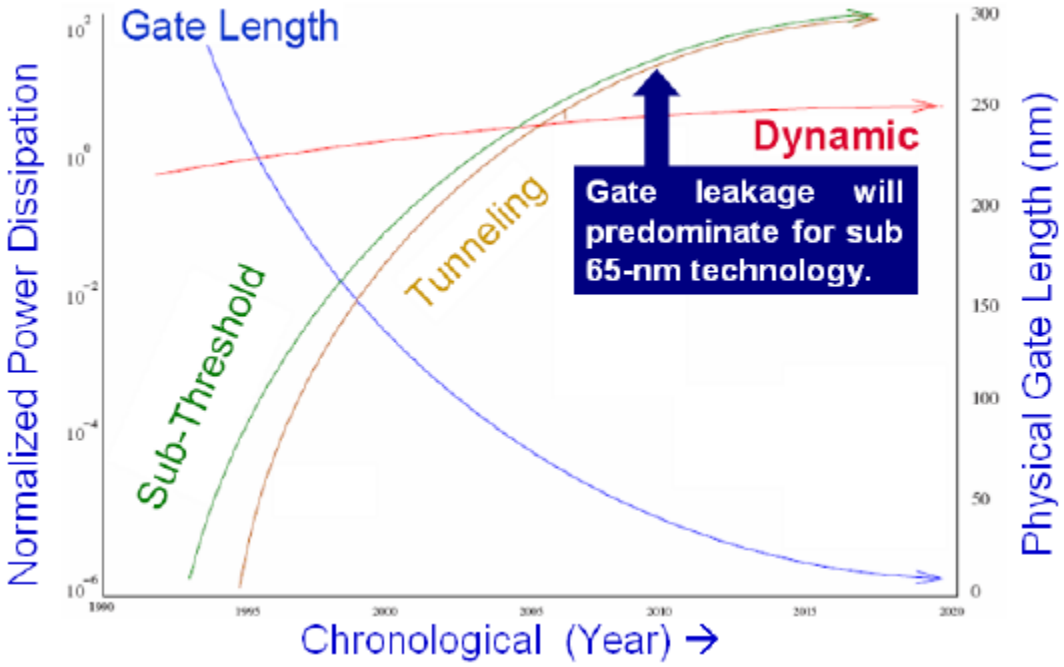



Figure 1.1. International Technology Roadmap for Semiconductors (ITRS) prediction of power dissipation sources in nano-CMOS transistors [61].

1.1.1.2. High-κ/Metal-oxide

The building blocks for transistors have followed Moore's law for more than a decade. Benchmark companies like Intel have come up with transistor gate dielectric scaling over the last fifteen years using silicon dioxide ($SiO_2$). The fact remains that, as transistor size decreases, leakages through gate dielectric may increase. Therefore, proper handling of leakage and reliable high-speed operations are very important and crucial in chip design. Intel has been somewhat successful in solving the power consumption problem for a chip by identifying new materials: "High-κ" (Hi-k) in order to replace the transistor's silicon dioxide gate dielectric, and further to replace with new metals the polysilicon gate electrode of N-channel metal-oxide semiconductor (NMOS) and P-channel metal-oxide semiconductor (PMOS) transistors. These new materials have advantages that reduce the gate leakage more than 100 times along with delivering high performance.

The dielectric constant κ basically means how much charge a material can hold. Materials like hafnium dioxide ($HfO_2$) and titanium dioxide ($TiO_2$) have a dielectric constant or "κ" above 3.9, the κ of silicon dioxide. In this case the transistor performance is dependent on the dielectric constant. The higher κ increases the transistor capacitance which allows the transistor to switch properly between "ON" and "OFF" states, with very low current when "OFF" and very high current when "ON."

The advantages of high-κ as a gate dielectric in a nano-CMOS transistor are listed below:

- Overcomes gate-leakage by over 100 times, which results in the transistor running cooler.
- Current flow will be at the same level like older materials.
- Low manufacturing costs are achieved by reducing the thickness by one molecular level at a time.
- Helps in device scaling.
- High-κ/metal gate transistors serve as a good alternative to classical CMOS transistors in nanoscale technologies.

1.1.2. Digital Circuits

The static random access memory (SRAM) digital circuit is used as an example case study throughout this dissertation. SRAM is a volatile memory that retains data as long as power is being supplied. It is more reliable and faster as compared to other memory devices. SRAM circuits are extensively used as cache memory in microprocessors and many application-specific circuits [66]. Although the increasing complexity of ICs has resulted in the still-pertaining issue of power dissipation and reliability, especially for battery-powered portable applications such as personal computing devices and wireless communication systems, in processor-based systems-on-chip (SoCs), the memories occupy an increasing part of the area budget and are the main contributor of the power dissipation.

Academics and industrial practitioners are continuously exploring different topologies for SRAM design in nanoscale processes for higher performance and minimum sized transistors. For robust design of nano-CMOS based SRAM, several issues including the total power consumption, process and supply variation, thermal aspects, and acceptable noise margins, need to be taken into consideration. The trend in nanoscale technologies is toward an increased contribution of the total power consumption, which is a major problem for the most frequent SRAM application, cache memories. In order to improve the reliability and portability of SoCs, the aim should be to build low-power memory circuit design. It has always been the aim of semiconductor industries to minimize the size of the components as much as possible while simultaneously improving their reliability and performance. A large number of integrated components are being implemented on the same die with limited area in order to achieve this goal.

An attempt has been made to systematically address the root causes of SRAM cell instability in scaled-downed technologies. Therefore, the objective of this dissertation is to bridge the gap between the challenges faced by technology scaling for SRAM circuit design, in particular. This dissertation gives an overview of various optimization techniques in order to control the most crucial and unresolved issues related to SRAM cell design. The optimization techniques are applied exhaustively on different topologies of SRAM cells using a seven transistor SRAM cell as the baseline design.

## 1.2. Nanoscale-CMOS Circuit Issues

The main issues faced by nano-CMOS circuits of today are discussed in the following subsections.

### 1.2.1. Power Dissipation and Leakage

Embedded systems, especially the real-time embedded systems which are targeted towards small-size or portable devices, for example, cell phones, PDAs, and MP3 players, must consume low power because they are battery powered. A major amount of power is

consumed in these devices during the memory accesses and that determines the battery life. Therefore, for the above mentioned reasons, efficient, active and leakage power saving SRAM designs are a must to explore for longer operation of battery powered applications while maintaining reliability.

As process technologies are scaled aggressively with the intense demand for increased integration, density and improved device performance are also increasing. This has resulted in very fast and high power dissipation computation models such as arithmetic units, etc.

High-speed computation units increase the chip temperature by developing local hot-spots. The power consumption and fluctuations present in power of a circuit affect its operational attributes. Increase in power dissipation is a disadvantage for battery life and further degrades the reliability because of the reduced efficiency. Another important parameter here is the power fluctuation which leads to power supply noise due to mutual inductance and capacitance (cross-talk). High heat dissipation may also lead to failure on the die. The magnitude of the leakage components depends on the following factors [61]:

- Device geometry
- Doping profiles
- Oxide thickness
- Supply voltage
- Temperature aspects

1.2.2. Propagation Delay

The delay of a CMOS device is approximately calculated by the following expression [72]:

(1) $$\tau_D = k \times \left( \frac{C_L V_{DD}}{\mu \left(\frac{\epsilon_{ox}}{T_{ox}}\right)\left(\frac{W_{eff}}{L_{eff}}\right)(V_{DD} - V_{Th})^{\alpha}} \right)$$

where $\tau_D$ is the delay, $k$ is a technology-dependent constant, $C_L$ is the load capacitance at the output of the transistor, $V_{DD}$ is the power supply voltage, $\mu$ is the electron surface mobility,

$\epsilon_{ox}$ is the permittivity of the oxide material, $T_{ox}$ is the thickness of the gate oxide, $L_{eff}$ and $W_{eff}$ are the device effective length and width, respectively, $V_{Th}$ is the threshold voltage of the transistor and $\alpha$ is the velocity saturation index, which varies from 1.4 to 2 for nanometer CMOS. The delay is generally calculated from the 50% level of the input swing to the 50% level of the output swing.

1.2.3. Performance

The latest SRAM cell designs strive to increase the bit count while maintaining low power dissipation and high performance. In order to achieve this, continuous scaling is required. The static noise margin (SNM) can serve as a figure of merit in performance evaluation of SRAM circuits [66]. The SNM is projected to reduce by 4x as scaling progresses in deeper nanometer regions. Therefore, accurate estimation of SRAM data storage stability in the pre-silicon design stage and verification of SRAM stability in the post-silicon design stage are equally important in overall SRAM design flow. As SRAM is scaled, sufficient SNM becomes difficult to maintain mainly due to increased variability. In particular, mismatch between voltage transfer characteristics (VTCs) of the two halves of the cell is enhanced because of dopant fluctuations.

1.2.4. Process and Supply Variation

Variability is another serious issue which occurs from the fluctuation of device characteristics due to process variation and is continuously growing in nanometer technologies [61, 55]. The parametric variations are combination of wafer, reticle, and local variations. The process variations are diverse in nature and originated from various sources including the following: ion implantation, chemical mechanical polishing (CMP), chemical vapor deposition (CVD), subwavelength lithography, lens aberration, materials flow, gas flow, thermal processes, spin processes, microscopic processes, and photo processes [55]. Process variations are classified into inter-die and intra-die variation for the purpose of analysis and characterization. Both intra-die and inter-die variations have significant

importance when analyzing the performance of a circuit and while predicting the yield of a chip. Modeling the intra-die process variation is essential to device and interconnect extraction tools for accurate timing and power analysis. Therefore the design cycle must include process variation to construct variation-tolerant digital designs.

1.2.5. Thermal or Temperature Variation

Aggressive scaling in nano-CMOS technology has resulted in increased chip density and maintaining the stability of the cells is becoming a major issue. Power dissipation and performance worsen at high temperatures because of the increase of leakage currents and the reduction of carrier mobility. As a result of the on-chip heat generation, the different portions of a SRAM circuit may experience different temperature profile depending on their proximity to other logic units. High computation logic units may be used to lessen this but they also increase the temperature of components by creating hot-spots due to on-chip temperature [53]. The maximum temperature that can be reached by a chip during its operation is increasing. These factors affect reliability and cooling costs. Ambient temperature analysis explored in this dissertation, has observed how the SRAM cell behaves in operating or environment temperature conditions.

1.2.6. Parasitics

The parasitics are the circuit elements which are manifested as due to non-ideal device and interconnect materials and structure. These include resistors (*R*), inductors (*L*), capacitors (*C*), mutual inductors (*K*). It is evident from the prior research and also from current literature that parasitics should be considered at the beginning of the design of any circuit because they cause degradation in performance [60]. In other words, parasitics along with process variation can lead to severe degradation in circuit performance. The design cycle must include process variations along with parasitics to produce variation-tolerant circuits. It affects the performance of some global interconnects in modern designs. On the basis of

design and technology specifications a physical design is converted into a netlist composed of resistors, capacitors and inductors.

### 1.2.7. Reliability

Circuit reliability is decreasing exponentially because of the combined effects from variations, power, and thermals. As the technology is scaling deeper into the nano regime along with the lower power voltages, and higher frequencies, it negatively affects the reliability of circuits and as a result increases the number of occurrences of transient faults. The small interconnect sizes and high operating frequencies increase the number of errors and affect stability significantly. It is noted that the reliability of processors and memories improved during the first half of the decade and a significant decrease of failure rates for SRAMs has been achieved during the past four years. The two most common features used to increase the reliability of VLSI circuits are fault avoidance and fault tolerance. These two important issues depend on improved materials, manufacturing processes, and circuit design.

### 1.2.8. Yield

Yield is an indicator of the number of healthy chips coming out of a manufacturing plant. This has strong impact on the cost of the chip. Circuit yield is decreasing due to increased variability. As the technology is scaling deeper, the reliable patterning and fabrication of shrinking device features is becoming more difficult with time.

## 1.3. Contribution and Organization of this Dissertation

The contributions of this dissertation are to present accurate power analysis, characterization and estimation in a case study SRAM circuit considering state-of-the-art CMOS processes. Novel Optimization methodologies are presented as integral parts of the proposed design flows. This dissertation is organized as follows: Prior research related to SRAM topologies is discussed and various optimization techniques are presented and compared with the work done in this dissertation and are summarized in Chapter 2. Chapter 3 discusses alternate topologies of SRAM cell design. Chapter 4 explores accurate leakage and power

analysis in nano-CMOS digital circuits and describes how performance metrics are defined and calculated. In Chapter 5 low power design techniques are presented followed by novel optimization methodologies (introduced in this dissertation) to achieve low power and high performance SRAM circuits. Chapter 6 presents the process-voltage variation analysis and process variation optimization using a 7-Transistor SRAM cell as a case study. This chapter also includes simultaneous optimization of power, performance and process variation awareness. Chapter 7 presents the process-voltage-temperature (PVT) tolerant SRAM design for a sample circuit using one of the novel optimization techniques. The results are summarized and the dissertation is concluded with a discussion of future directions of research in Chapter 8. Abbreviations and symbols used in this dissertation are given in Table 1.1 and Table 1.2.

Table 1.1. Abbreviations

| | |
|---|---|
| CMOS | Complementary metal oxide semiconductor |
| DOE | Design of experiments |
| ILP | Integer linear programming |
| ITRS | International technology roadmap |
| MOS | Metal oxide semiconductor |
| NMOS | N-channel MOS transistor |
| P2 | Power and performance |
| P3 | Power, performance and process variation |
| PMOS | P-channel MOS transistor |
| SRAM | Static random access memory |
| VLSI | Very large scale integration |
| SNM | Read static noise margin |
| SoC | System-on-chip |

Table 1.2. List of Symbols

| | |
|---|---|
| $C_L$ | Load capacitance |
| $I_{dynamic}$ | Dynamic current |
| $I_{gate\text{-}oxide}$ | Gate oxide current |
| $I_{subthreshold}$ | Subthreshold current |
| $I_{total}$ | Total current |
| $L$ | Device length |
| $P_{dynamic}$ | Dynamic power consumption |
| $P_{gate\text{-}oxide}$ | Gate-oxide power |
| $P_{subthreshold}$ | Subthreshold power |
| $P_{total}$ | Total power consumption |
| $V_{Th}$ | Threshold voltage |
| $V_{DD}$ | Supply voltage |
| $W$ | Device width |
| $\alpha$ | Velocity saturation index |
| $\epsilon_{ox}$ | Oxide permittivity |
| $\mu$ | Mean of distribution or mobility |
| $\sigma$ | Standard deviation of distribution |
| $\tau_D$ | Delay time |

# CHAPTER 2

# RELATED PRIOR RESEARCH IN SRAM DESIGN

This chapter comprises of current and prior research related to static random access memory (SRAM) design, simulation, and optimization techniques. Section 2.1 deals with literature on power minimization and discusses one of the approaches of power reduction techniques, that is, dual-$V_{Th}$ which is thoroughly investigated in this dissertation in later chapters. Section 2.2 reviews ongoing research in performance related areas of SRAM and in Sections 2.3 and 2.4 process and temperature variation related research results are reviewed. This discussion is followed by current literature related to SRAM circuits discussed in Section 2.5. Some of the most significant ones related to the research done in this dissertation are presented in Table 2.1.

## 2.1 Power and Leakage Dissipation Minimization

The existing literature is rich on the design of the SRAM for low power operation in the deep sub-micron and nanometer technology ranges. Standby power reduction is achieved in [42] for SRAM cell optimization for 65 nm complementary metal-oxide semiconductor (CMOS) technology. The research in [50] achieves 31.9 nW (accounting only leakage) power consumption (which shows 22.9% reduction) and 300 mV of static noise margin (SNM) (which is 50% increase in read stability margin). Liu et al. [49] achieve this target by using a separate data access mechanism. In [38], the leakage consumed by the SRAM is quoted as 0.11 μW (accounting leakage dissipation only) and the SNM is 78 mV which accounts for a 58% increase; this is achieved by using a Schmitt-trigger mechanism. A 9-transistor SRAM cell is proposed in [48], which increases the stability and reduces power consumption compared to a traditional 6-transistor SRAM. A design of experiments-integer linear programming (DOE-ILP)-based approach presented in [83, 82] accounts for total power reduction and static noise margin (SNM) improvement.

Table 2.1. Summary of Related Prior Research in SRAM.

| SRAM research | Power value ($\mu W$ or $nW$) | SNM value ($mV$) | Technology node | Temperature | Research Techniques |
|---|---|---|---|---|---|
| Agrawal [3] | - | 160 mV (approx.) | 65 nm | - | Modeling based approach |
| Liu [49] | 31.9 nW (leakage) | 300 mV | 65 nm | 70° C | Separate data access mechanism |
| Kulkarni [38] | 0.11 µW (leakage) | 78 mV | 130 nm | 27° C | Schmitt-trigger |
| Lin [48] | 4.95 nW (standby) | 310 mV | 32 nm | 25° C | Separate read mechanism |
| Bollapalli [11] | 10 mW (total) | - | 45 nm | - | Separate word line groups |
| Azam [7] | 63.9 µW vs 44.4 µW | 299 mV (total) | 45nm | - | Separate read/write assist circuitry |
| Singh [73] | - | -(total) | 65 nm | 27° C | Two-port 6TSRAM with multiport capabilities |
| Thakral [82] | 314.5 nW | 295 mV | 32 nm high-κ/metal gate | 27° C | Conjugate gradient approach |
| Nalam [64] | - | -(leakage) | 45 nm | - | Two-phase write and split bitline differential sensing |
| Amelifard [5] | - | - | 65 nm | | Dual $V_{Th}$ and $V\varepsilon_{ox}$ |
| Singh [75] | - | 305 mV | 65 nm | 27° C | Subthreshold 7T SRAM |
| Tavva [80] | - | 400 mV | 65 nm | 125° C | Novel 9T SRAM cell topology |
| Thakral [83] | 113.6 nW (total) | 303.3 mV | 45 nm | 27° C | Statistical DOE-ILP for dual-$V_{Th}$ |
| Thakral [85] | 100.5 nW (total) | 303.3 mV | 45 nm | 27° C | Combined DOE-ILP |

In [83], the authors have applied a statistical DOE-ILP algorithm which leads to 113.6 nW, that is, a 44.2% decrease in total power and leakage and 303.3 mV (43.9% increase) in SNM by using a 7-transistor SRAM circuit and also prove that the SRAM cell is process variant tolerant. Further, in [82], 86% reduction in total power and 8% SNM improvement are achieved. This is obtained using a 10-transistor SRAM sample circuit. The research in [11] quoted 10 mW of total power, that is, 53.4% reduction, by applying separate word line groups using the 45 nm CMOS technology node. In [85], a DOE-ILP-based methodology is proposed for dual-$V_{Th}$ assignment where the total (accounting dynamic and leakage) power dissipation achieved is 100.5 nW (which is 53.5% decrease) and SNM of 303.3 mV (which is 43.0% improvement). In [43], gate-oxide leakage current is studied with respect to logic gates and further introduces a pin ordering management scheme to reduce this gate leakage current. In [63], Mukhopadhyay et al. modeling and estimating total leakage current components of CMOS devices is investigated by considering the effects of parameter variation which will test the robustness of a circuit.

Multiple threshold CMOS have been used in [65] for subthreshold current reduction. In [36] the authors introduce metrics to quantify steady and transient gate leakage in nanoscale transistors [59]. A novel low-power SRAM cell design with SOI (silicon-on-insulator) technology is presented in [86]. In [68], leakage current mechanisms are studied and have present techniques for leakage reduction. This research emphasizes intrinsic features of transistor leakage mechanisms such as weak inversion, drain-induced barrier lowering and gate oxide tunneling. Furthermore, various mechanisms are investigated in order to achieve reduced leakage power consumption. Source [51] discusses leakage power estimation in SRAMs. In [8], the authors propose an asymmetric SRAM cell to lower gate leakage.

In [5] and [6], the authors have implemented a combined dual-$V_{Th}$ and dual-$\epsilon_{ox}$ assignment where leakage power reduction is 53.5% and SNM increase is 43.8%. However, in this

research, dynamic power is not considered during optimization. The desired results are obtained by using both dual-$V_{Th}$ and dual-$\epsilon_{ox}$ assignment, which will need more number of masks during fabrication of the SRAM chip. Similarly, in [81] a low-power and robust 7-transistor SRAM circuit is shown using the dual-$V_{Th}$ technique. In [84], a DOE-ILP based methodology is proposed for dual-$V_{Th}$ assignment where the total (which dynamic and leakage) power dissipation achieved is 100.5 nW (which is 53.5% decrease) and the SNM is 303.3 mV (which is 43.0% increase).

In [40], a dual-$V_{Th}$ assignment is applied over logic elements of a novel field-programmable gate array (FPGA) architecture. By setting some of the transistors to high $V_{Th}$ and low $V_{DD}$ allows maintaining overall performance while reducing leakage current [77, 39]. In [76], multiple channel lengths and multiple gate oxide thicknesses are used for reduction of leakage. Dual-$V_{DD}$ and dual-$V_{Th}$ designs are becoming increasingly popular because of the rising leakage current levels of ultra-small metal-oxide semiconductor field-effect transistors (MOSFETs) [18, 78]. To maintain the speed and performance of a circuit, a reduction in $V_{Th}$ is needed which, however, causes tremendous increase in leakage current.

Source [24] analyzes dual-$V_{Th}$ SRAM cells with single-ended bit line sensing for on-chip cache. In [34], subthreshold leakage current has been reduced by applying a dual-$V_{Th}$ technique. The aim is to target least used modules as the candidates for leakage optimization. In [22, 23] subthreshold current has be kept under control by using the multi-threshold CMOS approach. They propose binding algorithms for power, delay, and area trade-off and achieve this by a clique partitioning approach in [23]. In addition, a knapsack-based binding algorithm is explored in [22] for the same purpose. A pre-defined dual-$V_{DD}$ and dual-$V_{Th}$ fabric is used to present a low-power FPGA in [46].

## 2.2. Performance Maximization

Because of scaling of process technology nodes, the demand for high stability of SRAM and other digital circuits is a growing concern. The SRAM stability margin or SNM is

defined as the maximum amount of direct current (DC) noise voltage that a cell can handle [66]. Throughout this research, the SNM is considered as the primary figure of merit for the performance of SRAM circuits. In [80] the authors have proposed a novel 9-transistor SRAM cell which aims to reduce the bitline leakage power consumption while enhancing data stability, resulting in a 400 mV SNM. In [7], 299 mV of SNM is achieved at 45 nm CMOS using separate read/write circuitry.

In [3], the SRAM stability margin is quoted as approximately 160 mV through a modeling based approach. In addition, the stability of SRAM cells is analyzed in the presence of random fluctuations using a modeling based approach. The authors in [75] state that the SNM is 305 mV using a subthreshold 7-transistor SRAM cell. The research in [83] reports a 303.3 mV SNM using a 7-transistor SRAM circuit. In [82] it is given as 295 mV (which 8% increase) and in [85] the SNM is an improvement of 43.9%. In [9], authors discuss static and dynamic stability improvement strategies for 6-transistor CMOS low-power SRAMs. In [69], authors present a design and analysis of a 32 nm CMOS power-voltage-temperature (PVT)-tolerant CMOS SRAM cell for low leakage and high stability.

In [35] a Schmitt-trigger has been applied in a 130 nm CMOS technology node to attain an SNM of 78 mV. The research in [49] reports 300 mV which represents a 50% increase in SNM. The authors in [17] have presented an SNM model and discuss how to measure and calculate the stability of SRAM cells. In [4] the impact of random device variation on SRAM cell stability in sub-90 nm CMOS technologies is discussed. The authors in [71] present a novel high write speed, low power, read-SNM-free, 6-transistor SRAM cell. In [16], an accurate analytical SNM modeling technique is presented for SRAMs based on a Butterworth filter function. In [31], a low leakage and SNM free SRAM cell is presented in deep submicron CMOS technology.

2.3. Process and Supply Variation

The scaling of device technology is down to the nanometer region. As a consequence, the effect of intra-die variations such as gate oxide thickness ($\epsilon_{ox}$) and threshold voltage ($V_{Th}$) variations have become as important as inter-die variation when analyzing circuit performance and predicting the yield of a chip [10, 12]. In [30] 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. As the concern for intra-die variation is growing, some works which model process variations and perform timing analysis have been proposed. In [38], authors present a process variant tolerant SRAM cell array for ultra-low voltage applications. In [74], a nano-CMOS process variation induced read failure tolerant SRAM Cell is presented. In [60], the authors have shown the effect of simultaneous variation of supply and process parameters on power consumption of data-path components. In [59], the authors have developed process variation aware component libraries. This research focuses on behavioral (or high-level or register-transfer level) synthesis. An increase in the number of variation sources has led to even more corner cases that need to be simulated for each design. Future processes can have even larger amounts of variation.

In [28], authors have presented a voltage-scalable and process variation resilient hybrid SRAM architecture for MPEG-4 video processors. The research in [70] has successfully explored the process variation awareness in cache design for aggressive voltage-frequency scaling. The authors in [44] consider process-induced $V_{Th}$ variations using a dual-$V_{Th}$ technique. Authors in [19] implement probabilistic analysis to $V_{Th}$ variation. In [45], a low leakage SRAM array is presented which is robust to process variation by applying a dual-$V_{Th}$ technique.

2.4. Temperature Analysis

Different sections of an SRAM circuit may experience different temperature profiles depending on their proximity to other modules. The leakages present because of power

dissipation of the circuits cause self-heating, which affects the stability and performance of the SRAM. A variety of current and prior research has been done in thermal analysis alone, because it is one of the crucial issues for nano-CMOS circuits. For example, in [53] thermal analysis is carried under the influence of hotspots using an 8-transistor SRAM cell as a sample circuit. In [80] the authors have explored the stability and power consumption of an SRAM cell at high temperatures (125°C). In [38], authors also consider the power dissipation and stability figure of merits taking into account the temperature effect at 70°C. The research done in [47] elaborates the effects on energy efficiency and thermal nature of design style and clock-gating in different structures. It presents the power and thermal effects of SRAM versus latch-multiplexer design styles and clock gating choices.

### 2.5. SRAM Circuits

Current literature on different topological circuits of SRAMs is discussed in this section. The main circuits which are widely used because of their stability and robust functionality at nano-scale technologies are compared. Three main SRAM circuits are widely used for various applications in the electronics industry today, which are the following:

6-transistor SRAM (6T-SRAM)

7-transistor SRAM (7T-SRAM)

10-transistor SRAM (10T-SRAM)

#### 2.5.1. 6-transistor SRAM

Six-transistor SRAM circuit is the standard design which has been used for several decades because of its high stability performance. In [38], 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. In [73], a slightly different topology in standard 6-transistor SRAM is proposed in which the 6T-SRAM design is a 2-port bitcell design with multi-port capabilities at reduced area overhead. In [29], the authors have introduced a single-ended, 6T-SRAM cell design for ultra-low-voltage applications. In [9],

static and dynamic stability improvement strategies are explored for 6T-SRAM. The authors in [2] discuss a 6-transistor hybrid SRAM cell in sub-32 nm double-gate CMOS technology. The authors in [52] explain the limits of bias-based assist methods in nanoscale CMOS taking a 6-transistor SRAM as the sample circuit.

2.5.2. 7-transistor SRAM

In [75], a 7-transistor read-failure tolerant SRAM topology is introduced, which is suitable for low voltage applications. In [7], the authors explored variability resilient low-power 7T-SRAM design for nanoscale technologies. The research proposed in [83] and [85] is based on 7-transistor SRAM designs where the figures of merit taken into account are total power and SRAM performance. In [81], the authors have optimized power and achieved robust 7T-SRAM using a dual-$V_{Th}$ SRAM circuits. In [54], a look-ahead dynamic threshold voltage control scheme is investigated for improving the write margin of SOI-7T-SRAM.

2.5.3. 10-transistor SRAM

In [20], the authors have explored the 10T-SRAM topology using column-assist scheme. In [82] an approach is proposed to optimize the total power and SNM of a 10T-SRAM using DOE assisted conjugate gradient method. In [74], a nano-CMOS process variation induced read failure tolerant SRAM cell is introduced. In [27], a 10-transistor non-precharge 2-port SRAM is proposed for 74% power reduction for video processing.

# CHAPTER 3

# TOPOLOGIES FOR SRAM CIRCUITS

Researchers in academia and industry are continuously exploring alternative topologies for static random access memory (SRAM) design in nanoscale regions for high performance, fault-tolerant, highly stable, and robustness with minimum sized transistors. To discuss these topologies in detail and explore the features of SRAM cell designs, this chapter presents different circuit topologies used SRAM circuits. Transistor-level designs are constructed for nanoscale complementary metal-oxide semiconductor (CMOS) technologies. For each of the circuits, different modes of operation are presented. The current flow paths which help in accurate power analysis are identified. In these sample circuits different technologies are explored to study the behavior of SRAM circuit on different process nodes.

## 3.1. Six-Transistor Nano-CMOS SRAM Circuit

### 3.1.1. Logical Design of the 6-Transistor SRAM

SRAM is the main building block of cache memory in modern microprocessors. A significant area of the system-on-a-chip (SoC) is occupied with cache. Cache memory itself consists of more than 90% of transistors in some designs [66]. Therefore, understanding the SRAM design along with its current paths during operation is crucial for proper functioning of chip design and manufacturing. An SRAM cell stores binary information in the form of two states denoted by “0” and state “1.” These bits are stored on four transistors that form two cross-coupled inverters in SRAM. Two additional access transistors help controlling the access to the cross coupled unit formed by the inverters during Write (W) and Read (R) operations. Primarily, six transistors are used to store one memory bit. An SRAM circuit is intended to provide nondestructive writing ability, read access, and storage or holding data as long as the cell is powered. A cell must be designed in such a way that it strikes a balance among cell area, speed (access time), and leakage (which contributes to the total power consumption). The main objective is to minimize the cell area; by achieving this, a cell

allows a larger number of bits per unit area and decreases the cost per bit. By reducing the cell area speed is automatically improved, but at the same time reducing transistor dimensions cannot be achieved by compromising other parameters.

The logic design of the 6-transistor (6T) SRAM cell is shown in Figure 3.1. The naming convention is as follows:

WL : Wordline

BL : Bitline

$\overline{BL}$: Negated bitline

P1 : PMOS transistor of inverter 1

N1 : NMOS transistor of inverter 1

P2 : P-channel metal-oxide semiconductor (PMOS) transistor of inverter 2

N2 : N-channel metal-oxide semiconductor (NMOS) transistor of inverter 2

N3 : Access transistor

N4 : Access transistor

The function of the wordline (WL) of the SRAM is to grant access to the cell when it is asserted high; it controls the two access transistors N3 and N4. Further, this allows the access of the memory cell to the bit lines: BL and $\overline{BL}$. For both Write and Read operations these are used to transfer data. The presence of dual bit lines, i.e., BL and $\overline{BL}$, improves noise margins over a single bit line. This structure makes the SRAM function faster comparatively. Standard 6-transistor SRAM design accepts all address bits at the same time whereas dual random access memory (DRAM) has the address multiplexed in two halves, i.e., higher bits followed by lower bits; this contributes to the fact that SRAM is faster than DRAM. The operation of a CMOS SRAM cell is described in terms of three states, *viz*., Write, Read, and Idle or Hold operations. The 6T-SRAM is known as the standard and traditional SRAM design because of its low-power, low voltage operation nature. Transistor size is one of the most crucial parameters which should be considered in order to have successful write and

read operation in SRAM cell.

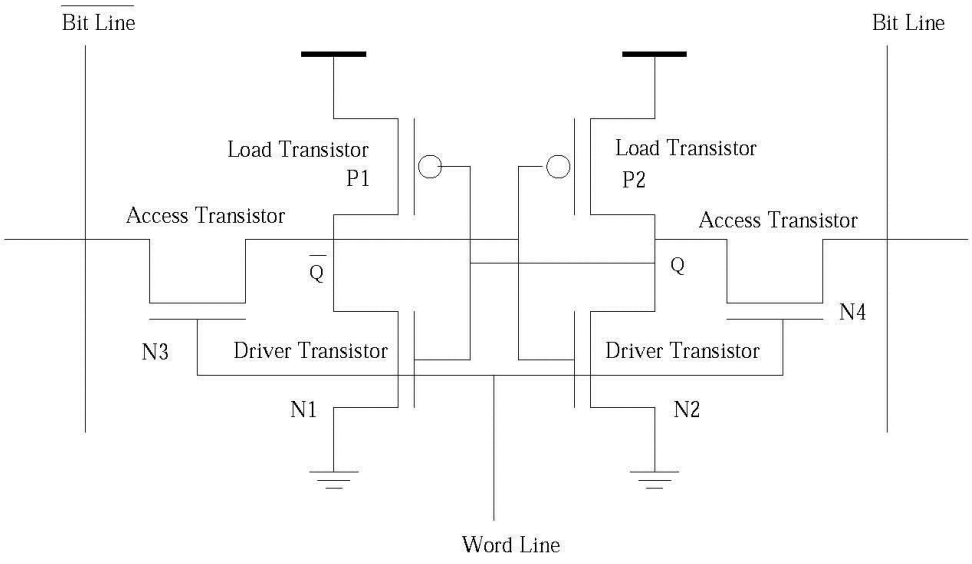



Figure 3.1. Circuit diagram of the 6-transistor (6T) CMOS SRAM cell.

3.1.2. Operation Modes of the 6-Transistor SRAM

The three states of the SRAM circuit, such as “Write,” “Read,” and “Hold,” are discussed below.

- Write operation (refer to Figure 3.2):

The Write cycle of this design is initiated by applying the value to be written to the bit lines. To write a “0,” a “0” is applied to the bit lines, i.e., setting $\overline{BL}$ to “1” and BL to “0.” To write a “1” the values of bitlines are just inverted. The word line (WL) is kept high and the value to be stored is latched in. The bitline is driven from the precharged value ($V_{DD}$) to ground potential by a write driver through transistor N4. The sizes of transistors P2 and N4 should be proper so that the cell is flipped and the data is effectively overwritten. The write margin is a property which explains the statistical measure of SRAM cell writeability. It is defined as the minimum voltage of bitline which is required to flip the state of an SRAM cell [66]. The bitline transistors are stronger than the weak transistors in the cell so that they can easily override the last state of cross-coupled inverters. Because of this design, careful sizing of transistors is a must in order to attain proper operation of an SRAM cell.

- Read operation (refer to Figure 3.3):

The Read cycle is started by activating the word line WL which means enabling both access transistors N3 and N4. Next stage comes when the values stored in Q and $\overline{Q}$ are transferred to the bit lines BL and $\overline{BL}$ through transistors N1 and N3. Whereas, on the BL side, the transistors P2 and N4 pull the bit line towards $V_{DD}$ (when a “1” is stored at Q), depending on the content of the memory wherever a “0” appears, the reverse would happen: BL will be pulled toward "1” and $\overline{BL}$ toward “0.”

- Idle or Hold operation (refer to Figure 3.4):

For the Idle state, the function of access transistors N3 and N4 is to disconnect the cell from the bit lines and word line is on the other hand is not asserted. Cross coupled inverters N1, N2, N3, and N4 will continue to re-enforce each other as long as they are disconnected from any outside circuit element.

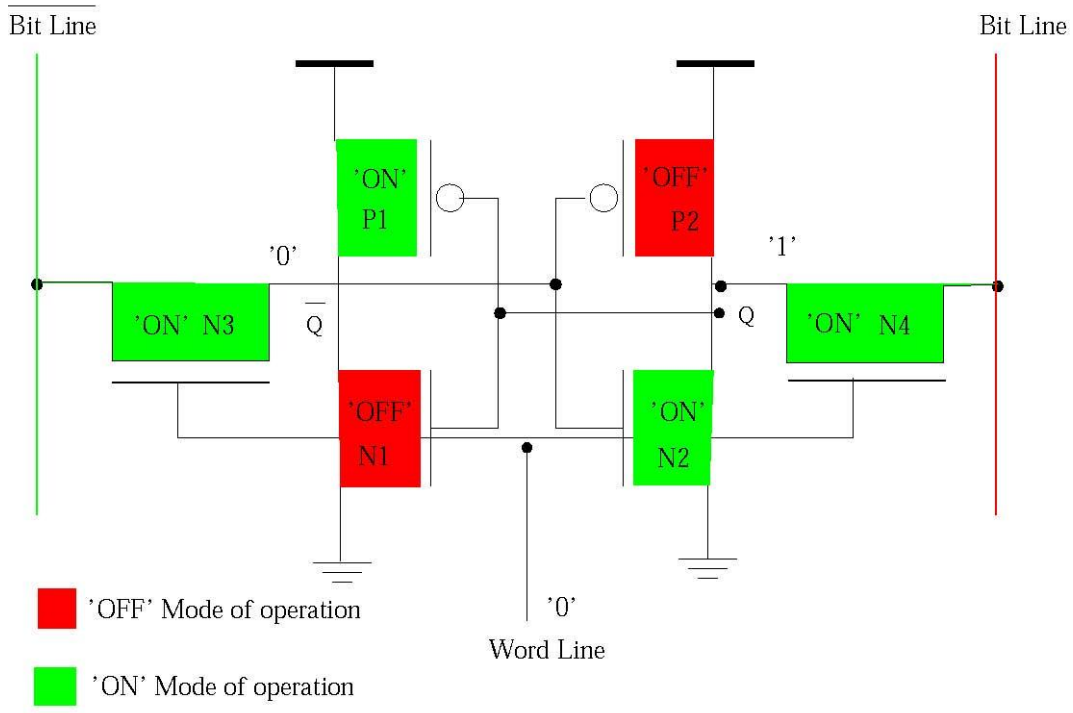
3.1.3. Current Flow Paths in 6-Transistor SRAM

A nano-CMOS device in digital circuits operates in three regions as follows:

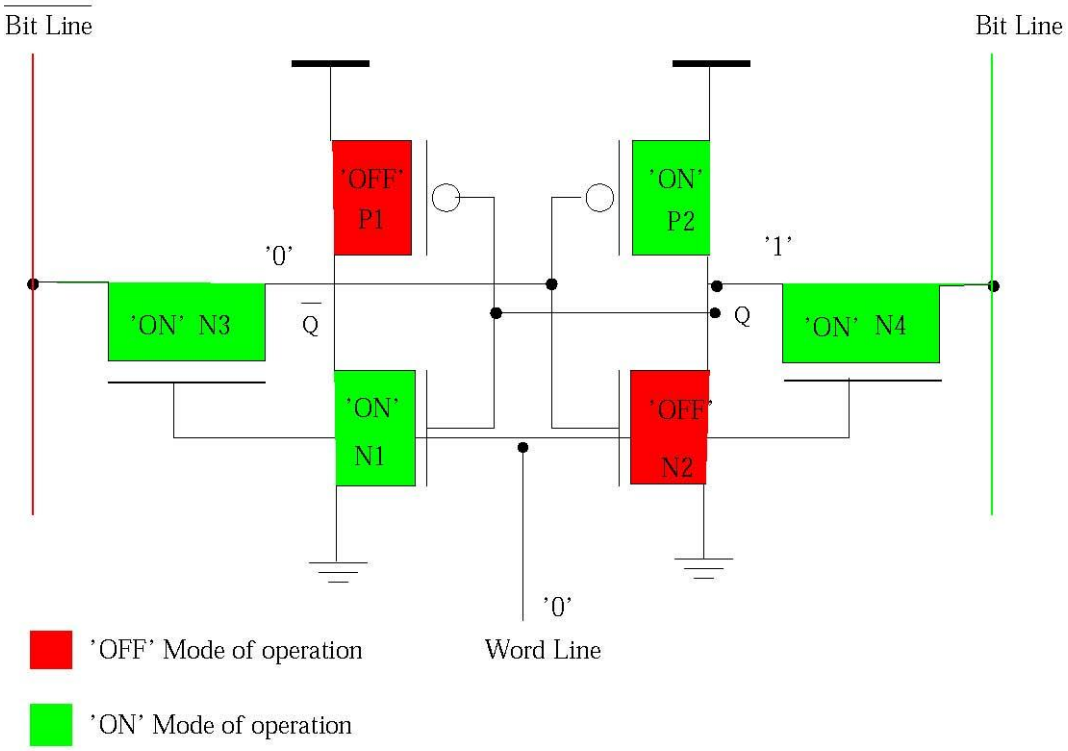
ON State

OFF State

Transition State, i.e., ON to OFF and OFF to ON.

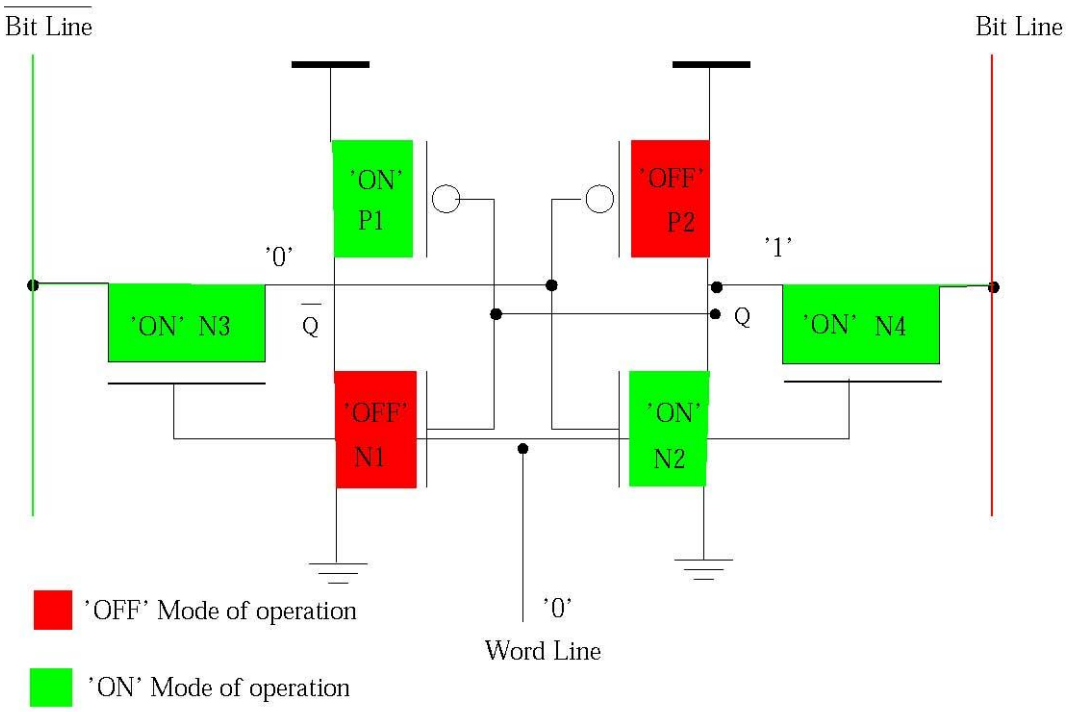



(a) Write "0" operation

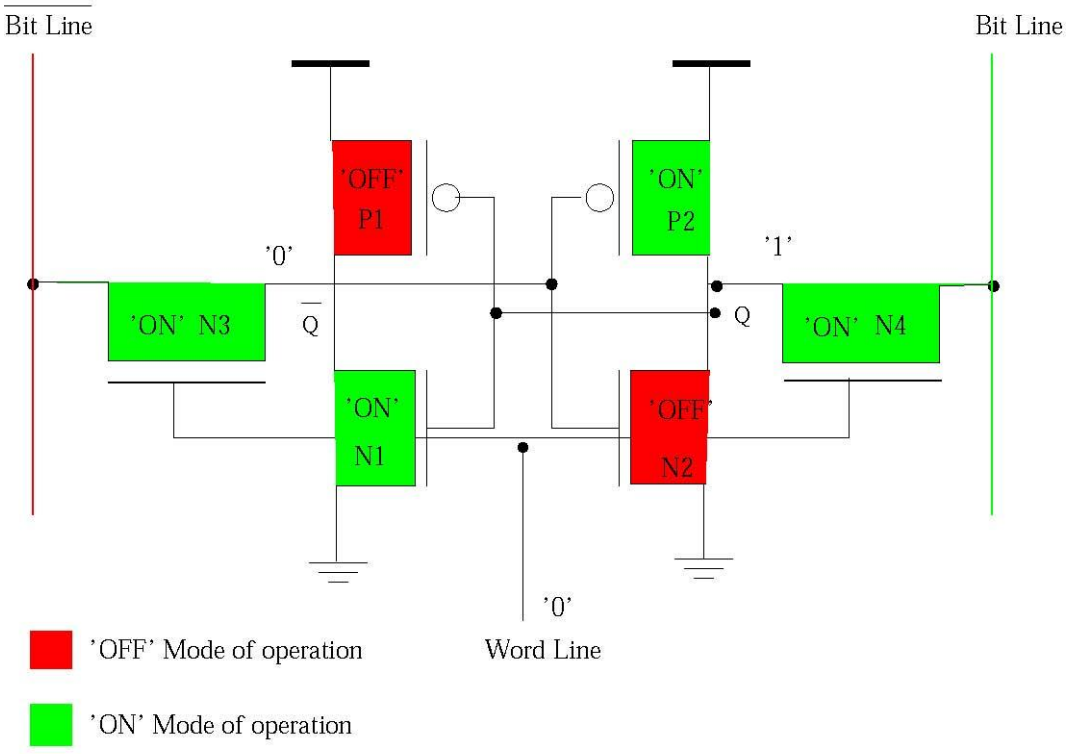



(b) Write "I" operation

Figure 3.2. Write Mode of operation for the 6T-SRAM cell.




(a) Read "0" operation




(b) Read "I" operation

Figure 3.3. Read Mode of operation for the 6T-SRAM cell.

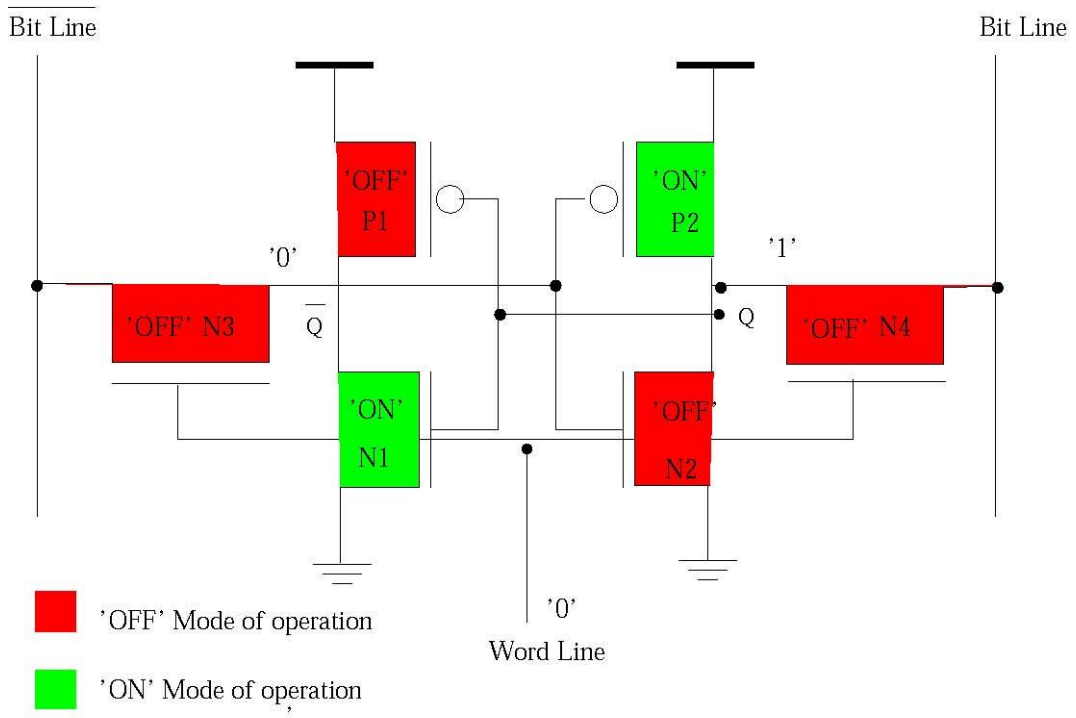



Figure 3.4. Hold mode of operation for the 6T-SRAM cell.

When the transistor is in the ON state, it conducts dynamic, as well as gate-oxide leakage current. The details are discussed in Chapter 4. For power consumption, the current flow in the SRAM cell depends on the location of the individual transistors in the circuit and the operation (write, read or hold) being performed. The modes of operation for all three states, Write, Read, and Idle, are shown in Figures 3.2, 3.3, and 3.4. The corresponding current flow paths are analyzed and illustrated with the help of schematic diagrams. The solid arrows indicate dynamic current flow. The dashed arrows represent gate-oxide leakage current and the dotted arrows indicate subthreshold leakage current. When the transistor (PMOS or NMOS) is in the ON state, it will conduct dynamic and gate-oxide leakage current. On the other hand, when the transistor (PMOS or NMOS) is in the OFF state, it will conduct subthreshold leakage current as well as gate-oxide leakage current. It is important to study each mode of operation separately as discussed below.

