ROBUST OPTIMIZATION OF NANOMETER SRAM DESIGNS

A Thesis

by

AKSHIT DAYAL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2009

Major Subject: Computer Engineering

ROBUST OPTIMIZATION OF NANOMETER SRAM DESIGNS

A Thesis

by

AKSHIT DAYAL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Peng Li
Committee Members,	Jiang Hu
	Duncan M. Walker
Head of Department,	Costas Georghiades

December 2009

Major Subject: Computer Engineering

ABSTRACT

Robust Optimization of Nanometer SRAM Designs. (December 2009)

Akshit Dayal, B.E., Delhi College of Engineering, University of Delhi;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Peng Li

Technology scaling has been the most obvious choice of designers and chip manufacturing companies to improve the performance of analog and digital circuits. With the ever shrinking technological node, process variations can no longer be ignored and play a significant role in determining the performance of nanoscaled devices. By choosing a worst case design methodology, circuit designers have been very munificent with the design parameters chosen, often manifesting in pessimistic designs with significant area overheads.

Significant work has been done in estimating the impact of intra-die process variations on circuit performance, pertinently, noise margin and standby leakage power, for fixed transistor channel dimensions. However, for an optimal, high yield, SRAM cell design, it is absolutely imperative to analyze the impact of process variations at every design point, especially, since the distribution of process variations is a statistically varying parameter and has an inverse correlation with the area of the MOS transistor. Furthermore, the first order analytical models used for optimization of SRAM memories are not as accurate and the impact of voltage and its inclusion as an input, along with other design parameters, is often ignored.

In this thesis, the performance parameters of a nano-scaled 6-T SRAM cell are modeled as an accurate, yield aware, empirical polynomial predictor, in the presence of intra-die process variations. The estimated empirical models are used in a constrained non-linear, robust optimization framework to design an SRAM cell, for a 45 nm CMOS technology, having optimal performance, according to bounds specified for the circuit performance parameters, with the objective of minimizing on-chip area. This statistically

aware technique provides a more realistic design methodology to study the trade off between performance parameters of the SRAM.

Furthermore, a dual optimization approach is followed by considering SRAM power supply and wordline voltages as additional input parameters, to simultaneously tune the design parameters, ensuring a high yield and considerable area reduction. In addition, the cell level optimization framework is extended to the system level optimization of caches, under both cell level and system level performance constraints.

DEDICATION

To My Parents

ACKNOWLEDGEMENTS

This thesis arose as a part of my Master of Science program in Computer Engineering at Texas A&M Engineering. The two years that I have spent here have given me a wonderful opportunity to interact with a great number of people who have been repositories of ideas and advice which infused in me a passion towards research. It is a pleasure to convey my gratitude to them in my humble acknowledgement.

I would like to express gratitude to Dr. Peng Li for his supervision advice and guidance from the very early stage of this research as well as giving me extraordinary experiences throughout the work. Above all and the most needed, he provided me unflinching encouragement and support in various ways. His intuition in the subject has made him a constant oasis of ideas and passion in VLSI CAD, which exceptionally inspired and enriched my growth as a student and a researcher. I am indebted to him more than he knows.

I am also thankful to the committee members, Dr. Jiang Hu and Dr. Duncan Walker, for their suggestions and the wonderful courses offered by them. I would also like to thank the faculty members in the computer engineering group for the courses offered by them. The courses that I have taken in graduate school enabled me to build a solid foundation in VLSI and have been an integral part in my research exploration.

I would also like to thank my friends for being the sounding board of my ideas, and offering me their insights. This enabled me to explore an idea based on unbiased opinion and a fresh approach. I would also like to thank them for being supportive in bad times and an integral part of my good times.

Where would I be without my family? My parents deserve a special mention for their inseparable support, prayers and unflinching love. My father, Ashok Dayal, in the first place is the person who helped instill in me a competitive spirit, showing me the joy of intellectual pursuit ever since I was a child. My mother, Neelam Dayal, is the one who sincerely raised me with her caring and gentle love. To my brother Amit, bhabhi Aashita and their son, Devansh, thanks for being ever caring and supportive, especially during my rough times.