- Current path for Write "0," Figure 3.5

In case of writing a "0" bit to the cell, $\overline{BL}$ will be pulled high and BL will go low. As

transistor N3 gets input “1” from $\overline{BL}$ it turns ON. Similarly P1 gets input “1” and the output of the second inverter will be bit “0” which will be written on the cell. P1, N3, N2, and N4 are ON, carrying dynamic and gate-oxide leakage currents. N1 and P2 are OFF so only subthreshold current and gate-oxide leakage current will flow through them. Note that despite the transistor being ON or OFF, gate-oxide current flows through all transistors.

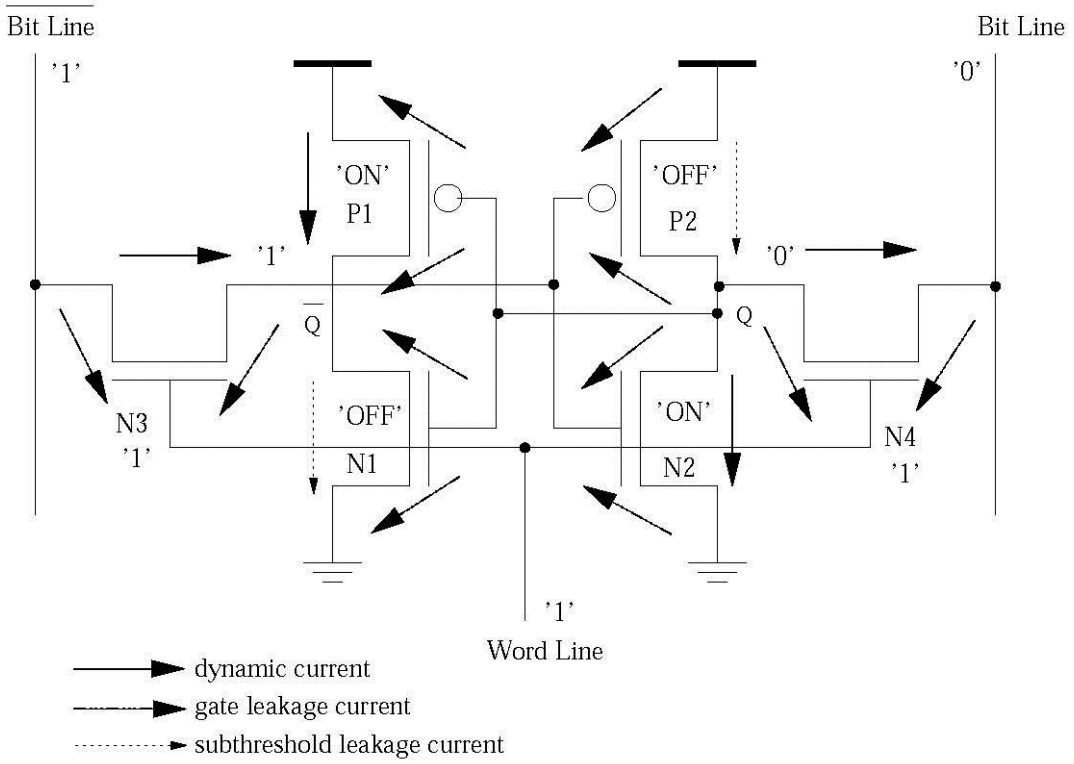



Figure 3.5. Write “0” Current paths for the 6T-SRAM cell.

- Current path for Write “1,” Figure 3.6

In the case of writing “1”, all the transistors will behave opposite since BL will now be pulled high with “1” written on it. Thus, transistors N3, N4, N1 and P2 will be ON and will conduct dynamic current and gate-oxide leakage current while transistors P1 and N2 will have subthreshold current and gate-oxide leakage current as they are OFF. Note that whatever the content of the bit is, “1” or “0,” access transistors N3 and N4 will be ON because the SRAM cell is being accessed or written and also the wordline will be pulled high whether it is a write or read operation for both “1” and “0.”

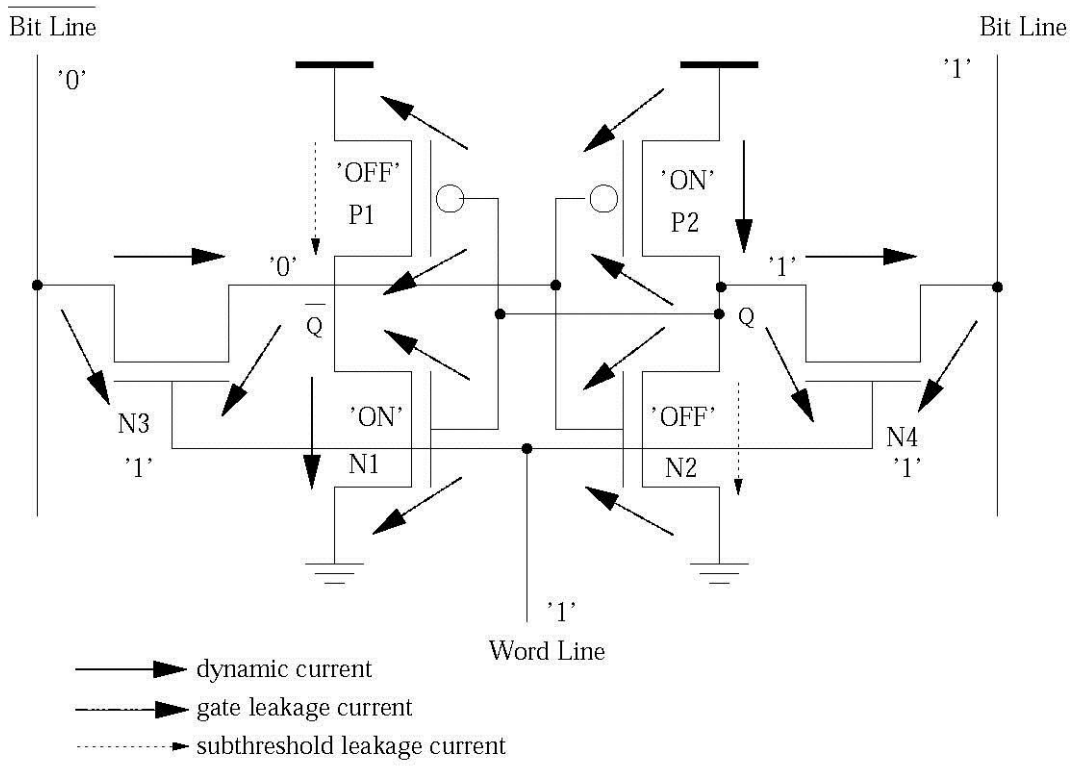



Figure 3.6. Write "1" Current paths for the 6T-SRAM cell.

- Current path for read "1," Figure 3.7

The Read "1" operation is the same as Write "1" operation because once the bit is written on the SRAM cell this operation requires accessing the information so at this point, the access time will be important.

- Current path for Read "0," Figure 3.8

The Read "0" operation is the same as Write "0" operation.

- Current path for Hold "0" or "1," Figure 3.9

During the Hold operation there is no active transistor because the SRAM cell is now only holding the bit information or is in the idle state. Hence all the transistors will have subthreshold and gate-oxide leakage current flowing through.

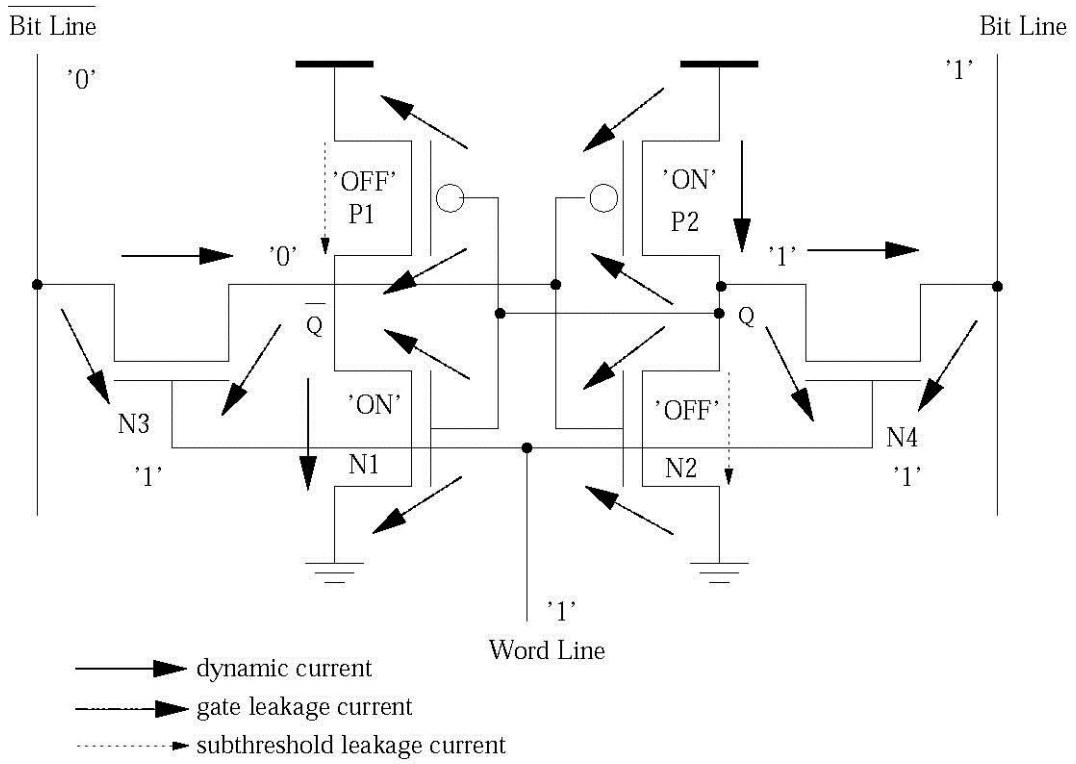



Figure 3.7. Read "1" Current paths for the 6T-SRAM cell.

## 3.2. Seven Transistor (7T) Nano-CMOS SRAM Design

### 3.2.1. Logical Design or the 7-Transistor SRAM

Embedded systems are memory dominated. Therefore, low energy consumption is a crucial requirement of such systems. Amongst the current techniques of low power, reducing the supply voltage is the most popular and well known technique. It is extremely challenging to build ultra-low-power designs for high density cache, where the operating voltage is lower than the threshold voltage. Therefore, for these reasons, the standard 6T-SRAM fails to cope with the density requirement and yield of modern designs circuits [75]. The advantages of 7-transistor SRAM (7T-SRAM) over 6T-SRAM are as follows:

- The 7T-SRAM circuit functions in ultra-low voltage regions and allows subthreshold operation, unlike 6T-SRAM.
- It has better read and write stability.
- Compared to 6T-SRAM, it has better process variation tolerance.

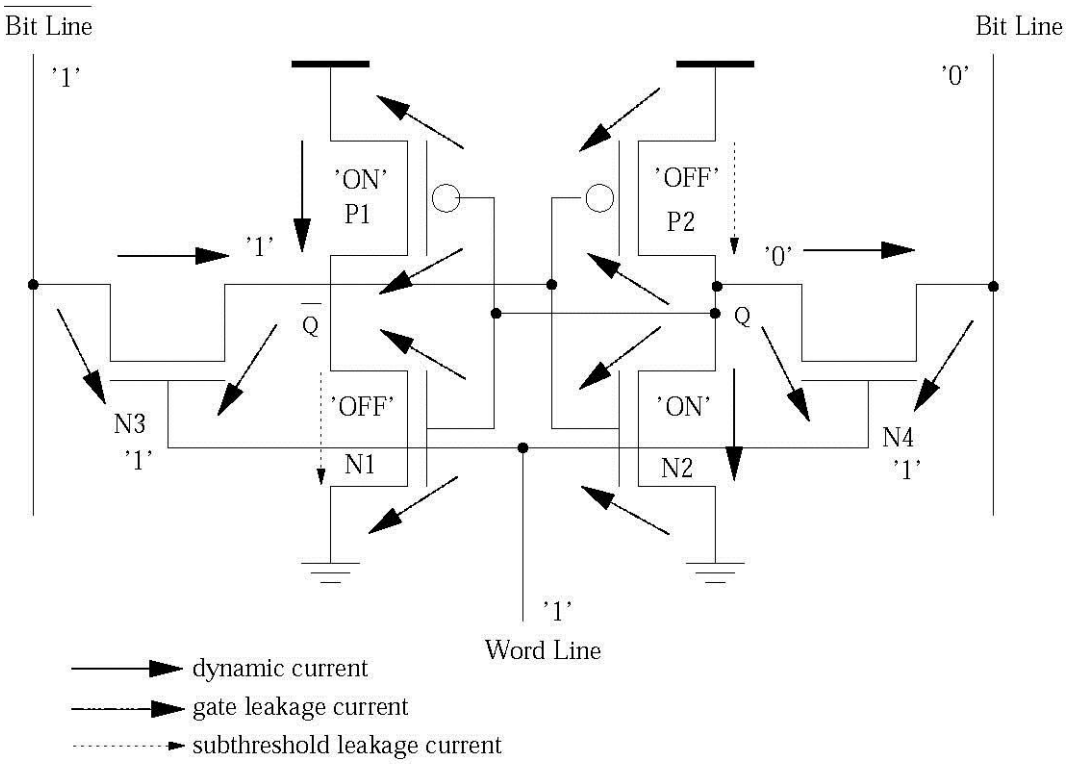



Figure 3.8. Read "0" Current paths for the 6T-SRAM cell.

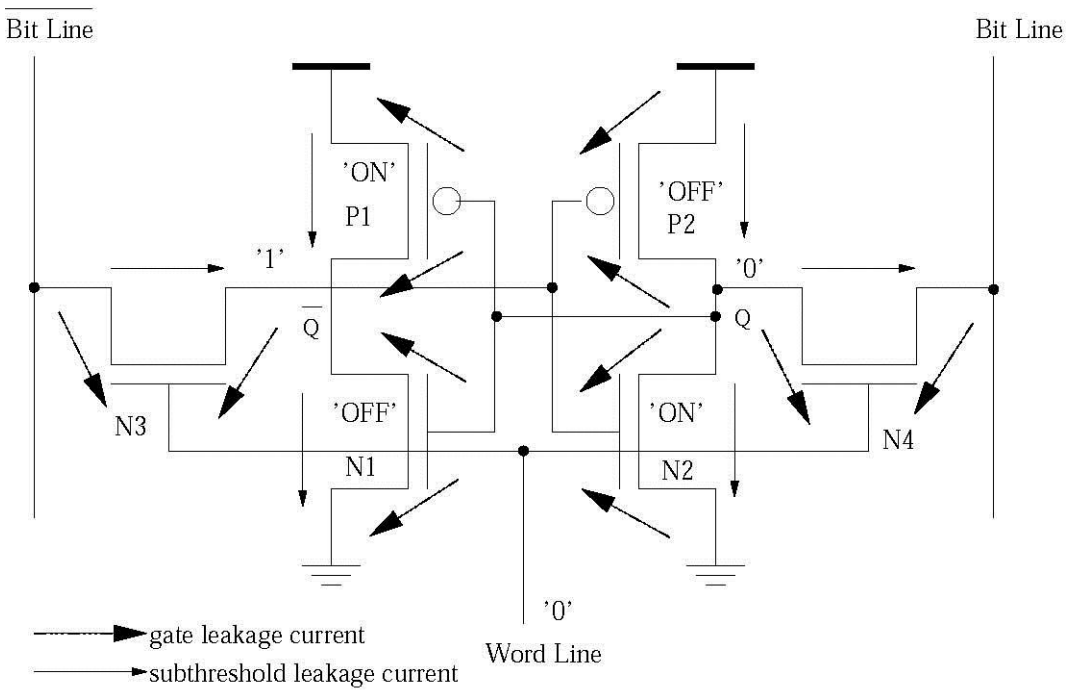



Figure 3.9. Hold "0" or "1" Current paths for the 6T-SRAM cell.

This section describes the logical design of 7T-SRAM in the nano-CMOS regime. Figure 3.10 shows the schematic diagram of a 7T-SRAM circuit design. This topology has been shown to be fit for the ultra-low voltage regime [75].

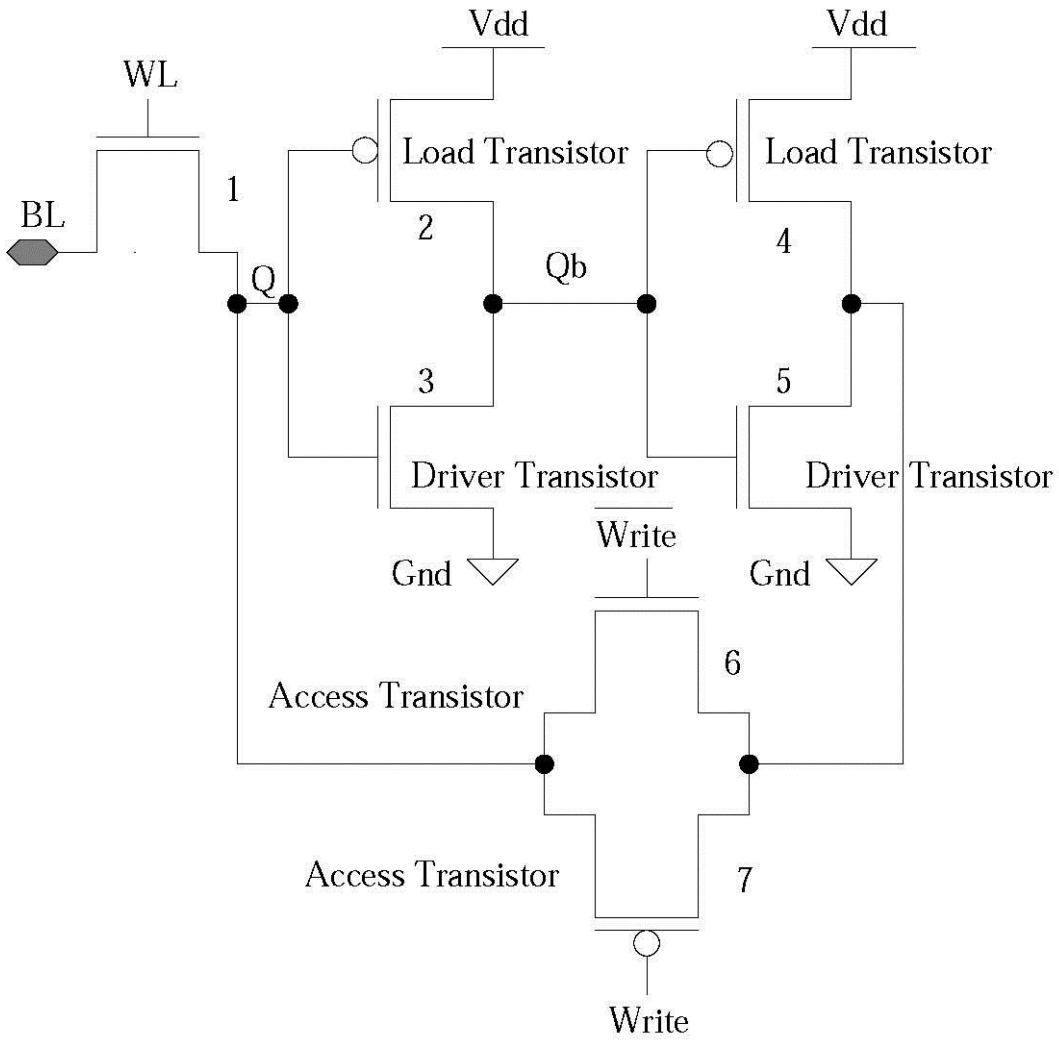



Figure 3.10. Seven-Transistor (7T) CMOS SRAM cell.

As noted, the 7T-SRAM cell operates on a single bit line instead of the traditional two bit lines as in the case of 6T-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 operations, which is similar to the standard 6-transistor SRAM cell. In case of Hold operations, the word line is kept low and, with the help of transmission gates, a

strong feedback is given to the cross-coupled inverters. The 7T-SRAM has been exhaustively characterized with different technology nodes and different sizing of transistors, such as minimum-sized transistors, as well as standard ratio of transistors has been considered. Figure 3.11 shows the functional simulation diagram of 7T-SRAM during various modes of operation, i.e., Write, Read, and Hold.

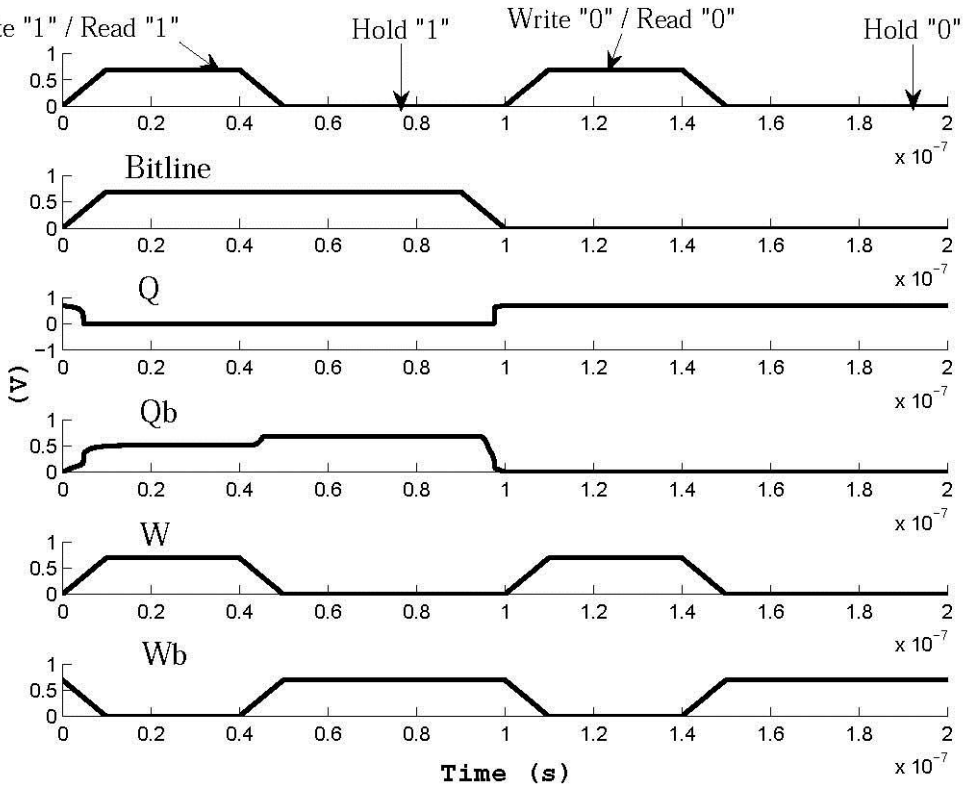



Figure 3.11. Functional simulation diagram for 7T-SRAM cell.

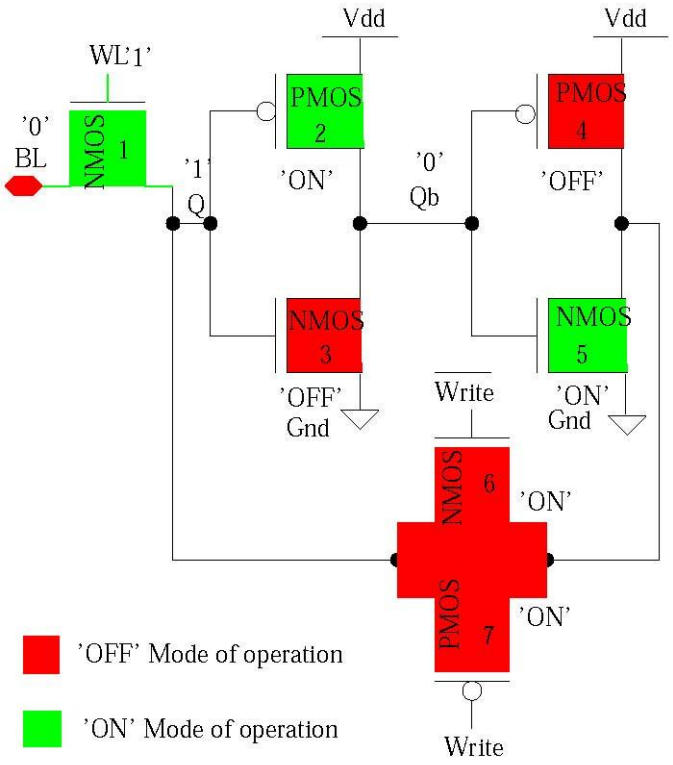
Each operating mode (Read-Write-Hold) of the 7T-SRAM is described in detail below.

3.2.2. Modes or Operations of 7-Transistor SRAM

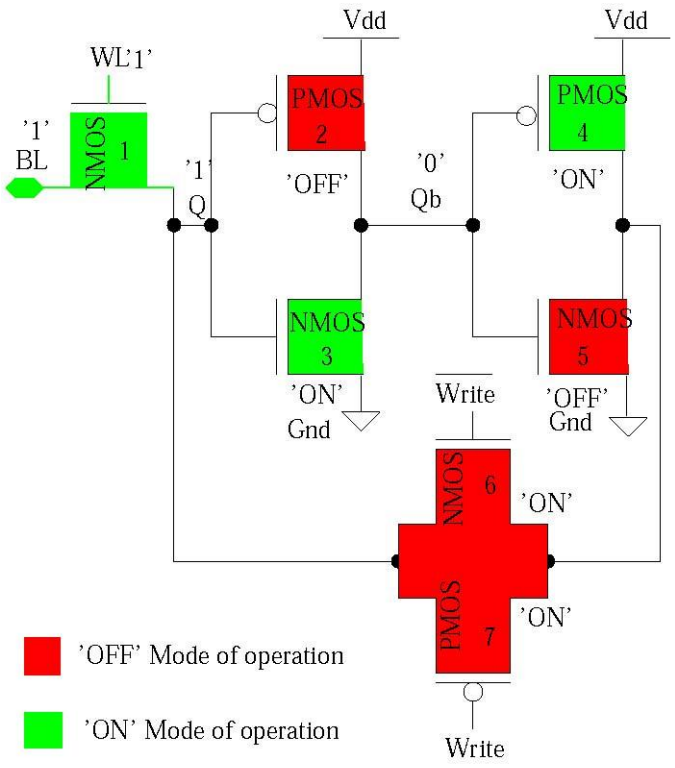
- Write operation (Refer to Figure 3.12):

In case of Write operation of 7T-SRAM, one of the bit lines, BL, is driven from precharged value ($V_{DD}$) to the ground potential by a write driver. The Write operation can be studied from Figures 3.15 and 3.16. If transistor 6, transistor 7, and the second inverter are properly sized, then the cell is flipped and its data is effectively overwritten. A statistical measure of SRAM cell writeability is defined as the write margin. It is defined as the minimum bit line voltage required to flip the state of an SRAM cell [66]. The write margin

value and variation is a function of the cell design, SRAM array size, and process variation. A cell is considered not writeable if the worst-case write margin becomes lower than the ground potential. Note that the Write operation is applied to the node storing a "1." The function of the pull-up transistors is only to maintain the high level on the "1" storage node and prevent its discharge by the of-state leakage current of the driver transistor during data retention and to provide the low-to-high transition during overwriting [66].




(a) Write "0" operation

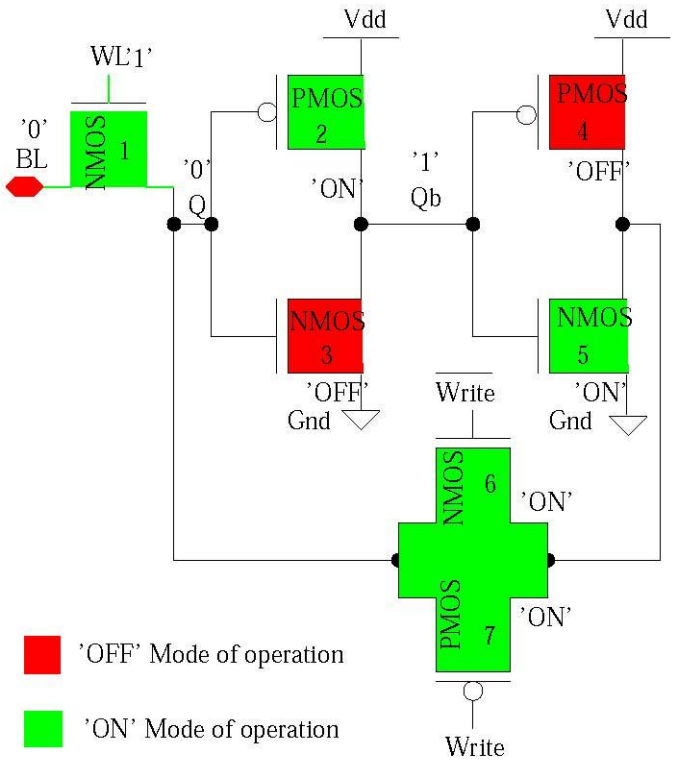



(b) Write "1" operation

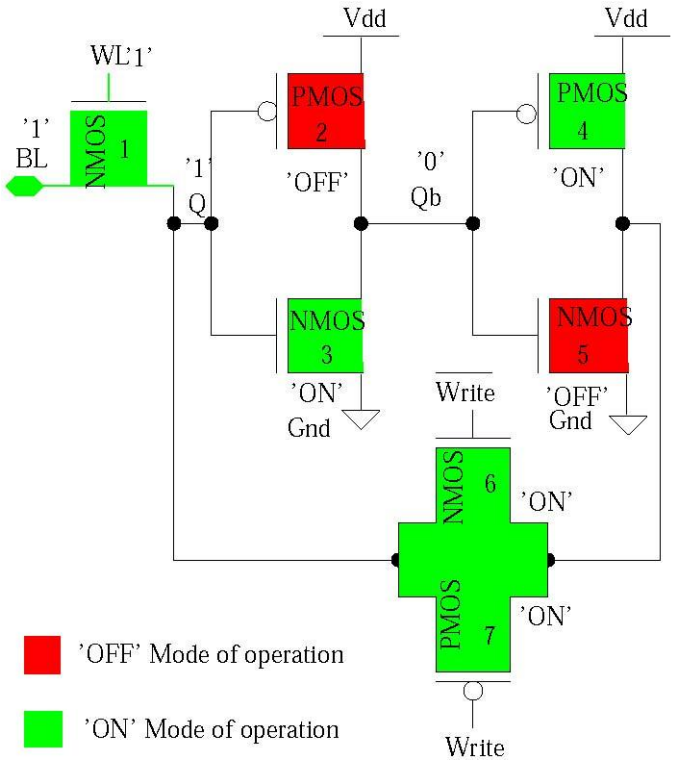
Figure 3.12. Write Mode of operation for the 7T-SRAM cell.

- Read operation (Refer to Figure 3.13):

Prior to start a Read operation, the bit lines are precharged to $V_{DD}$. The Read operation is initiated by enabling the word line and connecting the precharged bit line, BL, to the internal nodes of the cell. Upon Read access the bit line voltage remains at the precharge level. Effectively, transistor 1 and transistor 3 form a voltage divider whose output is now no longer at zero volts and is connected to the input of second inverter. Sizing of transistor 6 and transistor 7 should ensure that the second inverter does not switch causing a destructive read. In other words, $0 + \Delta V$ should be less than the switching threshold of the second inverter plus some noise margin.




(a) Read "0" operation




(b) Read "1" operation

Figure 3.13. Read modes of operation for the 7T-SRAM cell.

- Hold operation (Refer to Figure 3.14):

As discussed in the Read and Write operations of the 7T-SRAM, the word line is low and strong feedback loop is provided to the cross coupled inverters with the help of transistor 6 and transistor 7 of the transmission gate.

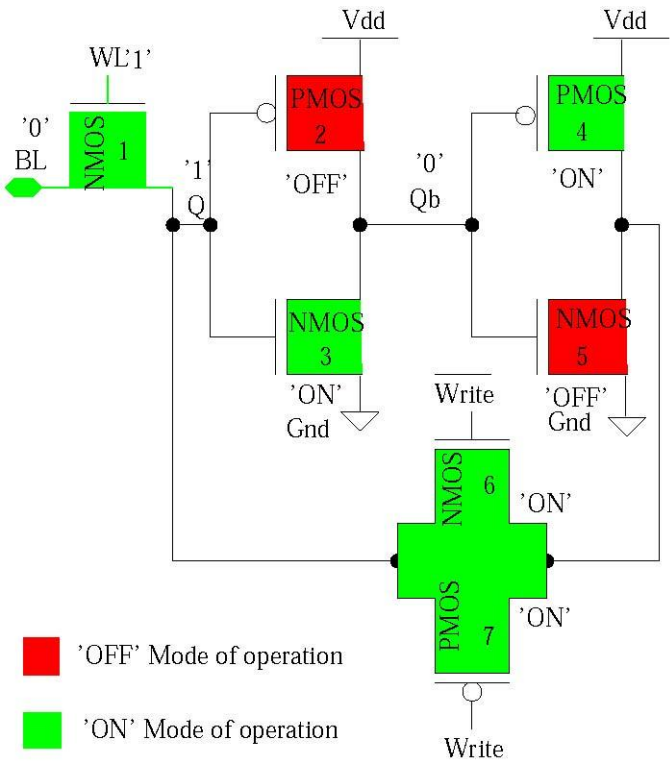



Figure 3.14. Hold modes of operation for the 7T-SRAM cell.

3.2.3. Current Flow Paths 7-Transistor SRAM

The current flow in each transistor of the 7T-SRAM cell depends on its location and the operation (Read, Write, or Hold) being performed. The current paths for Read and Write operation are analyzed and identified in the following subsection. The solid arrows shown in the figure indicate dynamic current. The dashed arrow represents gate-oxide leakage current; the subthreshold leakage current, shown by dotted arrows, 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 along with the gate-oxide leakage current and when the transistor is OFF state it will have gate-oxide leakage current as well as subthreshold leakage current.

- Current path for Write "0" (refer to Figure 3.15):

In this case the bit line will be “0” and the 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 OFF. Transistors 6 and 7 will be OFF during the write operation, and hence will have subthreshold leakage current and gate-oxide leakage current.

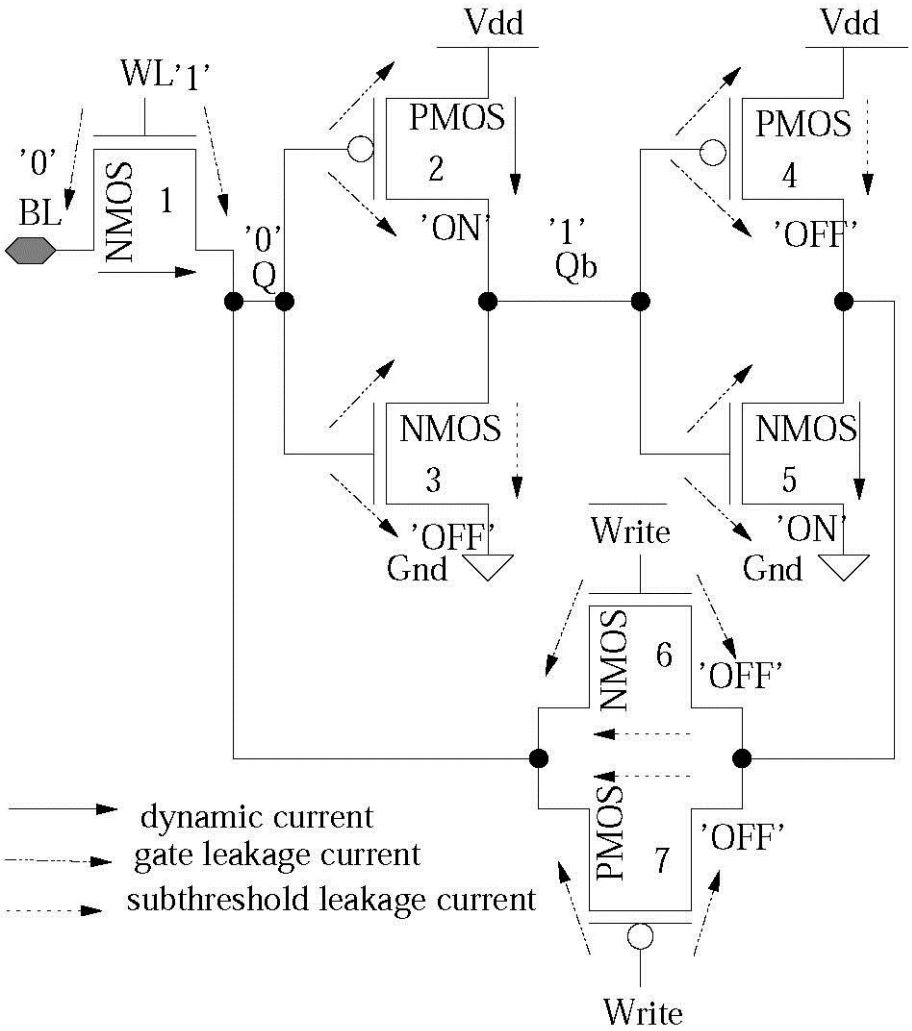



Figure 3.15. Current paths for write “0” of the 7T-SRAM cell.

- Current path for Write “1” (Refer to Figure 3.16):

In this case the bit line will be “1” and the WL is precharged to level “1.” In order to write “1” on the SRAM cell, Q will be “1.” Transistors 2 and 5 are OFF so they will have subthreshold current and gate-oxide leakage current. Transistors 3 and 4 will have dynamic

leakage current and gate-oxide leakage current as they are ON. Transistors 6 and 7 will be OFF during the Write operation, and hence will have subthreshold leakage current and gate-oxide leakage current.

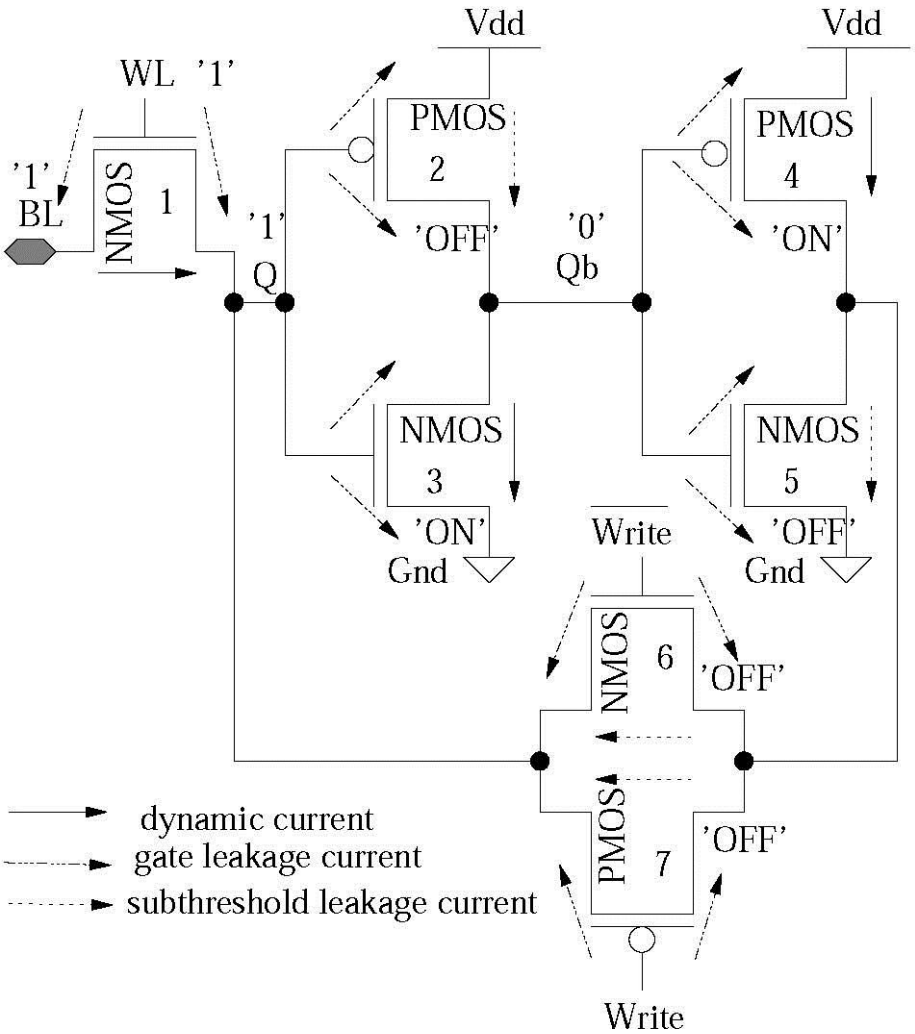



Figure 3.16. Current paths for write "1" of the 7T-SRAM cell.

- Current path for Read "1" (Refer to Figure 3.17):

In this case, WL and BL will be at high level in order to read a value. The Q node will have "1" and transistors 2 and 5 will be OFF, carrying gate-oxide leakage current and subthreshold leakage current. Transistors 3 and 4 will have dynamic current along with gate-oxide leakage current, as they are ON. Qb will be "0". In the read operation, transistors 6 and 7 of the transmission gate will be ON, hence carrying dynamic current and gate-oxide leakage current.

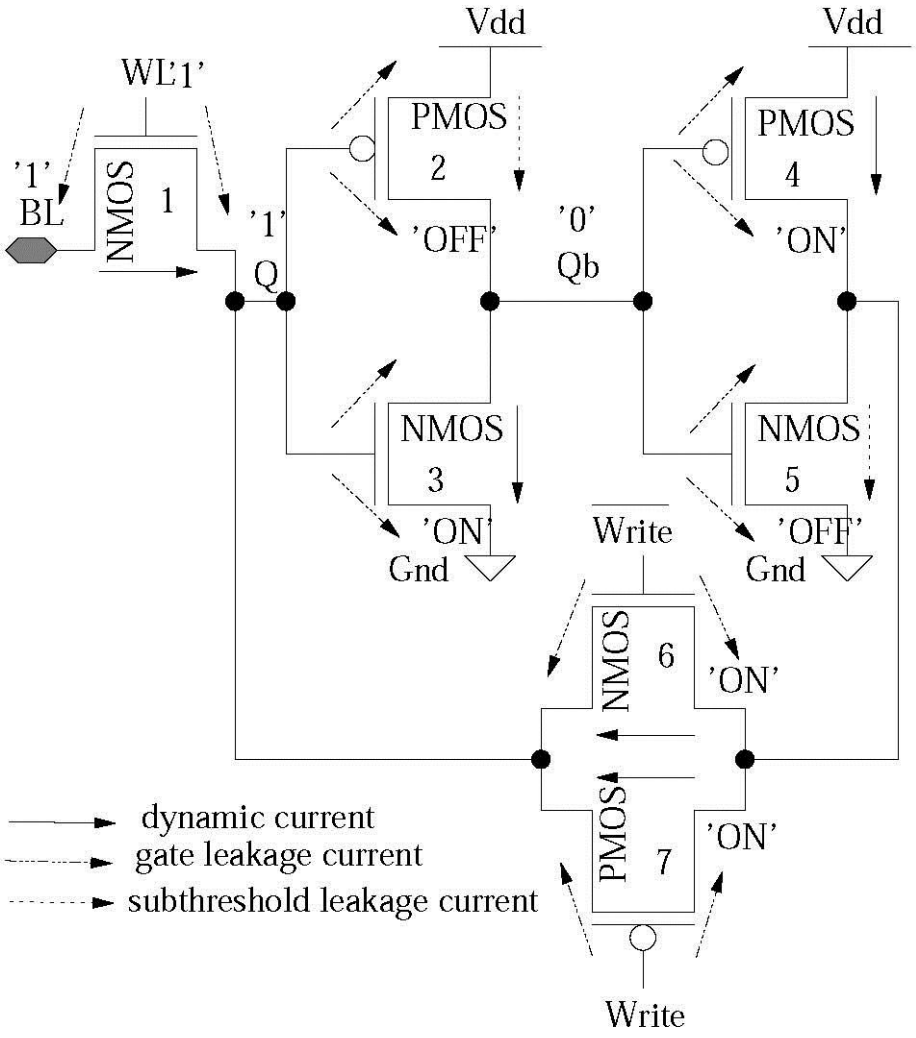



Figure 3.17. Current paths for Read "1" of the 7T-SRAM cell.

- Current path for Read "0" (Refer to Figure 3.18):

In this case, WL and BL will be at high level in order to read a value. The Q node will have "0" and transistors 2 and 5 will be ON, carrying gate-oxide leakage current and dynamic current. Transistors 3 and 4 will have subthreshold leakage along with gate-oxide leakage current, as they are OFF. The BL will be "0." In the read operation, transistors 6 and 7 of the transmission gate will be ON, hence, carrying dynamic current and gate-oxide leakage current.

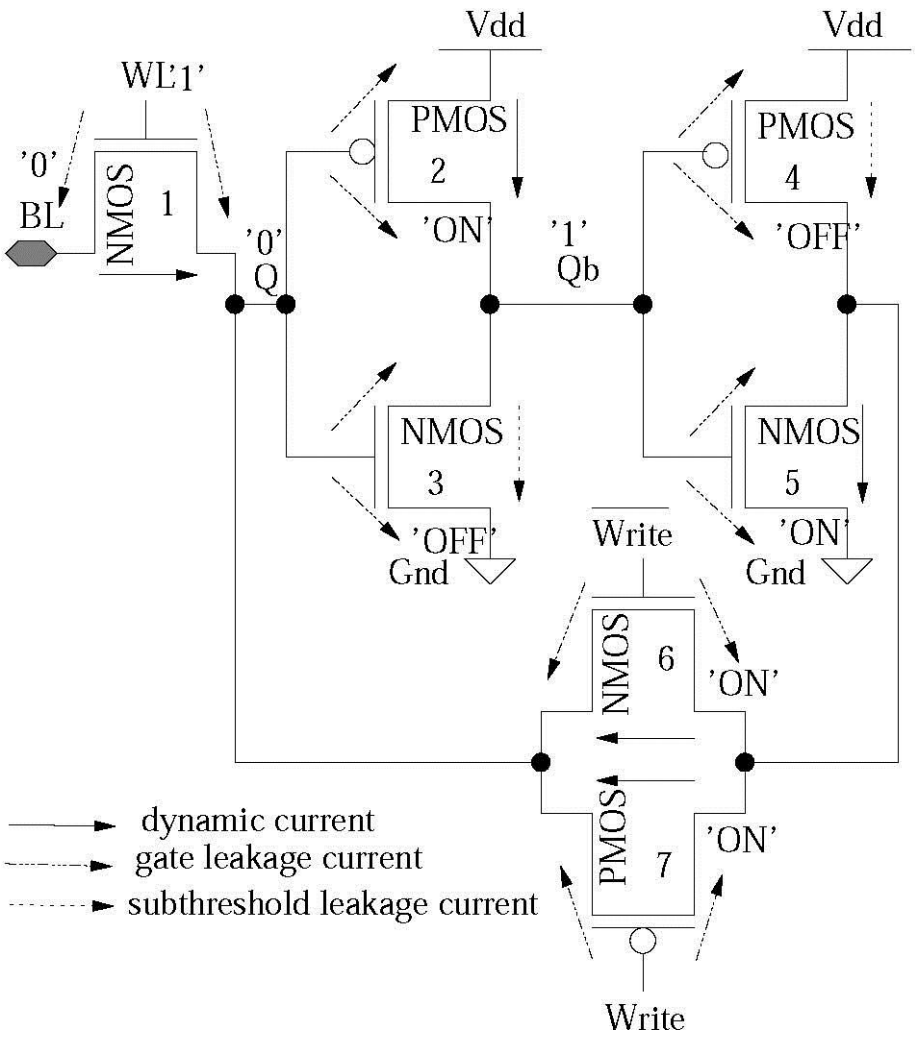



Figure 3.18. Current paths for read "0" of the 7T-SRAM cell.

- Current path for Hold "0" or "1" (Refer to Figure 3.19):

In the case of Hold operation there is no active transistor because the SRAM cell is now only holding the bit information or is in idle state. All transistors will have subthreshold and gate-oxide leakage current flowing through.

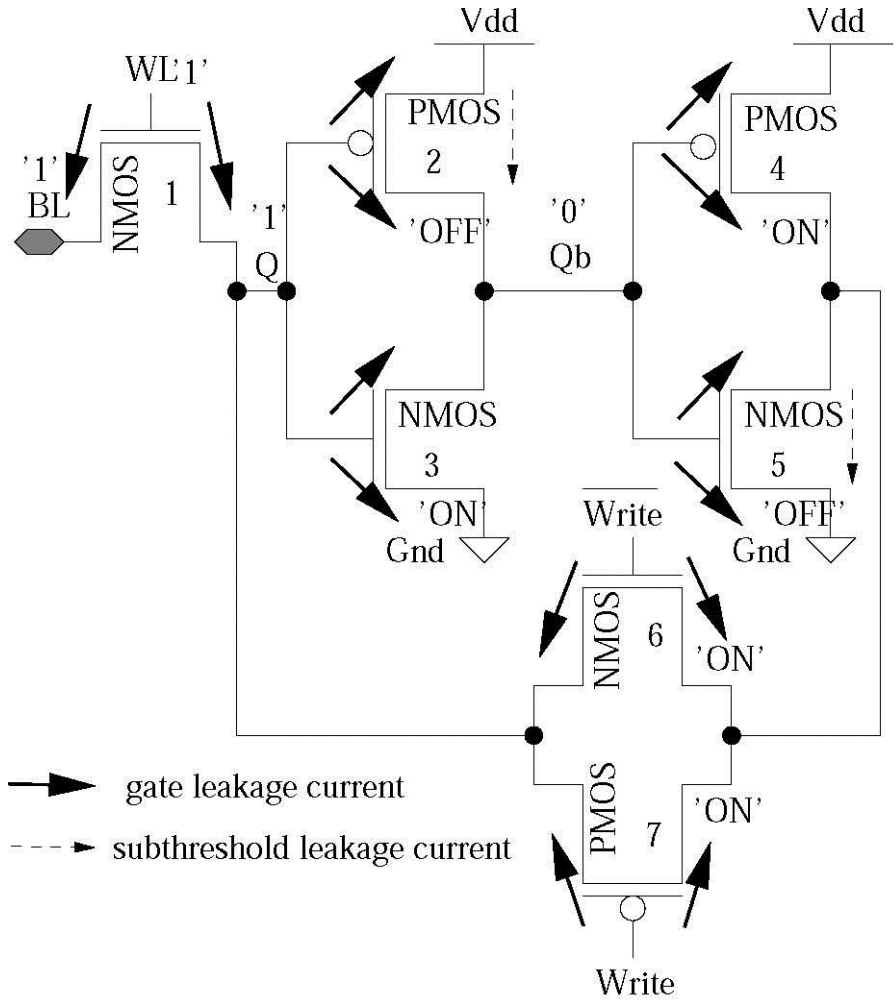



Figure 3.19. Current paths for Hold “0” or “1” of the 7T-SRAM cell.

3.2.4. 7-Transistor Array Organization

Array of a 7-transistor SRAM cells is constructed by stitching the individual cells in a matrix fashion. Figure 3.20 shows the organization of a 1 x 8 7T-SRAM array design.

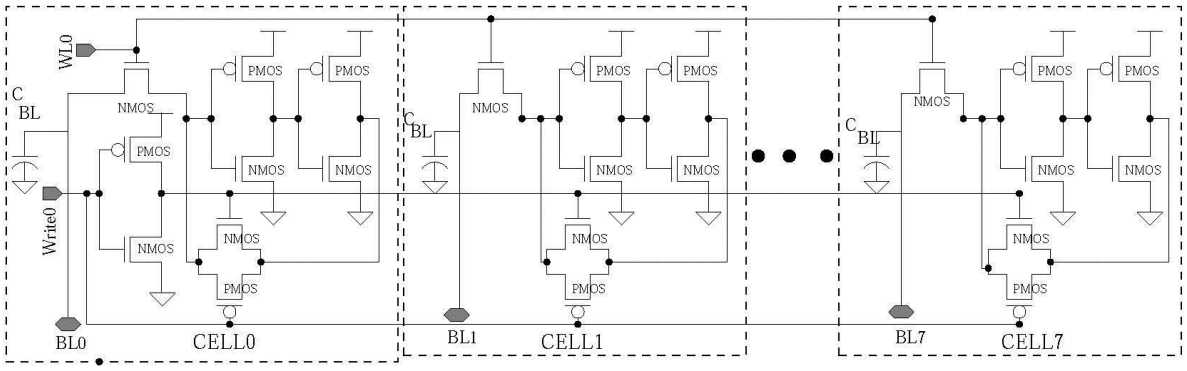



Figure 3.20. 7T-SRAM array organization.

3.3. High-κ/Metal-Gate 10-Transistor (10T) SRAM Design

3.3.1. Logical Design of the 10-Transistor SRAM

The baseline 10T-SRAM cell consists of 10 transistors as shown in Figure 3.21. This topology has been shown to be process variation tolerant in [74]. The SRAM cell is composed of two inverters connected back to back in a closed loop fashion in order to store the 1-bit information, three transmission gates for the Read, Write, and Hold states, instead of the access transistors used in traditional 6T SRAM design. Transmission gates carefully input and output the data to/ from the cell node Q at full logic level. This provides full swing during Write and Read operation. This feature of the 10T-SRAM design eliminates the use of sense amplifier and pre-charging circuitry for pre-charging of the BL and $\overline{BL}$ lines prior to Read and Write operations.

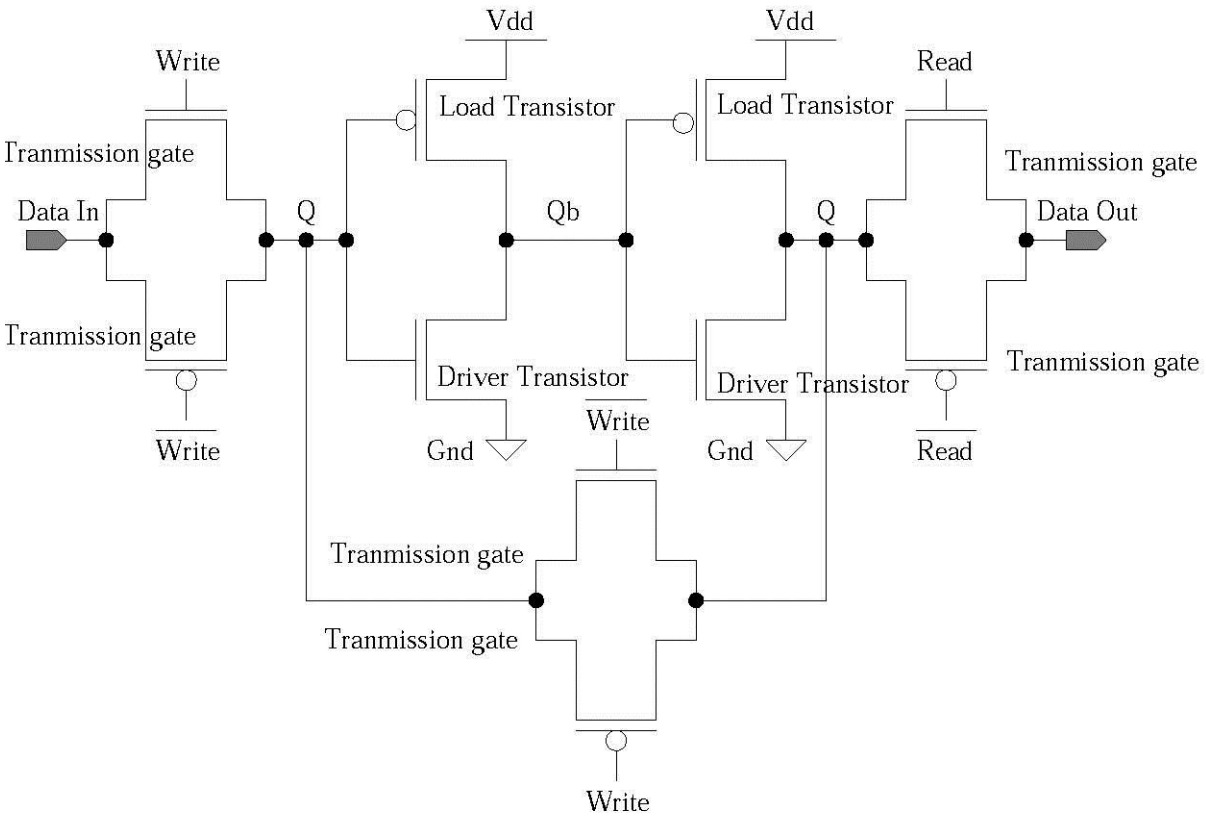



Figure 3.21. 10-Transistor (10T) CMOS SRAM cell.

3.3.2. High-κ/Metal Gate CMOS Compact Model

• Read operation (Refer to Figure 3.23):

The design presented in this research uses a 32 nm high-κ/metal-gate CMOS predictive technology model (PTM) [87]. The simulation results obtained are highly accurate and the calculated data are of comparable accuracy to technology computer-aided design (TCAD) simulations which are typically time and computation intensive. The PTM is based on *BSIM*4/5, hence two methods are used as follows:

(i) The model parameter in the model file that denotes relative permittivity (EPSROX) is changed

(ii) The equivalent oxide thickness (EOT) for the dielectric under consideration is calculated

Using these steps, the EOT is calculated so as to keep the ratio of relative permittivity over dielectric thickness constant. The EOT ($T_{ox}^{*}$) is calculated by using the following expression:

(2) $$T_{ox}^{*} = \frac{\kappa_{SiO_2}}{\kappa_{gate}} \times T_{gate}$$

where $\kappa_{gate}$ is the relative permittivity and $T_{gate}$ is the thickness of the gate dielectric material other than $SiO_2$, while $\kappa_{SiO_2}$ is the dielectric constant of $SiO_2$ (= 3.9). We have taken $\kappa_{gate}$ = 21 to emulate a $HfO_2$ based dielectric. The EOT is calculated to be 0.9 nm.

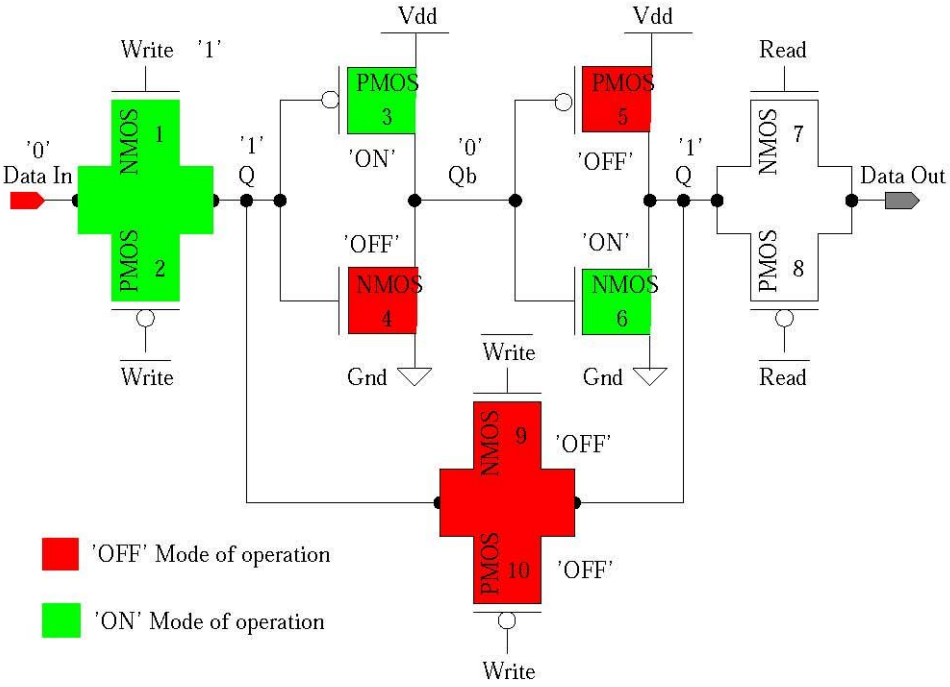
3.3.3. Modes or Operation of 10-Transistor SRAM

The write and read modes of operations of the 10-transistor SRAM is now discussed. The modes are analyzed for “1” and “0” data as the states of the transistors are different.

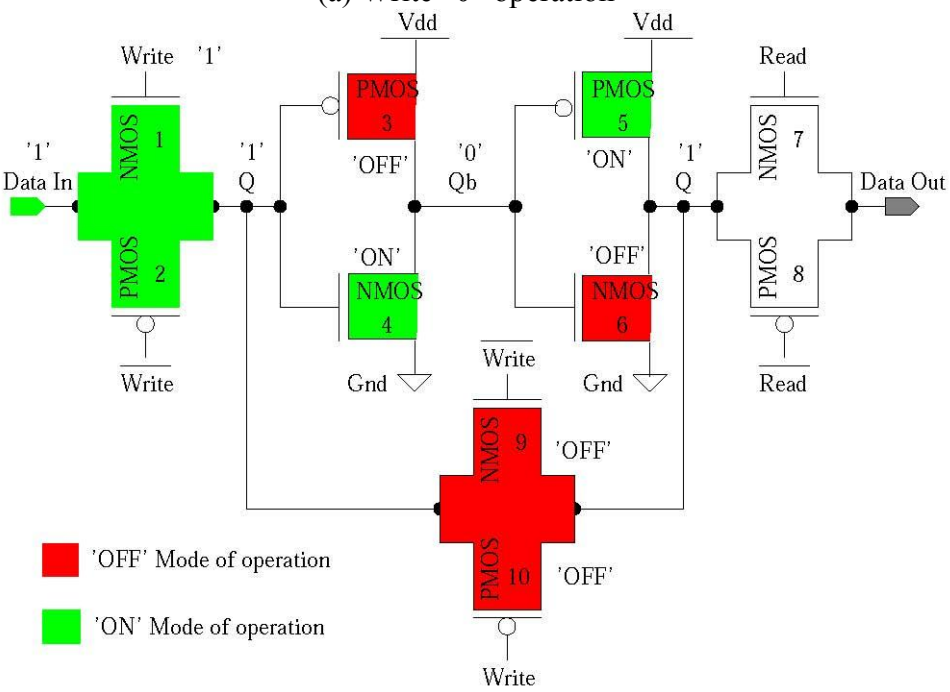
• Write operation (Refer to Figure 3.22):

The Write (0,1) operation of the 10T SRAM cell is now discussed. In the case of Write “1” operation, the write signal goes high and the transmission gate connects the data in node to the node Q of the cell. When the write node goes high the transmission gate forces node Q to the same level of the data line. The transistors which are in ON state will have

dynamic current whereas the transistors which are OFF conduct subthreshold current [61]. Transistors 3 and 6 are OFF carrying subthreshold current whereas transistors 4 and 5 carry dynamic current. Similarly in the Write 0 operation, dynamic current flows in transistors 3 and 6 whereas transistors 4 and 5 have subthreshold current. None of the transistors exhibit any appreciable gate leakage current because of the presence of the high-κ dielectric material.



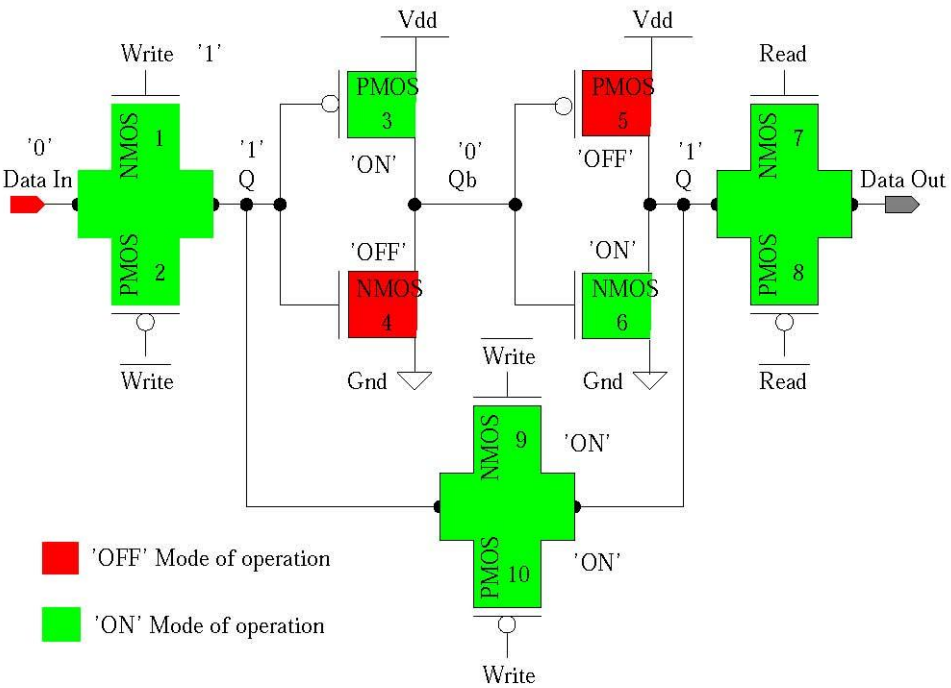

(a) Write "0" operation



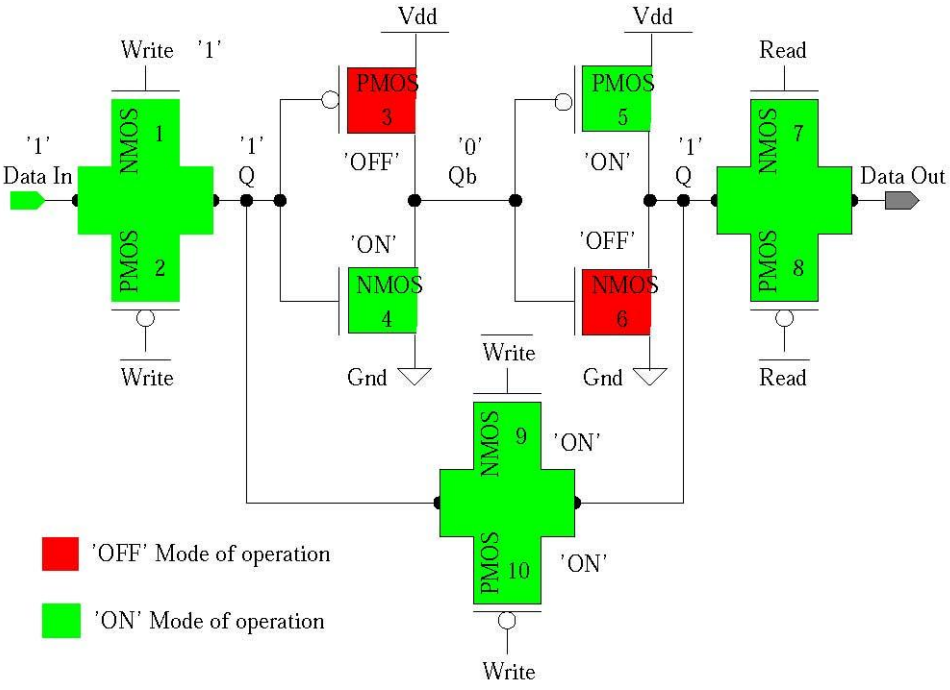

(b) Write "1" operation

Figure 3.22. Write modes of operation for the 10T-SRAM cell.

The 10T-SRAM cell initiates the read operation with the READ and READ nodes. In the Read "1" operation, the READ node will connect the NMOS transistor of the transmission gate that provides path for the Q and further to data out node. The READ node goes to high level and so does node Q. In the case of the read operation, transistors 9 and 10 are ON, thus carrying dynamic current. The transmission gate at the read side is also ON hence will carry dynamic current flow. Transistors 4 and 5 will have dynamic current being in the OFF state and transistors 3 and 6 will be having subthreshold current flow as they are in the ON state. During the Read "0" operation transistors 3 and 6 will be carrying dynamic current flow while transistor 4 and 5 will have subthreshold current. Subthreshold current will also flow at the transmission gate at the Write node and the transmission gate of the Read side will have dynamic current.




(a) Read "0" operation




(b) Read "1" operation

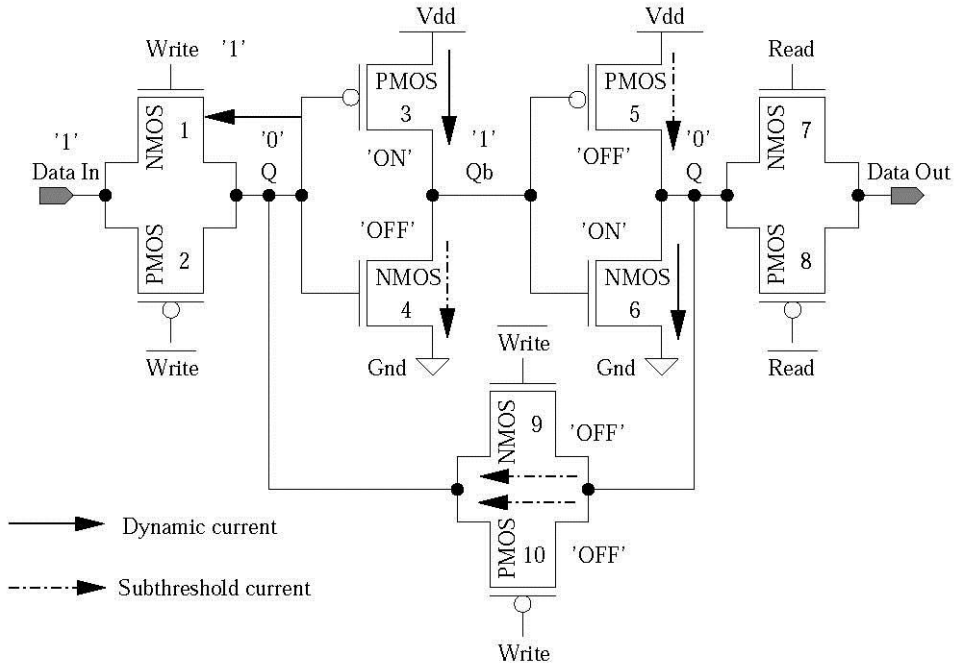
Figure 3.23. Read modes of operation for the 10T-SRAM cell.**Error! Bookmark not defined.**

3.3.4. 10-Transistor Current Flow Paths

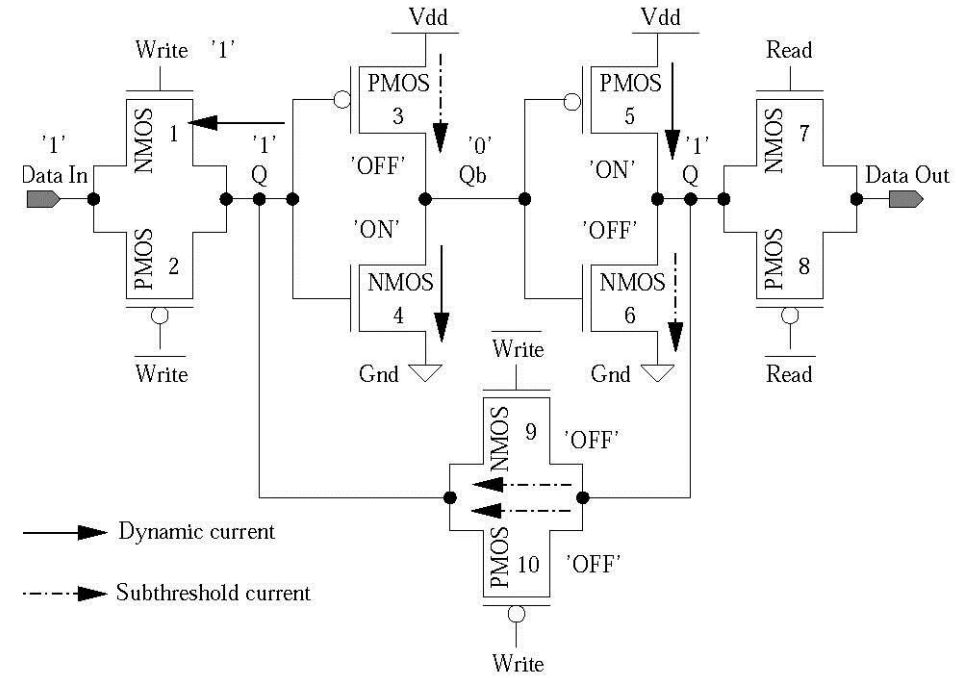
The current paths for the 10T-SRAM cell are identified for different modes of its operation and discussed below. This identification helps in accurately analyzing power estimation of the SRAM for its operations.

Current path for Write "0" (Refer to Figure 3.24[a]):

During the Write "0" operation the Write signal goes high, and the transmission gate connects the data in node to the node Q. When the Write node goes high the transmission gate forces node Q to the same level as the data line. Transistors 4, 5, and 7, will be OFF carrying subthreshold leakage current whereas transistors 1, 3, and 6 will be ON so will have dynamic current only. Transistors 9 and 10 will be OFF and will carry subthreshold leakage current.




(a) Write "0" operation




(b) Write "I" operation

Figure 3.24. Write operation current flow paths for the 10T-SRAM cell.

- Current path for write “1” (Refer Figure 3.24[b]):

During the write “1” operation the write signal goes high, and the transmission gate connects the data in node to the node Q. When the write node goes high the transmission gate forces node Q to the same level as the data line. Transistors 4, 5, and 7, will be ON carrying

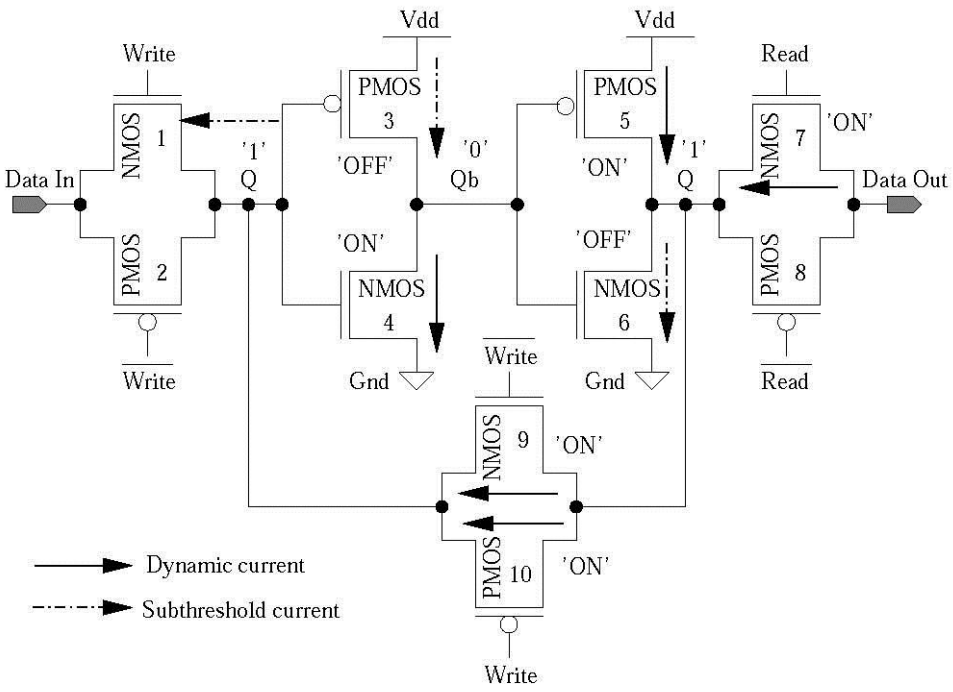
dynamic current whereas transistors 1, 3, and 6 will be OFF hence will have subthreshold leakage current only. Transistors 9 and 10 will be OFF and will carry subthreshold leakage current and gate-oxide current.

- Current path for read “1” (Refer Figure 3.25[a]):

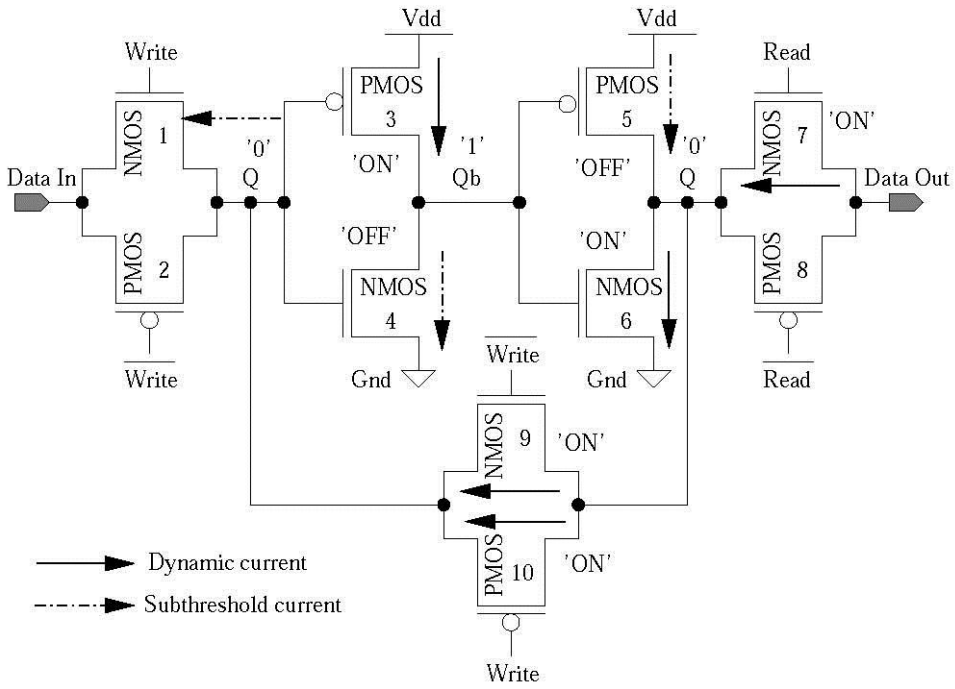
  Reading “1” operation is the same as writing “1” in a 10T-SRAM.

- Current path for read “0” (Refer Figure 3.25[b]):

  Reading “0” operation is the same as writing “0” in a 10T-SRAM.




(a) Current path for write "0"




(b) Current for path write "I"

Figure 3.25. Read operation current flow paths for the 10T-SRAM cell.

3.3.5. Array Organization or the 10-Transistor SRAM

Figure 3.26 shows the array organization of a 1x8 10T-SRAM Design. The array is functionally simulated for it operations.

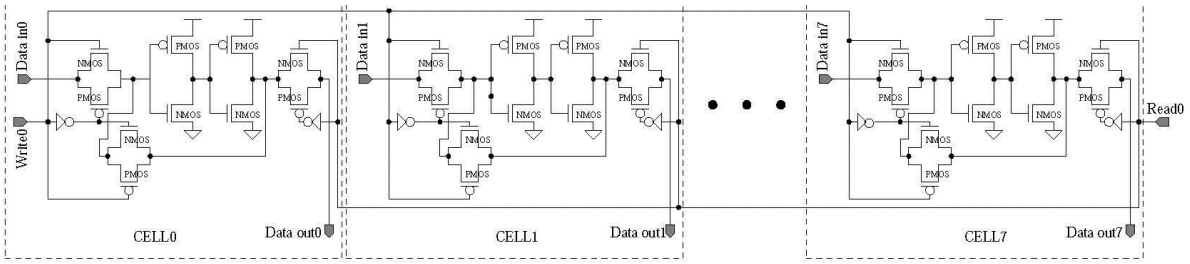



Figure 3.26. 10T-SRAM array organization.

# CHAPTER 4

## POWER, LEAKAGE AND STATIC NOISE MARGIN ANALYSIS

Power consumption of a chip and leakage currents are some of the most crucial design constraints faced by designers in order to achieve high performance process andsystems-on-chip (SoC) design as compared to already existing low-power targeted devices. The demand is ever increasing for such low-power consuming circuits along with the requirement for personal computing devices. The demands as a result give rise to battery life concerns, heat consumption, fabrication costs, packaging costs and reliability. The analysis of growing technology trends indicate that the dominant components of total power dissipation in complementary metal-oxide semiconductor (CMOS) circuit are dynamic current, gate-oxide and subthreshold leakage current, therefore one component cannot be compromised with respect to others. The reason is that low threshold voltage and gate-oxide thickness scale together in nanoscale regime, along with the reduction in supply voltage.

Static random access memory (SRAMs) are important building blocks for SoCs and high performance processors and also consume a large amount of the total power [51]. SRAM structures are used for cache memory and comprise a large number of the on-chip transistors in bulk-CMOS as well as SoCs. Thus, it is essential to study the power consumption profile of SRAMs in order to estimate and minimize their power consumption, specifically when they are implemented in nanoscale CMOS transistors. Design engineers are facing a huge challenge in order to meet the increasing demand of low power nano-CMOS circuits with simultaneously enhanced performance. They are implementing high order scaling in process technology and in the design parameters of the CMOS transistors. Therefore, there is a major need for analysis, explanation, and characterization of the various sources of power dissipation mechanisms in SRAMs.

An SRAM consumes dynamic (which is originated by capacitive switching) power only when the bitline or wordline are switching their level from low-to-high or high-to-low

for Write, Read or Hold operations. On the other hand, at all times, including the Idle/Hold state, power dissipation happens in the form of gate oxide leakage and subthreshold leakage. It is also important to note that the dynamic power is due to switching capacitance and transient gate oxide leakage current. Static power, on the other hand, is due to the steady gate oxide leakage and subthreshold leakage. Therefore, both dynamic and static components add up together to form the total power dissipation for the nano-CMOS transistor. Each one of them has several forms and origins; the components flow between different terminals and in different operations conditions of the transistor. Power dissipation is identified in individual transistors in the SRAM cell and for the overall SRAM array during different states in Write, Read and Hold modes. Eventually, this analysis and observations can be used in exploring new techniques for estimating power in large SRAM arrays and to design lower power SRAM structures.

In this chapter, the total power dissipation profile of the SRAM cell is thoroughly estimated, characterized and explained, and different components of total power, such as dynamic current, subthreshold leakage current, and gate leakage current in all three states of the SRAM cell operation (that is Write state, Read state and Hold state) are emphasized. In the following sections, the model used for analysis and simulation is the predictive technology model (PTM) 45 nm BSIM4 model. The analysis and characterization can be used in exploring new techniques for estimating power in large SRAM arrays and to design low power structures for SRAMs.

A nano-CMOS device in digital circuits operates in three regions: ON state, OFF state and transient state, i.e., ON to OFF and OFF to ON. A nano-CMOS device acts like a switch [61]. Figure 4.1 shows different current components for a nano-CMOS transistor as listed below:

- $I_1$: Drain-to-source active current (ON state). When $V_G$ (gate voltage) exceeds $V_{Th}$ (threshold voltage), active current flows between source and drain.

- $I_2$: Drain-to-source short circuit current (ON state). The origin of the short-circuit current l2 arises from various sources.
- $I_3$: Subthreshold leakage (OFF state).
- $I_4$: Gate-leakage current (both ON and OFF states).
- $I_5$: Gate current due to hot-carrier injection (both ON and OFF states).
- $I_6$: Channel punch-through current (OFF state).
- $I_7$: Gate-induced drain leakage current (OFF state).
- $I_8$: Band-to-band tunneling current (OFF state).

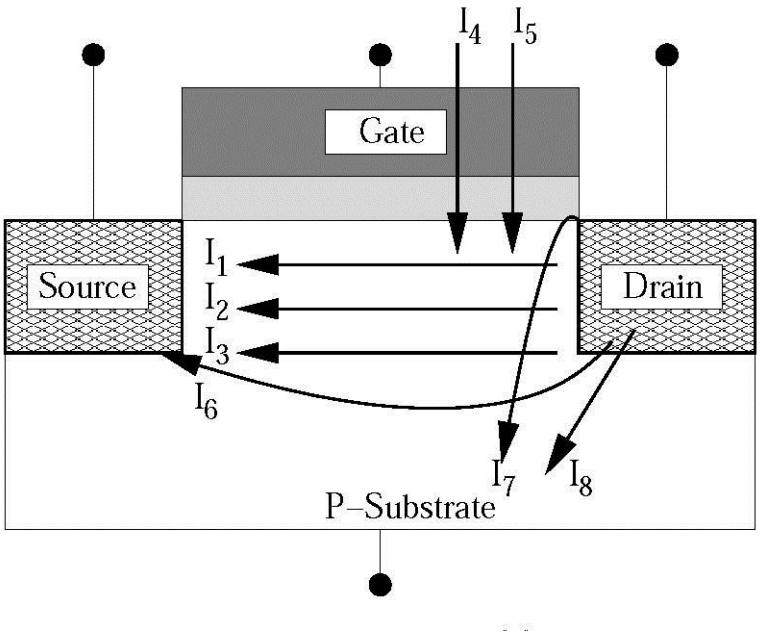



Figure 4.1. Current flow paths in a nano-CMOS transistor. The path and magnitude of these depend on the state of its operation [61].

4.1. Power and Leakage Models

Power dissipation of nano-CMOS circuits has become a major technical problem and challenge for the semiconductor industry. In order to reduce the power dissipation, extensive research is ongoing in this area. The major sources of power dissipation for a nano-CMOS circuit, as opposed to a transistor, are due to capacitive switching, subthreshold leakage, and gate leakage. They have diverse characteristics, origins, and paths.

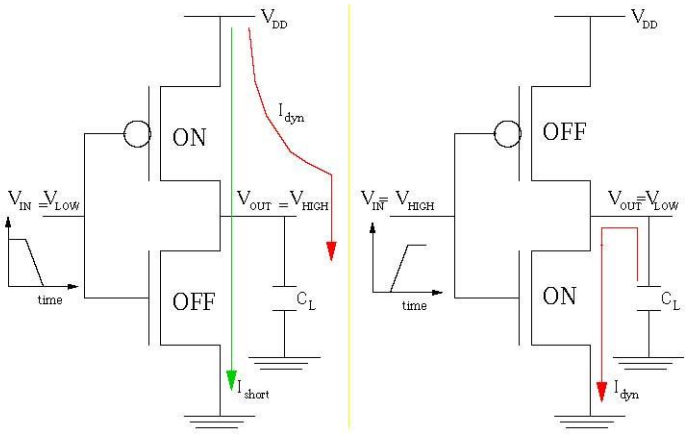
Figure 4.2 explains the different forms of current in the context of an inverter [68, 57, 13] dynamic power, subthreshold leakage and gate leakage. Gate oxide tunneling current ($I_{ox}$) has several components and they are all modeled in BSIM4 [14]. $I_{gs}$ and $I_{gd}$ are the components because of the overlap of gate and diffusions, $I_{gcs}$ and $I_{gcd}$ are the components due to tunneling from the gate to the diffusions via the channel and $I_{gb}$ is the component due to tunneling from the gate to the bulk via the channel. Similar components flow in a P-channel metal-oxide semiconductor (PMOS) device, with relatively smaller, yet of comparable magnitude, currents.

4.1.1. Total Power Dissipation

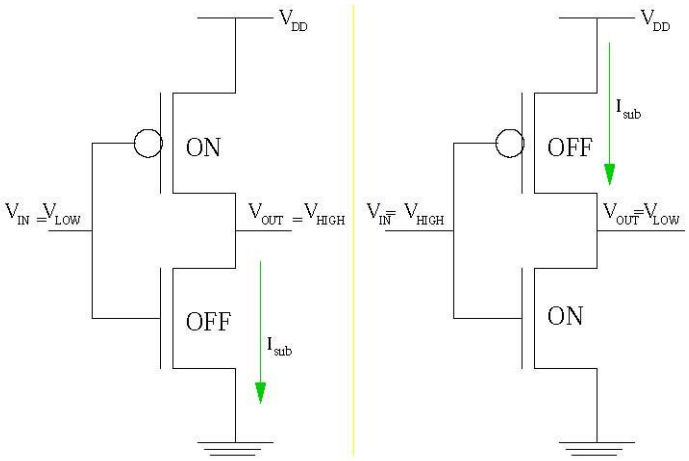
It is observed that the analysis of a complete SRAM cell would need the basic analysis of its building blocks, NMOS and PMOS transistors. The regions of operation are identified for NMOS and PMOS transistors, distinguishing their transient and steady states. Different modes of operations contribute to the overall current during different phases of the switching cycle. There are two regions of operations namely steady-state region (ON or OFF) and transient state (during Low-to-High and High-to-Low transition) [58]. In the case of steady state ON region, the gate and drain of the transistor are at high with the source being grounded forming a channel with three separate components of gate tunneling current $I_{gs}$, $I_{gcs}$, and $I_{gcd}$ are being in active mode. Here, the electric field in the oxide region is considered to be zero and so the component from gate to drain overlap ($I_{gd}$ ) is absent. The current flow path is from gate to source and channel, opposite to the flow in the OFF state.

In the other region of operation, the steady-state OFF region, both gate and source are at ground keeping the drain at high ($V_{DD}$) voltage. The only active component here is $I_{gd}$ because no channel is formed. When the device changes from ON to OFF or OFF to ON state, it is said to be in transient state. During low-to-high (LH) and high-to-low (HL), four components of the gate tunneling current become active as shown in Figure 4.3. In this case the source is at ground, the drain is at $V_{DD}$ and the gate is switched from low to high or high

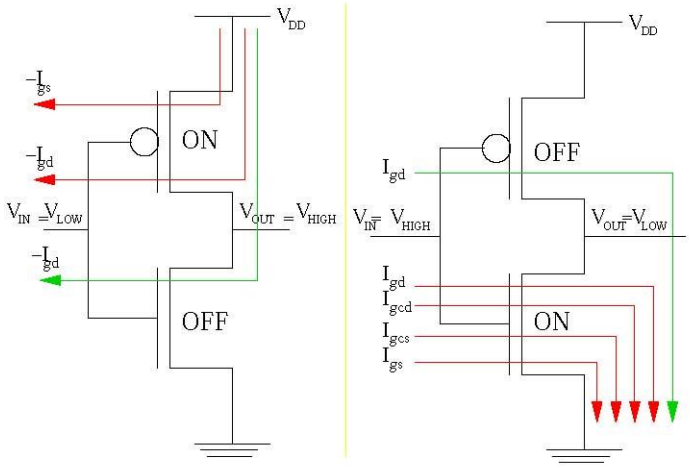
to low. In the LH transition, the channel is originating from the course and extends to the drain and the components $I_{gs}$, $I_{gcs}$, and $I_{gcd}$ start becoming significant. The field across the oxide region over the drain is reduced and thus $I_{gd}$ decreases. A study of this state is important for SRAMs because it can show the effect that transition from one of the states to the other has on the gate leakage.




(a) Capacitive switching power.

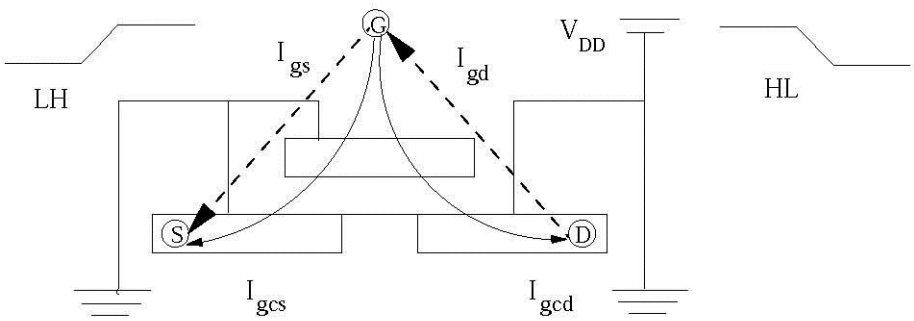



(b) Subthreshold leakage.

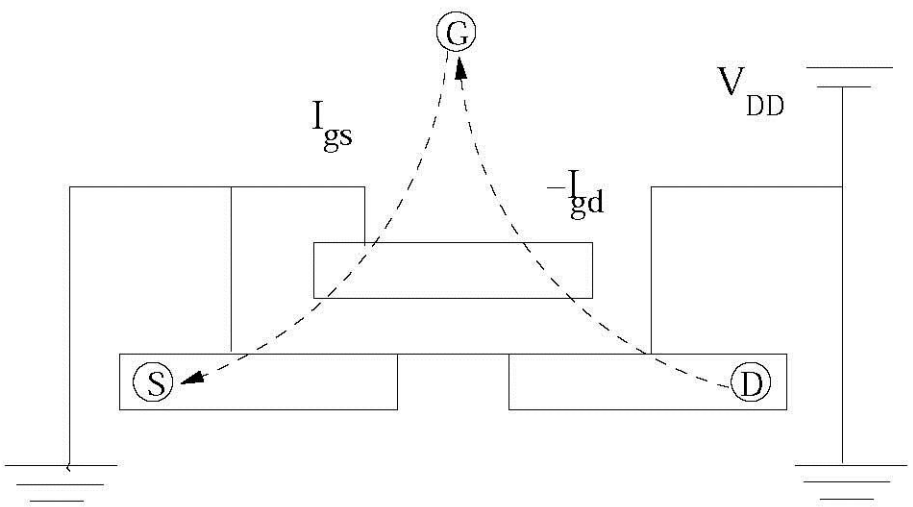



(c) Gate oxide leakage.

Figure 4.2. Major sources of power dissipation in a nano-CMOS circuit.




(a) NMOS.




(b) PMOS

Figure 4.3. Gate tunneling current flow for PMOS and NMOS transistors.

The total power consumption in an SRAM cell can be defined as the summation of transitions, all dynamic power, and static power using the following expression:

(3) $$P_{total} = P_{dynamic} + P_{static}$$

where $P_{dynamic}$ is the *dynamic power* which is consumed by the transistor as it makes transition from ON to OFF and vice versa. It includes the charging and discharging of capacitances. The *static power consumption* $P_{static}$ is due to leakage and standby components. Leakage consists of subthreshold leakage current, gate oxide leakage current and diode leakage flowing in the transistors of the cell. *Standby power* is present when NMOS and PMOS transistors are ON in a pseudo-NMOS inverter [56], when the drain of this NMOS transistor is driving the gate of another NMOS transistor or when the inputs of a CMOS gate leak a value between $V_{DD}$ and ground.

4.1.2. Dynamic Power

$P_{dynamic}$ is the dynamic power consumed by the circuit while there is a transition, that is when the circuit transits from ON to OFF or OFF to ON. The metric for dynamic power dissipation of the nano-CMOS SRAM circuit is quantified using the following expression [36]:

(4) $$P_{dynamic} = P_{trans-tunn} + P_{cap-switch} + P_{short-circuit}$$

where $P_{trans\text{-}tunn}$ is introduced in this research for accurate quantification of dynamic power consumption. It is the gate leakage power dissipation due to gate oxide tunneling during the device transition from ON to OFF or vice-versa [36]. $P_{cap\text{-}switch}$ is defined as the capacitive switching power which originates because of charging and discharging and $P_{short\text{-}circuit}$ is called the short-circuit power dissipation that occurs because of conduction of N-channel metal-oxide semiconductor (NMOS) and PMOS simultaneously. This results in a conduction path from the power supply to the ground for a short duration of time.

The above components are modeled using the following expression:

(5) $$P_{dynamic} = C_{eff-tunn}\left(\frac{V_{DD}^2}{t_r}\right) + \alpha C_{L-switch} V_{DD}^2 f + \left(\frac{\beta}{12}\right)(V_{DD} - 2V_{Th})^3 \left(\frac{t_r}{t_p}\right)$$

where α is the switching activity factor (between 0 and 1), $C_{L\text{-}switch}$ is the switching load capacitance, $V_{DD}$ is the supply voltage, and $f$ is the operating frequency. $C_{eff\text{-}tunn}$ is the loading effect of tunneling current [36] which quantifies the gate oxide leakage because of tunneling when the device makes transitions. β is the transistor gain factor, $V_{Th}$ is the threshold voltage, $t_r$ is the rise/fall time, and $t_p$ is the period of the input waveform.

Switching capacitance: Capacitance switching power is due to the charging and discharging of parasitic capacitance [56]. $P_{cap\text{-}switch}$ is proportional to the transistor widths in the circuit. $C_{L\text{-}switch}$ is related to dynamic energy consumption by:

(6) $$\text{Energy}_{dynamic} = C_{L\text{-}switch} V^2_{DD}$$

In Equation 6 we define the load capacitance as the total capacitance connected to the output, which includes the capacitance of the wires and transistors.

Short circuit current: The stand-by current is due to the DC voltage path between the supply and ground during the output transitions. In other words, the short circuit current is inducted when the input voltage transition is completed and the PMOS is turned off. The short-circuit current is dominant when the output load capacitance is small and when the input signal rise and fall times are large. Because of the fact that the PMOS and NMOS transistors are ON for a short period of time during the transition from 0 to 1 and 1 to 0, there is flow of current from $V_{DD}$ to ground (short-current pulse). Hence, we can define the short-circuit power consumption as [56]:

(7) $$P_{short\text{-}circuit} = \frac{\beta}{12}(V_{DD} - 2V_{Th})^3 \frac{t_r}{t_p}$$

where β is the transistor gain factor, $V_{DD}$ is the supply voltage, $V_{Th}$ is the threshold voltage, $t_r$ is the rise/fall time, and $t_p$ is the input waveform period.