TABLE OF CONTENTS

Page

ABSTRACT iii

DEDICATION v

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS vii

LIST OF FIGURES x

LIST OF TABLES xiii

I. INTRODUCTION 1

1. Motivation 1
   1.1 Technology Scaling 1
   1.2 Negative Impact of Technology Scaling on SRAM Performance 2
2. Prior Work 3
   2.1 Nominal/Deterministic Design Techniques 4
   2.2 Worst Case Analysis 4
   2.3 Monte-Carlo Analysis 5
   2.4 Statistical Analysis 6
3. Research Objectives 7

II. DESIGN PARAMETERS FOR SRAM MEMORY CELLS 10

1. Standby Leakage Power 10
   1.1 Subthreshold Leakage 11
   1.2 Gate Leakage 12
2. Static Noise Margin 14
   2.1 Static Read Noise Margin (RNM) 14
   2.2 Static Write Noise Margin (WNM) 16
3. Read Access Time 17
4. Write Access Time 18

III. PROCESS VARIATIONS AND YIELD AWARE RESPONSE SURFACE MODELING 19

1. Modeling Process Variations 19

Page

2. Variation Aware Response Surface Modeling 21

IV. STATISTICAL MODELING OF SRAM PERFORMANCE PARAMETERS. 26

1. Cell Level SRAM Performance Modeling 26
1.1 Standby Leakage Current 26
1.1.1 Subthreshold Leakage 27
1.1.2 Gate Leakage 29
1.1.3 Total Leakage 30
1.2 Static Read Noise Margin (RNM).. 32
1.3 Static Write Noise Margin (WNM)... 35
1.4 Read Access Time 38
1.5 Write Access Time 40
2. Statistical Modeling of Performance Parameters for Caches 41
2.1 Cache Leakage Power 43
2.2 Cache Access Time 47

V. IMPACT OF DESIGN PARAMETERS ON SRAM PERFORMANCE........... 49

1. Sizing the Pull-up Transistor (Wp) 49
2. Sizing the Access Transistor (Wax) 50
3. Sizing the Pull-down Transistor (Wn) 51
4. Supply Voltage (V_{dd}).. 51
5. Wordline Voltage (V_{wl}).. 51

VI. PROBLEM FORMULATION AND DESIGN PROCEDURE......................... 53

1. Problem Description.. 53
2. Stochastic Design Optimization 56
2.1 Cell Level Optimization for a 6-T SRAM Memory Cell 56
2.2 Dual-Optimization of a 6-T SRAM Memory Cell with Voltage Tunability... 61
2.3 System Level Optimization of Caches 62
2.3.1 SRAM System Level and Cell Level Performance Parameter Interactions... 63
2.3.2 System Level Optimization Methodology 64

VII. EXPERIMENTAL RESULTS 66

1. Sensitivity Analysis of SRAM Performance Yields 66
2. Optimization Results 69
2.1 Cell Level Optimization 69
2.2 Optimization of a Voltage Tunable SRAM Cell 73

Page

2.3 System Level Optimization of Caches 76

VIII. FUTURE WORK 78

1. Dynamic vs Static 78
2. Dynamic Noise Margin 80
   2.1 Read Dynamic Noise Margin 80
   2.2 Write Dynamic Noise Margin 81
3. Preliminary Optimization Results 82
   3.1 SRAM Cell Level Optimization: Dynamic Perspective 82

IX. CONCLUSIONS 84

REFERENCES 86

VITA 90

LIST OF FIGURES

Page

Figure 2.1 Subthreshold leakage components in a 6-T SRAM cell in the retention mode 11

Figure 2.2 Gate tunneling components in a MOSFET 12

Figure 2.3 Gate leakage components in the retention mode of the SRAM 13

Figure 2.4 Two cross-coupled inverters with static-noise voltage source V_{n1} and V_{n2} for a 6-T SRAM cell........ 15

Figure 2.5 Read static noise margin (RNM) for a symmetric 6-T SRAM cell .. 15

Figure 2.6 SRAM during a write operation 16

Figure 2.7 Write noise margin 17

Figure 3.1 Flow chart depicting the response surface methodology flow used for circuit parameters........ 25