4.1.3. Subthreshold Leakage

Static power is defined using the following expression, followed by a detailed discussion of each component:

$$P_{static} = P_{steady\text{-}tunn} + P_{subthreshold} + P_{reverse\text{-}biased}$$

where $P_{steady\text{-}tunn}$ is the gate leakage power originated from gate oxide tunneling during the steady state, $P_{subthreshold}$ is the subthreshold leakage, and $P_{reverse\text{-}biased}$ is the reversed-biased diode leakage.

Subthreshold leakage current: Subthreshold leakage is defined as the drain-source current present in the transistor in its OFF state. We denote the subthreshold leakage current by $I_{subthreshold}$. It is given by the following expression [25]:

$$I_{\text{subthreshold}} = A \exp\left[\left(\frac{1}{mV_t}\right)(V_G - V_S - V_{Th0} - \gamma V_S + \eta V_{DS})\right] \times \left(1 - \exp\left(\frac{-V_{DS}}{V_{Th}}\right)\right) \quad (9)$$

$V_{Th}$ where $A$ is a technology-dependent constant given by equation 10, $m$ is a doping-dependent coefficient, $V_t$ is the thermal voltage, $V_G$ is the gate voltage, $V_S$ is the source voltage, $V_{Th0}$ is the zero bias threshold voltage, γ is the body effect coefficient, η is the drain-induced barrier lowering (DIBL) coefficient, $V_{DS}$ is the voltage drop from drain to source, and $V_{Th}$ is the threshold voltage.

The constant $A$ is given by the following expression [25]:

$$A = e^{1.8}\,\mu_0 C_{ox}\left(\frac{W_{eff}}{L_{eff}}\right) V_t^2 \exp\left[\frac{-\Delta V_{Th}}{\eta V_{Th}}\right] \quad (10)$$

where $\mu_0$ is the zero bias mobility, $C_{ox}$ is the gate oxide capacitance, $W_{eff}$ and $L_{eff}$ are the effective transistor channel width and length, respectively. The term $\Delta V_{Th}$ accounts for

transistor-to-transistor leakage variations. Equation 9 takes into account the entire sub threshold leakage including weak inversion, DIBL and body effect.

Reverse-biased diode leakage current: Reverse-biased diode leakage current basically originates from the diodes formed between the diffusion regions of a device and the substrate consumes power in the form of reverse bias current. This reverse biased current is from the power supply, denoted by $V_{DD}$. Junction leakage originates due to band-to-band tunneling (BTBT), which means electron tunneling from band of p-side to the conduction band of n-side [67].

In case of inverter, depending on the input high or low, the NMOS device is turned ON and the output voltage will be pulled down to zero. Reverse potential difference of $V_{DD}$ is formed which in turn give rise to diode leakage current to flow through the drain junction. The reverse-biased diode leakage current is given by the following expression [67]:

$$I_{\text{reverse-biased}} = AJ_s \exp\left[\left(\frac{V_{bias}}{V_t}\right) - 1\right] \tag{11}$$

where $A$ is the junction area, $J_S$ is the reverse saturation current density and $V_{bias}$ is the reverse bias voltage across the junction.

4.1.4. Gate Oxide Leakage

Transient gate oxide tunneling current: In Equation 5, $C_{eff\text{-}tunn}$ is the effective capacitance that is due to tunneling. It is a crucial factor considering the loading of transistor due to tunneling [36]. During the transition of a transistor from ON to OFF (high to low) and OFF to ON (low to high), there is a flow of transient gate oxide leakage current. This happens when both NMOS and PMOS are ON for a very short duration of time and it results in flow of current from $V_{DD}$ to ground (short current pulse). Figure 4.3 shows the transient states HL (high to low) LH (low to high) of the CMOS cell. A gate tunneling current flow component is seen in the various regions of operation of a MOS. $I_{gb}$ (gate to bulk current) is zero throughout all regions of operation [36]. We can highlight that different mechanisms contribute to the overall current during different phases of the switching cycle. During tran-

sitions the components $I_{gs}$ , $I_{gd}$ , $I_{gcs}$ and $I_{gcd}$ are active. During HL and LH transitions we can define $C_{eff\text{-}tunn}$ as follows:

(12) $$\left| \frac{(I_{ON} - I_{OFF})}{\frac{dV_g}{dt}} \right|$$

where $I_{0N}$ and $I_{0FF}$ are the gate leakage components in the ON and OFF states, respectively, and $V_g$ is the gate voltage.

Steady-state gate oxide tunneling current: Steady state region refers to the OFF or ON region which means both gate and source are at ground while the drain is high (at $V_{DD}$) voltage. In this region is the active component is $I_{gd}$ (as no channel is formed here) and the direction of current flow is from diffusion to gate.

4.2. Power Consumption Analysis in different States of the SRAM Cell

The accurate power analysis in different states of the SRAM is explored using a sample standard 6T SRAM cell from Chapter 3. Figure 4.4 shows the basic SRAM cell. When the word line and bitline is asserted high, these both control the two access transistors N3 and N4 and as a result providing access to the cell. Further, these transistors will allow access to the memory cell via bitlines: BL and $\overline{\text{BL}}$. The function of BL and $\overline{\text{BL}}$ is used to transfer the data for Write and Read mode of operations. Also, BL and $\overline{\text{BL}}$ are used to improve the noise margins over a single bit line.

An analysis of the SRAM would need the basic analysis of its building blocks, NMOS and PMOS transistors. We identify the regions of operations of an NMOS device (which can then be extrapolated for a PMOS), distinguishing its transient and steady states. Different mechanisms contribute to the overall current during different phases of the switching cycle. The physical mechanism of the tunneling current comprises of steady-state region and transient state.

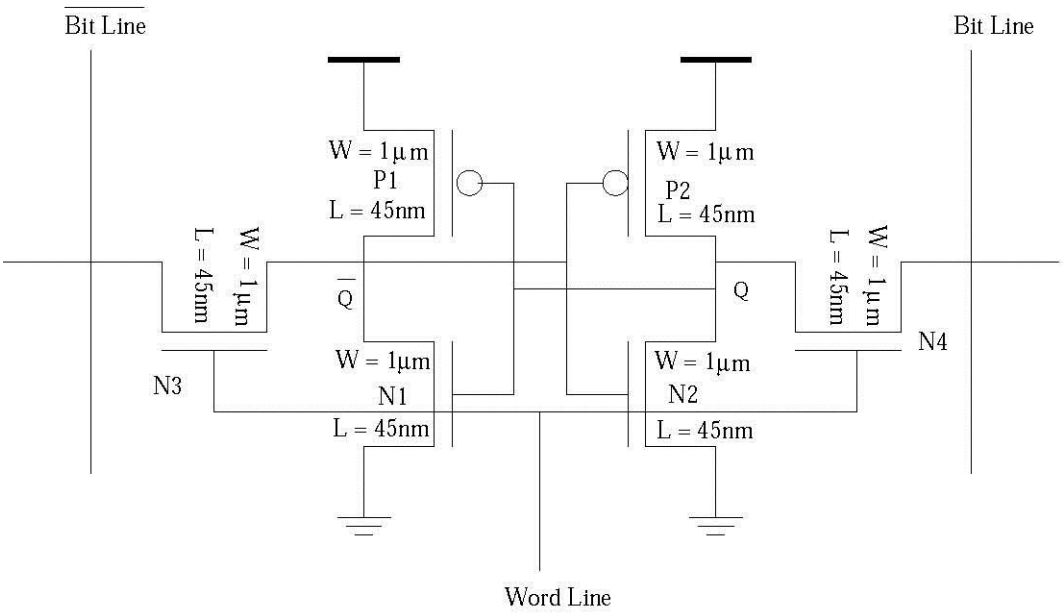



Figure 4.4. A 6T-SRAM cell with transistor sizes shown for 45nm CMOS technology.

As discussed above, the SRAM can operate in three modes, *viz*., Write, Read and Idle. These modes have different states and different combinations of transistors are active during these states leading to the prevalence of different leakage components and different values of gate leakage. The test bench for the analysis of gate leakage in the 6T SRAM cell is shown in Figure 4.5 and is based on the circuit in [1].

• Write State: The analysis is started with the WRITE operation. In case of writing a "1" the BL is held high and the $\overline{BL}$ is held low. In this state N2 and P2 leak the most as they are connected to the BL which is high. Also the access transistor on the BL side, N4 leaks significantly as both BL and the WL are high. These set of transistors are ON and provide a path for the gate leakage to flow. In Figure 4.5, P3 and P4 are known as the precharge circuit. In other words, PMOS transistors P3 and P4 are used for precharging bitlines BL and $\overline{BL}$. Drain is connected to $V_{DD}$ and source is connected to Bitlines. P1 and P3 are turned ON, too, and leak in this state. N1 and N3 are OFF and leakage is low in their case. The case for writing a "1" is exactly the reverse of this case where the transistors that are OFF while writing a "0" are ON during writing a "1" because of the symmetry in the SRAM cell. This research introduces terminologies like average power in Write "1" and Write "0" operation.

Average power over Write, Read, and Idle operations are described as the power averaged over the time which includes a complete ON-OFF cycle. The average power of the write operation can be mathematically expressed as:

(13) $$\text{AvgPower}_{\text{Write}} = \frac{1}{T_{\text{write}}} \int_0^{T_{\text{Write}}} p(t)dt$$

(14) $$= \left(\frac{V_{DD}}{T_{Write}}\right) \int_0^{T_{Write}} \left(I_{\text{gate}} + I_{\text{dynamic}} + I_{\text{subthreshold}}\right) V_{DD} dt$$

where $V_{DD}$ is the supply voltage and $I_{gate}$ is the current associated with gate-tunneling leakage whereas $I_{dynamic}$ contributes to the capacitive switching and $I_{subthreshold}$ is the subthreshold leakage flowing in the transistors.

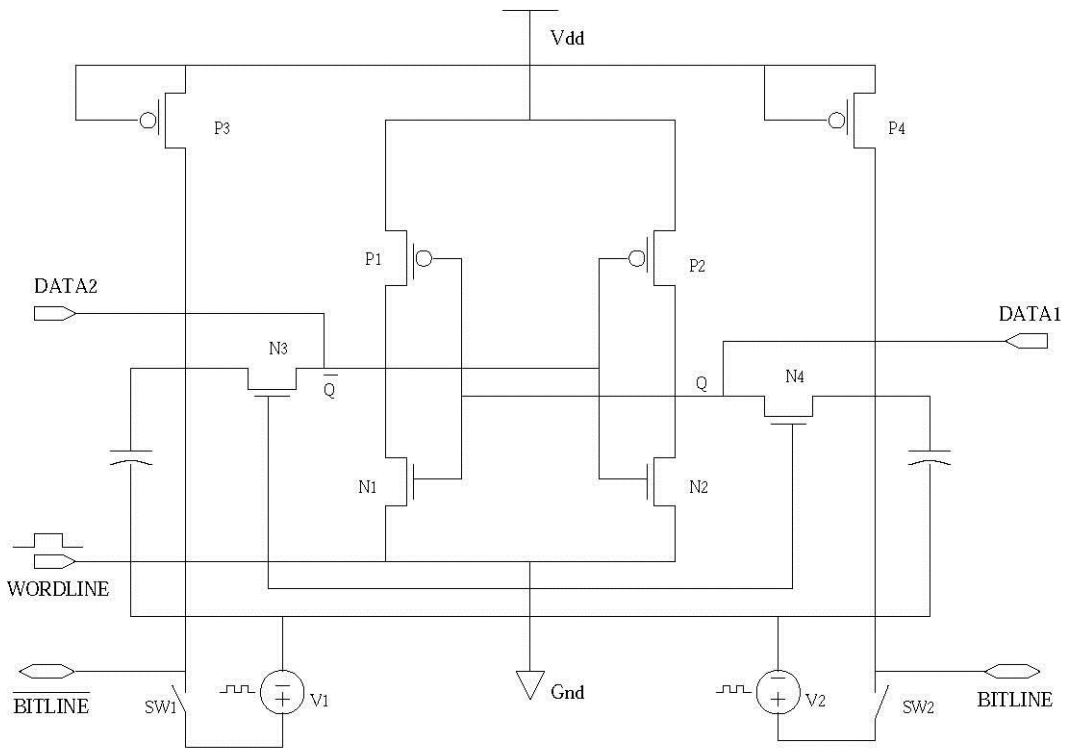



Figure 4.5. 6T-SRAM test bench used for power dissipation analysis.

• Read state: During the Read operation, the supplies $V_1$ and $V_2$ shown in Figure 4.5 are not required and hence disconnected from the circuit by turning of switches $SW_1$ and $SW_2$. Let us consider that the memory content is "1" (stored at Q) (for the sake of analysis). By making the BL high, read cycle is started which in turn enables both access transistors N3 and N4, which causes gate leakage created during the Write operation. BL is at its precharged value as a consequence of the Read operation and $\overline{\text{BL}}$ is discharged through N1 and N3.

Whereas, for BL side, transistors P2 and N4 pull the bit line towards $V_{DD}$ and as they are both ON, contribute to leakage. Depending on the content, if memory had a "0," the opposite action will take place and $\overline{\text{BL}}$ will be pulled towards "1" and BL towards "0." So the same set of transistors that contribute to the leakage in the Write operation contribute to the leakage during the Read state. Hence the leakage profile for a Read operation is expected to the be same as in the case of writing, and reading a "0" or "1" does not make any difference.
The average power in the case or read "1" or read "0" operation can be mathematically defined as follows:

(15) $$P_{\text{Read}} = \frac{1}{T_{\text{Read}}} \int_0^{T_{\text{Read}}} p(t)dt$$

(16) $$= \left(\frac{V_{DD}}{T_{\text{Read}}}\right) \int_0^{T_{\text{Read}}} \left(I_{\text{gate}} + I_{\text{dynamic}} + I_{\text{subthreshold}}\right)dt$$

• Hold state: The memory is in Hold state, or as some call it Idle state, when the word line WL is kept low, the function of pass transistors N3 and N4 is to disconnect the cell from the bit lines. The two cross coupled inverters N1-P1 and N2-P2 will continue to be active and hence leakage will take place in them even if they are disconnected from any external circuit. In this state the transistors connected to the power supply, i.e., P1 and P2 will still leak and hence a considerable gate leakage is still expected even if not of the same order as the Read or Write states. The average power consumed by the SRAM cell is defined as follows:

(17) $$P_{\text{Hold}} = \frac{1}{T_{\text{Hold}}} \int_0^{T_{\text{Hold}}} p(t)dt$$

(18) $$= \left(\frac{V_{DD}}{T_{\text{Hold}}}\right) \int_0^{T_{\text{Hold}}} \left(I_{\text{gate}} + I_{\text{ds}}\right)dt$$

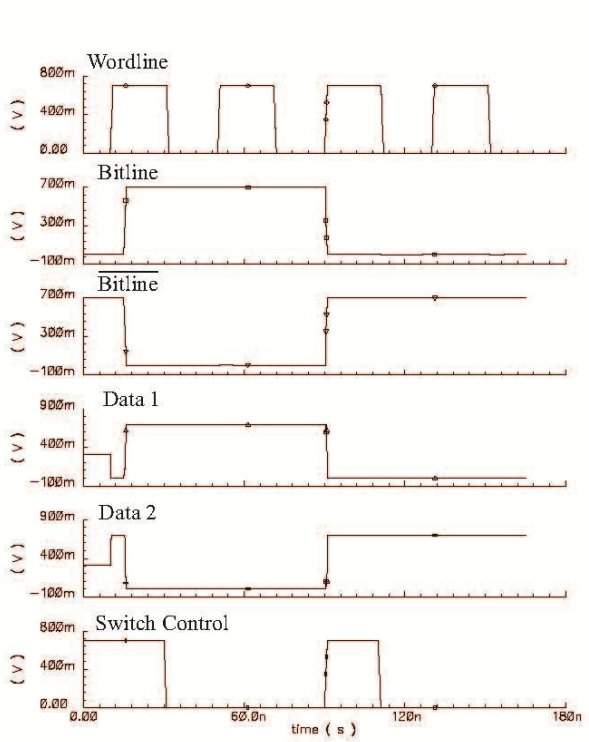
where $I_{gate}$ is the current associated with steady-state gate oxide tunneling and $I_{ds}$ is the drain to source current flowing in the transistors.

4.3 Accurate Power Analysis

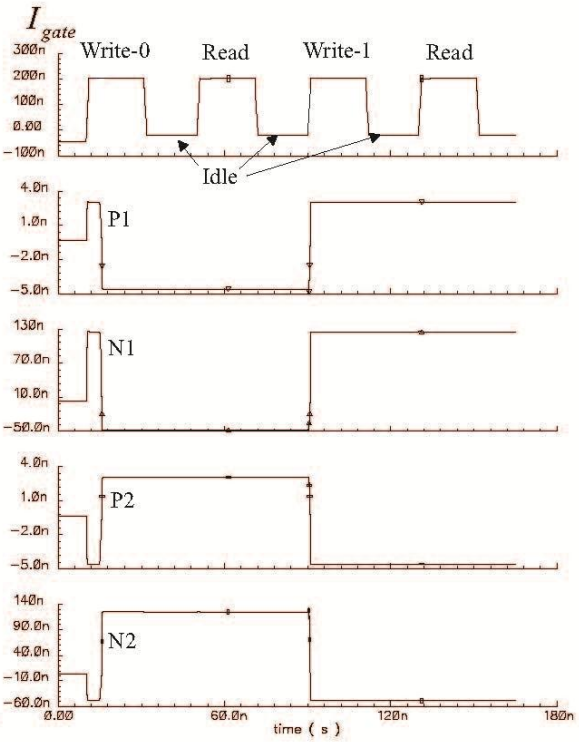
The state of the different inputs given to the SRAM circuit which result in various operations (Write-Read-Idle) are shown in Figure 4.6. Figure 4.6(a) shows the transient response of the input to the 6T SRAM cell and its corresponding gate leakage plot is given in

Figure 4.6(b). In the gate leakage plot, we have shown all different modes of operations: Write “0”-Idle-Read “0”-Idle-Write “1”-Idle-Read “1”-Idle. Along with this, the response of difference transistors PMOS1 (P1), NMOS1 (N1), PMOS2 (P2), and NMOS2 (N2) is also plotted; the subthreshold leakage current and the dynamic current are shown in Figure 4.6(d) and Figure 4.6(c).

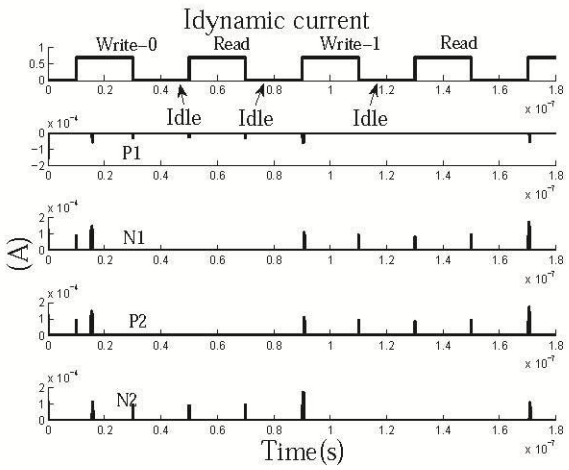
Transient analysis is done and shown in Figure 4.6; using the simulation results we show here the rise and fall time of the bitline switching and its corresponding behavior of dynamic and subthreshold current. Figure 4.7 represents the dynamic and subthreshold current during the rise and fall time in the active switching of, say, the write operation, whereas Figure 4.8 shows the gate-oxide current during the rise and fall time of the bitline switching.



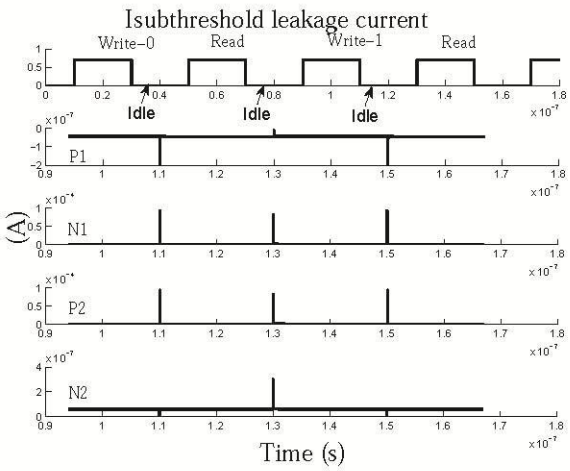

(a) The transient of total current during bitline switching.




(b) The transient of gate current during bitline switching.

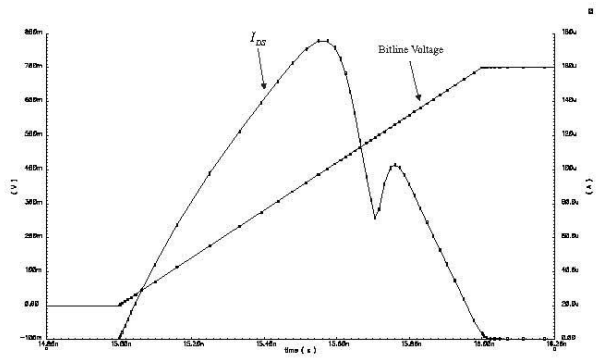



(c) The transient of dynamic current during bitline sitching.

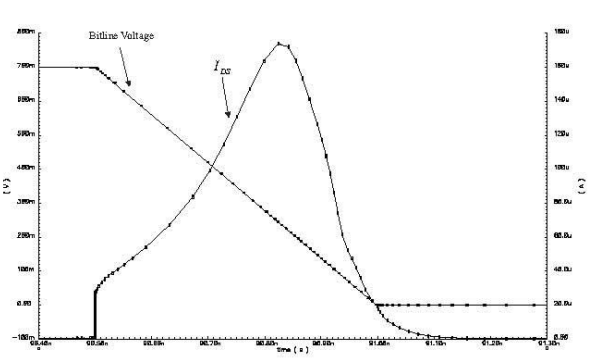



(d) The transient of subthreshold current during bitline switching.

Figure 4.6. 6T-SRAM transient analysis.

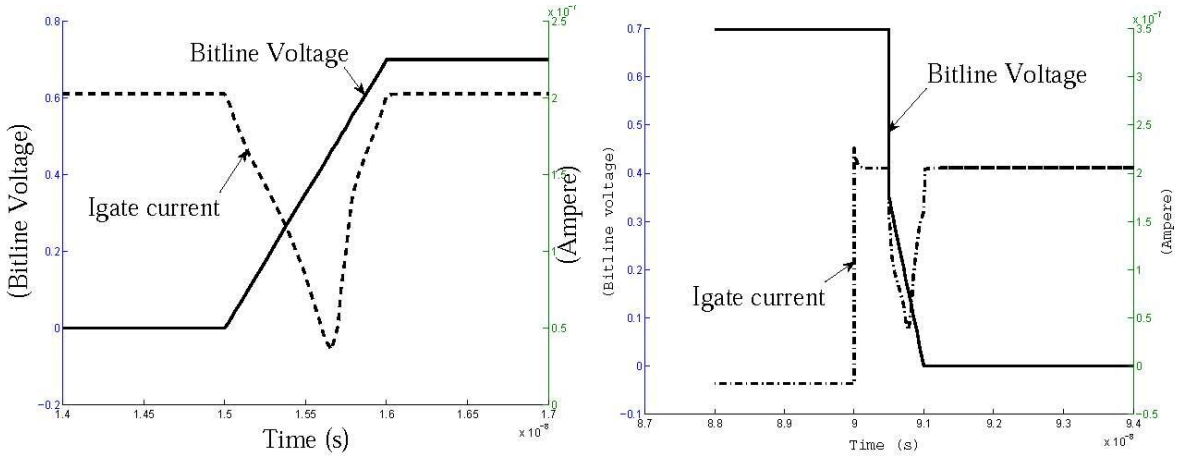



(a) The transient of $I_{ds}$ current during the rise time of bitline sitching.




(b) $I_{ds}$ current during the fall time of bitline switching.

Figure 4.7. $I_{ds}$ current (dynamic and subthreshold) during rise and fall times.




(a) The transient of $I_{gate}$ current during the rise time of bitline switching.

(b) $I_{gate}$ current during the fall time of bitline switching.

Figure 4.8. $I_{gate}$ current during rise time and fall time.

4.4. Statistical Process Variation Analysis of Total Power Dissipation

At the device level, we use the predictive technology model (BPTM) for a 45 nm device technology node with $T_{ox}$ = 1.4 nm, threshold voltage $V_{Th}$ = ±0.22V, and supply voltage $V_{DD}$ = 0.7 V. For design and simulation of the SRAM cell, the tool used here is Cadence Design Systems Analog Design Environment and Spectre (CDSADES) circuit simulator. Gate leakage is analyzed as the gate direct tunneling current by evaluating all components (source, drain and bulk) in each of the transistors of the cell BSIM 4.4.0 model [15]. The various states of the SRAM were simulated using the same test bench by including a pair of switches implemented in Verilog-A. The switches $SW_1$ and $SW_2$ connected to the BL and $\overline{BL}$ during the Write operation [and?] to the inputs and sources $V_1$ and $V_2$ help in precharging them. These were turned off during the Read and the Idle operations when the bitlines were only sensed. So, using a combination of piece-wise linear inputs for the wordline and the bit lines, a sequence of input and outputs states were achieved. The sequence comprised of Write “1” -Idle -Read “1”-Idle -Write “0” -Idle -Read “0” -Idle operations. We have calculated gate leakage currents over all the states. The average overall operations are then taken and Monte Carlo simulations are run. The 12 process parameters considered for variability are as follows:

- $T_{oxn}$: NMOS gate oxide thickness (nm)
- $T_{oxp}$: PMOS gate oxide thickness (nm)
- $L_{na}$: NMOS access transistor channel length (nm)
- $L_{pa}$: PMOS access transistor channel length (nm)
- $W_{na}$: NMOS access transistor channel width (nm)
- $W_{pa}$: PMOS access transistor channel width (nm)
- $L_{nd}$ : NMOS driver transistor channel length (nm)
- $W_{pd}$ : NMOS driver transistor channel width (nm)
- $L_{pl}$ : PMOS load transistor channel length (nm)
- $W_{pl}$ : PMOS load transistor channel width (nm)
- $N_{chn}$: NMOS channel doping concentration ($cm^{-3}$)
- $N_{chp}$: PMOS channel doping concentration ($cm^{-3}$)

Amongst the above parameters some are independent and others are correlated, which is taken into account during the simulation. The parameters discussed above are considered to have a Gaussian distribution with mean (μ) specified in the PTM and standard deviation (σ) taken as 10% of the mean. Thus we calculate the average gate leakage power and average dynamic plus subthreshold power for all three conditions, i.e., Write, Read and Idle, by recording the values of the mean and standard deviation respectively.

This section discusses the distribution plots for all three modes of operation for the 6T SRAM cell. Average gate leakage power, average dynamic power consumption, and average subthreshold leakage is calculated using $N$ = 1,000 Monte Carlo runs. Their respective means and standard deviations are recorded in Table 4.1 and also shown in the distribution plots in Figures 4.9, 4.10, 4.11, 4.12, 4.13, and 4.14. In all plots, the random variable is taken to be the natural logarithm of the power dissipation component investigated. Since these follow obvious normal distributions, the power leakage components themselves are lognormally distributed.

Figures 4.9(a), 4.9(b), and 4.9(c) show the distribution plots for read “0” for all three current components, namely, gate oxide current, dynamic current, and subthreshold current. Similarly, Figures 4.10(a), 4.10(b), and 4.10(c) show the distribution plots for read “1.” Figures 4.12(a), 4.12(b), and 4.12(c) show the distribution plots for Write “1” ; Figures 4.11(a), 4.11(b), and 4.11(c) show the distribution plots for Write “0”; Figures 4.14(a) and 4.14(b) show the distribution plots for idle “1” for gate oxide and subthreshold current. Figures 4.13(a) and 4.13(b) show the distribution plots for idle “0” for gate oxide and subthreshold current. Note that when the SRAM cell is idle, there is no dynamic current flowing, hence only gate oxide and subthreshold leakage currents pass through this mode. This simulates all possible scenarios involved in the SRAM where there is a gap between the various operations and different values are written to and read from the SRAM.

We consider that the SRAM cell initially writes a “1,” which is a bit information, and then it stores and reads this value using idle “1” and read “1” modes. Value “1” will be the output at Data1. Thereafter, in order to write a “0,” the procedure is the same except for just reversing the bitlines, and we get value “0” at Data1.

Table 4.1. Average dynamic, subthreshold power and average gate leakage power in Write, Read and Idle operations

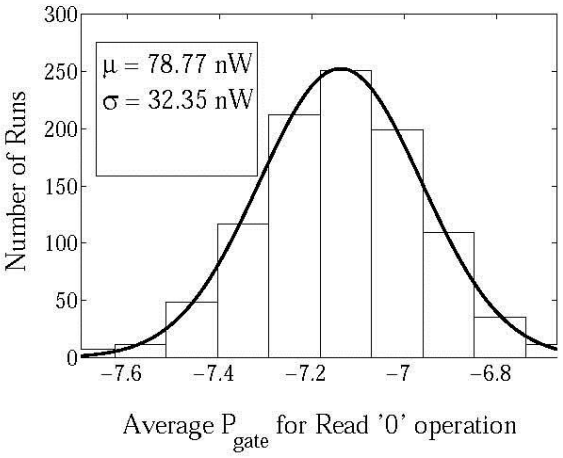
| Average Power | Operation | Mean (.) | Standard Deviation (0) |
|---|---|---|---|
| Average Gate Leakage Power | WRITE 1<br>WRITE 0 | 221.27nW<br>221.98nW | 94.73nW 95.07nW |
| | READ 1<br>READ 0 | 129.36nW<br>78.77nW | 54.43nW 32.35nW |
| | IDLE 1<br>IDLE 0 | 28.53nW<br>8.48nW | 18.59nW 5.09nW |
| Average Dynamic Power | WRITE 1<br>WRITE 0 | 76.90nW<br>80.5nW | 58.78nW 49.36nW |
| | READ 1<br>READ 0 | 76.56nW<br>77.49nW | 58.86nW 57.41nW |
| | IDLE 1<br>IDLE 0 | -- | -- |
| Average Subthreshold Leakage | WRITE 1<br>WRITE 0 | 144.17nW<br>106.7nW | 343.66nW 35.05nW |
| | READ 1<br>READ 0 | 78.56nW<br>65.58nW | 58.86nW 224.32nW |
| | IDLE 1<br>IDLE 0 | 65.87nW<br>79.86nW | 46.97nW 50.42nW |
| Average Total Power | WRITE 1<br>WRITE 0 | 226.77nW<br>123.56nW | 340.21nW 42.18nW |
| | READ 1<br>READ 0 | 92.68nW<br>93.48nW | 65.20nW 63.77nW |
| | IDLE 1<br>IDLE 0 | 82.37nW<br>96.13nW | 58.06nW 57.53nW |

We notice from Table 4.1 that write “1” and write “1” consume similar average gate power. While in the read operation, we consider that at this point one bit information is already written to the SRAM cell. Thus the read operation consumes less power in average gate leakage, dynamic power and subthreshold leakage as compared to that of the write operation. Also from Table 4.1 we observe that the average power consumption of the read operation regardless of “1” or “0” is similar. After the data is written or read it may be stored to the SRAM cell and thus idle “1” is the state where data “1” is stored and idle “0” is where data “0” is stored. In this state there is no transition in the transistors hence there will be no dynamic current flowing so it only consumes subthreshold and gate leakages which are shown in Table 4.1. The average power consumption in the case of idle “1” or “0” is similar.

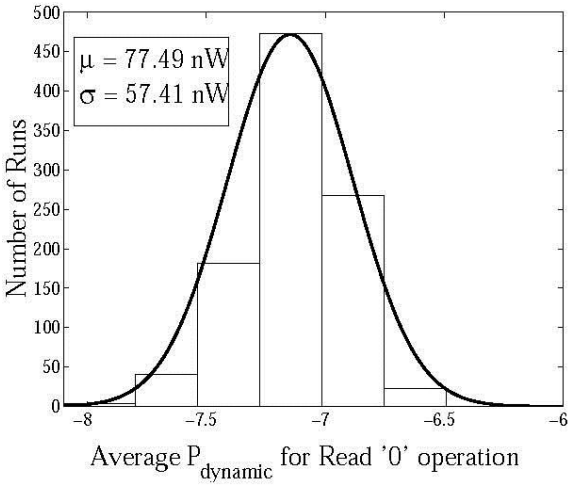
The average power for all leakage and power consumption is calculated using $N$ = 1000 Monte Carlo runs and then the mean and standard deviation values are calculated in Table 4.1. We consider that the SRAM cell initially writes a ",1" which is a bit information, and then it stores and reads this value using idle "1" and read "1" modes. Value "1" will be the output at Data1. Thereafter, in order to write a "1," the procedure is the same and by just reversing the bitlines, we get value "0" at Data1.

We notice from Table 4.1 that write "1" and write "0" consume similar average gate power. It is observed that the write operation results in a lot of subthreshold leakage compared to the dynamic power because in the subthreshold state the transistors are OFF for a longer duration of time as compared to that of dynamic state, thus resulting in more power. While in the read operation, we consider that at this point one bit information is already written to the SRAM cell. Thus as a result the read operation consumes less power in average gate leakage, dynamic power, and subthreshold leakage as compared to that of the write operation. Also from Table 4.1 we observe that the average power consumption of read operation whether "1" or "0" is similar.

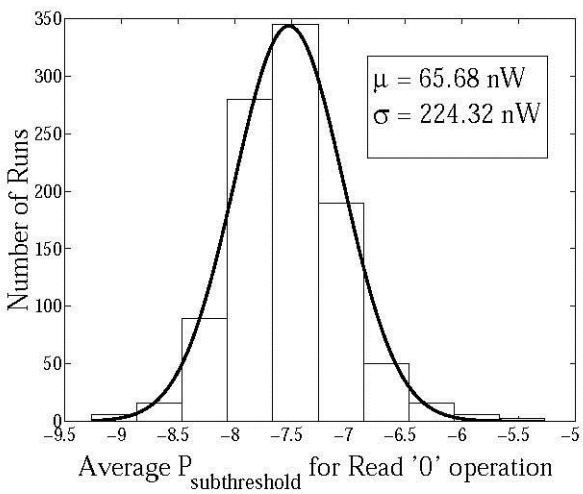
After the data is written or read it can be stored to the SRAM cell and thus idle "1" is the state where data "1" is stored and idle "0" is where data "0" is stored. In this state there are no transitions in the transistors, hence there will be no dynamic current flowing so it only gives the subthreshold and gate leakages, which are shown in Table 4.1. The average power consumption in the case of idle "1" or "0" is similar.




(a) Distribution of average gate leakage power for Read "0" operation.

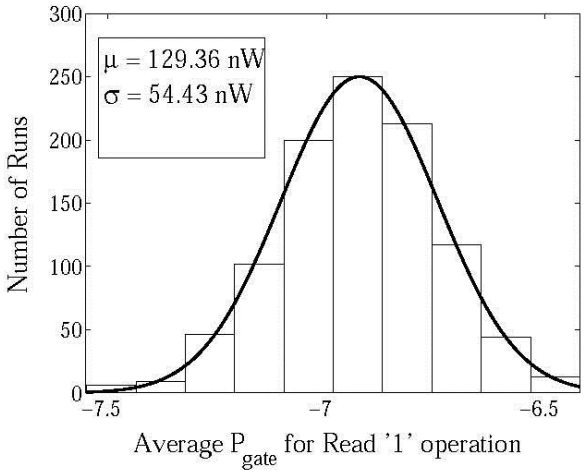



(b) Distribution of average dynamic power for Read"0" operation.

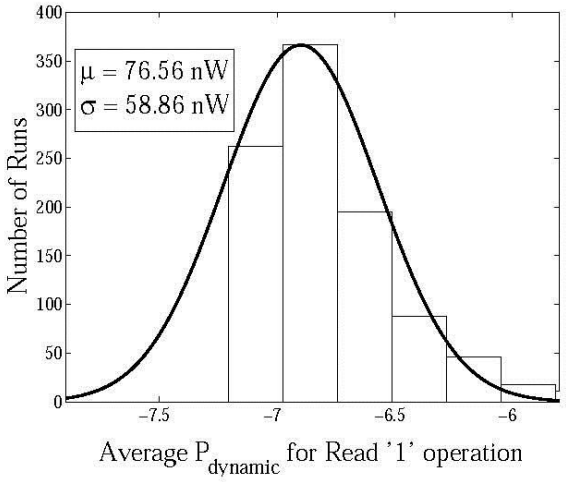



(c) Distribution of average subthreshold leakage power for Read "0" operation.

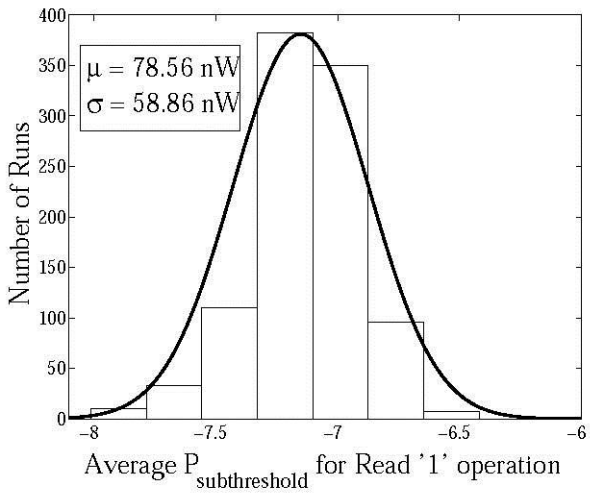
Figure 4.9. Statistical distributions for total power (including leakage) during Read “0” operation of the 6T-SRAM cell.




(a) Distribution of average gate leakage power for Read "I" operation.

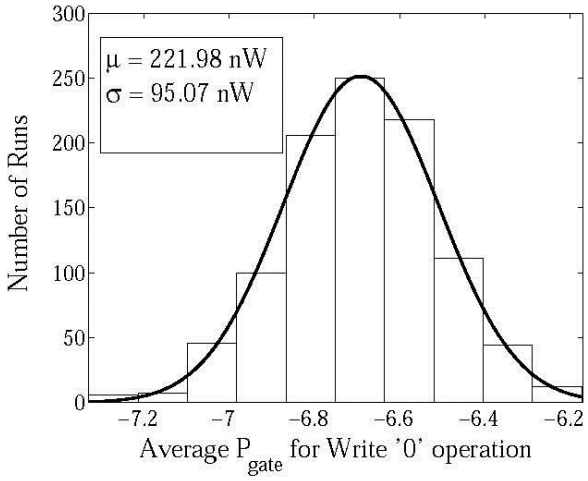



(b) Distribution of average dynamic power for Read "I" operation.

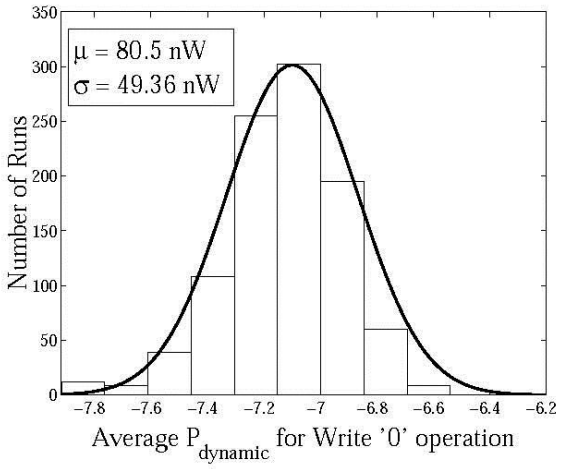



(c) Distribution of average subthreshold leakage power for Read "I" operation.

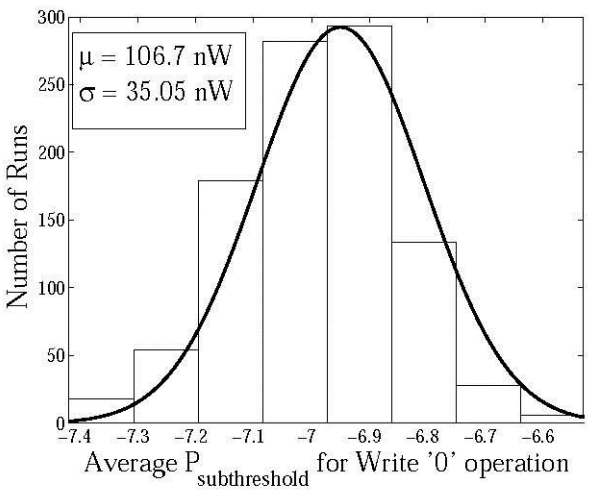
Figure 4.10. Statistical distributions for total power (including leakage) during Read "1" operation of the 6T-SRAM cell.




(a) Distribution of average gate leakage power for Write "0" operation.

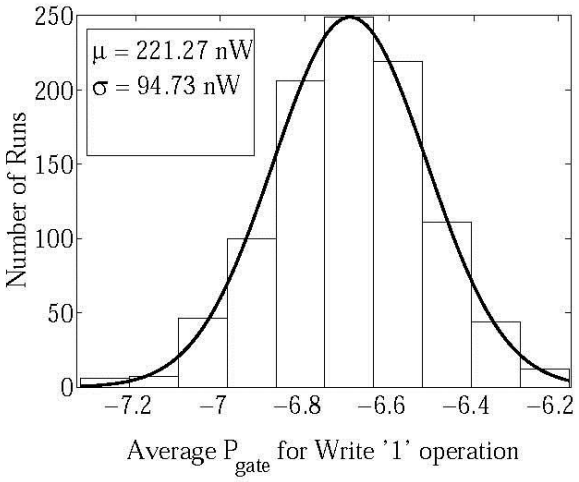



(b) Distribution of average dynamic power for Write"0" operation.

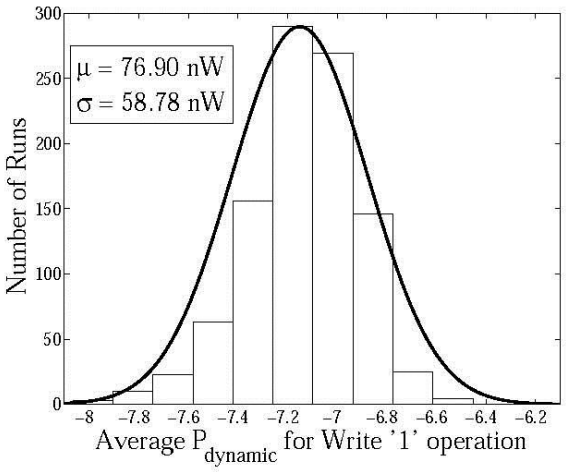



(c) Distribution of average subthreshold leakage power for Write "0" operation.

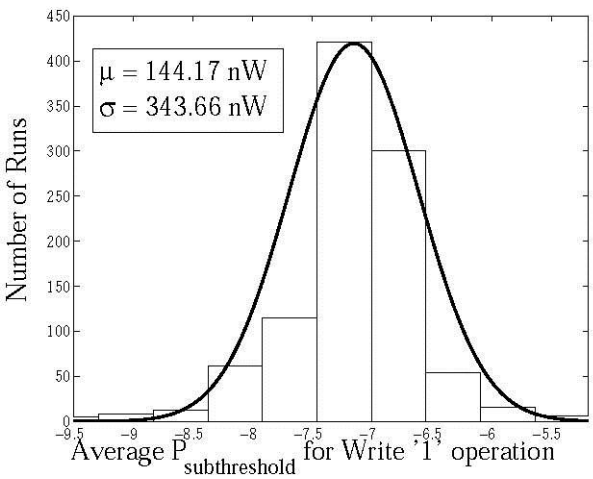
Figure 4.11. Statistical distributions for total power (including leakage) during Write "0" operation of the 6T-SRAM cell.




(a) Distribution of average gate leakage power for Write "I" operation.

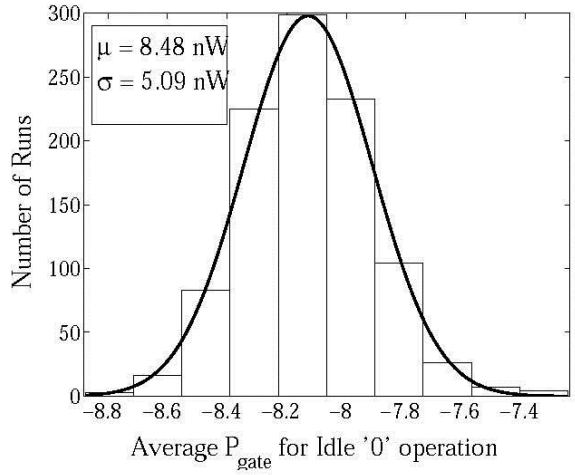



(b) Distribution of average dynamic power for Write "I" operation.

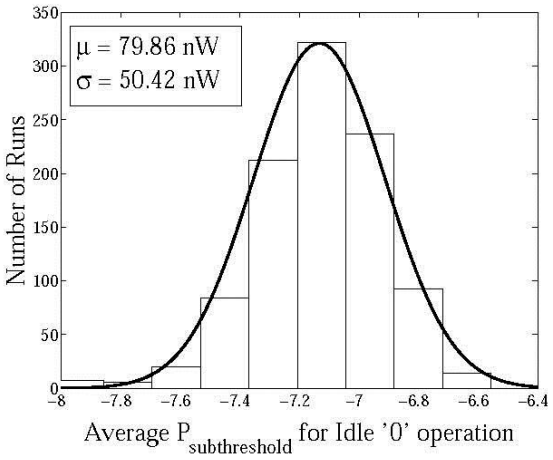



(c) Distribution of average subthreshold leakage power for Write "I" operation.

Figure 4.12. Statistical distributions for total power (including leakage) during Write "1" operation of the 6T-SRAM cell.

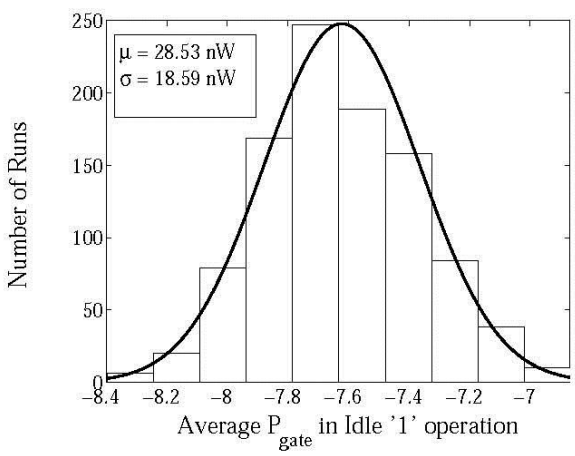



(a) Distribution of average gate leakage power for Idle power for Idle "0" operation.

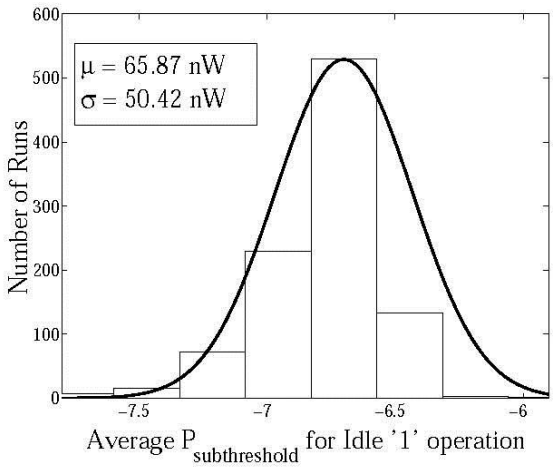



(b) Distribution of average subthreshold leakage "0" operation.

Figure 4.13. Statistical distributions for total power (including leakage) during Idle "0" operation of the 6T-SRAM cell.




(a) Distribution of average gate leakage power for Idle power for Idle "I" operation.




(b) Distribution of average subthreshold leakage "I" operation.

Figure 4.14. Statistical distributions for total power (including leakage) during Idle "1" operation the of 6T-SRAM cell.

4.5. Static Noise Margin (SNM) Analysis

The objective of functionality of SRAM in modern-day electronics is to increase the bit counts stored while maintaining the low power dissipation and high performance of the circuit. Low power consumption and high performance of a nano-CMOS circuit can be

achieved by scaling down the supply voltage. Another aspect of scaling down the supply voltage is that it presents a challenge for design engineers and process engineers to achieve reliable data storage in SRAM arrays. This might be a difficult task as SRAM arrays contain millions of bits.