Figure 4.1 SRAM cell during the data retention mode........ 27

Figure 4.2 Non-central distribution of the logarithm of leakage current derived by varying process variables, in HSPICE simulation........ 32

Figure 4.3 Distribution of RNM on varying process variables 33

Figure 4.4 Characterization of RNM on varying Vth and Tox........ 33

Figure 4.5 Monte-Carlo sampling of WNM on varying process variables 36

Figure 4.6 Characterization of WNM on varying Vth and Tox 36

Figure 4.7 Comparison of HPSICE predicted and polynomial predicted values using sample test points........ 37

Figure 4.8 (a) Non-central distribution of read access time and (b) Gaussian distribution of inverse read access time 39

Page

Figure 4.9 Characterization of read access time on varying V_{th} and T_{ox} of all the 6 Transistors 39

Figure 4.10 Comparison of HPSICE predicted and polynomial predicted values using sample test points and the distribution of relative error 40

Figure 4.11 Model vs HSPICE fit for write access time. 41

Figure 4.12 Total cache leakage power as evaluated by a modified CACTI 5.2 under the influence of process variations 45

Figure 4.13 Total cache leakage power, a comparison between model evaluated and modified HSPICE-CACTI setup 46

Figure 4.14 Fit between RSM predicted model and CACTI-5.2 for Cache access time 48

Figure 5.1 A 6-T SRAM cell storing a “1” 49

Figure 6.1 A schematic of a 6-T SRAM memory cell 54

Figure 6.2 Top level representation of the optimization loop of the NLP problem formulated 62

Figure 7.1 Experimental setup of the generalized optimization flow methodology 67

Figure 7.2 Effect of transistor widths on yield of SRAM performance 68

Figure 7.3 Yield values for every iteration of the optimization process 71

Figure 7.4 Cell area for every iteration of the optimizer 71

Figure 7.5 SRAM transistor widths: pull-up (Wp), access (Wax) and pull-down (Wn) at every iteration of the optimizer run 72

Figure 7.6 Yield values for every iteration of the optimization process of a tunable SRAM cell 74

Figure 7.7 Cell area for every iteration of the optimizer for a tunable SRAM cell 75

Page

Figure 7.8 SRAM transistor widths: pullup (Wp), access (Wax) and pull-down (Wn) at every iteration of the optimizer run for a tunable SRAM cell.. 75

Figure 7.9 SRAM voltages: supply voltage (Vdd) and wordline voltage (Vwl) 75

Figure 8.1 (a) Read dynamic noise margin, (b) Write dynamic noise margin .. 81

LIST OF TABLES

Page

Table I Fitted Parameters Using Non Linear Regression 31

Table II Leakage Power of a 45nm L-2 Cache 46

Table III Performance Results of Cell Level Optimization 70

Table IV Optimized Design Parameters of a SRAM Cell 70

Table V Performance Results of a Voltage Tunable SRAM Cell 73

Table VI Optimized Design Parameters of a Voltage Tunable SRAM Cell 74

Table VII Performance Results of a L-2 Cache RAM 76

Table VIII Optimized Design Parameters of a L-2 Cache 76

Table IX Performance Threshold of a 6-T SRAM Cell 82

Table IX Optimization Results for Dynamic Constraints 82

I. INTRODUCTION

1. MOTIVATION

In 1965 Intel co-founder Gordon E. Moore had described a law which would be a long term indicator of the technological development of the functional blocks of digital circuits. One of the most popular formulation of Moore's law pertains to the transistors per IC and the projection that it would double every two years. Astoundingly, that law, according to recent [1] articles published by Intel, has maintained its relevance even till 2007 and Moore's Law remains an upper bound on the number of transistors that can be packed into a micro-processor. It is also the basis for International Technology Roadmap for semiconductors (ITRS) [2] to identify the challenges and demands the semiconductor industry is going to face over the next 15 years. Furthermore, with the advancement in lithographic techniques, the die-size of the microprocessors has progressively reduced over the years. The sacrosanctity of moore's hypothesis is maintained mainly due to the advancement in techniques used to scale the channel dimensions of MOS transistors, a technique known as *Technology Scaling,* which makes it possible to manufacture commercially viable, high performance microprocessors, containing millions of transistors and clocking at speeds, in excess of 3 GHZ.