Another factor affecting the reliability of the SRAM cell is the SRAM stability margin or static noise margin (SNM). Read static noise margin is defined as the minimum DC noise voltage which is required to flip the state of the SRAM cell during the read operation [5]. 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. The SNM can serve as a figure of merit in stability evaluation of SRAM cells; thus in this dissertation we treat the read SNM as a measure of performance. The SNM of even defect-free cells is declining with technology scaling. 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 [66]. Detection and correction/repair of such cells in modern scaled-down SRAMs becomes a necessity.

The stability of memory is a growing concern ever since, but this issue becomes more serious when it comes to nano-CMOS devices in fabrication because of process variation. These variations in device parameters translate into variations in SRAM areas like power and performance. Under worse conditions such SRAMs may even corrupt data. Thus, maintaining acceptable performance of SRAM along with minimum feature sizes and supply voltages of SoCs is an ongoing challenge.

This increased contribution becomes extremely important in applications with long idle modes, as in the case of wireless micro sensor systems, in which the standby period is much longer than the active mode. Thus with the progression of technology scaling into the nanometer region, the stability of embedded SRAMs is a growing concern. Large SRAM arrays occupy a large portion of the die area and are used mainly as cache memory in

application-specific integrated circuits (ICs) and microprocessors. As a result large SRAM arrays have a huge effect on all aspects of chip design and manufacturing.

The read static noise margin measurement model is described in this section. It is defined using the input voltage to output voltage transfer characteristics (VTC). Noise is present long enough for the circuit to react, i.e., the noise may be static or DC. A static noise margin is implied if the noise is a DC source.

The first model for SNM measurement was discussed by Hill [26], [66]. This method can be automated using a circuit simulator. Figure 4.15 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 static noise sources. Static noise source is defined as DC disturbances and mismatches because of variations and processing in operating conditions of the cell. 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 [17]. In order to obtain the butterfly curves as shown in Figure 4.16, the voltages are varied to and from node Q and Qb alternatively. The SRAM cell is simulated at 45nm CMOS technology using the PTM model [82] with supply voltage $V_{DD}$ of 0.7 V and with standard sized transistors. Given in Equation 19 is the standard SNM model for a 6-Transistor SRAM cell [17]:

(19) $$\text{SNM} = V_{Th} - \left(\frac{1}{k+1}\right)\left[\left(\frac{V_{DD} - \left(\frac{2r+1}{r+1}\right)V_{Th}}{1+\left(\frac{r}{k(r+1)}\right)}\right)\right.$$

(20) $$\left. - \left(\frac{V_{DD} - 2V_{Th}}{1+\left(\frac{r}{q}k\right)+\sqrt{\left(\frac{r}{q}\right)\left(1+2k+\frac{r}{q}k^2\right)}}\right)\right]$$

where $r$ is the ratio of $\left(\frac{\beta d}{\beta a}\right)$, which is the SRAM cell ratio or in other terms the ratio of the driver transistors $\left(\frac{W}{L}\right)$ to the access transistors $\left(\frac{W}{L}\right)$. Similarly, $q$ is the ratio of $\left(\frac{\beta p}{\beta a}\right)$, i.e., the ratio of the load transistors $\left(\frac{W}{L}\right)$ to the access transistors $\left(\frac{W}{L}\right)$. $V_{Th}$ is the threshold voltage; $k$ is

defined as follows:

(21) $$k = \left(\frac{r}{r+1}\right)\left[\sqrt{\frac{r+1}{r+1-V_s^2/V_r^2}} - 1\right]$$

$V_s$ is ($V_{DD}$ - $V_{Th}$), and $V_r$ is calculated using the following expression:

(22) $$V_r = \left(V_s - \left(\frac{r}{r+1}\, V_{Th}\right)\right)$$

The research done in this dissertation is a simulation based approach and is based on the same concept for a 7-Transistor SRAM. The SNM measured for the baseline 7T SRAM is shown in Figure 4.16(a) and optimized SNM or optimized butterfly curve is shown in Figure 4.16(b). The details on optimization are discussed in Chapter 1.

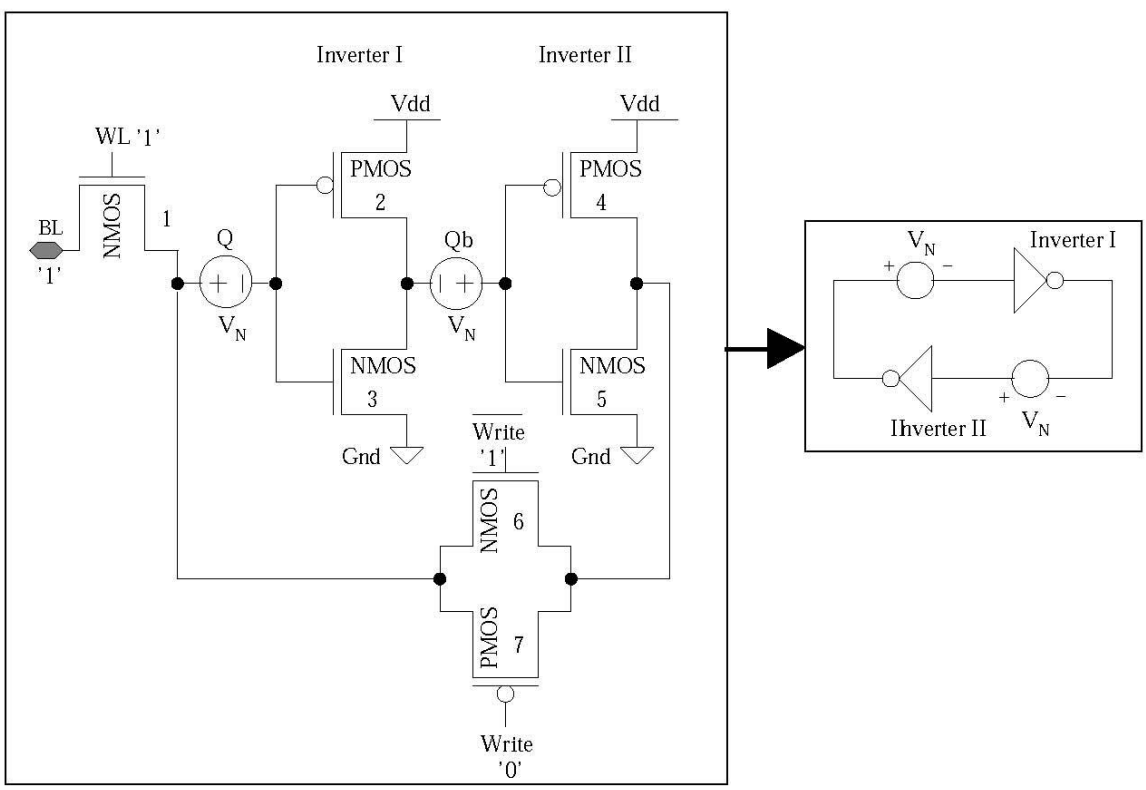
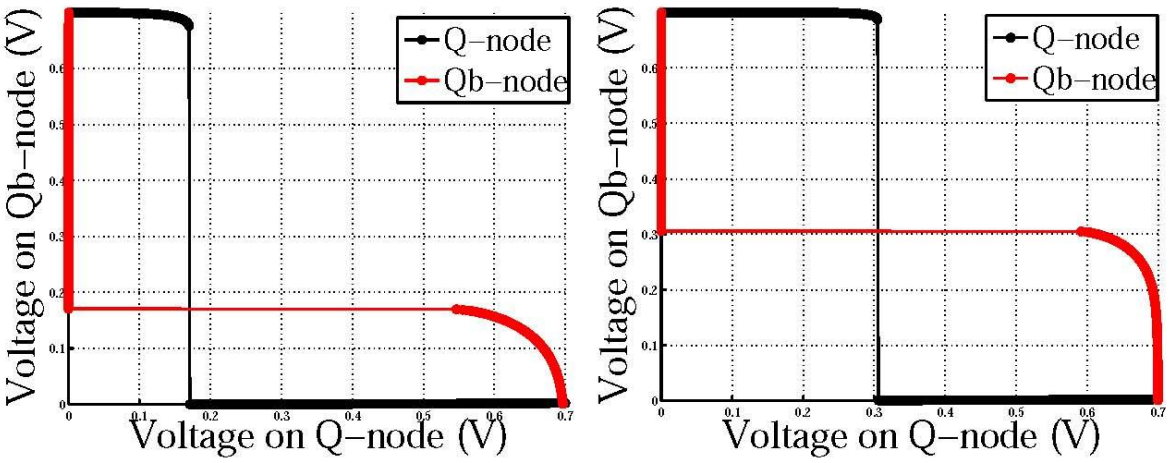



Figure 4.15. SNM model setup.




(a) Butterfly curve for baseline design (b) Butterfly curve for optimized design

Figure 4.16. Butterfly curves for baseline and optimized 7T SRAM.

# CHAPTER 5

# LOW-POWER SRAM DESIGN AND PROPOSED OPTIMIZATION METHODOLOGIES

This chapter constitutes one of the significant contributions of this dissertation. The issue of power dissipation and its reduction techniques are discussed first. This is followed by an explanation of optimization methodologies for power and performance optimal designs in general for digital circuits at nanoscale.

## 5.1. Why Low-Power?

The low-power design of complementary metal-oxide semiconductor (CMOS) circuits in general and static random access memory [SRAM] in particular as it occupies significant space budget of a chip) is as follows [61]:

- Decreases leakage which is prevailing with technology scaling
- Maintains supply voltage levels
- Increases system reliability
- Uses smaller heat sinks
- Makes packaging cheaper
- Increases battery life time
- Enhances noise margin
- Decreases energy costs.
- Increases robustness
- Decreases in natural resources
- Reduces power supply noise
- Reduces cross-talk and electromagnetic noise

## 5.2. Low-Power Techniques

From the technology point of view, the various leakage mechanisms, especially subthreshold leakage current, will continue to dominate the overall power dissipation. Similarly, from an optimal power point of view, the optimum energy during active or

dynamic modes of operations will correspond to a larger subthreshold leakage component [32]. Currently, leakage currents have become extremely important, especially for systems where the majority of time they are in idle or sleep mode, which means not actively working. The best example would be a cell phone; it is in active mode only when we are talking or texting or using some application on it, other than that it goes into sleep mode. For such systems, it may be acceptable to consume leakage current during active mode but during the sleep mode it is considered as waste because no work is being done. There are a large number of proposed techniques to help in reducing subthreshold leakage currents during sleep modes. Some of them are listed in Table 5.1. However, this dissertation concentrates on two low-power techniques: dual $V_{Th}$ and transistor sizing. They are described below as separate subsections.

Table 5.1. Different Power Reduction Techniques

| Traditional Techniques | Dynamic Power Reduction | Leakage Power Reduction | Other low power Techniques |
|---|---|---|---|
| Clock gating | Clock gating | Minimum usage of low $V_{Th}$ cells | Multi oxide devices |
| Power gating | Power efficient circuits | Power gating | Minimize capacitance by custom design |
| Variable frequency | Variable frequency | Back biasing | Power efficient circuits |
| Variable voltage supple | Variable voltage supply | Reduced oxide thickness | |
| Variable device threshold | | Uses FinFETs | |

### 5.2.1. Dual Threshold (dual $V_{Th}$)

There are typically two $V_{Th}$ process options provided by foundries: high $V_{Th}$ and low $V_{Th}$. With the limited number of options the designers are forced to optimize either power of performance of the circuits. The options are to use different $V_{Th}$ transistors for different logic gates within a device. So, for critical paths where performance is required, they can use low $V_{Th}$ leakage transistors, and where main concern is leakage and not performance, designers can use high $V_{Th}$ transistors. Increasing the threshold voltage would decrease the leakage current abundantly as the leakage is exponentially dependent on threshold voltage. Thus, the

dynamic approach results in reducing standby leakage power along with the performance degradation.

Another technique known as the dual-$V_{Th}$ dual-threshold complementary metal-oxide semiconductor (DTCMOS) design technique, is commonly known as a static approach. It is widely used for reducing leakage power (both active and standby leakage) in custom very large scale integration (VLSI) circuit designs without any performance degradation. Among the different kinds of transistors used in the circuit, some transistors have a high threshold voltage (having less subthreshold leakage-power dissipation but also have a larger delay), and other transistors have a low threshold voltage which are faster.

5.2.2. Transistor Sizing

Transistor sizing means increasing or reducing the width of the channel of a transistor and is used to improve the delay of a nano-CMOS circuit. Increasing the width of the channel results in increasing current drive capability of the transistor, which further reduces the rise and fall times at the output gate terminal of the device. Resizing the transistors has a tremendous impact on the power dissipation of a nano-CMOS circuit. There has been a lot of work explored in transistor sizing and gate sizing, which also involves several assumptions. The implementation of this area in real-life circuits involves proper cost models and accurate algorithms. This dissertation includes some of these algorithms. Before proceeding further, the theory behind this approach is discussed.

If Equation 23 is studied carefully, it is noticed that the static and dynamic characteristics of CMOS device are dependent on the transistor properties β and $V_{Th}$. For proper functioning of transistors, designers should have control over these parameters. However, in reality there is only a limited amount of control available. The threshold voltage may not be modified by the designer and is determined by the process: $V_{Th}$ is directly proportional to the thickness of gate-oxide and potential of the substrate where the doping in the substrate and the oxide thickness are parameters determined by process designers. The potential of the

substrate can be varied by the designer, which in turn effects $V_{Th}$, a process known as body effect. The other parameter, β, is given by the following expression:

$$\beta = \left(\frac{\mu\epsilon_{ox}}{T_{ox}}\right)\left(\frac{W}{L}\right) \quad (23)$$

It is observed that the circuit designer has no control over the factor $\left(\frac{\mu\epsilon_{ox}}{T_{ox}}\right)$ because the carrier mobility μ is a fixed property, $\epsilon_{ox}$ is a fixed parameter, and $T_{ox}$ is set by the process. Therefore, it is concluded that the designer has a two-dimensional parameter space as an option, *W* and *L*, both of which can be altered, provided they stay within the design rules for any feature size technology node. However, the smaller the *L*, the larger the β, and the smaller the gate capacitance. This translates to a faster circuit; therefore circuit designers mostly opt for minimum *L* feature size.

Table 5.2. Parameter Sets

| Geometric parameter set | Process parameter set | On-chip parameter set |
|---|---|---|
| $L_{na}$: NMOS access transistor | $V_{thn}$: NMOS threshold voltage (V) channel length (nm) | $V_{DD}$: Supply voltage |
| $W_{na}$: NMOS access transistor | $V_{thp}$: PMOS threshold voltage (V) channel width (nm) | |
| $L_{pl}$: PMOS load transistor channel length (nm) | $N_{gatep}$: NMOS gate doping concentration ($cm^{-3}$) | |
| $W_{pl}$: NMOS load transistor channel width (nm) | $N_{gatep}$: PMOS gate doping concentration ($cm^{-3}$) | |
| | $N_{chn}$: NMOS channel doping concentration ($cm^{-3}$) | |
| | $N_{chp}$: PMOS channel doping concentration ($cm^{-3}$) | |
| | $T_{ox}$: Gate oxide thickness (nm) | |
| | $N_{sdn}$: NMOS source/drain doping concentration ($cm^{-3}$) | |
| | $N_{sdp}$: PMOS source/drain doping concentration ($cm^{-3}$) | |

The research in this dissertation explores a geometric parameter set and investigates the sizing of NMOS, PMOS, and access transistors of the SRAM cell. The parameter sets are categorized in Table 5.2.

## 5.3. Proposed Optimization Methodologies

In this section various novel optimization methodologies are discussed.

### 5.3.1. Combined Design or Experiments-Integer Linear Programming (DOE-ILP) Based Algorithm

This section discusses a novel optimization methodology which is applied on one of the low-power techniques already discussed in Section 5.2. A novel design approach for simultaneous optimization of two figures of merit, that is, power and stability (static noise margin [SNM]), applied to a 7T SRAM is presented, as shown in Figure 5.1. In this work, a 45 nm 7T SRAM (discussed in Chapter 3) is used as a case study. To show the robustness of the design, the process variation analysis of the optimal SRAM cell is performed.

The input of the flow is a baseline 7T SRAM, which refers to the implementation of design with nominal sized transistors for a 45 nm technology. In embedded SRAMs it is a challenge for designer engineers to simultaneously maintain read SNM while reducing power consumption of the cell. This work explores dual-$V_{Th}$, which is a process level technique already discussed in Section 5.2. The proposed algorithm will help in finding solutions to the research problem, i.e., finding appropriate transistors for high-$V_{Th}$ assignment. This technique decreases the optimization search space, converges solutions faster, and also maintains the accuracy of ILP; this is the advantage of using this technique. This approach may be used to build large circuits for optimization in less time.

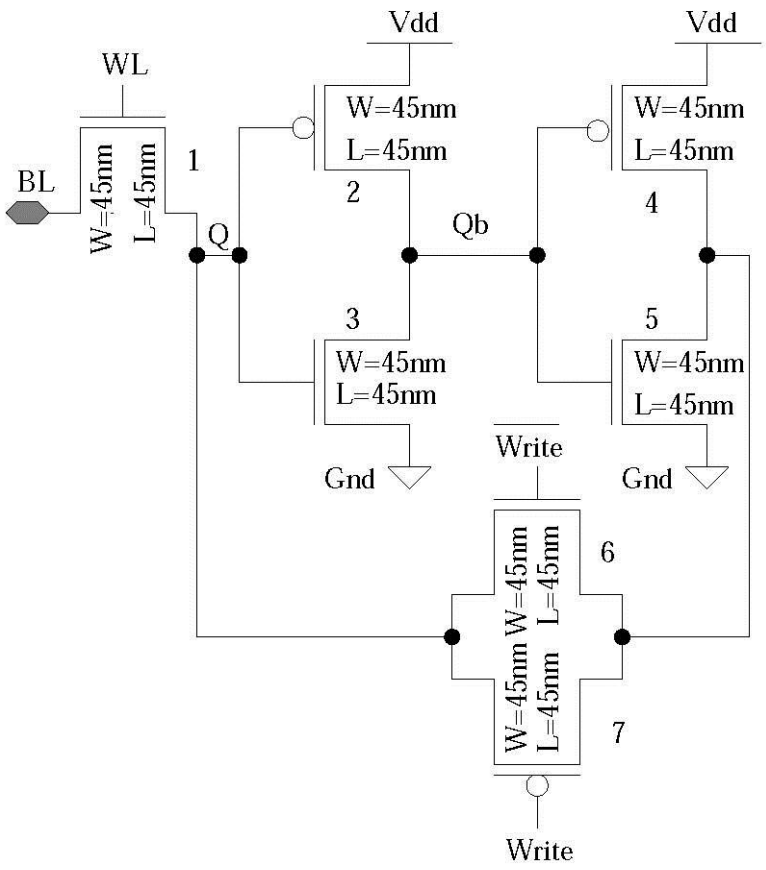


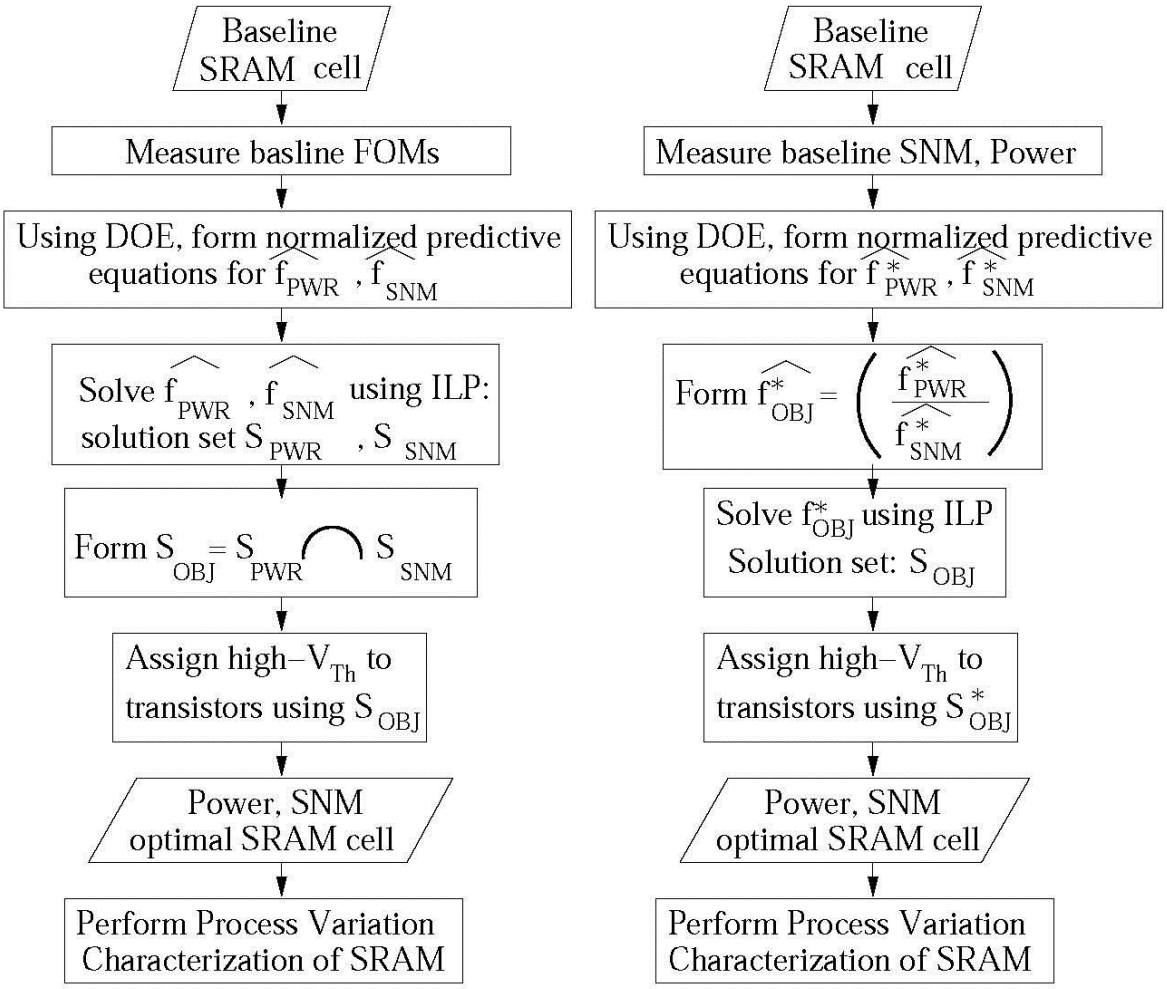

Figure 5.1. A single ended 7-T SRAM cell with transistor sizes shown for a 45nm baseline design.

The combined DOE-ILP approach explores two design flows, shown in Figure 5.2.

Figure 5.2(a) shows the optimal design flow 1. In this flow predictive equations are formulated for two figures of merit, power, and SNM $\widehat{f_{PWR}}$ and $\widehat{f_{SNM}}$. The equations are assumed to be linear and will be based on the experiments performed on $V_{Th}$ states (high or low). Solution variables will be 0 (for nominal $V_{Th}$) and 1 (for high-$V_{Th}$). Solution set for power minimization is $S_{PWR}$ and the solution set for SNM maximization is $S_{SNM}$ . The goal is to minimize $S_{PWR}$ and maximize $S_{SNM}$ (i.e., minimize power and maximize SNM); the overall objective function formed is formulated as $S_{PWR} \cap S_{SNM}$ where $\cap$ is the intersection operator.

Figure 5.2(b) presents the optimal design flow 2. In this flow the predictive equations are formulated for the same figures of merit, that is, power and SNM $\widehat{f_{PWR}}*$ and $\widehat{f_{SNM}}*$. They are normalized based on experiments performed for transistors assigned to either high

or nominal $V_{Th}$ states. Note that, because power and SNM have different units, normalization allows us to formulate a combined objective function, $\widehat{f_{OBJ}}*$, which is formulated by division of $\widehat{f_{PWR}}*$ and $\widehat{f_{SNM}}*$.




(a) Combined DOE-ILP optimal design flow 1. (b) Combined DOE-ILP optimal design flow 2.

Figure 5.2. Combined DOE-ILP optimal design flows for 7%-SRAM.

The optimized solution set is denoted by $S_{OBJ*}$. After achieving the target of optimal solution set $S_{OBJ*}$ or in other words, optimal dual-$V_{Th}$ assignment either from flow 1 or flow 2, the new SRAM configuration is re-simulated for power and SNM and then undergoes process variation in order to study the robustness of the design.

Figure 5.2 provides an overview of the two optimal design flows. In order to study the details of how the technique works, novel algorithms are proposed.

Algorithms 1 and 2 provide two optional optimization algorithms. The two combined DOE-ILP algorithms differ in the way the power and SNM objectives are simultaneously handled.

A 2-Level Taguchi L-8 array is used to run simulations for each of 8 experiments and the values for both power and SNM are recorded. In Algorithm 1, experimental analysis is performed for the transistors of the SRAM using a 2-Level Taguchi L-8 array [62]. The linear predictive equations are formulated using DOE. A full factorial experiment will take $2^7 = 128$ runs, whereas a 2-Level Taguchi L-8 array requires 8 runs. The 2-Level Taguchi L-8 array approach of DOE is a better choice compared to other techniques as it is fast and efficient.

Algorithm 1 for simultaneous power and SNM optimization

1: Input: Baseline $P_{sram}/SNM_{sram}$, Nominal/High-$V_{Th}$ models.

2: Output: Objective set $S_{OBJ} = [f_{PWR}, f_{SNM}]$ with transistors identified for high $V_{Th}$ assignment.

3: Setup experiment for transistors of SRAM cell using 2-Level Taguchi L-8 array, where the factors are the transistors and the responses are average $P_{sram}$ and read $SNM_{sram}$.

4: for Each 1:8 experiments of 2-Level Taguchi L-8 array do

5: Perform simulations and record $P_{sram}$ and $SNM_{sram}$.

6: end for

7: Form predictive equations: $\widehat{f_{PWR}}$ for power, $\widehat{f_{SNM}}$ for SNM.

8: Solve $\widehat{f_{PWR}}$ using ILP. Solution set: $S_{PWR}$.

9: Solve $\widehat{f_{SNM}}$ using ILP. Solution set: $S_{SNM}$ .

10: Form $S_{OBJ} = S_{PWR} \cap S_{SNM}$.

11: Assign high $V_{Th}$ to transistors based on $S_{OBJ}$.

In Algorithm 2, equations for power ($\widehat{f_{PWR}}*$) and SNM ($\widehat{f_{SNM}}*$) are obtained by normalization. Further, the objective function ($\widehat{f_{OBJ}}*$) is formed as the division of ($\widehat{f_{PWR}}*$) and ($\widehat{f_{SNM}}*$). Minimization of this expression will lead to simultaneous power minimization

and SNM maximization. This objective function is further solved to get the solution set $S_{OBJ}$, which is the target.

The factors considered for the DOE experiment taken are the threshold voltages of the 7 transistors of the SRAM cell, and the responses are the average-power consumption ($\widehat{f_{PWR}}$) and SNM ($\widehat{f_{SNM}}$). Each factor can take a high $V_{Th}$ state (1) or a nominal $V_{Th}$ state (0).

Algorithm 2 for simultaneous power and SNM optimization

1: Input: Baseline $P_{sram}$/$SNM_{sram}$, Nominal/High-$V_{Th}$ models.

2: Output: Objective set $S_{OBJ}$* = [$f_{PWR}$*, $f_{SNM}$*] with transistors identified for high $V_{Th}$ assignment.

3: Setup experiment for transistors of SRAM cell using 2-Level Taguchi L-8 array, where the factors are the transistors and the responses are average $P_{sram}$ and read $SNM_{sram}$.

4: for Each 1:8 experiments of 2-Level Taguchi L-8 array do

5: Perform simulations and record $P_{sram}$ and $SNM_{sram}$.

6: end for

7: Form normalized predictive equations: ($\widehat{f_{PWR}}*$) and ($\widehat{f_{SNM}}*$)

8: Form $f_{OBJ}$* = $\left(\frac{\widehat{f_{PWR}*}}{\widehat{f_{SNM}*}}\right)$

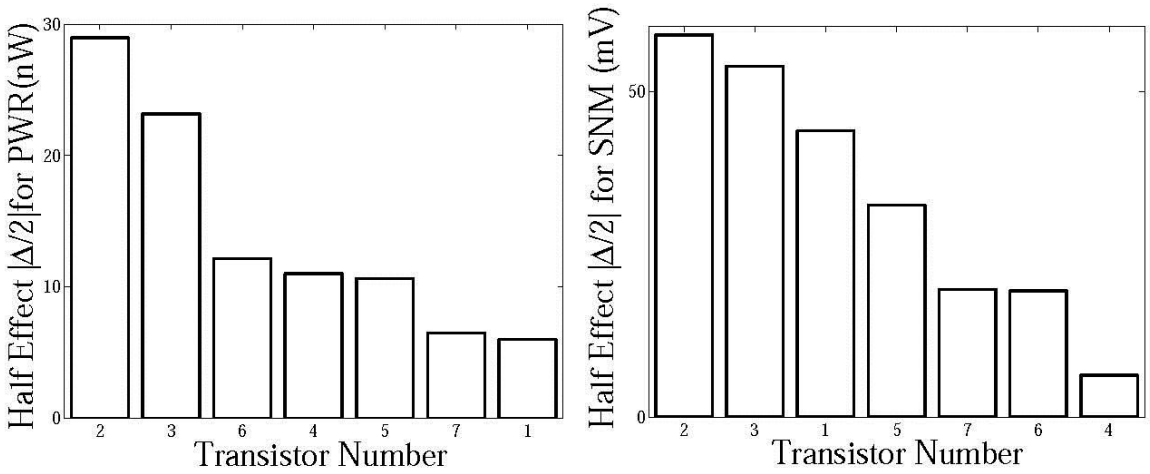
9: Solve $\widehat{fOBJ}*$ using ILP. Solution set: $S_{OBJ}$*.

10: Assign high $V_{Th}$ to transistors based on $S_{OBJ}$*.

Figures 5.3(a) and 5.3(b) show the Pareto plots of the half-effects for the transistors from which predictive equations are then obtained for the response $\hat{f}$ as follows:

$$\hat{f} = \bar{f} + \sum_{n=1}^{7}\left(\frac{\Delta(n)}{2} \times x_n\right) \quad (24)$$

where $\bar{f}$ is the average response, and $x_n$ is the $V_{Th}$ state of transistor $n$.




(a) Power (b) Read SNM

Figure 5.3. Pareto plots for power and SNM of the 7T-SRAM cell.

5.3.1.1 Power minimization: $S_{PWR}$.

The predictive equation for average power consumption is given by the following expressions:

(25) $$\widehat{f_{PWR}}(nW) = 118.2075 - 5.975 \times x_1 - 28.955 \times x_2 - 23.1625 \times x_3 - 10.995 \times x_4 - 10.6375 \times x_5 - 12.1425 \times x_6 + 6.475 \times x_7$$

where $x_1$ represents the $V_{Th}$-state of transistor $i$. The ILP formulation for average power minimization is given by the following expression:

(26) $$\min \widehat{f_{PWR}}$$
$$\text{s.t.} \quad f_{SNM} > T_{SNM} \text{ and } x_1 \; \forall \; I \; 1 \rightarrow 7 \text{ either 0 or 1}$$

where $T_{SNM}$ is a designer defined constraint on SNM. Solving the ILP problem, the optimal solution is obtained as follows:

(27) $$S_{PWR} = \{x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 1,$$

$x_6 = 1,$

$x_7 = 1\}$

This is interpreted as transistors 1, 2, 3, 4, 5, 6 are of high $V_{Th}$, and transistor 7 is of nominal $V_{Th}$. The results for this configuration are shown in Table 5.3.

Table 5.3. Optimization results for different objectives.

| Design alternative | Parameter | Value | Change |
|---|---|---|---|
| Baseline | Power | 203.6 nW | - |
| | SNM | 107.0 mV | - |
| $S_{PWR}$ | Power | 26.32 nW | 87.1% decrease |
| | SNM | 231.9 mV | 26.7% increase |
| $S_{SNM}$ | Power | 113.6 nW | 44.2% decrease |
| | SNM | 303.3 mV | 43.9% increase |
| $S_{OBJ}$ | Power | 113.6 nW | 44.2% decrease |
| Approach 1 | SNM | 303.3 mV | 43.9% increase |
| $S_{OBJ}$* | Power | 100.5 nW | 50.6% decrease |
| Approach 2 | SNM | 303.3 mV | 43.9% increase |

5.3.1.2. SNM maximization: SSNM.

The predictive equation for the read SNM is given by the following expression:

$$\widehat{f_{SNM}}\ (mV) = 156.675 - 44.025 \times x_1 + 58.725 \times x_2 - 53.925 \times x_3 - 6.425 \times x_4 + 32.575 \times x_5 + 19.375 \times x_6 - 19.625 \times x_7 \quad (28)$$

The ILP formulation for SNM maximization is given by the following expression:

$$\max \ \widehat{f_{SNM}} \quad (29)$$

$$\text{s.t.} \quad f_{PWR} < \tau_{PWR} \text{ and } \widehat{x_1 \ \forall \ i}\ 1 \rightarrow 7 \text{ either 0 or 1}$$

where $\tau_{PWR}$ is the designer-defined power constraint.

ILP yields the optimal solution as follows:

$$S_{SNM} = \{x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 0, x_5 = 1, x_6 = 1, x_7 = 0\}. \quad (30)$$

The results for the corresponding SRAM configuration are shown in Table 5.3.

5.3.1.3 Combined power minimization and SNM maximization

- Using Approach 1: The objective set $S_{OBJ}$ for simultaneous optimization of power and *SNM* is formed as the following expression:

$$S_{OBJ} = S_{PWR} \cap S_{SNM} \quad (31)$$

where ∩ is the intersection of the two solution sets $S_{PWR}$ and $S_{SNM}$ . This equation is derived from the set theoretical domain where logical AND translates to set intersection. Hence, in order to obtain low power and high SNM for the SRAM cell, we use the set intersection operator to achieve $S_{OBJ}$. The constraints are the same as the above ILP formulations. The ILP optimization results in the following solution:

$$S_{OBJ} = \{x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 0, x_5 = 1, x_6 = 1, x_7 = 0\}. \quad (32)$$

This leads to the configuration of Figure 5.4(a) and results in Table 5.3.

• Using Approach 2: In this approach, normalized forms of $\widehat{f_{PWR}}$ and $\widehat{f_{SNM}}$, denoted as $\widehat{f_{PWR}}*$ and $\widehat{f_{SNM}}*$, are used. The normalization is performed by division of each data by the maximum value of the corresponding data set. Normalization enables directly accommodating different units, while forming the objective functions as follows:

$$f_{PWR}* = 0.58 - \widehat{0.03} \times x_1 - 0.14 \times x_2 - 0.11 \times x_3 - 0.05 \times x_4 - 0.05 \times x_5 - 0.06 \times x_6 + 0.03 \times x_7 \quad (33)$$

$$f_{SNM}* = 0.52 - \widehat{0.15} \times x_1 + 0.19 \times x_2 - 0.18 \times x_3 - 0.02 \times x_4 + 0.11 \times x_5 + 0.06 \times x_6 - 0.06 \times x_7 \quad (34)$$

The combined objective function is formed as follows:

$$\widehat{f_{OBJ}}* = \left(\frac{\widehat{f_{PWR}*}}{\widehat{f_{SNM}*}}\right) = 0.18 \times x_3 - 0.02 \times x_4 + 0.11 \times x_5 + 0.06 \times x_6 - 0.06 \times x_7 \quad (35)$$

Equation 35 is obtained from the division of normalized Equation 33 and normalized Equation 34. Through normalization, we eliminate the problem of different units for power and SNM and hence we get the ratio as Equation 35. The ILP formulation for this combined method is obtained as:

$$\min \ \widehat{f_{OBJ}}* \quad (36)$$

$$\text{s.t. } f_{PWR} < \tau_{PWR}, f_{SNM} > \tau_{SNM,} \, x_1 \, \forall \ i \in 0 \text{ or } 1$$

From this optimization problem, the optimal solution is obtained as follows:

$$S_{OBJ} = \{x_1 = 0, x_2 = 1, x_3 = 0,$$

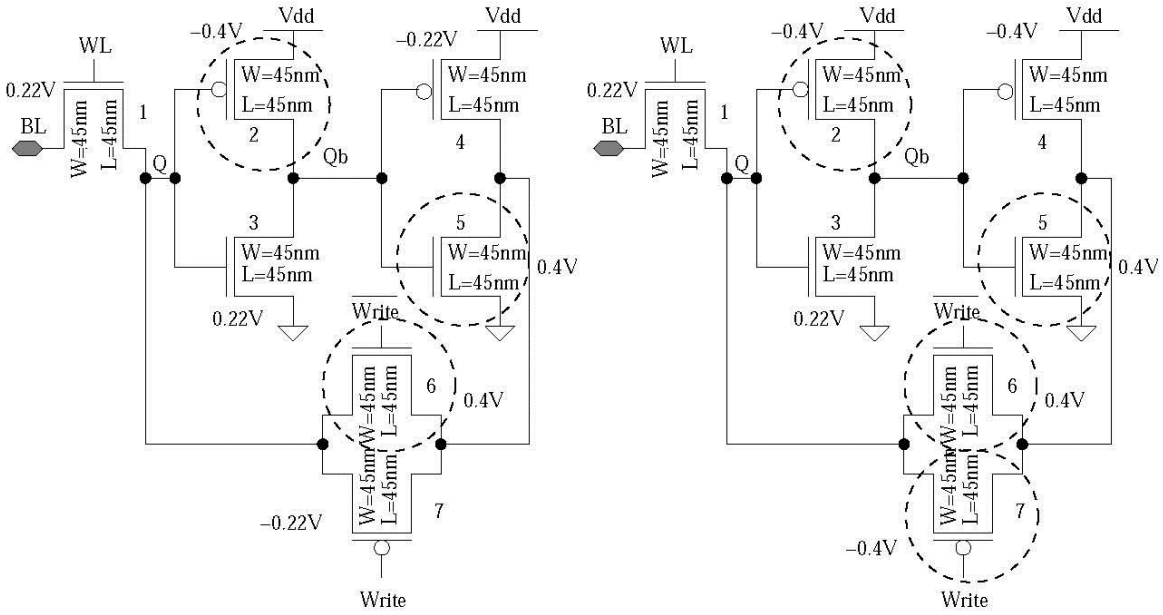
(37) $$x_4 = 0,$$

$$x_5 = 1,$$

$$x_6 = 1,$$

$$x_7 = 1\}.$$

The corresponding SRAM configuration is shown in Figure 5.4(b) and results in Table 5.3.




(a) Configuration for flow-1 (b) Configuration for flow-2

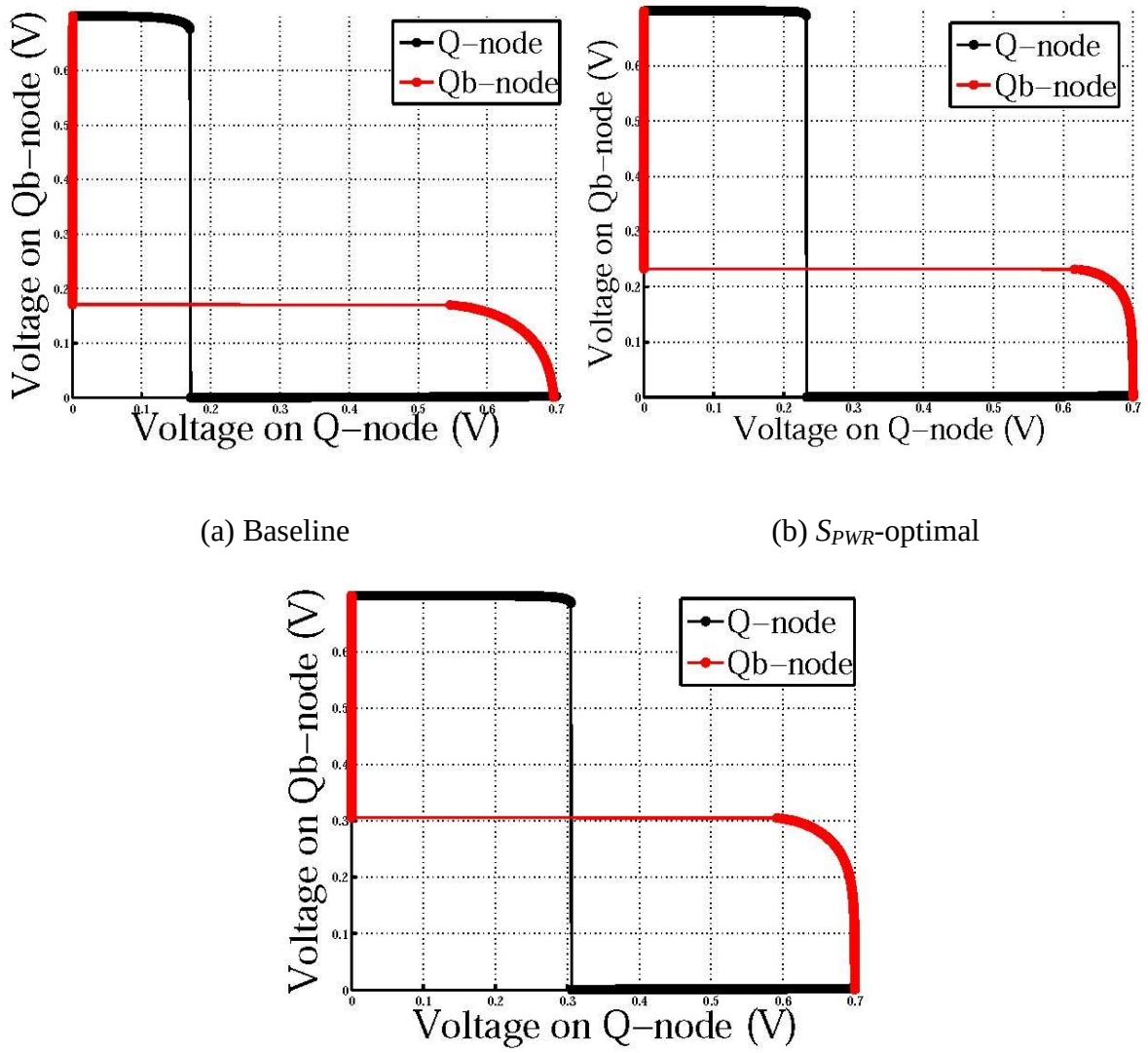
Figure 5.4. Alternative dual-$V_{Th}$ configurations of the SRAM cell with high $V_{th}$ transistors circled; high $V_{Th}$ = 0.4 V and nominal $V_{Th}$ = 0.22 V.

Similarly, Figure 5.5 presents three alternative butterfly curves for the baseline (5[a]), optimized power configuration (5[b]) and finally, objective function, i.e., $S_{obj}$ (5[c]) designs.

To study the power and SNM of the optimal SRAM, simulations are performed for various $V_{DD}$ values and the results are shown in Figure 5.6. It is observed that both power and SNM increase with $V_{DD}$. For $V_{DD}$ = 0.7 V, power has been reduced by 44.2% and SNM has increased by 43.9% compared to the baseline design using Approach 1, and power has been reduced by 50.6% and SNM has increased by 43.9% using Approach 2.

As per the proposed design flow, an 8 x 8 array was constructed using the optimized SRAM cells, shown in Figure 5.7. The average power consumption of the array is 4.5 μW.

Intuitively and from the exhaustive experimental observations, it is concluded that Approach 2 is better in order to achieve our objective, that is, reduced power and high SNM.




(a) Baseline (b) $S_{PWR}$-optimal

(c) $S_{OBJ}$-optimal

Figure 5.5. Butterfly curve for three SRAM alternatives to measure their SNM.

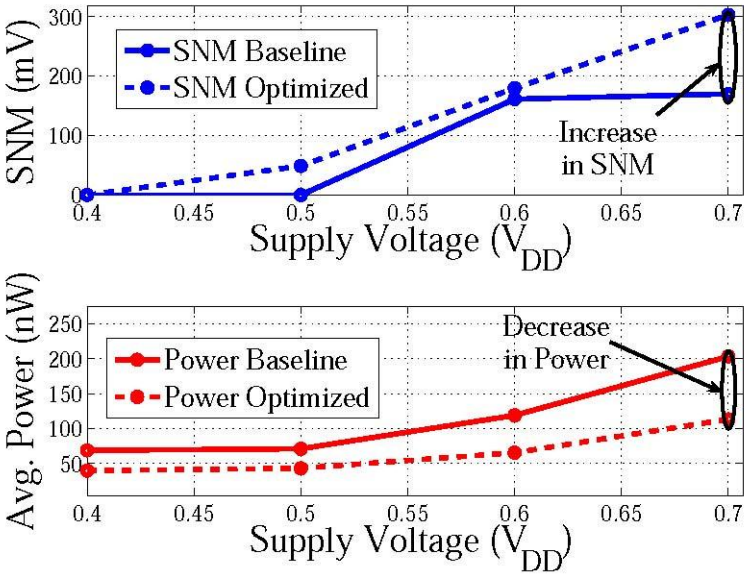



Figure 5.6. Power and SNM comparison of optimal and baseline 7T SRAM.

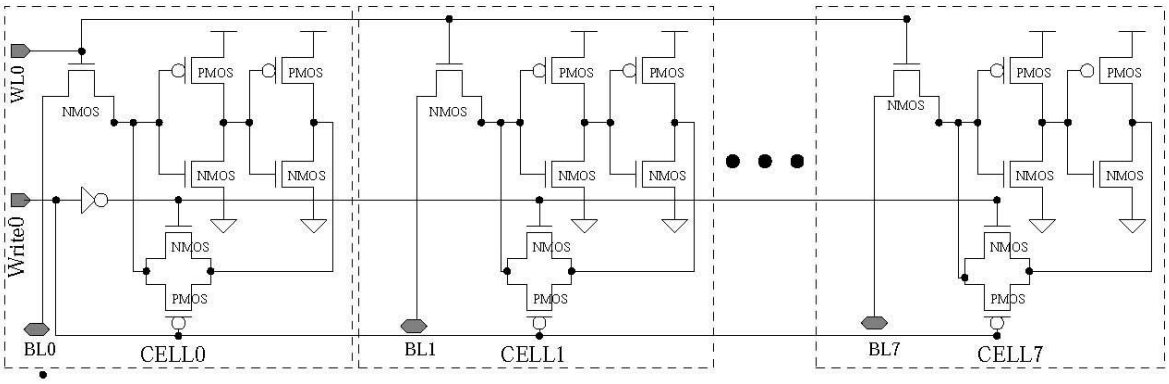
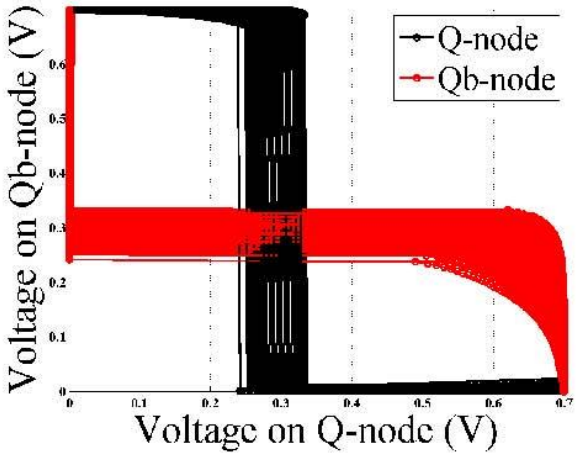


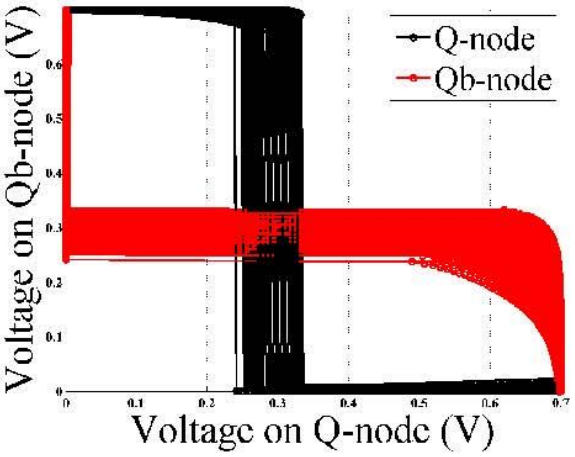

Figure 5.7. One row of the 8 x 8 array constructed using optimal cells.

5.3.1.4. Statistical variability analysis or the SRAM

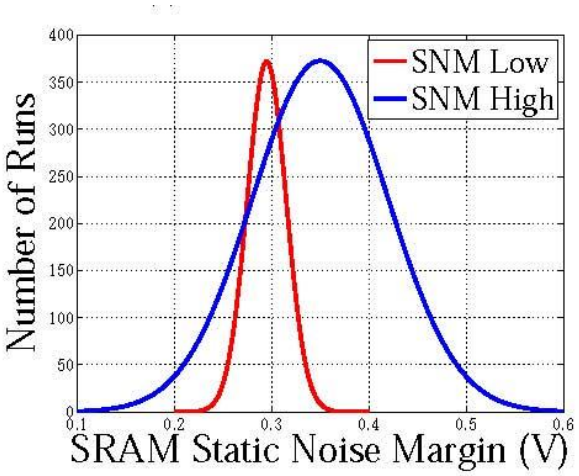
The SNM is studied through 1,000 Monte Carlo simulations to ensure there is no process variation induced failure in the SRAM. A correlation coefficient of 0.9 between $T_{oxn}$ and $T_{oxp}$ is assumed. Monte Carlo simulation is an efficient approach as it does not require relating the output to input which otherwise would have been cumbersome for the large number of parameters involved [33]. The process parameters is said to have a Gaussian distribution with mean (μ) which is taken as the nominal values specified in the PTM [87] and standard deviation (σ) as 5% of the mean. Results are shown in Figure 5.8 and Table 5.4.



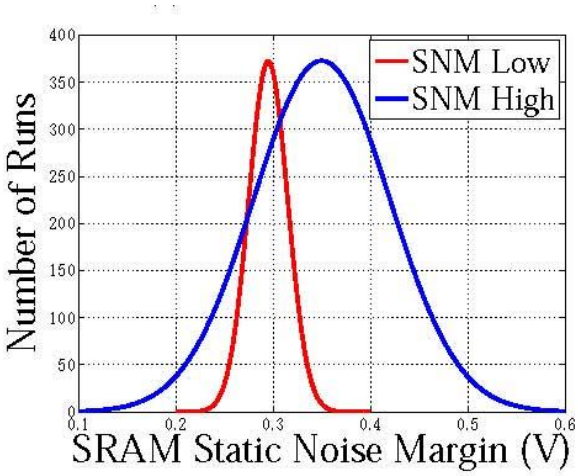

(a) Flow-I butterfly curve



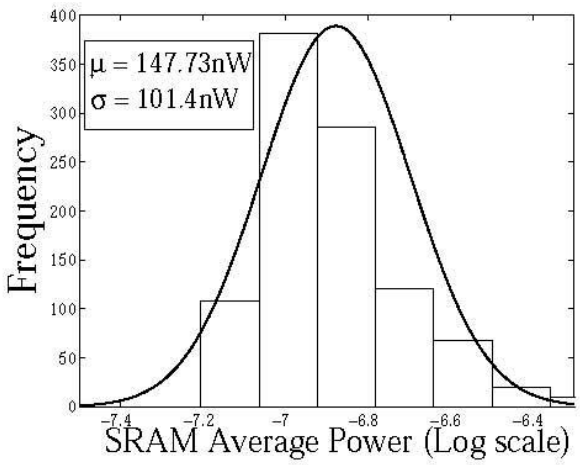

(b) Flow-2 butterfly curve



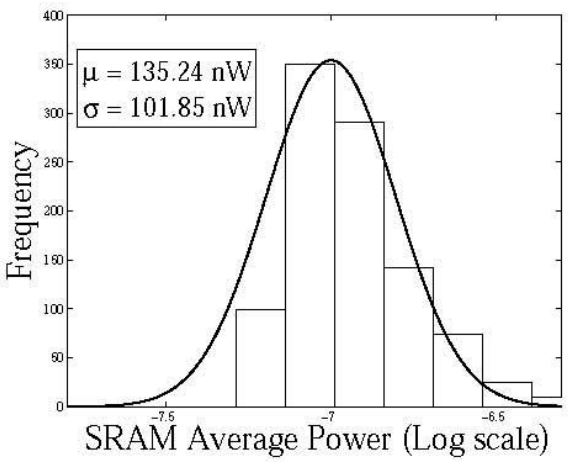

(c) Flow-I SNM distribution




(d) Flow-2 SNM distribution




(e) Flow-I power distribution




(f) Flow-2 power distribution

Figure 5.8. Process variation study of the 7T-SRAM.

Table 5.4. Statistical process variation effects on SRAM power and SNM.

| Optimization | Parameter | μ | σ |
|---|---|---|---|
| $S_{PWR}$ | Power | 28.91 nW | 8.26 nW |
| | SNM | 180 mV | 30 mV |
| $S_{SNM}$ | Power | 147.73 nW | 101.4 nW |
| | SNM | 295 mV | 28 mV |
| $S_{OBJ}$: Approach 1 | Power | 147.73 nW | 101.4 nW |
| | SNM | 295 mV | 28 mV |
| $S_{OBJ}$ *: Approach 2 | Power | 135.24 nW | 101.85 nW |
| | SNM | 295 mV | 28 mV |

5.3.2. Design or Experiments (DOE) assisted conjugate gradient approach

This is an alternate optimization methodology for power minimization, performance maximization, and process variation tolerant SRAM cell. A 32 nm high-κ/metal gate SRAM has been used as example circuit. During this methodology, the baseline 10T-SRAM cell (shown in Figure 5.10) is taken as a sample circuit from Chapter 3. This circuit is subjected to power minimization using dual-$V_{Th}$ assignment (discussed in Section 5.2) along with a novel combined design of experiments-integer linear programming (DOE-ILP) approach. However, this leads to a 15% reduction in the SNM of the SRAM cell. It is further improved by using a conjugate gradient optimization, while maintaining the minimum power consumption. The final SRAM design shows 86% reduction in power (including leakage) consumption and 8% increase in the SNM compared to the baseline design. The variability analysis of the optimized cell is performed considering the effect of 12 parameters. An 8 x 8 array is constructed to demonstrate the feasibility of the proposed SRAM cell. The proposed design flow is shown in Figure 5.9.

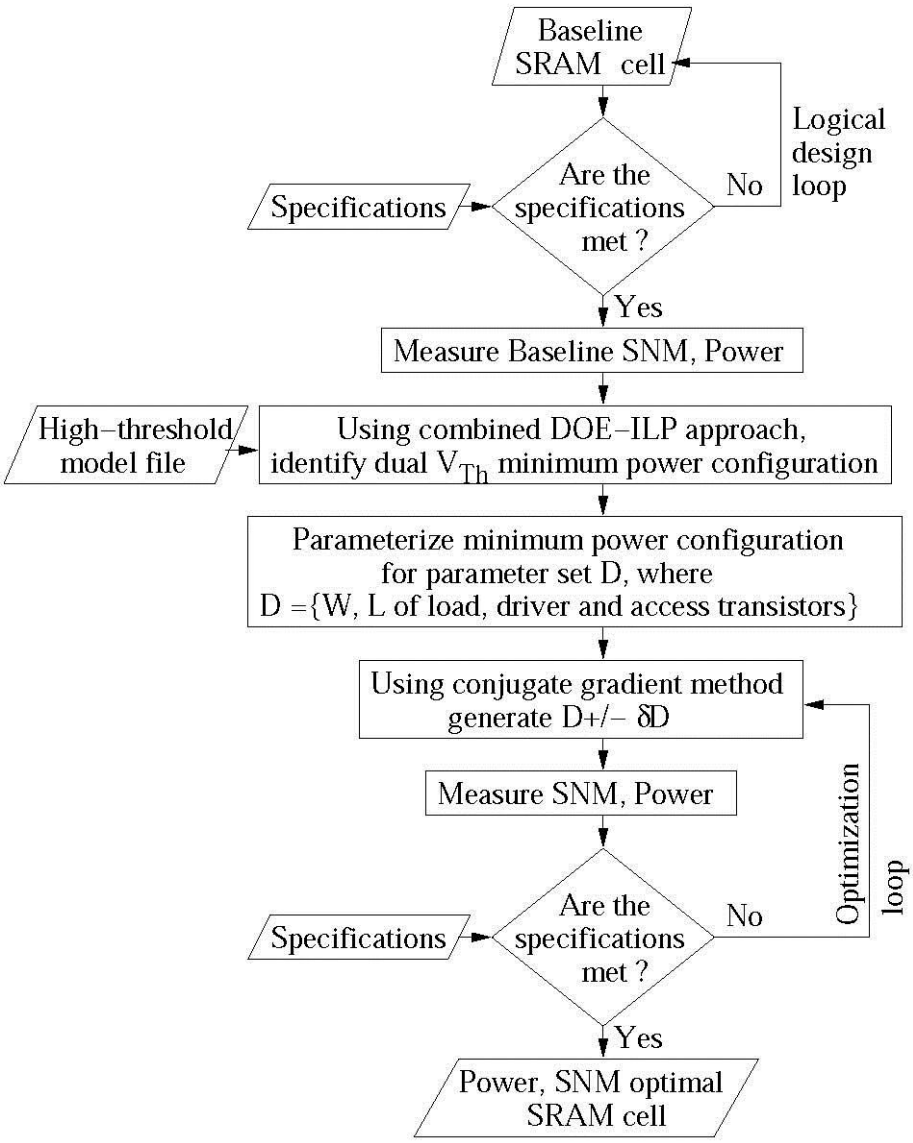



Figure 5.9. DOE-ILP assisted conjugate gradient design flow.

The input to the flow is a baseline SRAM cell. This baseline design is required to meet specifications with minimum sized transistors. The SRAM design optimization specifications are as follows:

- Minimum power consumption with all sources of leakage,
- when SNM ≥ a designer defined value.

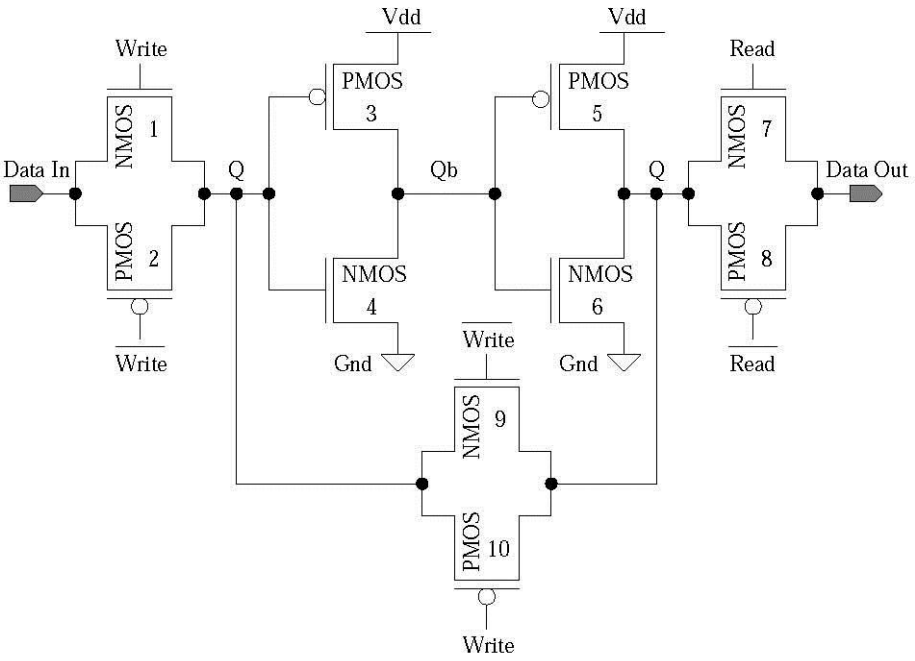



Figure 5.10. Schematic representation of baseline 10 transistor static random access memory (10T SRAM) cell with transistors numbered.

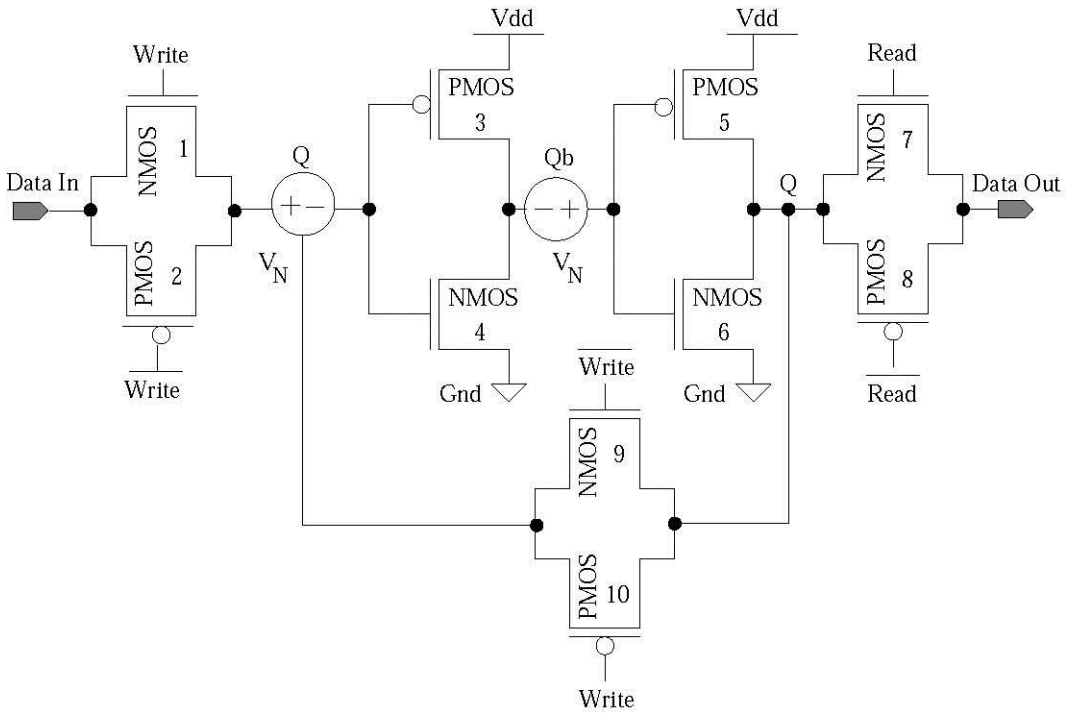



Figure 5.11. Simulation set-up for SNM measurement.

This 10T-SRAM design (Figure 5.10) eliminates the use of sense amplifier and pre-charging circuitry for pre-charging of the bitlines prior to read and write operations. The supply voltage $V_{DD}$ = 0.7 V. Figure 5.11 shows how the SNM is calculated through

simulating the two inverters through two voltage sources. The SNM setup is similar to that of the 7T-SRAM explained in Chapter 3.

The following results are obtained for the baseline design:

Table 5.5. Baseline Results

| Parameter | Value |
|---|---|
| Average Power | 2.27 μW |
| SNM | 271 mV (Figure 5.14[a]) |

5.3.2.1. High-κ/metal-gate CMOS compact model

For the design presented in this work, a 32 nm high-κ/metal-gate CMOS PTM [87] is used already discussed in Chapter 1. The simulation results obtained are highly accurate and the calculated data are of comparable accuracy to technology computer-aided design (TCAD) simulations which are typically time and computation intensive. The PTM is based on BSIM 4/5, hence two methods are used as follows:

(i) The model parameter in the model file that denotes relative permittivity (EPSROX) is changed, and

(ii) The equivalent oxide thickness (EOT) for the dielectric under consideration is calculated.

Using these steps, the EOT is calculated so as to keep the ratio of relative permittivity over dielectric thickness constant. The EOT is calculated as using the following expression:

(38) $$T_{ox}^{*} = \left(\frac{\kappa_{SiO_2}}{\kappa_{gate}}\right) \times T_{gate}$$

where $\kappa_{gate}$ is the relative permittivity and $T_{gate}$ is the thickness of the gate dielectric material other than $SiO_2$, while $\kappa_{Si0_2}$ is the dielectric constant of $SiO_2$(= 3.9). We have taken $\kappa_{gate}$ = 21 to emulate a $HfO_2$-based dielectric. The EOT is calculated to be 0.9 nm.

5.3.2.2. Combined DOE-ILP approach for minimum power and leakage configuration in 10T SRAM

The baseline 10T SRAM cell is subjected to a design of experiments [21, 37] based approach using a 2-level Taguchi L-12 array. The factors considered here are the 10 transistors of the SRAM cell (Figure 5.10), and the response under consideration is the average power consumption of the cell ($f_{P_{SRAM}}$). Each factor can take a high $V_{Th}$ state (+1) or a nominal $V_{Th}$ state (-1). After running the experiments, the half-effects are recorded using the following expression:

(39) $$\left(\frac{\Delta(n)}{2}\right) = \left(\frac{\text{avg}(+1) - \text{avg}(-1)}{2}\right)$$

where $\left[\frac{\Delta(n)}{2}\right]$ is the half-effect of the *n*-th transistor, avg(+1) is the average value of power when transistor *n* is in high-$V_{Th}$ state, and avg(-1) is the average value of power when transistor *n* is in nominal $V_{Th}$ state. Figure 5.12 shows a Pareto plot constructed using the half-effects obtained.

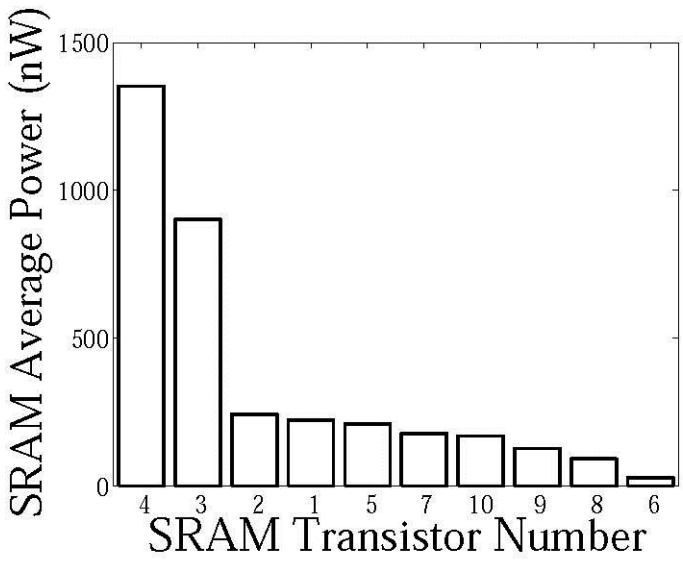



Figure 5.12. Pareto plot for 10T SRAM power.

From this data, we form a predictive equation of the form:

(40) $$\widehat{f_{P_{SRAM}}} = \overline{f_{P_{SRAM}}} + \sum_{n=1}^{10}\left(\frac{\Delta(n)}{2}\right) \times x_n$$

where $\widehat{f_{P_{SRAM}}}$ is the response, $\overline{f_{P_{SRAM}}}$ is the average, $\left[\frac{\Delta(n)}{2}\right]$ is the transistor, and $x_n$ is the $V_{Th}$

state of the *n*-th transistor. Equation 41 shows the predictive equation obtained.

$$\begin{aligned}\widehat{f_{P_{SRAM}}}(nW) = {} & 2192.4 + 223.9 \times x_1 + 243.7 \times x_2 \\ & + 902.8 \times x_3 - 1352.5 \times x_4 + 211.9 \times x_5 \\ & - 29.2 \times x_6 - 179.1 \times x_7 + 92.6 \times x_8 \\ & - 128.2\, x_9 - 170.72 \times x_{10}\end{aligned} \tag{41}$$

$x_1$ represents the $V_{Th}$ state of transistor *i* Figure 5.10. From this, the ILP problem is formulated as follows:

$$\begin{aligned}&\min \widehat{f_{P_{SRAM}}} \\ &\text{s.t.} \quad 1 \le x_1 \le +1\end{aligned} \tag{42}$$

The constraints '+1' and '-1' represent coded values for high $V_{Th}$ and nominal $V_{Th}$ states, respectively. The predictive equations are formed for power ($f_{PWR}$) and read SNM ($f_{RSNM}$) based on the experiments performed on the $V_{Th}$ state (high or nominal) of the transistors in the cell. Predictive equations and constraints are considered to be linear during these experiments. In solving the ILP problem, we get the optimal solution as the following:

$$\begin{aligned}P_{SRAM} = \{&x_1 = 0, \\ &x_2 = 0, \\ &x_3 = 0, \\ &x_4 = 1, \\ &x_5 = 0, \\ &x_6 = 1, \\ &x_7 = 1\}.\end{aligned} \tag{43}$$

This can be interpreted as transistors 4, 6, 7, 9, 10 are high $V_{Th}$ transistors, and transistors 1, 2, 3, 5, 8 are nominal $V_{Th}$ transistors. Figure 5.13 shows the SRAM cell with the high $V_{Th}$ transistors circled.

The following results are obtained from the minimum power configuration:

Table 5.6. Minimum power configuration results

| Parameter | Value |
|---|---|
| Average Power | 314.5 nW |
| SNM | 230.4 mV |

It shows 86.15% power reduction over the baseline design. However, it also results in 15% degradation in SNM, shown in Figure 5.14(b).

5.3.2.3. Conjugate gradient-based SNM maximization.

The algorithm is shown as Algorithm 3. Table 5.7 shows the final values of the parameter set for SNM optimal SRAM. The optimization algorithm converged in 9 iterations with each iteration lasting approximately 4 minutes. The results obtained after the optimization are given in Table 5.8.

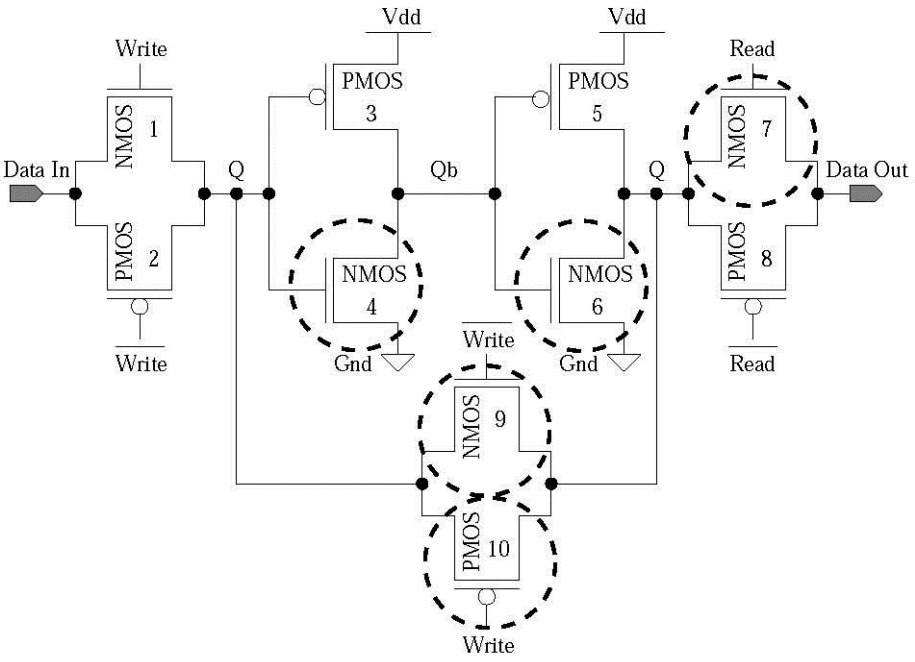



Figure 5.13. Minimum power configuration SRAM cell. The circled transistors are the high $V_{Th}$ transistors. The rest are nominal $V_{Th}$ transistors.

Table 5.7. Optimized values of the parameter set

| $D$ | $C_{low}$ | $C_{up}$ | $D_{opt}$ |
|---|---|---|---|
| $W_{pl}$ | 64 nm | 1.28 μm | 1.18 μm |
| $L_{pl}$ | 64 nm | 1.28 μm | 1.28 μm |
| $W_{nd}$ | 64 nm | 1.28 μm | 1.28 μm |
| $L_{nd}$ | 64 nm | 1.28 μm | 32.28 nm |
| $W_{pa}$ | 64 nm | 1.28 μm | 1.28 μm |
| $L_{pa}$ | 64 nm | 1.28 μm | 74.8 nm |
| $W_{na}$ | 64 nm | 1.28 μm | 1.28 μm |
| $L_{na}$ | 64 nm | 1.28 μm | 32 nm |

This approach achieves 86.15% power reduction over the baseline design and 8% improvement in SNM (Figure 14[c]), shown in Figure 5.15. Also, the read access time for the optimized 10T-SRAM cell is 7 ns.

Algorithm 3 for SNM optimization using conjugate gradient approach for 10TSRAM

1: Input: Minimum power configuration SRAM, Baseline model file, High-threshold model file, Objective Set $F = [SNM\ P_{SRAM}]$, Stopping criteria $S$, parameter set $D = [W_{pl}, L_{pl}, W_{nd}, L_{nd}, W_{pa}, L_{pa}, W_{na}, L_{na}]$, Lower parameter constraint $C_{low}$, Upper parameter constraint $C_{up}$.

2: Output: Optimized objective set $F_{opt}$, Optimal parameter set $D_{opt}$ for $S \leq \pm\beta$. {where $1\% \leq \beta \leq 5\%$}

3: Run initial simulation with initial guess of $D$.

4: while ($C_{low} < D < C_{up}$) do

5: Use conjugate gradient method to generate new set of parameters $D' = D \pm \delta D$.

6: Compute $F = [SNM, P_{SRAM}]$.

7: if ($S \leq \pm\beta$) then

8: return $D_{opt} = D'$

9: end if

10: end while

11: Using $D_{opt}$ , simulate SRAM cell.

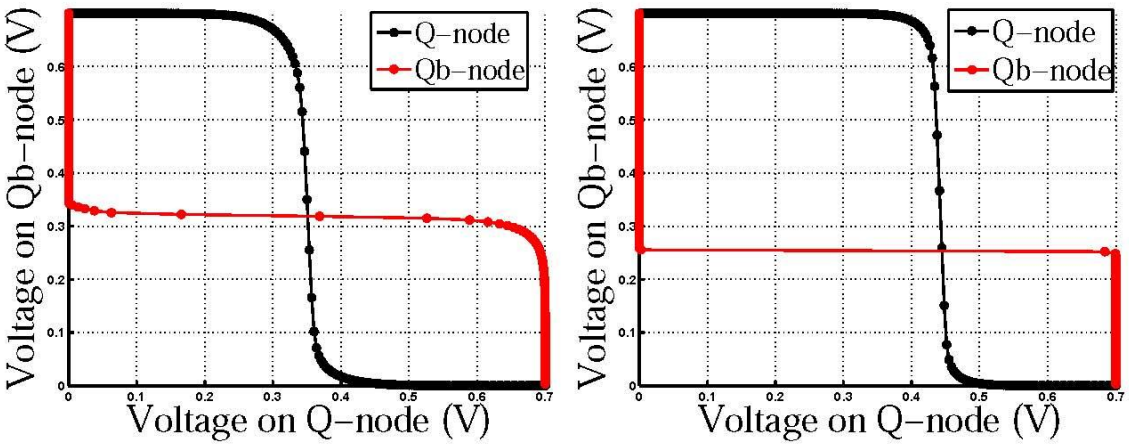
12: Record $F_{opt}$ .

---

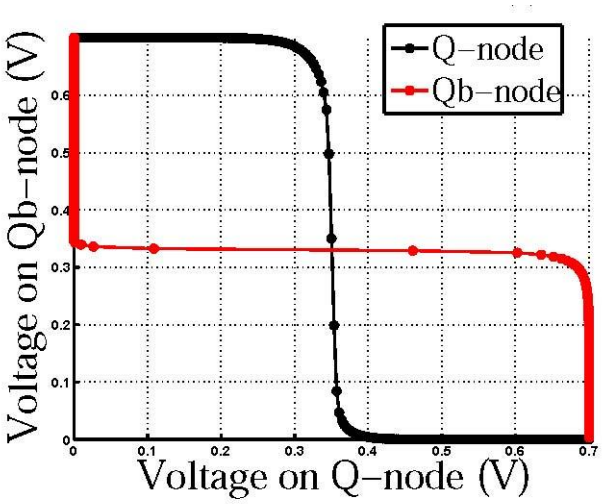
Table 5.8. Optimization results

| Parameter | Value |
|---|---|
| Average Power | 314.5 nW |
| SNM | 295 mV |

As per the design flow, we then construct an 8 x 8 array using the optimized cell, shown in Figure 5.16. The average power consumption of the array is 1.2 µW. The worst case read access time for the array is found to be 52.4 ns.




(a) Baseline (b) Power optimal




(c) Power SNM optimal

Figure 5.14. Butterfly curve for three SRAM alternative designs.

5.3.2.4. Process variation analysis or 10T-SRAM cell. The device parameters are exhaustively evaluated for stability through 1,000 Monte Carlo simulations to ensure there is no process variation induced failure in the SNM. Twelve process parameters, which are dependent on the threshold voltage variation, are considered for variability:

- NMOS/PMOS channel length ($T_{oxn}$, $T_{oxp}$),
- NMOS/PMOS channel length ($T_{oxn}$, $T_{oxp}$),
- access-transistor length and width ($L_{na}$, $L_{pa}$, $W_{na}$, $W_{pa}$),

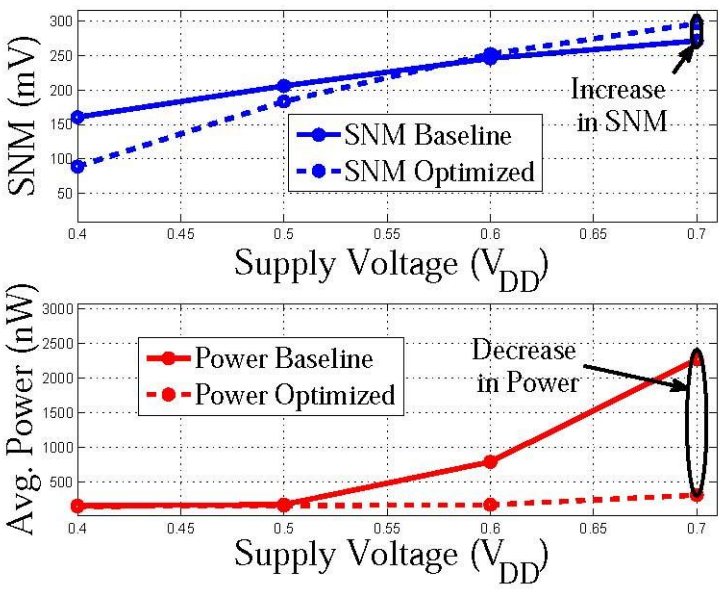



Figure 5.15. SNM, power (including leakage) comparison of optimized, baseline 10T-SRAM.

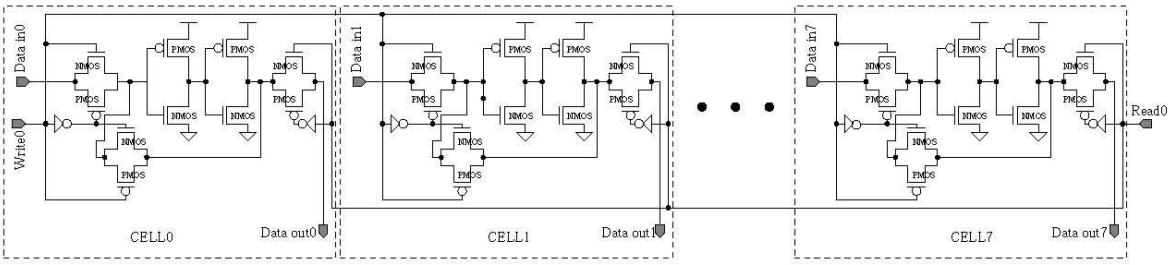


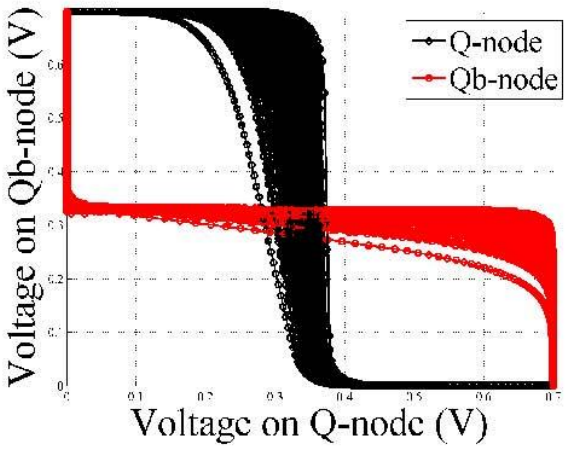

Figure 5.16. Schematic showing one row of the 8 x 8 array constructed using proposed 10T SRAM cells.

- Driver-transistor length and width ($L_{nd}$, $W_{nd}$ ),
- Load-transistor length and width ($L_{pl}$, $W_{pl}$ ).

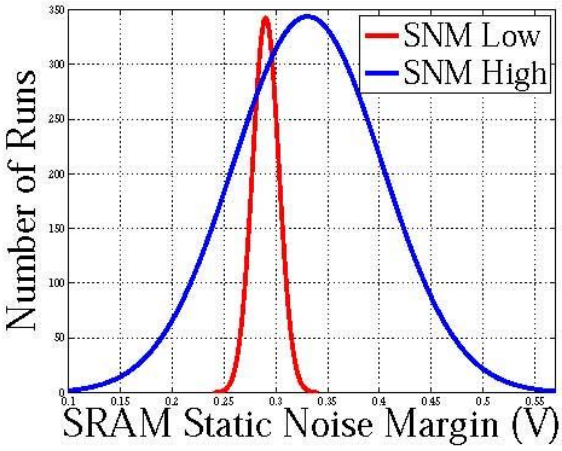
Each of these process parameters is assumed to have a Gaussian distribution with mean (μ) taken as the nominal values specified in the PTM [87] and standard deviation (σ) as 10% of the mean. Figure 17(a) shows the effect of process variations on the butterfly curve of SRAM. Figure 17(b) shows the distributions for "SNM High" and "SNM Low" extracted from the Monte Carlo simulations, where "SNM High" is the higher SNM and "SNM Low" is the lower SNM due to asymmetry in the cell, for each Monte Carlo run. However, "SNM Low" is treated as the actual SNM. Table 5.6 shows the corresponding statistical data. Figure 17(c) shows the distribution of average power of the SRAM cell. The average power distribution is observed to be lognormal in nature.

Table 5.9. Statistical Data for SNM

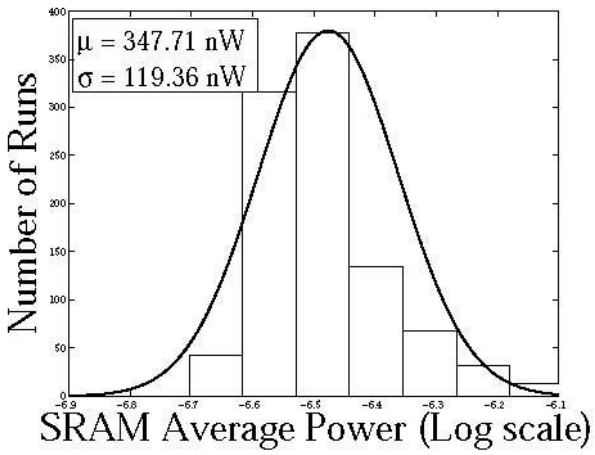
| SNM | μ ($mV$) | σ ($mV$) |
|---|---|---|
| SNM High | 330.7 | 71.9 |
| SNM Low | 290.3 | 12.7 |




(a) Butterfly curve




(b) SNM distribution




(c) Power distribution

Figure 5.17. Butterfly curve, SNM distribution and power distribution for the optimal SRAM under process variation.

# CHAPTER 6

# PROCESS AND SUPPLY VOLTAGE VARIATION AWARE OPTIMIZATION OF SRAMS

This chapter covers the process-voltage variation aspects of static random access memory (SRAM) design. A novel optimization technique (statistical P3 optimal technique) is provided using the 7T-SRAM as a case study. The results demonstrate that the optimized SRAM achieves minimum power, better it is a high performance and process variation tolerant design.

## 6.1. Process (*P*) and Voltage (*V*) Variation Characterization in SRAM

The variations can be process variations, voltage variations, or temperature variation. Each one of these has a different origin and consequently affects SRAM perform differently. They need to be characterized in different ways for accurate accounting during the design space exploration by the designers. Process variation is one of the most important and serious problems present in nano-CMOS technology. Nanoscale complementary metal-oxide semiconductor (CMOS) SRAM memory design faces several challenges like reducing noise margins along with fluctuation of device characteristics by process variation, because of the continued technology scaling. Random variations cause a high mismatch in neighboring devices and therefore are largely responsible for the poor yield of SRAM arrays in nanoscale CMOS technologies [4]. Specifically for SRAMs, it is observed that as the supply voltage is reduced, the sensitivity of the circuit parameters to the process variation increases [38]. Every process inherits variation capabilities and it becomes the responsibility of process engineers to control and at the same time minimize the sources of external variation. Variability in process and design parameters has increased and affects design decisions, yield, and circuit performance. In spite of all these preventive measures, every process has generic or intrinsic variations, as in nano-CMOS processes. Variation in processes will eventually translate to variability in power, performance, and reliability of the entire chip.

Process variations are of two types: interdie and intradie [61]. Interdie variation, which is present in lot-to-lot, wafer-to-wafer, and within-wafer, affects all devices on a single chip. On the other hand, intradie variation is due to device characteristics such as device geometry change, dopant density change, threshold voltage, and gate oxide thickness, which vary from device to device within the same chip. Some of the variations are random and some are systematic. In short, interdie variation in a parameter alters the value of that parameter for all the transistors in a die in the same direction whereas the intradie variations move the process parameters of different transistors in a die in different directions. Digital circuits are optimized only for speed and power while analog circuits are designed to meet a variety of performance metrics.

Variation has a tremendous impact on performance metrics of various circuits including digital, analog/mixed signals and RF circuits, which results in design yield loss. In an SRAM cell, random intradie variations in process parameters induce mismatch between strengths of different transistors, which results in functional failure and yield loss [66]. As process technology is scaling in nano-CMOS circuits, sufficient static noise margin (SNM) becomes difficult to maintain because of increased variability. The voltage transfer characteristics (VTCs) of two halves of the circuit have mismatch present in them, and are enhanced due to dopant fluctuations. Random intradie variations present in the process parameters cause mismatch between transistors in the circuit, which in turn results in functional failures, most commonly read failures [4], [80]. The threshold voltage variation caused due to random dopant fluctuations is the most significant originator among the various sources of random variations [4].

Cache design must be compatible and reliable with subthreshold combinational logic operating at ultra-low voltages. However, this increases the sensitivity of design and process parameter variability. This problem further enhances in nanometer regime with ultra-voltage operation, making SRAM circuit design and performance more challenging. The variations in

threshold voltage ($V_{Th}$) of SRAM cell transistors due to random dopant fluctuations are the main reason for parametric failures. The threshold voltage variation is related to the device geometry (length, width and oxide thickness) and doping profile. Equation 44 shows how the threshold voltage standard deviation ($\sigma_{V_{Th}}$) varies with gate oxide thickness ($T_{ox}$), the channel dopant concentration ($N_{ch}$), and the channel length ($L$) and width ($W$) [79]:

$$\sigma_{V_{Th}} = \left( \frac{\sqrt[4]{4q^3 \in_{S_1} \varphi_B}}{2} \right) \left( \frac{T_{ox}}{\epsilon_{ox}} \right) \left( \frac{\sqrt[4]{N_{ch}}}{\sqrt{WL}} \right) \quad (44)$$

where $\varphi_B = 2\ \kappa_B T\ \mathrm{In}(N_{ch}/n_i)$ with the following notations:

- $\kappa_B$ - Boltzmann's constant,
- $T$ - the absolute temperature,
- $n_i$ - the intrinsic carrier concentration,
- $q$ - the elementary charge,
- $\epsilon_{ox}$ - permittivity of oxide, and
- $\epsilon_{Si}$ - permittivity of oxide silicon.

The above expression is consistent with observation that $\sigma_{V_{Th}}$ is inversely proportional to the square root of the device area.

6.2. Statistical DOE-ILP Approach for Nano-CMOS SRAM Optimization

The statistical design-of-experiments-integer linear programming (DOE-ILP) approach aims to present simultaneous P3 (power, performance, and process variation tolerance) optimization of a nano-CMOS 7T-SRAM circuit using the 45 nm predictive technology model. The proposed design flow for P3 optimization of 7T-SRAM cell is presented in Figure 6.2. P3 optimization defines minimized power consumption, maximized performance (SNM), and process variation tolerant SRAM cell. The sample 7T-SRAM cell is subjected to the dual-$V_{Th}$ assignment (discussed in Chapter 5). This optimization technique results in 44.2% reduction in total power (including leakage) and 43.9% increase in performance (SNM), compared to the baseline design. The process variation analysis of the

optimized cell is performed considering the variability effect in 12 device parameters. An 8 x 8 array is constructed to show the feasibility of the optimized 7T-SRAM cell.

6.2.1. Proposed Flow for P3-Optimal SRAM Design

The process variation aware statistical optimization approach needs to be handled as a multicost objective function. As seen from Figure 6.1, the theory is that $\mu_{baseline}$ of the quantity (Power or SNM), which is under consideration, will be shifted toward the left or right wherever it needs to be minimized ($\mu_{minimized}$ ) or maximized ($\mu_{maximized}$). Furthermore, the $\sigma_{baseline}$ ($\sigma$ is the spread) also needs to be minimized ($\sigma_{baseline}$).

Figure 6.2 displays the optimal design flow. The optimization process involves the following steps. For each experiment, *N* number of Monte Carlo simulations are performed and the mean ($\mu$) and standard deviation ($\sigma$) are recorded (Gaussian distribution values) for average power and average SNM.

The mean and standard deviation of power will be: $\mu_{PWR}$, $\sigma_{PWR}$ and for SNM is: $\mu_{SNM}$ , $\sigma_{SNM}$ . They are organized as predictive equations which are assumed to be linear in nature. Integer linear programming (ILP) is used to solve (to minimize or maximize the quantity) these predictive equations. As a result, solution sets are obtained in the form of mean and standard deviation or power of 7T-SRAM cell and they are denoted by $S_{\mu PW}$ and $S_{\sigma PWR}$, and for SNM will be $S_{\mu SNM}$ and $S_{\sigma SNM}$. The target is to simultaneously achieve minimized power and maximized performance for 7T-SRAM cell. Therefore, the objective solution set $S_{obj}$ is formulated as follows:

(45) $$S_{obj} = S_{\mu PWR} \cap S_{\sigma PWR} \cap S_{\sigma PWR} \cap S_{\sigma SNM}$$

where $\cap$ is the intersection operator. Based on the values obtained from $S_{obj}$, the assignment of high-$V_{Th}$ will take place for selected transistors in the cell. This SRAM cell will then be re-simulated to obtain our objective that is P3 optimal design. Finally, an *M x N* size array will be constructed using this optimized cell; where *M* and *N* are any integer number depending on the desired array size.

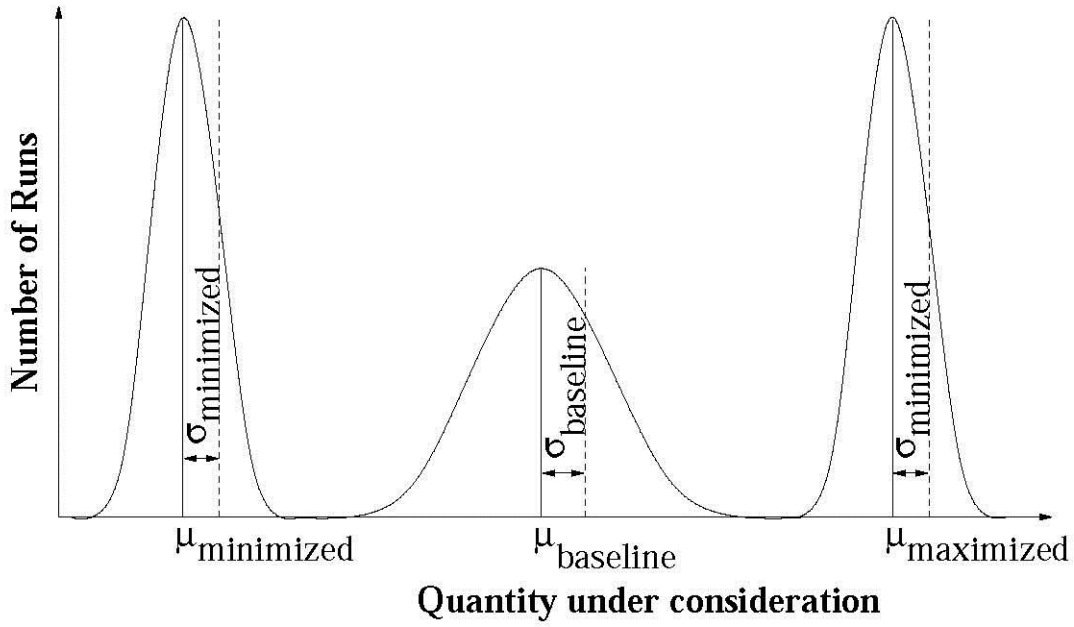



Figure 6.1. Theory behind ILP formulations.

6.2.2. The Sample Circuit or 7T-SRAM

The 7T-SRAM circuit with baseline sizes is shown in Figure 6.3. The functional simulation of the 7T-SRAM is shown in Figure 6.4 to demonstrate its operation. The power consumption and SNM of the baseline cell are measured, and are shown in Table 6.1. Designer-defined constraints are introduced and are denoted by $\tau_{PWR}$ and $\tau_{SNM}$ in the optimization methodology. This research work considers the parameters $\tau_{PWR}$ and $\tau_{SNM}$ as baseline values which are shown in Table 6.1.

Table 6.1. Power and SNM for baseline SRAM cell.

| Parameter | Value |
|---|---|
| $\tau_{PWR}$ | 203.6 nW |
| $\tau_{SNM}$ | 170 mV (Figure 6.12[a]) |

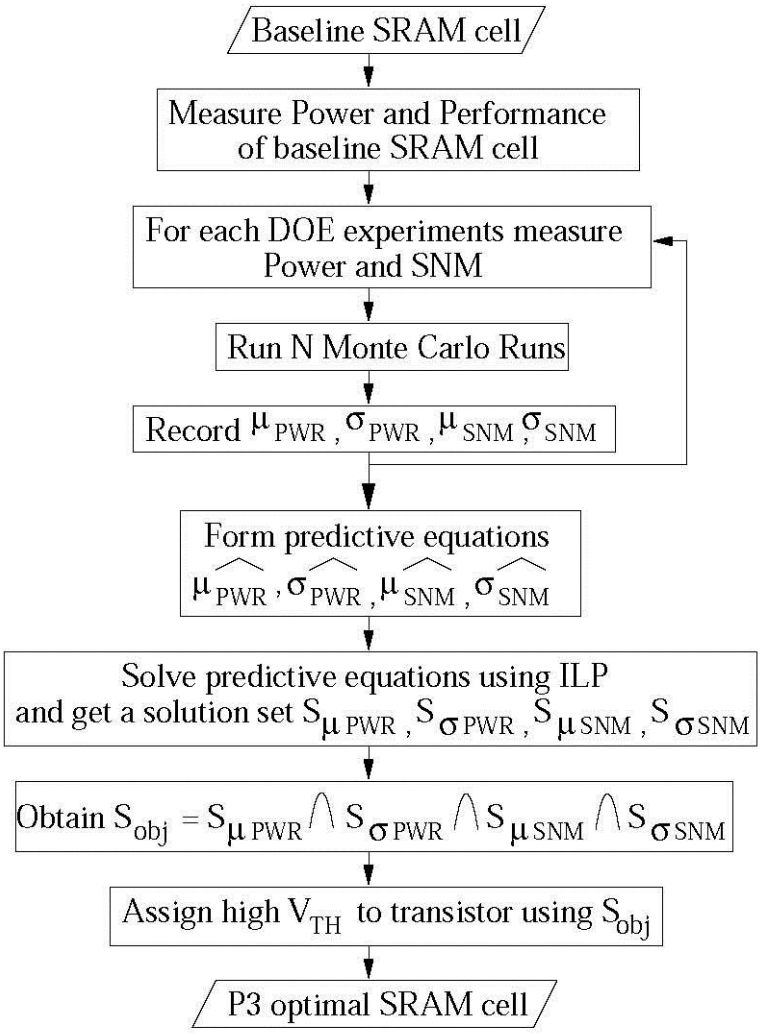



Figure 6.2. Statistical DOE-ILP approach for Nano-CMOS 7T-SRAM cell.

The instantaneous results are observed in Figure 6.5. The figure shows the instantaneous behavior of all current components during Write-Read-Hold mode, respectively. Average currents are noted in Table 6.2. With the help of this table we can easily analyze the importance of each current component: dynamic current ($I_{dynamic}$), subthreshold current ($I_{subthreshold}$), gate oxide current ($I_{gate\text{-}oxide}$) and total current ($I_{total}$). The results show that $I_{dynamic}$ is large because of the transition of bitlines from high to low or low to high, even if it is for a very brief period. On the other hand, $I_{gate\text{-}oxide}$ also contributes a significant amount. The total current, i.e., $I_{total}$ has the maximum average current because of the fact that along with the above-mentioned current components it consists of several origins, reverse-biased currents, short-circuit current, etc. The results have been quoted for

dynamic current, subthreshold current, gate-oxide current, and total current consumption during the different modes of operation of the SRAM cell, namely, Read-Write-Hold.

### 6.2.3. Proposed Statistical DOE-ILP Algorithm for P3-Optimal SRAM Design

Section 6.2.1 presents an overview of how the optimization methodology works. This section provides the detailed algorithm using the 7T cell as a test case. The schematic diagram of a baseline 7T-SRAM cell is shown in Figure 6.3. The supply voltage is $V_{DD}$ 6.2 = 0.7 V. The SRAM cell has been designed at the 45 nm node [87] with minimum-sized transistors. The sizes of all transistors (NMOS, PMOS and access transistors) in this section are considered to be the same as the technology node minimum feature ($L$ = 45 nm and $W$ = 45 nm). The algorithms for optimal design flow are presented in Algorithms 4 and 5.

Table 6.2. Average currents for 7T-SRAM cell.

| Parameter | Average value |
|---|---|
| Dynamic current | 155.14 nA |
| Subthreshold current | 101 nA |
| Gate-oxide current | 97.4 nA |
| Total current | 162.28 nA |

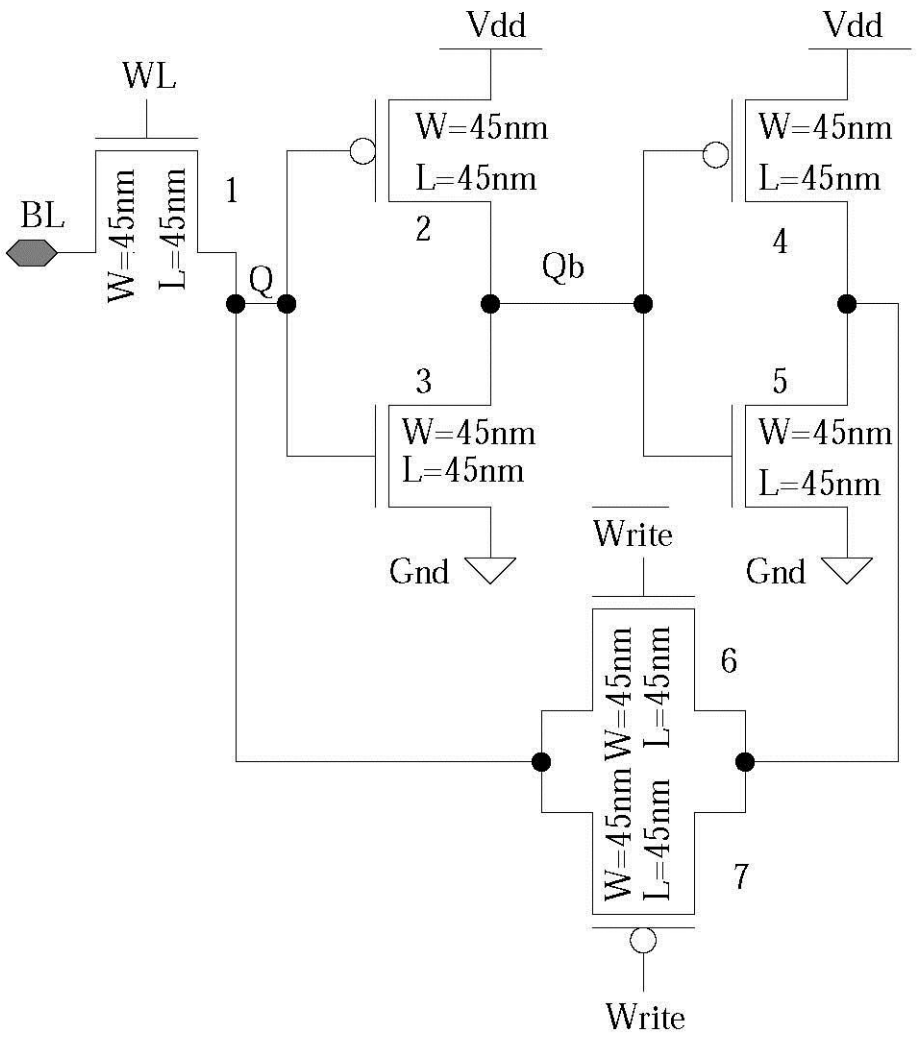



Figure 6.3. Baseline 7T-SRAM cell shown with transistor sizes.

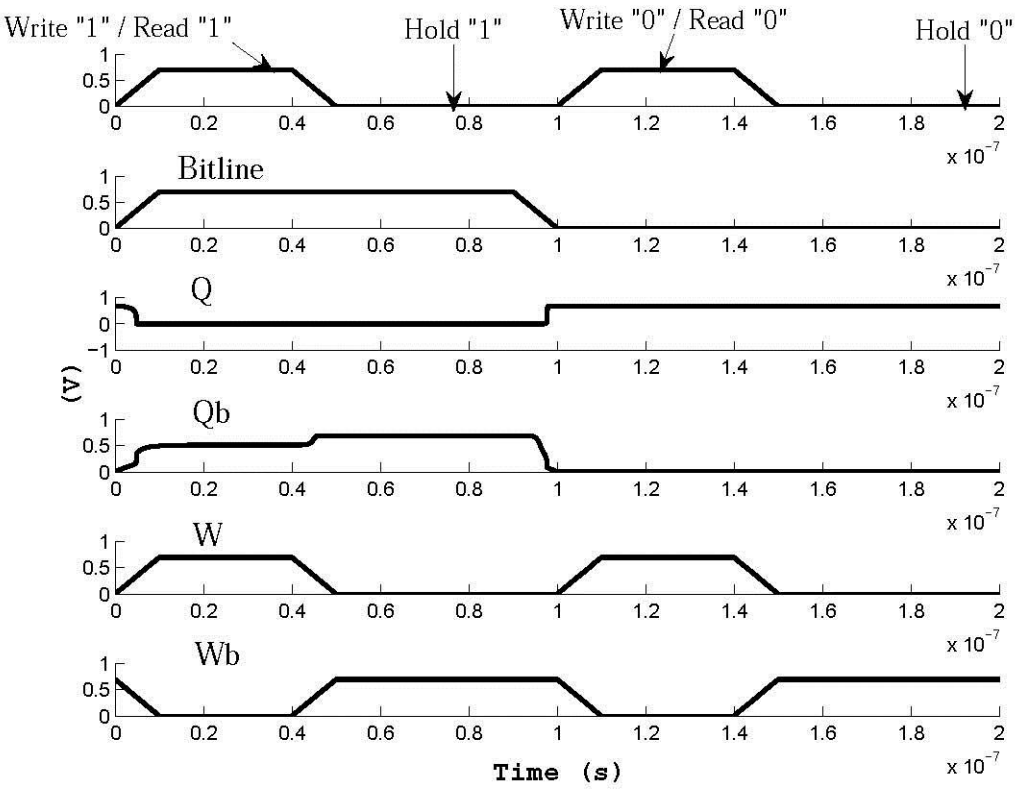



Figure 6.4. Functional simulation diagram for 7T SRAM cell.

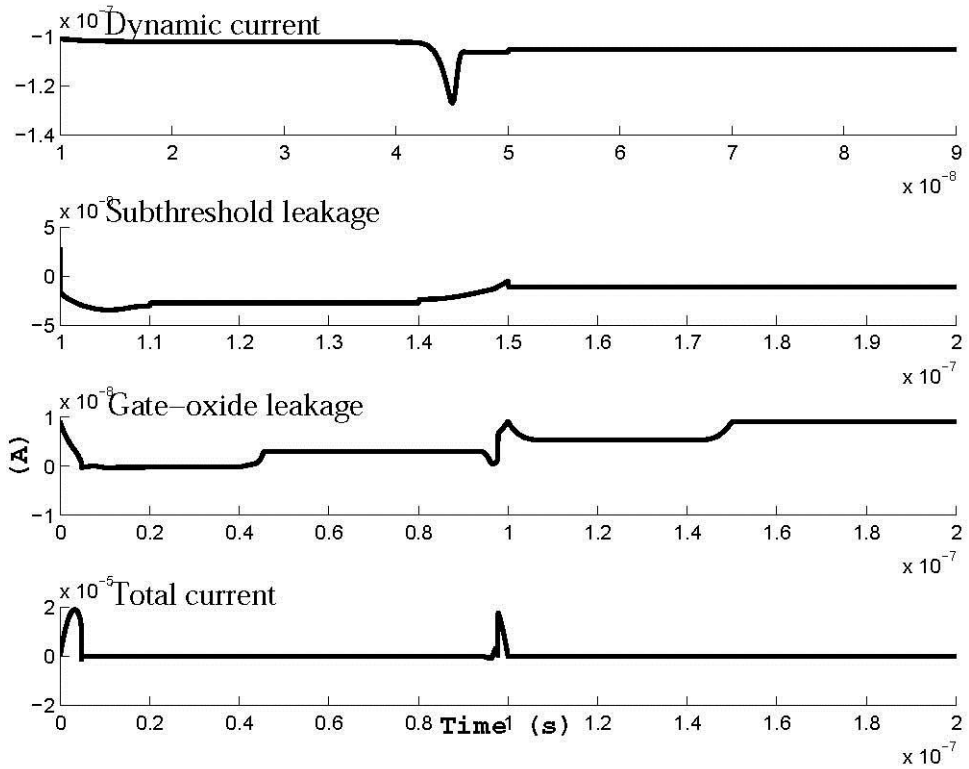



Figure 6.5. Curves for $I_{dynamic}$, $I_{subthreshold}$, $I_{gate}$, and $I_{total}$ for 7T-SRAM.

Algorithm 4 aims to achieve P3 optimization, i.e., minimized power consumption (P), maximized performance (P), and process variation (P) tolerant 7T-SRAM cell. The input to the algorithm is a baseline cell which refers to the 7T circuit with nominal-sized transistors. To maintain good performance while reducing the power consumed by the circuit in embedded SRAMs, along with scaling the minimum feature sizes of systems on chip (SoC), in itself is a very challenging task. The idea here is that leakage is a major component of the total power for the nano-CMOS. Hence, by reducing power through a dual-$V_{Th}$ technique it results in reduction of total power along with noticeable difference in performance. As discussed earlier, the dual-$V_{Th}$ technique is used to optimize the figures of merit. Through Algorithm 5 we achieve a pre-optimal 7T-SRAM cell which has P2 optimal design (power minimized and performance maximized). Once we get the P2 optimized SRAM circuit we perform process variation where the variability effect is considered in 12 device parameters.

Algorithm 4 for P3 optimization in nano-CMOS SRAM

1: Input: Baseline SRAM.

2: Output: Optimized P3: power minimization, performance maximization and process variation tolerant SRAM cell.

3: Measure power, performance of baseline SRAM cell.

4: Go to Algorithm 5 for optimizing baseline SRAM.

5: Re-simulate SRAM cell to obtain P2 (power minimization and performance maximization) SRAM cell.

6: Perform process variation characterization of SRAM cell using device parameters, in this case 12 device parameters.

7: Obtain P3 optimal SRAM cell.

8: Construct an 8 x 8 array to observe the feasibility of the SRAM cell.

In Algorithm 5, the baseline cell is taken as the input along with the baseline model file and high-threshold model file. We subject the baseline 7T-SRAM cell to a DOE [21, 37] based approach using a 2-level Taguchi L-8 array. The factors are the 7 $V_{Th}$ states of the 7 transistors of the cell (Figure 6.3). Each factor can take a high $V_{Th}$ state (1) or a nominal $V_{Th}$ state (0). The L-8 array has a total of 8 experiments. Power and performance are measured for the above-discussed baseline SRAM design. Minimum sized transistors are taken for the baseline design. Monte Carlo simulations for $N$ runs are performed for each experiment trial. Gaussian distribution values, mean (μ) and standard deviation (σ) values are recorded for average power and performance (SNM) of the cell. Thereafter, predictive equations are formed for μ and σ and are denoted by $\widehat{\mu_{PWR}}$, $\widehat{\sigma_{PWR}}$ for power and for SNM as $\widehat{\mu_{SNM}}$, $\widehat{\sigma_{SNM}}$. These predictive equations -- $\widehat{\mu_{PWR}}$, $\widehat{\sigma_{PWR}}$, $\widehat{\mu_{SNM}}$, and $\widehat{\sigma_{SNM}}$ -- 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. Using an L-8 array, we need perform a total of 8 experiments.

We obtain a solution set for the 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 $S\mu_{SNM}$, $S\sigma_{SNM}$. Since we are looking for power minimization and SNM maximization, we form our final objective $S_{obj}$ as the following:

(46) $$S_{obj} = S_{\mu_{PWR}} \cap S_{\sigma_{PWR}} \cap S_{\mu_{SNM}} \cap S_{\sigma_{SNM}}$$

where ∩ is defined as the intersection of the sets $S_{\mu_{PWR}}, S_{\sigma_{PWR}}, S_{\mu_{SNM}}, S_{\sigma_{SNM}}$.

For formation of the linear equations to be subjected to ILP, we use DOE. DOE-ILP is a much better approach compared to the other techniques because it is more efficient and faster. The proposed algorithm converges to a solution faster, using fewer resources. The complexity of the algorithm is $O(2^n)$ (where $n$ is the transistor number), or in other words

Algorithm 5 has exponential complexity. One hundred Monte Carlo simulations are run for each the experiments for a total of 800 Monte Carlo runs, taking 12 parameters in account. The 12 process parameters considered are the followings:

- $T_{oxn}$: NMOS gate oxide thickness (nm)
- $T_{oxp}$: PMOS gate oxide thickness (nm)
- $L_{na}$: NMOS access transistor channel length (nm)
- $L_{pa}$: PMOS access transistor channel length (nm)
- $W_{na}$: NMOS access transistor channel width (nm)
- $W_{pa}$: PMOS access transistor channel width (nm)

Algorithm 5 for 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 $f_{obj}$ = [$f_{PWR}$, $f_{SNM}$ ] 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 VT h 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 $\widehat{\mu_{SNM}}$, $\widehat{\sigma_{SNM}}$.

4: for Each 1:8 experiments of 2-Level Taguchi L-8 array do

5: Run 100 Monte Carlo runs

6: Record $\mu_{PWR}$, $\sigma_{PWR}$ and $\mu_{SNM}$ , $\sigma_{SNM}$

7: end for

8: Form linear predictive equations

$\widehat{\mu_{PWR}}$, $\widehat{\sigma_{PWR}}$ for power

$\widehat{\mu_{SNM}}$, $\widehat{\sigma_{SNM}}$ for SNM

9: Solve $\widehat{\mu_{PWR}}$ using ILP: Solution set $S_{\mu_{PWR}}$

10: Solve $\widehat{\sigma_{PWR}}$ using ILP: Solution set $S_{\sigma_{PWR}}$

11: Solve $\widehat{\mu_{SNM}}$ using ILP: Solution set $S_{\mu_{SNM}}$

12: Solve $\widehat{\sigma_{SNM}}$ using ILP: Solution set $S_{\sigma_{SNM}}$

13: Form $S_{obj} = S_{\mu_{PWR}} \cap S_{\sigma_{PWR}} \cap S_{\mu_{SNM}} \cap S_{\sigma_{SNM}}$

14: Assign high $V_{Th}$ to transistors based on $S_{obj}$

15: Re-simulate SRAM cell to obtain optimized objective set.

---

- $L_{nd}$ : NMOS driver transistor channel length (nm)
- $W_{nd}$ : NMOS driver transistor channel width (nm)
- $L_{pl}$ : PMOS load transistor channel length (nm)
- $W_{pl}$ : PMOS load transistor channel width (nm)
- $N_{chn}$: NMOS channel doping concentration ($cm^{-3}$)
- $N_{chp}$: PMOS channel doping concentration ($cm^{-3}$)

Amongst these parameters, some are independent and others are correlated, which is to be considered during the simulation. One of the other advantage of using DOE is that this method does not require the output to be related with the input, which otherwise would have been a cumbersome process. Each of these process parameters is considered to have a Gaussian distribution with mean (μ) taken as the nominal values specified in the PTM [87] and standard deviation 10% of the mean. A correlation coefficient of 0.9 between $T_{oxn}$ and $T_{oxp}$ is assumed. The responses under consideration are 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:

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

where $\left[\frac{\Delta(n)}{2}\right]$ is the half-effect of the *n*-th transistor, avg(1) is the average value of power when transistor n is in high-VT h state, and avg(0) is the average value of power when

transistor *n* is in 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 mV for SNM. The normalized predictive equations have been formed as follows:

$$\hat{f} = \bar{f} + \sum_{n=1}^{7} \left( \frac{\Delta(n)}{2} \times x_n \right) \tag{48}$$

Where $\hat{f}$ is the response, $\bar{f}$ is the response average, $\left[\frac{\Delta(n)}{2}\right]$ is the half effect of the *n*-th transistor, and $x_n$ is the $V_{Th}$ state of the *n*-th transistor.

Normalization is a mathematical process of smoothing or confining to a set range. Hence, we have used the normalization approach in order to get a specified set of values in a given range.

Equation 49 shows the predictive equation for mean of the average power consumption of the SRAM cell.

$$\begin{aligned} \widehat{\mu_{PWR}} = {} & 0.58 - 0.92 \times x_1 - 0.15 \times x_2 \\ & - 0.10 \times x_3 - 0.05 \times x_4 - 0.59 \times x_5 \\ & - 0.05 \times x_6 + 0.02 \times x_7 \end{aligned} \tag{49}$$

Figure 6.6(a) shows the Pareto plots of the half-effects of the transistors for $\widehat{\mu_{PWR}}$. Here, $x_i$ represents the $V_{Th}$ state of transistor *i* (Figure 6.3). From this, we formulate an ILP problem as the following expression:

$$\begin{aligned} & \min \quad \widehat{\mu_{PWR}} \\ & \text{s.t.} \quad x_n \in \{0, 1\} \; \forall \, n \\ & \qquad \mu_{SNM} > \tau_{SNM} \end{aligned} \tag{50}$$

Since we wish to minimize power consumption, we minimize $\widehat{\mu_{PWR}}$. The constraints "1" and "0" represent coded values for high $V_{Th}$ and nominal $V_{Th}$ states, respectively. ILP has been used for smaller circuits, but the methodology is automated, and hence can be used for larger circuits. Solving the ILP problem, we get the optimal solution as follows:

$$S_{\mu PWR} = \{x_1 = 1,$$
$$x_2 = 1,$$
$$x_3 = 1,$$
(51) $$x_4 = 1,$$
$$x_5 = 1,$$
$$x_6 = 1,$$
$$x_7 = 0\}.$$

This can also be interpreted as transistors 1, 2, 3, 4, 5, 6 are high $V_{Th}$ transistors, and transistor 7 is nominal $V_{Th}$ transistor.

The Pareto plot of the half-effects of the transistors for $\sigma_{PWR}$ is shown in Figure 6.6(b). Similarly, Equation 52 shows the predictive equation for the standard deviation of the average power consumption of the SRAM cell.

(52) $$\widehat{\sigma_{PWR}} = 0.61 + 0.07 \times x_1 - 0.18 \times x_2 - 0.11 \times x_3 - 0.06 \times x_4 - 0.11 \times x_5$$

From this, we formulate an ILP problem as the following expression:

(53) $$\min \quad \widehat{\sigma_{PWR}}$$
$$\text{s.t.} \quad x_n \in \{0, 1\} \forall n$$
$$\mu_{SNM} > \tau_{SNM}$$

As we seek to minimize the standard deviation (which is an indication of the spread) of power, we minimize $\widehat{\sigma_{PWR}}$. Solving the ILP problem, we get the optimal solution as the following expression:

$$S_{\sigma PWR} = \{x_1 = 0,$$
$$x_2 = 1,$$
$$x_3 = 1,$$
(54) $$x_4 = 1,$$
$$x_5 = 1,$$

$$x_6 = 1,$$

$$x_7 = 0\}.$$

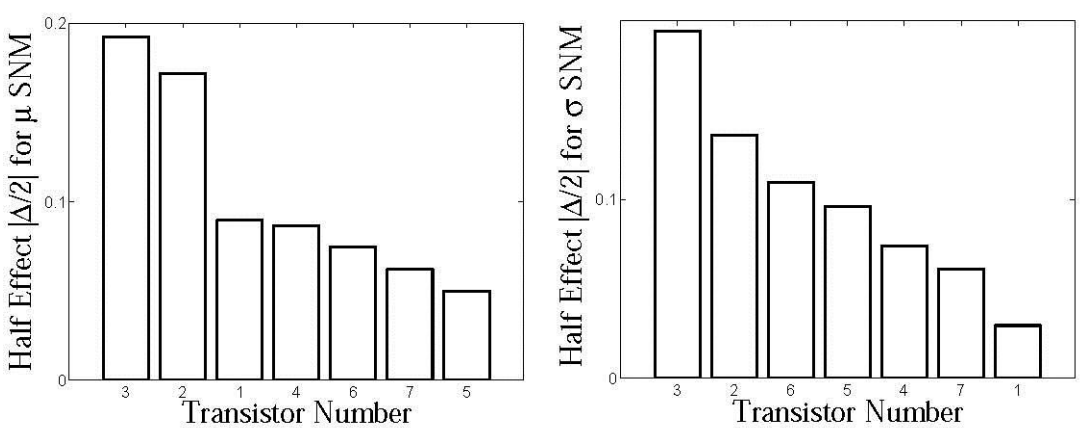
This can also be interpreted as transistors 2, 3, 4, 5, 6 are high $V_{Th}$ transistors, and transistors 1, 7 are nominal $V_{Th}$ transistor. Similarly, the predictive equation for $\mu_{SNM}$ is formed as follows:

$$\widehat{\mu_{SNM}} = 0.45 - 0.09 \times x_1 + 0.17 \times x_2 - 0.19 \times x_3 - 0.09 \times x_4 + 0.05 \times x_5 + 0.07 \times x_6 - 0.06 \times x_7 \quad (55)$$

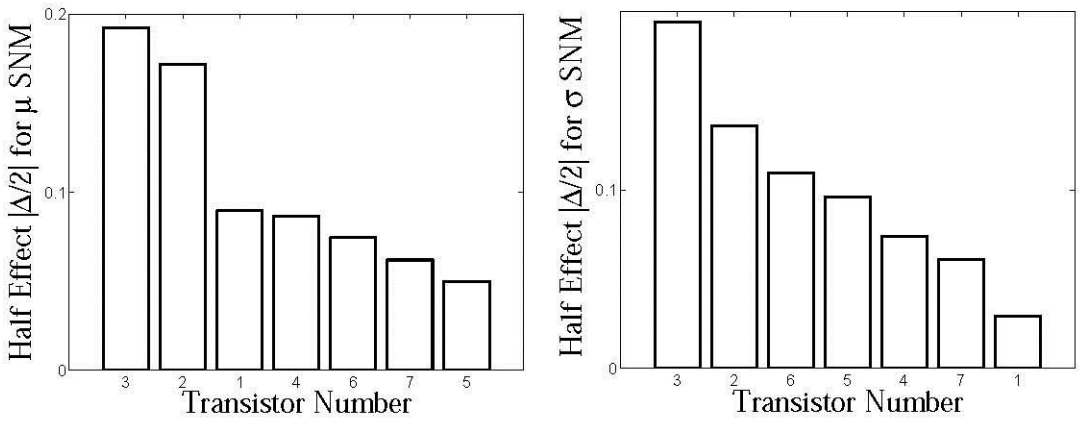
Figure 6.6(c) shows the Pareto plot of the half-effects of the transistors for $\sigma_{SNM}$. Equation 55 shows the predictive equation for the mean of the read SNM of SRAM cell. From this, we formulate an ILP problem:

$$\max \quad \widehat{\mu_{SNM}}$$
$$\text{s.t.} \quad x_n \in \{0,1\} \forall n \quad (56)$$
$$\mu_{PWR} > \tau_{PWR}$$

To maximize SNM, maximize $\widehat{\mu_{SNM}}$. Solving the ILP problem obtains the $S\mu_{SNM}$ = {$x_1$ = 0, $x_2$ = 1, $x_3$ = 0, $x_4$ = 0, $x_5$ = 1, $x_6$ = 1, $x_7$ = 0}. This 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.




(a) $\mu_{PWR}$ (b) $\sigma_{PWR}$




(c) $\mu_{SNM}$ (d) $\sigma_{SNM}$

Figure 6.6. Pareto plot for mean of SRAM read power ($\mu_{Power}$), SNM ($\mu_{SNM}$) and standard deviation of power ($\mu_{Power}$), read SNM ($\sigma_{SNM}$).

Figure 6.6(d) shows the Pareto plot of the half-effects of the transistors for $\sigma_{SNM}$. The predictive equation for $\sigma_{SNM}$ is formed as shown in Equation 57:

$$\begin{aligned}\widehat{\sigma_{SNM}} &= 0.35 + 0.03 \times x_1 - 0.13 \times x_2 \\ &+ 0.19 \times x_3 + 0.07 \times x_4 - 0.09 \times x_5 \\ &- 0.11 \times x_6 + 0.06 \times x_7\end{aligned} \tag{57}$$

From this, we formulate an ILP problem as presented below:

$$\begin{aligned}&\min \quad \widehat{\sigma_{SNM}} \\ &\text{s.t.} \quad x_n \in \{0,1\}\ \forall\, n \\ &\mu_{PWR} > \tau_{PWR}\end{aligned} \tag{58}$$

As we want to minimize the standard deviation of SNM, we minimize $\widehat{\sigma_{SNM}}$. Solving the ILP problem obtains the optimal solution as follows:

$$\begin{aligned}S_{\sigma SNM} = \{&x_1 = 0, \\ &x_2 = 1, \\ &x_3 = 0, \\ &x_4 = 0, \\ &x_5 = 1,\end{aligned} \tag{59}$$

$$x_6 = 1,$$

$$x_7 = 0\}.$$

This can also be interpreted as transistors 2, 5, and 6 are high $V_{Th}$ transistors, and transistor 1, 3, 4, and 7 are nominal $V_{Th}$ transistor.

Our final objective function $S_{obj}$ is formed as the following:

(60) $$S_{obj} = S_{\mu_{PWR}} \cap S_{\sigma_{PWR}} \cap S_{\mu_{SNM}} \cap S_{\sigma_{SNM}}$$

where $\cap$ is interpreted as the set intersection operator. Devices which are part of low-power and high-SNM solution sets are selected separately, followed by the formation of normalized equations for power and SNM so that there is no unit interference because the aim is to achieve low power and high stability in the proposed design. We get the following solution:

$$S_{obj} = \{x_1 = 0,$$

$$x_2 = 1,$$

$$x_3 = 0,$$

(61) $$x_4 = 0,$$

$$x_5 = 1,$$

$$x_6 = 1,$$

$$x_7 = 0\}.$$

In other words, transistors 2, 5, 6, are high $V_{Th}$ transistors, and transistors 1, 3, 4, 7 are nominal $V_{Th}$ transistors. Figure 6.7 shows the SRAM cell with the high $V_{Th}$ transistors circled.

Table 6.3. Statistical DOE-ILP results for 7T-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% |

Table 6.3 shows that DOE-ILP based dual-$V_{Th}$ assignment achieves 44.2% power reduction and 43.9% increase in read SNM over the baseline design. The optimized butterfly curve is shown in Figure 6.12(b). Figure 6.10 shows the comparison of baseline and P3

optimized SRAM cell power and SNM for various values of $V_{DD}$. Figure 6.8 shows the effect of process variations on the butterfly curve of the P3 optimized SRAM. Figure 6.9 shows the distributions for "SNM High" and "SNM Low" extracted from the Monte Carlo simulations. "SNM Low" is treated as the actual SNM. Table 6.4 shows the corresponding statistical data. Figures 6.12(a) and 6.12(b) show the baseline and optimized performance of the 7T-SRAM cell. As per the design flow, we then construct an 8 x 8 array using the optimized cell, shown in Figure 6.11. The average power consumption of the array is 4.5 μW.

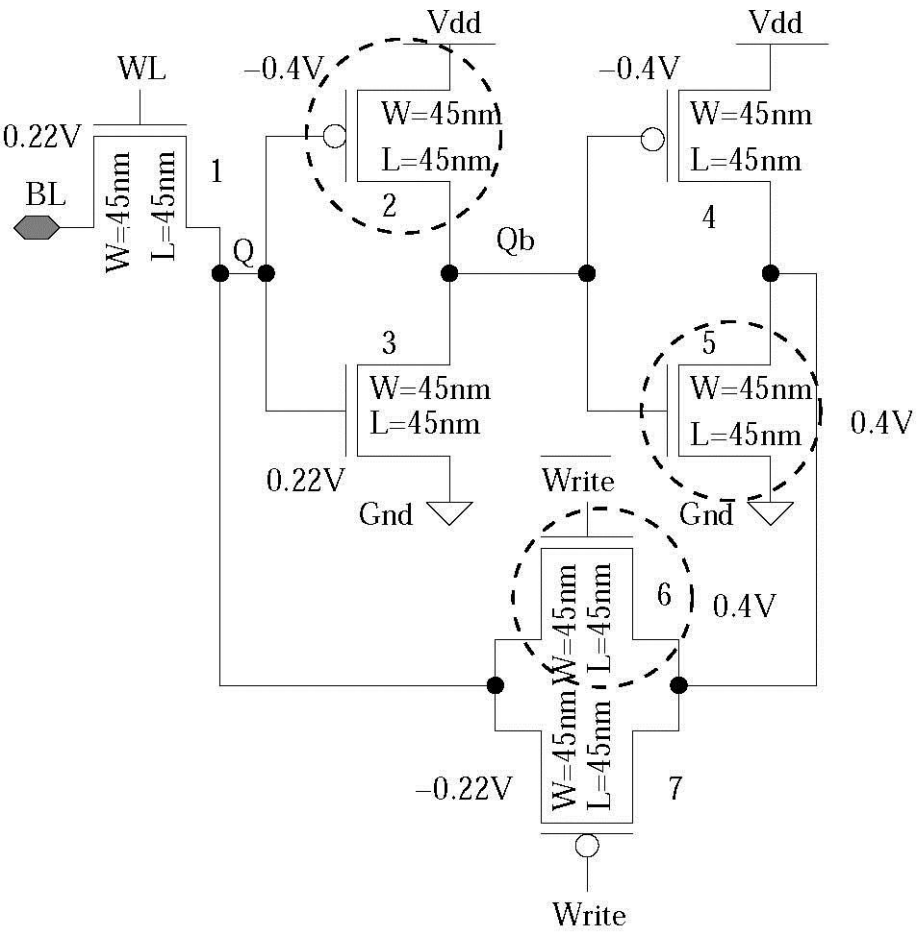



Figure 6.7. P3 optimized 7T SRAM cell with the circled transistors having high $V_{Th}$.

Table 6.4. Statistical experimental results for SNM

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

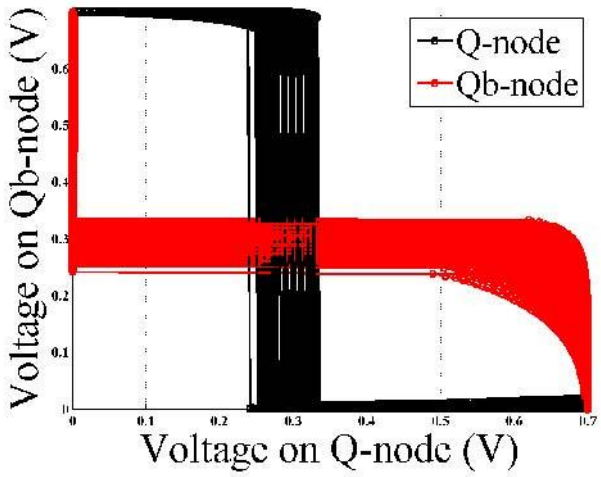



Figure 6.8. Process Variation of 7T-SRAM for SNM.

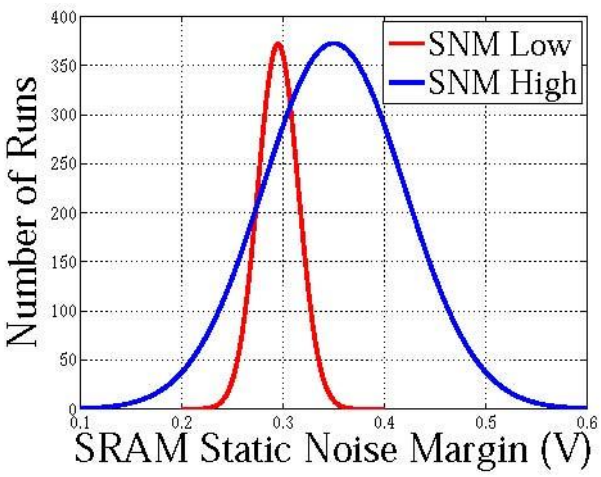



Figure 6.9. SNM distribution for optimized 7T-SRAM.

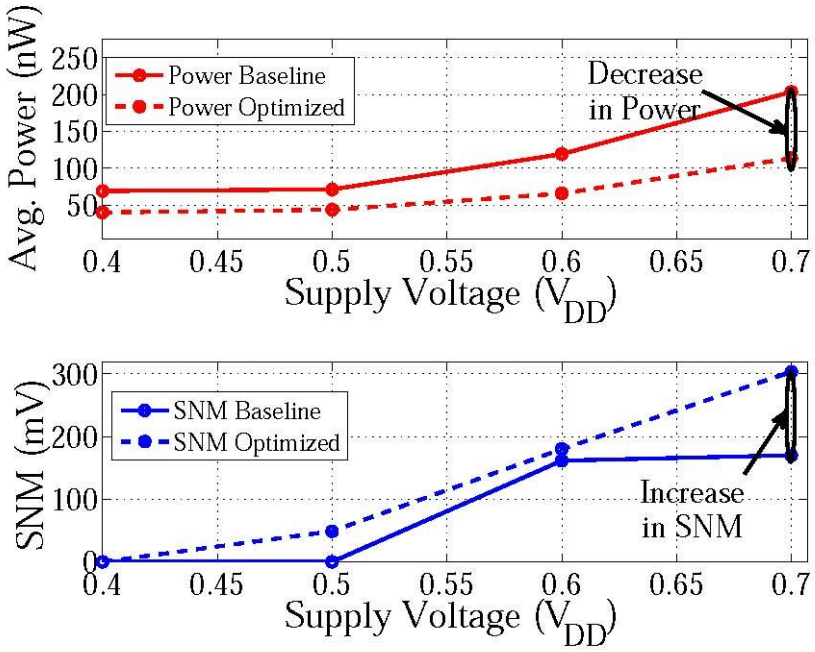



Figure 6.10. Power and read SNM comparison of the P3 optimized and baseline 7T-SRAM.

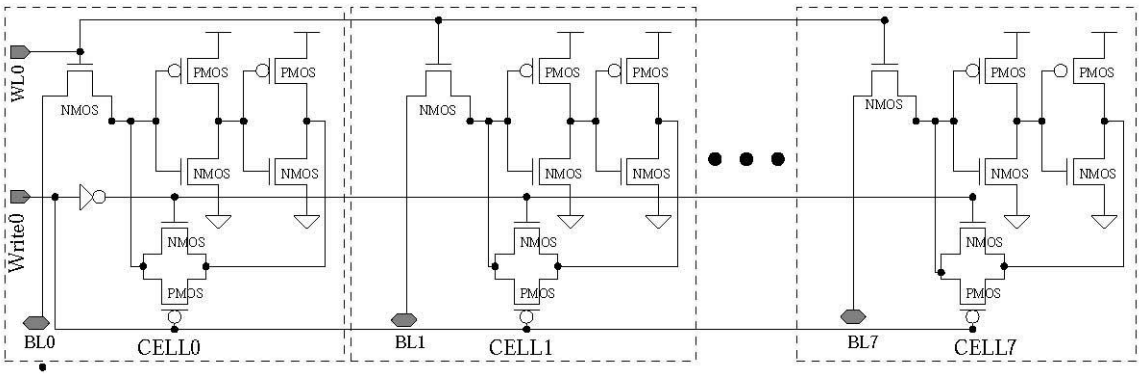
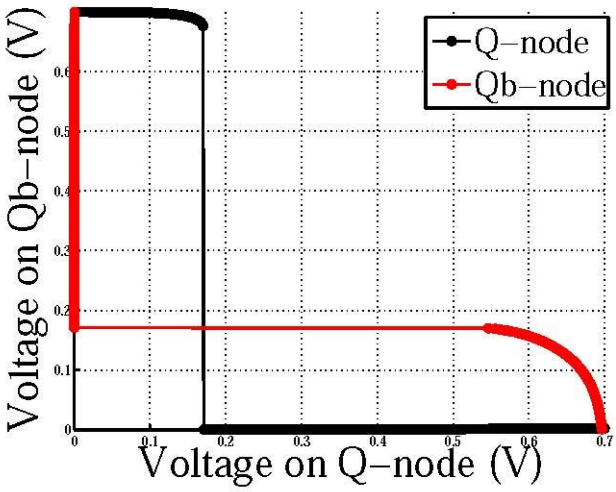
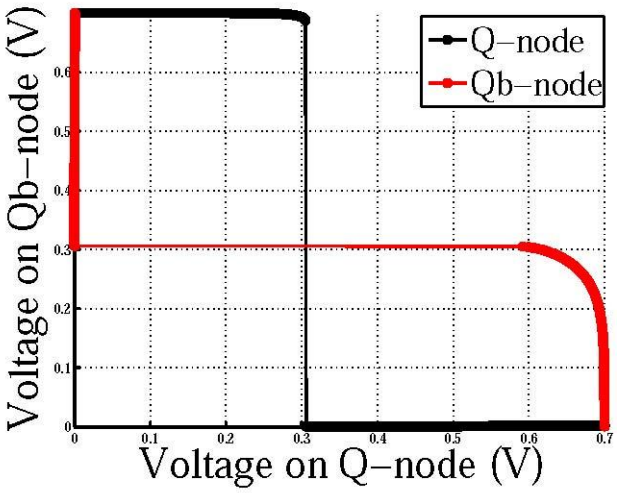



Figure 6.11. One row of the 8 x 8 array constructed using P3 optimized 7T-SRAM cells.




(a) Baseline




(b) Optimized SNM

Figure 6.12. Butterfly curve for (a) baseline and (b) optimized 7T-SRAM cell.

# CHAPTER 7

# PROCESS-VOLTAGE-TEMPERATURE (PVT) OPTIMIZATION OF THE SRAM

This chapter deals with the process-voltage-temperature (PVT) tolerant 7T-SRAM cell using a novel, polynomial regression based technique. Even though the design flow and the corresponding algorithm is presented in the case of a static random access memory (SRAM), it can be easily extended to any other nanoscale complementary metal-oxide semiconductor (CMOS) circuits.

## 7.1. PVT-Tolerant 7T-SRAM Design Flow

The optimal design flow for a 7T-SRAM circuit is shown in Figure 7.1. It accounts for power, performance, process, and temperature. Initially, the logical design is performed for a 7T-SRAM circuit (discussed in Chapter 3), shown in Figure 7.2. From this design, baseline parameters are measured such as average power consumption and SNM. The design is constructed using a 45 nm predictive technology model (PTM) library. Standard sizes are taken for the transistors in the baseline design where the length of the P-channel metal-oxide semiconductor (PMOS) and N-channel metal-oxide semiconductor (NMOS) transistors ($L_p$ and $L_n$) is 45 nm and the width of PMOS and NMOS ($W_p$ and $W_n$) is 8 x $L_p$ and 4 x $L_n$, respectively. The figures of merit considered are average power consumption of the SRAM. Static noise margin (SNM) is taken as the figure of merit (FOM) in stability evaluation and thus taken as a performance parameter. Both these figures of merit are optimized through a third parameter (which we introduce in this research work), called power over SNM (PSR). By minimizing PSR, the average power dissipation is minimized and performance (SNM) is maximized.

A worst-case ambient temperature analysis is performed on the baseline SRAM design (measured at 27°C, 50°C, 75°C, 100°C, and 125°C). Power and leakage models are identified to construct power the dissipation profile. Another important parameter along with the ambient temperature is the on-chip temperature. Ambient temperature provides

information on the operating conditions of transistors whereas on-chip temperature identifies the hot-spots for the circuit. The SRAM circuit is then subjected to process variation analysis for geometric, process, and on-chip parameter set, but the most significant parameter, that is, geometric parameter set, is considered and thus applied for process variation analysis. The device parameters taken in the geometric set are $W_n$ and $W_p$.

The baseline SRAM design observed for worst case ambient temperature is then subjected to a polynomial regression technique for all three FOMs. Simulations are run for these FOMs for a certain range and surface plots are developed. Through these surface plots, polynomial equations are generated and solved for optimizing each figure of merit. For example, minimizing the polynomial equation for average power generates the minimum value for $W_n$ and $W_p$; at these values the circuit consumes minimum power. Similarly the equations are solved for SNM and PSR.

Finally, we analyze and compare the baseline design and optimal design through process variation analysis using the geometric parameter set.

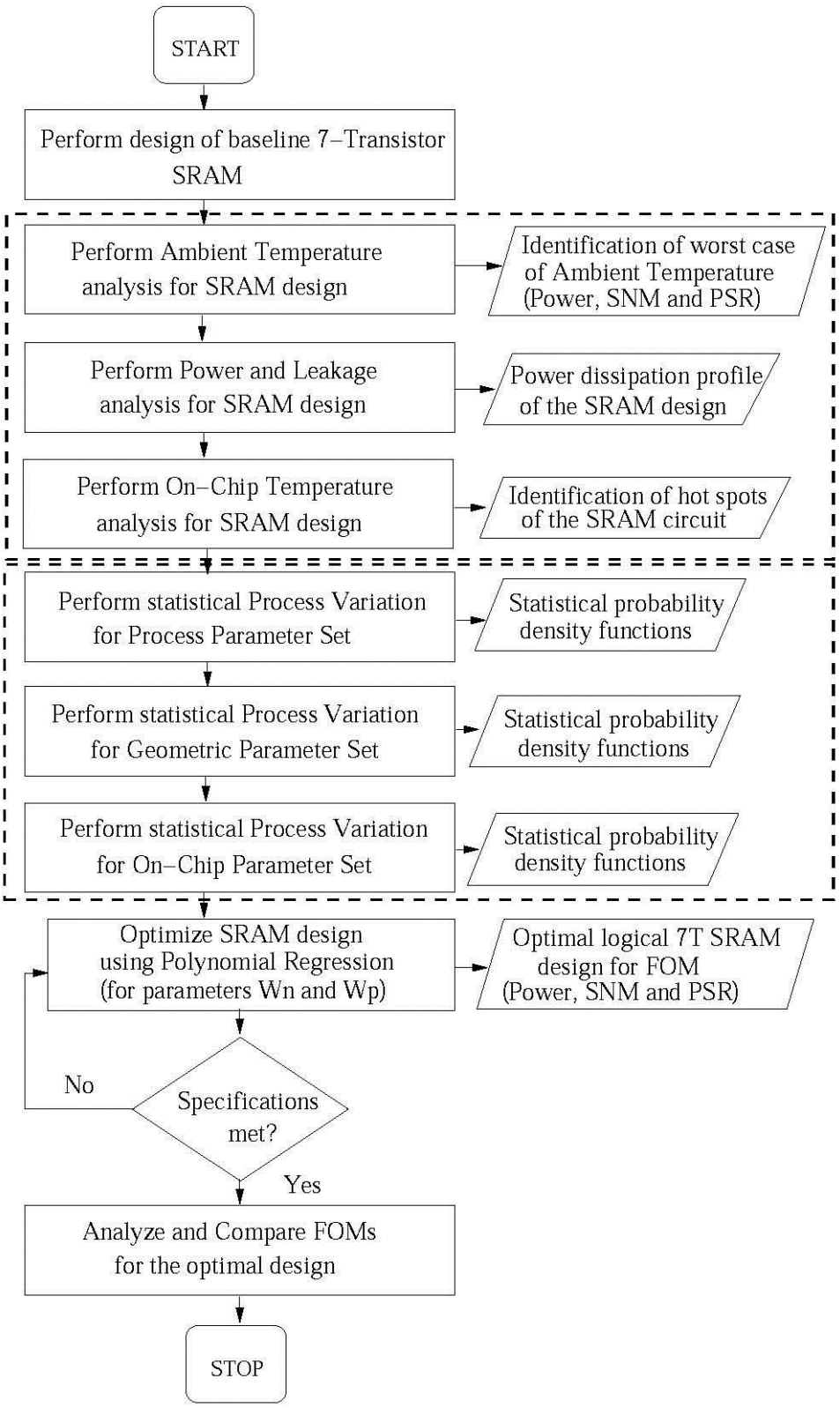



Figure 7.1. Design flow for PVT-optimal 7T SRAM.

7.2. The Sample Circuit of 7T-SRAM: Sample

Table 7.1 shows the power and SNM results for the baseline design. The butterfly curve is shown in Figure 7.3.

Table 7.1. Power and SNM for baseline SRAM cell.

| Parameter | Value |
|---|---|
| Average Power | 536.70 nW |
| SNM | 182.5 mV |

Power Leakage Measurement and SNM Measurement

The total power consumption is taken into consideration and the SNM is measured in the same way as in Chapter 5.

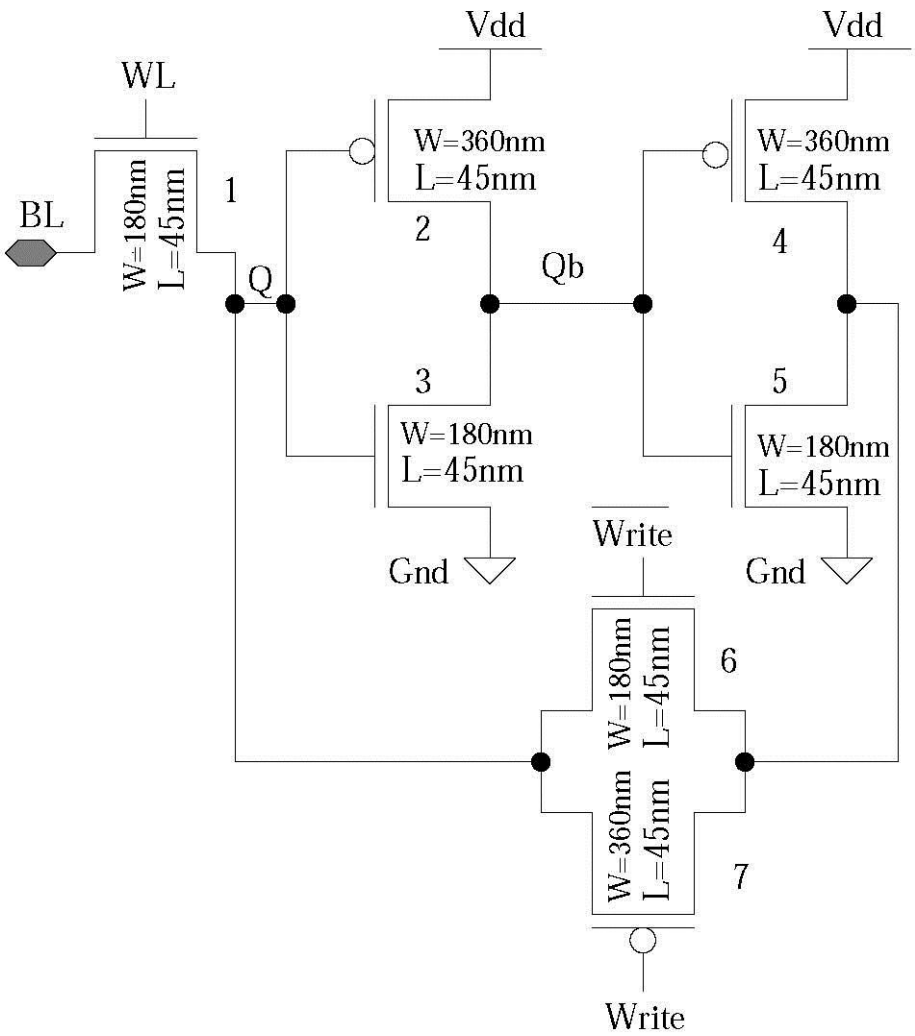



Figure 7.2. Baseline 7T-SRAM cell.

PSR Measurement

A novel parameter, PSR has been introduced in this research work. PSR is defined as the power over SNM ratio. To minimize the average power dissipation and improve and/or maximize SNM, we simply need to minimize the PSR parameter. It can be formulated as the following expression:

(62) $$PSR = \left(\frac{\text{Power}}{SNM}\right)$$

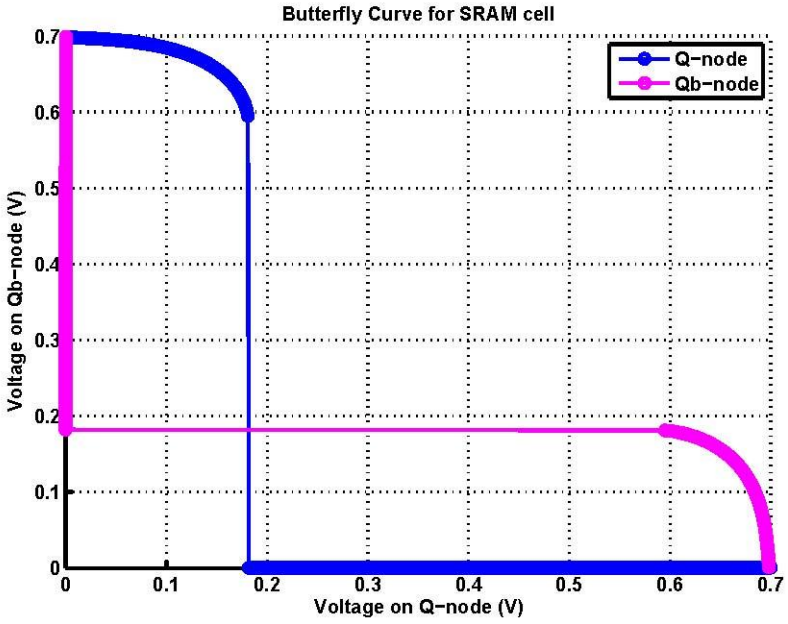



Figure 7.3 Butterfly curve for baseline 7T-SPRAM cell.

7.3. Temperature (T) Variation Characterization in SRAM

Fluctuations present in the die temperature affect the device characteristics, which in turn affect the performance of integrated circuits (ICs). The value of device parameters fluctuates because of increase in doping concentration in transistors. Therefore, the models for nano-CMOS circuits have to be carefully chosen for proper characterization. Power and performance of the nano-CMOS circuits are the important figure of merits that require accurate characterization. Both these parameters are results of the die temperature. The circuit is widely diversified at different sections of an IC and thus the temperature can vary significantly from one die to another die. Furthermore, environmental temperature

fluctuations may also cause high variations in the die temperature, fluctuations in the die temperature affect the device parameters affecting the performance and power dissipation of ICs [41].

Aggressive scaling in nano-CMOS technology has resulted in increased chip density while maintaining stability of the cells. Different sections of SRAM can experience different temperature profile depending on their proximity to other logic units. In order to achieve this density, high computation logic units are needed but they also increase the temperature (by creating hot-spots in on-chip temperature) [53]. The maximum temperature that can be reached by a chip during its operation is increasing. This in turn, affects reliability and cooling costs. In ambient temperature analysis we have observed how the SRAM cell behaves in operating or environment temperature conditions.

The impact of ambient temperature variation (measured at 27°C, 50°C, 75°C, 100°C, and 125°C) is observed in the 7-Transistor SRAM circuit for all three FOMs that is, average power, performance, and PSR in Figure 7.4. The increase in leakage in the SRAM cell increases the temperature (both ambient temperature and on-chip temperature) because of the strong dependence of subthreshold leakage current as shown in Equations 63 and 64 [53] below:

(63) $$I_{subthreshold} = I_o \exp\left(\frac{V_{gs} - V_{TH0} - \eta V_{ds} - \eta V_{sb}}{nV_0}\right) \times \left(1 - \exp\frac{-V_{ds}}{V_t}\right)$$

(64) $$I_o = \mu C_{ox}\left(\frac{W}{L}\right) \times V_t^2 e(1.8)$$

where $W$ and $L$ are the channel widths and lengths of the transistors, μ is the carrier mobility, $V_t = kT/q$ the thermal voltage, the drain induced barrier lowering (DIBL) coefficient is given by η and $n$ is the slope shape factor/subthreshold swing coefficient.

From Equations 63 and 64, we observe the following dependencies of subthreshold leakage on device parameters:

- Transistor width ($W$) is directly proportional
- Transistor length ($L$) is inversely proportional

- Temperature ($T$) exponential increase
- Input voltage $V_{gs}$ exponential increase

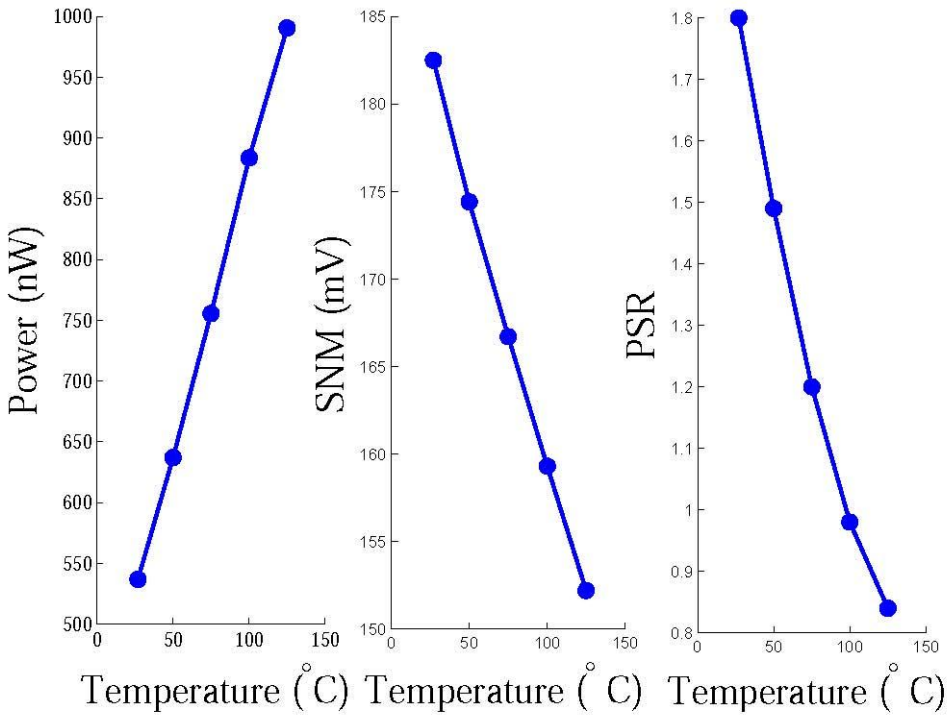



Figure 7.4. Worst case ambient temperature analysis for FOMs.

Hence, as seen from Figure 7.4, we observe that as leakage increases so will the total power dissipation of the circuit. The SNM will be degraded as the temperature increases. Therefore, as a result the PSR is analyzed on the basis of power and SNM. It may be noted that PSR is the ratio of power and SNM and so cannot be calculated directly because of different units for power and SNM ($nW/\mu W$ and $mV$), respectively. Therefore, we normalize the two quantities in order to analyze it.

7.4. Process Variation Analysis

This research highlights the geometric parameter set as it has the most significant effect. One thousand Monte Carlo simulations are run for each FOM, taking into account the geometric device parameters ($W_n$ and $W_p$). Each of these process parameters is considered to have a Gaussian distribution with mean ($\mu$) taken as the nominal values specified in the PTM and 3 x standard deviation (3 - $\sigma$) as 10% of the mean. The transistors in the 7T-SRAM topology are placed very close to each other therefore, all widths for the PMOS transistors

have been assigned $W_p$ while all the widths of the NMOS transistors have been assigned $W_n$. Future work will take into consideration other geometric parameters, that is $L_p$ and $L_n$. The process variation results for average power dissipation are shown in Figure 7.10(a) and the resultant plot for SNM is presented in Figure 7.8(b).

## 7.5. Polynomial Regression Optimization Technique

This section discusses the polynomial regression based optimization, which is the heart of the PVT optimal design flow. As shown in Algorithm 6, the baseline SRAM cell is taken as the input along with the baseline model file. The SRAM design, initially, is identified for worst case ambient temperatures. The three FOMs are measured at different temperatures, i.e. 27°C, 50°C, 75°C, 100°C, and 125°C. The baseline 7-T SRAM cell undergoes process variation at 125°C, identified as the worst-case ambient temperature. The proposed algorithm converges to a solution faster using fewer resources.

The formulation of the polynomial equations is done by taking a certain range of $W_n$ and $W_p$, and then SNM and PSR and simulated and surface plots are generated. The formulation of the objective function is given as the following:

(65) $$\widehat{f_{PSR}} = \left(\frac{\widehat{f_{PWR}}}{\widehat{f_{SNM}}}\right)$$

The three surface plots are polynomial in nature, and polynomial equations will be formulated. Polynomial regression is a very effective approach compared to other techniques because it is more efficient, reliable and faster. For instance, in [83] we have implemented a DOE-ILP approach, which requires a smaller number of runs than a full factorial, whereas in this research, just solving (minimize or maximize) the polynomial equation will result in optimal solutions. The advantages of using polynomial regression are that it requires few iterations, is accurate, and is faster (results are obtained in a few seconds).

Algorithm 6 for PVT tolerant optimization

1: Input: Baseline power and SNM of the SRAM cell, baseline model file.

2: Output: Optimized Figure of Merit: $\widehat{f_{PSR}} = \frac{\widehat{f_{PWR}}}{\widehat{f_{SNM}}}$ with transistors identified for optimized $W_n$ and $W_p$ assignment.

3: Identify worst case ambient temperature (measured at 27°C, 50°C, 75°C, 100°C, and 125°C) for defined FOMs (Power, SNM and PSR) of SRAM design.

4: Generate power dissipation profile of SRAM design by measuring average (total) power consumption and total leakages.

5: for Each range of $W_n$ and $W_p$ of transistors in SRAM do

6: Run simulations.

7: Record power, SNM and PSR.

8: end for

9: Generate surface plots using polynomial regression, for all three FOMs.

10: Form polynomial equations: $\widehat{f_{PWR}}$ for power, $\widehat{f_{SNM}}$ for SNM, and $\widehat{f_{PSR}}$ for PSR.

11: Minimize $\widehat{f_{PWR}}$ using second order differential equation.

12: Maximize $\widehat{f_{SNM}}$ using second order differential equation.

13: Minimize $\widehat{f_{PSR}}$ using second order differential equation.

14: Optimize $\widehat{f_{PSR}} = \frac{\widehat{f_{PWR}}}{\widehat{f_{SNM}}}$

15: Assign optimized values of $W_n$ and $W_p$ for the NMOS and PMOS transistors.

16: Re-simulate SRAM cell to obtain optimized objective $\widehat{f_{PSR}}$ .

The three surface plots are generated by fitting simulation data to quadratic polynomials of the form given below:

(66)

$$\widehat{f_x} = \sum_{i,j=0}^{2} \alpha_{ij} W_n^i W_{jp}$$

where X is PWR, SNM, or PSR and $\alpha_{ij}$ is the matrix of coefficients obtained during the polynomial regression. Once the analytical polynomials of the form 67 are obtained, optimal values of the vector $\mathbf{x} = [W_n\ W_p]^T$ are obtained from:

(67) $$\frac{\delta f_x}{\delta W_n} = \frac{\delta f_x}{\delta W_n} = 0$$

(68) $$\frac{\delta^2 f_x}{\delta W_n^2}, \frac{\delta^2 f_x}{\delta W_p^2} \begin{matrix} > \\ < \end{matrix} 0$$

where the > 0 criterion is used for minimization and the < 0 criterion is used for maximization.

7.5.1. Power Optimality: $\widehat{f_{PWR}}$

The surface plot general for power is shown in Figure 7.5.

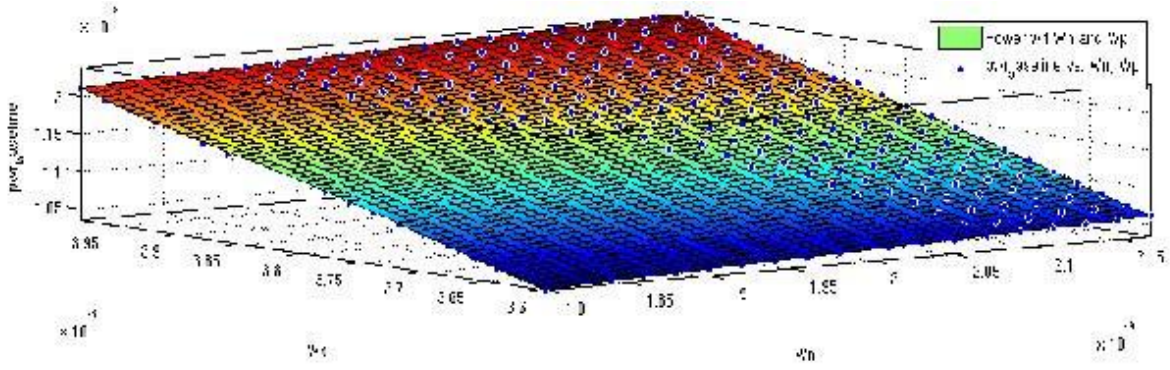


Figure 7.5. Surface plot for average power.

The matrix formed from the polynomial equation is shown below:

(69) $$\alpha_{ij} = \begin{bmatrix} 1.13 \times 10^{-6} & 5.02 \times 10^{-8} & -1.76 \times 10^{-9} \\ 7.81 \times 10^{-9} & 1.3 \times 10^{-10} & 1.33 \times 10^{-11} \\ -4.07 \times 10^{-9} & -8.54 \times 10^{-12} & 0 \end{bmatrix}$$

To minimize the power consumption, $\widehat{f_{PWR}}$ is minimized. The power optimal results are shown in Table 7.2.

7.5.2. SNM Optimality: $\widehat{f_{SNM}}$

Similarly, for SNM the surface plot is shown in Figure 7.6.

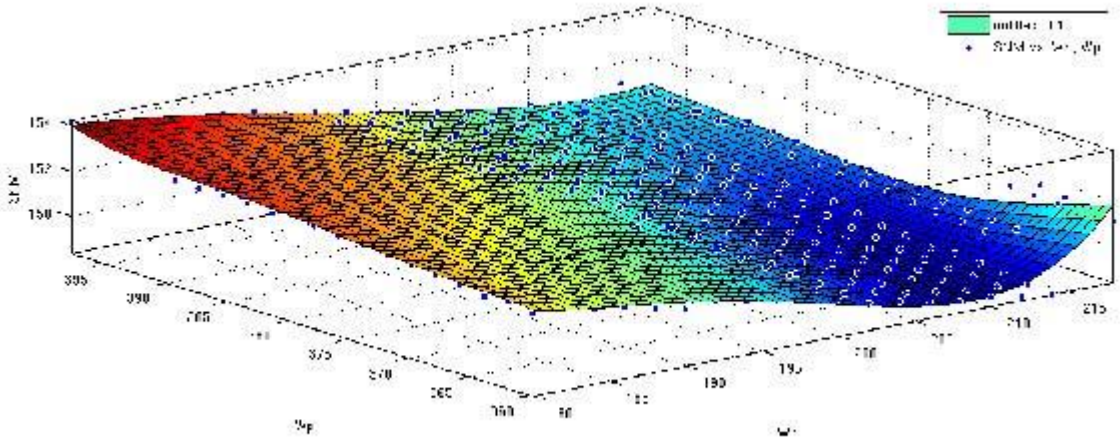


Figure 7.6. Surface plot for SNM.

The matrix formed from the polynomial equation of SNM is shown below:

(70) $$\alpha_{ij} = \begin{bmatrix} 150.9 & 0.73 & 0.06 \\ -1.67 & -0.07 & 0.07 \\ 0.21 & -0.15 & 0 \end{bmatrix}$$

To maximize the SNM, $\widehat{f_{SNM}}$ is maximized. The results for SNM optimality are shown in Table 7.2.

7.5.3. PSR Optimality: $\widehat{f_{PSR}}$

For PSR optimality, the method is similar to that of power optimality. The equations are formed by normalizing the values. The normalization is performed by division of each data by the maximum value of the data in the set. Normalized data enables directly accommodating different units.

The surface plot is shown in Figure 7.7. The matrix formed for PSR from the polynomial equation is shown below:

$$\alpha_{ij} = \begin{bmatrix} 0.94 & 0.05 & 0 \\ 0 & 0 & 0.01 \\ 0 & -0.01 & 0 \end{bmatrix}$$

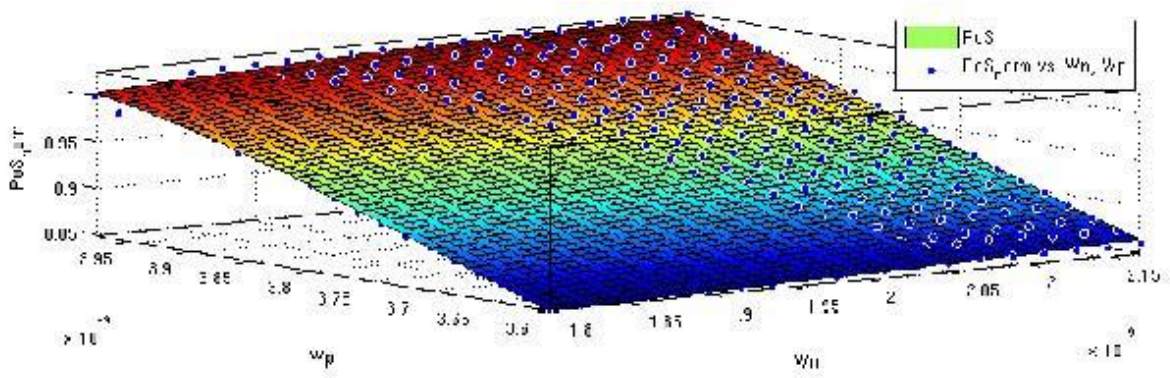


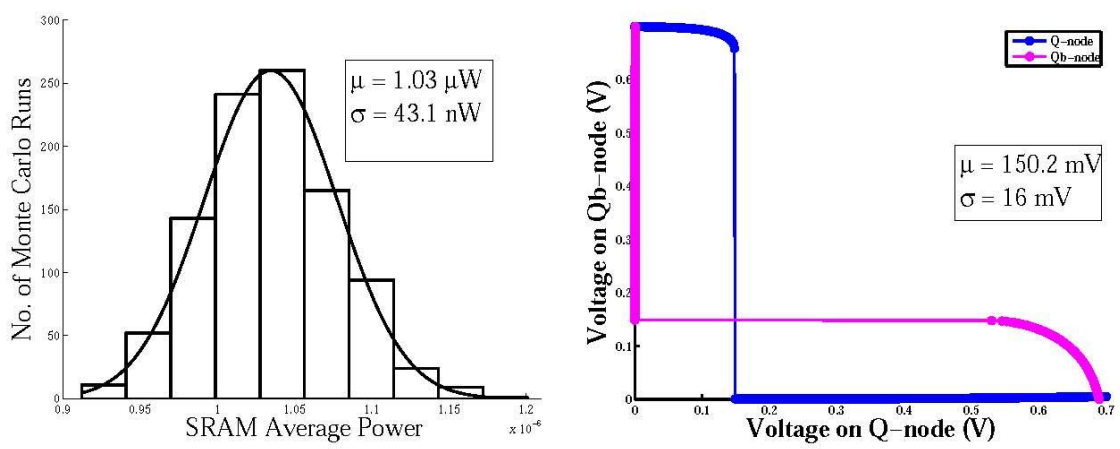
Figure 7.7. Surface plot for PSR.

Table 7.2. Average power, SNM, and PSR for optimal SRAM cell

| Parameter | Baseline design | Power optimality | SNM optimality | PSR optimality |
|---|---|---|---|---|
| Average Power | 1.03 µW | 1.03 µW | 1.23 µW | 1.03 µW |
| SNM | 150.1mV | 150.1mV | 154mV | 154mV |
| PSR | 18.94 | 18.94 | 20.84 | 18.94 |

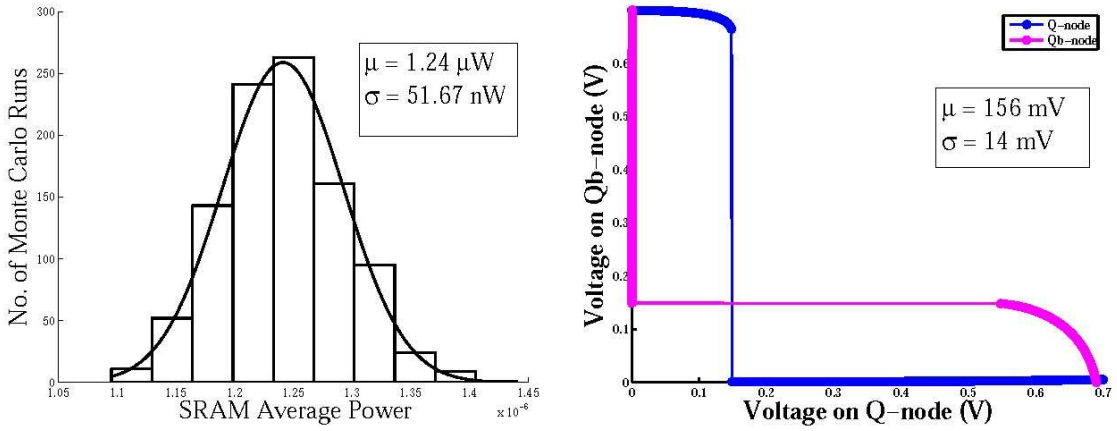
7.6. PVT-Tolerant SRAM design

This section includes the analysis and comparison of the baseline and optimal design for the 7T-SRAM sample circuit at worst case ambient temperature. Process variation is again conducted on the optimal design using device parameters $W_n$ and $W_p$ and the results are observed for the three FOMs in Figures 7.8, 7.9, and 7.10. Thus, it is observed that the optimal SRAM design is PVT-tolerant.



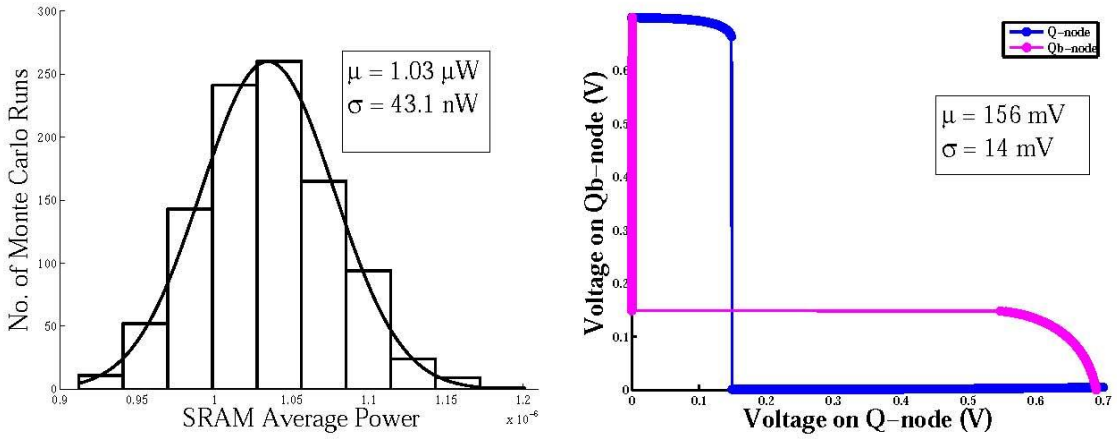

(a) Power optimality design (average power) (b) Power optimality design (SNM)

Figure 7.8. Power optimal design.




(a) Average power (b) SNM

Figure 7.9. SNM optimal design.




(a) Average power (b) SNM

Figure 7.10. PSR optimality design.

# CHAPTER 8

# CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

In this chapter we present a summary of the research investigated in this dissertation, followed by conclusions drawn from the research and experimental results. Finally, directions for future research are discussed which will serve as a road map advancement of the state-of-the art in nanoscale complementary metal-oxide semiconductor (CMOS) circuit design.

## 8.1. Summary of the Proposed Research

In this dissertation, low-power design techniques have been applied to specific nano-CMOS memory circuits, specifically static random access memory (SRAMs). A different topology is considered than the traditional 6-transistor SRAM (6T-SRAM), that is, 7T-SRAM, and is exhaustively explored. Comparison is made for 6T-SRAM, 7T-SRAM, and 10T-SRAM cells and discussion regarding the modes of operation, advantages and disadvantages of one over the other are discussed. Accurate power dissipation and leakage analysis is done to build the background for optimizing these digital circuits. Optimization approaches have been applied to these circuits to achieve low power designs which are also process variation tolerant. The optimization methodologies proposed are as follows:

- Design of experiments-integer linear programming (DOE-ILP) approach
- Design of experiments (DOE) assisted conjugate gradient approach
- Statistical design of experiments-integer linear programming (DOE-ILP) approach
- Polynomial regression-based technique

The following circuits have been subjected to these optimization methodologies:

- 45 nm 6-Transistor SRAM
- 45 nm 7-Transistor SRAM
- High-κ/Metal-Gate 32 nm 10-Transistor SRAM

## 8.2. Conclusions from the Experimental Results

The current section presents the conclusions drawn from the experimental results. Op-

timization methodologies have been presented for cell-level optimization of SRAM power and stability. A 32 nm high-κ metal gate 10T-SRAM is subjected to the proposed methodology which has shown 86% reduction in power and 8% increase in static noise margin (SNM). A novel DOE-ILP approach has been used for power minimization, and a conjugate gradient method is used for SNM maximization. The effect of process variation of 12 parameters on the proposed SRAM is evaluated. An 8 x 8 array has been constructed using the optimized cell and data for power and read access time are presented.

A methodology for simultaneous optimization of SRAM power and read stability is presented for a 45 nm single ended 7T-SRAM and the cell is subjected to the proposed methodology leading to 50.6% power reduction and 43.9% increase in read stability (read SNM). A novel DOE-ILP algorithm is used for power minimization and read SNM maximization. It is found to be process variation tolerant. An 8 x 8 array has been constructed using the optimized cells whose average power consumption is 4.5 μW. This methodology, which considers only dual-$V_{Th}$, has resulted in power reduction (accounting for all leakage components) of 50.6% and increase in read SNM of 43.9%.

A statistical DOE-ILP approach has been also presented for simultaneous P3 (power-performance-process) optimization of SRAM cells. The read SNM has been treated as the performance metric. The optimization has been performed at cell level. Towards this end, a single ended 7T-SRAM cell of 45 nm has been subjected to the proposed approach which leads to 44.2% power reduction (including leakage) and 43.9% increase in performance (Read SNM). Using the P3 optimized cell an 8 x 8 array is constructed and data are 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* stands for temperature.

A comparative study done for all optimization approaches is shown in Table 8.1.

Table 8.1. Comparison of results of optimization approaches

| Approach | Power ($nW/\mu W$) | SNM ($mV$) | Temperature | No. Transistors |
|---|---|---|---|---|
| Combined DOE-ILP | 100.5 nW | 303.3 mV | 27°C | 7T |
| DOE-ILP Assisted Conjugate-Gradient | 314.5 nW | 295 mV | 27°C | 10T |
| Statistical DOE-ILP | 113.6 nW | 303.3 mV | 27°C | 7T |
| Polynomial Regression | 1.03 μW | 154mV | 125°C | 7T |

8.3. Future Research Within the Scope of the Current Research

For different topologies of SRAM design, as part of future research, we propose to perform the complete design cycle including physical design at 32 nm and 45 nm. Alternative topologies for SRAM will be explored to achieve optimized SRAM design. Future research will be extended from Chapter 7, where only ambient temperature is considered. Taking into account the on-chip temperature which is responsible for hot spots in a circuit will be useful in exploring and building up a design which will be close to fabrication.

Also as part of future work we plan to implement the 7T-SRAM design at a lower nano-CMOS technology node, i.e., 32 nm. The physical design will be constructed. The impact of parasitics on the performance of the SRAM physical design using dual-$V_{Th}$ process kits may be studied. The scope of the research in this dissertation has been kept at cell-level optimization but future work includes the application of optimization methodologies to array level circuits. For array optimization, both mismatch and process variation will be considered as part of the design flow.

REFERENCES

[I] *Aries: An LSI Macro-Block for DSP Applications*, 1998, http://www.iis.ee.ethz.ch/-kgf/aries/all.html.

[2] A. Amara. B. Giraud and O. Thomas, "An Innovative 6T Hybrid SRAM Cell in Sub-32nm Double-Gate MOS Technology," *5th IEEE Int'l Symp. Electronic Design, Test and Applications*, 2010, pp. 241-244.

[3] K. Agarwal and S. Nassif, "Statistical Analysis of SRAM Cell Stability," *Proc. Design Automation Conference*, 2006, pp. 57-62.

[4] _____,"The Impact of Random Device Variation on SRAM Cell Stability in Sub-90-nm CMOS Technologies," *Proc. IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, 2008, p. 8697.

[5] B. Amelifard, F. Fallah, and M. Pedram, "Reducing the Sub-threshold and Gate-tunneling Leakage of SRAM Cells using Dual-$V_t$ and Dual-$T_{ox}$ Assignment," *Proc. Design Automation and Test in Europe*, 2006, pp. 1-6.

[6] _____, "Leakage Minimization of SRAM Cells in a Dual-$V_t$ and Dual-$T_{ox}$ Technology," *Proc. IEEE Trans. VLSI Systems*, 2008, pp. 851-860.

[7] T. Azam. B. Cheng. and D.R.S. Cumming, "Variability Resilient Low-power 7T-SRAM Design for Nano-Scaled Technologies," *Proc. 11th IEEE Int'l Symp. Quality Electronic Design (ISQED)*, 2010, pp. 9-14.

[8] N. Azizi. A. Moshovos and F.N. Najm, "Low-Leakage Asymetric-Cell SRAM," *Int'l Symp. Low Power Electronics and Design*, 2002, pp. 48-51.

[9] B. Alorda, G. Torrens, S.A. Bota, and J. Segura, "Static and Dynamic Stability Improvement Strategies for 6T CMOS Low-power SRAMs," *Proc. Int'l Symp. Circuits and Systems*, 2010, pp. 429-434.

[10] S. Basu, B. Kommineni, and R. Vemuri, "Mismatch Aware Analog Performance Macromodeling Using Spline Center and Range Regression on Adaptive Samples," *Proc. Int'l Conf. VLSI Design*, 2008, pp. 287-293.

[11] *Proc. 19th ACM Great Lakes Symp. on VLSI (GLSVLSI)*, 2009, pp. 441-444.

[12] K.A. Bowman and J.D. Meindl. "Impact of Within-die Parameter Fluctuations on Future Maximum Clock Frequency Distributions," *Proc. IEEE Custom Integrated Circuits Conf.*, 2001, pp. 229-232.

[13] J.A. Butts and G.S. Sohi, "A Static Power Model for Architects," *Proc. 33rd Annual IEEE/ACM Int'l Symp. on Microarchitecture (MICRO-33)*, 2000, pp. 191-201.

[14] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New Paradigm of Predictive MOSFET and Interconnect Modeling for Early Circuit Design," *Proc. IEEE Custom Integrated Circuits Conf.*. 2000, pp. 201-204.

[15] S. Bhardwaj, Y. Cao, S. Vrudhula, "Statistical leakage minimization of digital circuits using gate sizing, gate length biasing, and threshold voltage selection," *ASP Journal of Low Power Electronics*, vol. 2, no. 2, August 2006, pp. 240-250.

[16] Q. Chen, A. Guha, and K. Roy, "An Accurate Analytical SNM Modeling Technique for SRAMs Based on Butterworth Filter Function," *Proc. VLSI Design Conf.*, 2008, pp. 615-620.

[17] E. Seevinck et. al. "Static Noise Margin Analysis of MOS SRAM Cells," *IEEE J. Solid-State Circuits*, vol. 22, no. 5, 1987, pp. 748-754.

[18] L. Wei et. al. "Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications," *IEEE Transactions on VLSI Systems*, vol. 7, no. 1, 1999, pp. 16-24.

[19] M. Liu et. al. "Leakage Power Reduction by Dual-$V_{Th}$ Designs under Probabilistic Analysis of $V_{Th}$ Variation," *Proc. Int'l Symp. Low Power Electronics and Design*, 2004, pp. 2-7.

[20] S. Okumura et. al. “A 0.56-V l28kb l0T-SRAM Using Column Line Assist (CLA) Scheme,” *Proc. Int’l Symp. Quality Electronic Design,* 2009, pp. 659-663.

[21] 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,” *Proc. Int’l Symp. Quality Electronic Design,* 2009, pp. 172-178.

[22] C. Gopalakrishnan and S. Katkoori, “Knapbind: An Area Efficient Binding Algorithm for Low Leakage Datapaths,” *Proc. Int’l Conf. Computer Design,* 2003, pp. 430-435.

[23] _____, “Resource Allocation and Binding Approach for Low Leakage Power,” *Proc. Int’l Conf. VLSI Design,* 2003, pp. 297-302.

[24] F. Hamzaoglu, Y. Ye, A. Keshavarzi, K. Zhang, S. Narendra, S. Borkar, M.R. Stan, and V. De, “Analysis of Dual-$V_{Th}$ SRAM Cells with Full-swing Single-ended Bit Line Sensing for On-chip Cache,” *Proc. IEEE Trans. VLSI Systems,* 2002, pp. 91-95.

[25] J.G. Hansen, *Design of CMOS Cell Libraries for Minimal Leakage Currents,* Master's thesis, Dept. of Informatics and Mathematical Modeling, Computer Science and Engineering Technical University of Denmark, Fall, 2004.

[26] C.F. Hill, “Definitions of Noise Margin in Logic Systems,” *Mullard Technology Communications,* 1967, pp. 239-245.

[27] H. Noguchi, Y. Iguchi, H. Fujiwara, Y. Morita, K. Nii, H. Kawaguchi, and M. Yoshimoto, “A l0T Non-Precharge Two-Port SRAM for 74 Percent Power Reduction in Video Processing,” *Proc. IEICE Trans. (IEICET),* 2008, pp. 543-552.

[28] I.J. Chang, D. Mohapatra, and K. Roy, “A Voltage-scalable and Process Variation Resilient Hybrid SRAM Architecture for MPEG-4 Video Processors,” *Proc. Design Automation Conf. (DAC),* 2009, pp. 670-675.

[29] S. Hollis, J. Singh, D.K. Pradhan, and S.P. Mohanty, “A Single Ended 6T-SRAM Cell Design for Ultra-Low-Voltage Applications,” *IEICE Electronics Express,* vol. 5, no. 18, 2008, pp. 7.

[30] S. Jahinuzzaman, M. Sharifkhani, and M. Sachdev, “Investigation of Process Impact on Soft Error Susceptibility of Nanometric SRAMs Using a Compact Critical Charge Model,” *Proc. Int’l Symp. on Quality Electronic Design*, 2008, pp. 207-212.

[31] S.K. Jain and P. Agarwal, “A Low Leakage and SNM Free SRAM Cell Design in Deep Submicron CMOS Technology,” *Proc. VLSI Design Conf.*, 2006, pp. 495-498.

[32] J.T. Kao and A.P. Chandrakasan, “Dual-Threshold Voltage Techniques for Low-Power Digital Circuits,” *Processings of Solid-State Circuits*, vol. 35, no. 7, 2000, pp. 305-327.

[33] R. Keerthi and C.H Chen, “Stability and Static Noise Margin Analysis of Low-Power SRAM,” *Proc. Instrumentation and Measurement Technology Conf.,* 2008, pp. 1681-1684.

[34] K.S. Khouri and N.K. Jha, “Leakage Power Analysis and Reduction During Behavioral Synthesis,” *Proc. Int’l Conf. on Computer Design,* 2000, pp. 561-564.

[35] N.S. Kim, Y.J. Yoon, U.R. Cho, and H.G. Byun, “New Dynamic Logic-Level Converters for High Performance Applications,” *Proc. IEEE Int’l Symp. on Circuits and Systems*, 2003, pp. 93-96.

[36] E. Kougianos and S.P. Mohanty, “Metrics to Quantify Steady and Transient Gate Leakage in Nanoscale Transistors: NMOS Vs. PMOS Perspective,” *Proc. 20th IEEE Int’l Conf. on VLSI Design,* 2007, pp. 195-200.

[37] _____, “Impact of Gate-Oxide Tunneling on Mixed-Signal Design and Simulation of a Nano-CMOS VCO,” *Elsevier Microelectronics J.*, vol. 40,, no. 1, 2009, pp. 95-103.

[38] J.P. Kulkarni, K. Kim, S.P. Park, and K. Roy, “Process Variation Tolerant SRAM Array for Ultra Low Voltage Applications,” *Proc. IEEE Custom Integrated Circuits Conf.,* 2001, pp. 108-113.

[39] S.H. Kulkarni, A.N. Srivastava, and D. Sylvester, "A New Algorithm for Improved $V_{DD}$ Assignment in Low Power Dual $V_{DD}$ Systems," *Proc. Int'l Symp. on Low Power Electronics and Design,* 2004, pp. 200-205.

[40] A. Kumar and M. Anis, "Dual-$V_{Th}$ Design of FPGAs for Subthreshold Leakage Tolerance," *Proc. 7th Int'l Symp. on Quality Electronic Design,* 2006, pp. 735-740.

[41] R. Kumar and V. Kursun, "Reversed Temperature Dependent Propagation Delay Characteristics in Nanometer CMOS Circuits," *IEEE Trans. on Circuits and Systems-II,* vol. 53, no. 10, 2006, pp. 1078-1082.

[42] S. Lakshminarayanan, J. Joung, G. Narasimhan, R. Kapre, M. Slanina, J. Tung, M. Whately, C.-L. Hou, W.-J. Liao, S.-C. Lin, P.-G. Ma, C.-W. Fan, M.-C. Hsieh, F.-C. Liu, K.-L. Yeh, W.-C. Tseng, and S.W. Lu, "Standby Power Reduction and SRAM Cell Optimization for 65nm Technology," *Proc. Int'l Symp. on Quality Electronic Design,* 2009, pp. 471-475.

[43] D. Lee, D. Blaauw, and D. Sylvester, "Gate Oxide Leakage Current Analysis and Reduction for VLSI Circuits," *IEEE Trans. on VLSI Systems*, vol. 12, no.2, 2004, pp. 155-166.

[44] J. Lee and A. Davoodi, "Comparison of Dual-$V_{Th}$ Configurations of SRAM Cell Considering Process-Induced $V_{Th}$ Variations," *Proc. Int'l Symp. on Circuits and Systems*, 2007, pp. 3018-3021.

[45] J. Lee, L. Xie, and A. Davoodi, "A Dual-$V_{Th}$ Low Leakage SRAM Array Robust to Process Variations," *Proc. Int'l Symp. on Circuits and Systems,* 2008, pp. 580-583.

[46] F. Li, Y. Lin, L. He, and J. Cong, "Low Power FPGA Using Pre-defined Dual-$V_{DD}$/Dual-$V_t$ Fabrics," *Proc. Int'l Symp. on FPGAs,* 2004, pp. 42-50.

[47] Y. Li, M. Hempstead, P. Mauro, D. Brooks, Z. Hu, and K. Skadron, "Power and Thermal Effects of SRAM vs. Latch-Mux Design Styles and Clock Gating Choices," *Proc. Int'l Symp. Low Power Electronics and Design (ISLPED),* 2005, pp. 173-178.

[48] S. Lin, Y.B. Kim, and F. Lombardi, "A Low Leakage 9T-SRAM Cell for Ultra Low Power Operation," *Proc. ACM Great Lakes Symp. on VLSI,* 2008, pp. 123-126.

[49] Z. Liu and V. Kursun, "High Read Stability and Low Leakage Cache Memory Cell," *Proc. Int'l Symp. on Circuits and Systems,* 2007, pp. 2774-2777.

[50] _____, "High Read Stability and Low Leakage Cache Memory Cell," *Proc. Int'l Symp. on Circuits and Systems,* 2009, pp. 2774-2777.

[51] M. Mamidipaka, K. Khouri, N. Dutta, and M. Abadir, "Leakage Power Estimation in SRAMS," *CECS Technical Report 03-32*, Center for Embedded Computer Systems, U. California, Irvine, 2003.

[52] R.W. Mann, S. Nalam J. Wang, and B.H. Calhoun, "Limits of Bias-based Assist Methods in Nanoscale," *Proc. 11th IEEE Int'l Symp. on Quality Electronic Design (ISQED),* 2010, pp. 1-8.

[53] M. Meterelliyoz, J.P. Kulkarni, and K. Roy, "Thermal Analysis of 8T-SRAM for Nano-Scaled Technologies," *Proc. 13th Int'l Symp. on Low Power Electronics and Design,* 2008, pp. 123-128.

[54] M.Iijima, K. Seto, M. Numa, Tada, and T. Ipposhi, "Look-Ahead Dynamic Threshold Voltage Control Scheme for Improving Write Margin of SOI-7T-SRAM," *Proc. IEICE Trans. (IEICET),* 2008, pp. 2691-2694.

[55] S.P. Mohanty, "Unified Challenges in Nano-CMOS High-Level Synthesis," *Proc. 22nd Int'l Conf. on VLSI Design,* 2009, pp. 531-531.

[56] _____, *Energy and Transient Power Minimization During Behavioral Synthesis*, Ph.D. dissertation, U. South Florida, October, 2003.

[57] S.P. Mohanty and E. Kougianos, "Modeling and Reduction of Gate Leakage During Behavioral Synthesis of Nano-CMOS Circuits," *Proc. 19th Int'l Conf. on VLSI Design,* 2006, pp. 83-88.

[58] _____, “Steady and Transient State Analysis of Gate Leakage Current in Nanoscale CMOS Logic Gates,” *Proc. 24th IEEE Int’l Conf. on Computer Design (ICCD),* 2006, pp. 210-215.

[59] _____, “Simultaneous Power Fluctuation and Average Power Minimization During Nano-CMOS Behavioral Synthesis,” *Proc. 20th IEEE Int’l Conf. on VLSI Design (VLSID),* 2007, pp. 577- 582.

[60] S.P. Mohanty, E. Kougianos, D. Ghai, and P. Patra, “Interdependency Study of Process and Design Parameter Scaling for Power Optimization of Nano-CMOS Circuits under Process Variation,” *Proc. 16th ACM-IEEE Int’l Workshop on Logic and Synthesis (IWLS),* 2007, pp. 207-213.

[61] S.P. Mohanty, N. Ranganathan, E. Kougianos, and P. Patra, *Low-power High Level Synthesis for Nanoscale CMOS Circuits*, Springer Science and Business Media, 2008.

[62] D.C. Montgomery, *Design and Analysis of Experiments,* John Wiley and Sons, 6th edition, 2005.

[63] S. Mukhopadhayay and K. Roy, “Modeling and Estimation of Total Leakage Current in Nanoscaled CMOS Devices Considering the Effect of Parameter Variation,” *Proc. Int’l Symp. on Low Power Electronics and Design*, 2003, pp. 172-175.

[64] S. Nalam, V. Chandra, C. Pietrzyk, R.C. Aitken, and B.H. Calhoun, “Asymmetric 6T-SRAM with Two-phase Write and Split Bitline Differential Sensing for Low Voltage Operation,” *Proc. 11th IEEE Int’l Symp. on Quality Electronic Design (ISQED),* 2010, pp. 139-146.

[65] P. Pant, R.K. Roy, and A. Chatterjee, “Dual-Threshold Voltage Assignment with Transistor Sizing for Low Power CMOS Circuits,” *IEEE Trans. on VLSI Systems,* vol. 9, no. 2, 2001, pp. 390-394.

[66] A. Pavlov and M. Sachdev. *CMOS SRAM Circuit Design and Parametric Test in Nano-scaled Technologies: Process-aware SRAM Design and Test,* Springer Science and Business Media, 2008.

[67] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *CMOS: Digital Integrated Circuits: A Design Perspective*, 2nd edition, Prentice Hall, 2005.

[68] K. Roy, S. Mukhopadhyay, and H.M. Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proc. IEEE,* vol. 91, no. 2, 2003, pp. 305-327.

[69] F. Lombardi, S. Lin, Y.B. Kim, "Design and Analysis of a 32 nm PVT Tolerant CMOS SRAM Cell for Low Leakage and High Stability," *Integration, the VLSI Journal*, vol. 43, no. 2, 2010, pp. 176-187.

[70] A. Sasan, H. Homayoun, A.M. Eltawil, and F.J. Kurdahi, "Process Variation Aware SRAM/Cache for Aggressive Voltage-frequency Scaling," *Proc. Design Automation and Test in Europe*, 2009, pp. 911-916.

[71] A. Sil, S. Ghosh, N. Gogineni, and M. Bayoumi, "A Novel High Write Speed, Low Power, Read-SNM-Free 6T-SRAM Cell," *Proc. Midwest Symp. on Circuits and Systems*, 2008, pp. 771-774.

[72] F. Sill, J. You, and D. Timmerman, "Design of Mixed Gates for Leakage Reduction," *Proc. 17th Great Lakes Symp. on VLSI,* 2007, pp. 263-268.

[73] J. Singh, D.S. Aswari, S.P. Mohanty, and D.K. Pradhan, "A 2-Port 6T-SRAM Bitcell Design with Multi-Port Capabilities at Reduced Area Overhead," *Proc. 11th IEEE Int'l Symp. on Quality Electronic Design (ISQED)*, 2010, pp. 131-138.

[74] J. Singh, J. Mathew, S.P. Mohanty, and D.K. Pradhan, "A Nano-CMOS Process Variation Induced Read Failure Tolerant SRAM Cell," *Proc. Int'l Symp. on Circuits and Systems*, 2008, pp. 3334-3337.

[75] J. Singh, J. Mathew, D.K. Pradhan, and S.P. Mohanty, "A Subthreshold Single Ended I/O SRAM Cell Design for Nanometer CMOS Technologies," *Proc. IEEE Int'l SOC Conf. (SOCC),* 2008, pp. 243-246.

[76] N. Srisantana and K. Roy, "Low Power Design Using Multiple Channel Lengths and Oxide Thicknesses," *IEEE Design and Test of Computers*, vol. 21, no. 1. 2004, pp. 56-63.

[77] A. Srivastava, "Simultaneous $V_{Th}$ Selection and Assignment for Leakage Optimization," *Proc. Int'l Symp. on Low Power Electronics and Design,* 2003, pp. 146-151.

[78] A. Srivastava, D. Sylvester, and D. Blaauw, "Power Minimization Using Simultaneous Gate Sizing, Dual-$V_{DD}$ and Dual-$V_{Th}$ Assignment," *Proc. Design Automation Conf. (DAC),* 2004, pp. 783-787.

[79] P.A. Stolk, F.P. Widdershoven, and D.B.M. Klaassen, "Modeling Statistical Dopant Fluctuations in MOS Transistors," *IEEE Trans. on Electron Devices*. vol. 45, no. 9, 1998, pp. 1960-1971.

[80] S. Tavva and D. Kudithipudi, "Variation Tolerant 9T-SRAM Cell Design," *Proc. ACM Great Lakes Symp. on VLSI,* 2010, pp. 55-60.

[81] S.A. Tawfk and V. Kursun, "Low Power and Robust 7T Dual-$V_{Th}$ SRAM Circuit," *Proc. Int'l Symp. on Circuits and Systems,* 2008, pp. 1452-1455.

[82] G. Thakral, S.P. Mohanty, D. Ghai, and D.K. Pradhan, "A DOE-ILP Assisted Conjugate-Gradient Approach for Power and Stability Optimization in High -κ/Metal-Gate SRAM," *Proc. 20th ACM-IEEE Great Lakes Symp. on VLSI (GLSVLSI),* 2010, pp. 323-328.

[83] _____, "P3 (Power-Performance-Process) Optimization of Nano-CMOS SRAM Using Statistical DOEILP," *Proc. 11th IEEE Int'l Symp. on Quality Electronic Design (ISQED),* 2010, pp. 176-183.

[84] G. Thakral, S.P. Mohanty, D.K. Pradhan, and E. Kougianos, "DOE-ILP Based Simultaneous Power and Read Stability Optimization in Nano-CMOS SRAM," *ASP J. Low Power, Electronics (JOLPE),* vol. 6, no. 3, 2010, pp. 1-11.

[85] 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," *Proc. 23rd IEEE Int'l Conf. on VLSI Design,* 2010, pp. 45-50.

[86] Y. Yang and Z. Qian, *A Novel Low Power SRAM/SOI Cell Design*, 2008, http://www.paper.edu.cn/.

[87] W. Zhao and Y. Cao, "New Generation of Predictive Technology Model for Sub-45 nm Design Exploration," *Proc. Int'l Symp. on Quality Electronic Design,* 2006, pp. 585-590.