1.1 TECHNOLOGY SCALING

One of the most obvious ways of etching more number of transistors on a fixed size wafer is to reduce the dimensions of the transistors, or in popular lexicon, by *technology scaling.* Scaling of CMOS transistors in next generation technologies leads to improved performance, lower power consumption and increased transistor density. Typically, technology scaling has the following goals; these are well established in theory [3] and are not ad hoc.

This thesis follows the style of *IEEE Transactions on Circuits and Systems*

Reduce the delay by 30%, resulting in an increase in operating frequency of about 43%;

a. Double transistor density; and
b. Reduce energy per transition by about 65%, saving 50% of power, at a 43% increase in frequency.

Motivated by the significant CMOS performance improvements enabled by technology scaling, has resulted in a larger on-chip memory for micro-processors, reducing read/write latency and faster, compact computing systems. SRAMS, being faster and less power hungry than DRAMS, have always been the primary choice of primary on-chip memory like L1 and L2 caches in microprocessor. It is therefore used where either bandwidth or low power, or both, are principal considerations. SRAM is also easier to control (interface to) and generally more truly *random access* than modern types of DRAM [4]. Moreover SRAM based caches occupy nearly 50 % of on chip area in modern micro-processors, therefore, scaling of SRAMS greatly affects the performance of caches and is an important criterion, while evaluating the overall performance of micro-processors.

Ideally, to ensure a high performing processor, on-chip cache should be as large as possible, capacity wise. Technology scaling is a godsend to designers and customers, which enables packing millions of transistors on a chip, the size of a small postal stamp, to ensure a high performing micro-processor and a significant on-chip memory storage capability. Unfortunately, SRAM design is becoming increasingly challenging with each new technology node its impact on SRAM performance is described subsequently.

1.2 NEGATIVE IMPACT OF TECHNOLOGY SCALING ON SRAM PERFORMANCE

The scaling to deep sub-micron CMOS technologies brings with it many pitfalls which can potentially offset the gains due to scaling. One of the preferred methods of scaling is the constant electric field (supply voltage scaling) [3] as it gives lower-energy delay product thus the trade-off's between performance and power consumption are minimal. However, this reduction in supply voltage results in increased sensitivity of

circuit parameters to process variations [5]. According to ITRS [2], the threshold voltage V_{th} variation for a 45 nm CMOS is going to be nearly 42 %, channel length variation , L_g, of 12 % and gate oxide thickness, T_{ox}, variation of 4%.

The dependence on process variations, degrades and limits circuit operations in the low voltage regime, particularly for SRAM cells where minimum sized transistors are used. The reduced geometry transistors are vulnerable to inter-die and intra-die process variations. Intra-die process variations include random dopant fluctuations (RDF), line edge roughness (LER) etc. This results in threshold voltage mismatch between the adjacent transistors in a memory cell giving asymmetric characteristics to the SRAM cell.

As mentioned in [3], voltage scaling accompanies the reduction in MOS transistor dimensions, with the ever decreasing technological node. The reduction in voltage with the reducing technological node requires the commensurate scaling of threshold voltage, thus reducing the turn-on voltage for the MOS transistors. The cumulative effect of reduced voltage, threshold voltage and large process variations leads to increased memory failures such as *read failure*, *write failure*, *hold failure*, and *access time failure*. In addition, the reduced oxide thickness and the reduced Subthreshold voltage values, along with process variations contribute immensely towards the Subthreshold and gate leakage components of the total leakage in an SRAM cell, greatly degrading memory performance. Since, most of the time SRAM memories are used to store and sustain data without being in the active mode of operation; it becomes imperative to control leakage power of SRAM's in the static mode of operation, to limit the total power dissipated by on-chip memories. Thus, increasing battery life of mobile electronic gadgets, of which SRAM's are an integral memory storage component.

2. PRIOR WORK

Significant time and effort has gone into optimization of digital/analog circuits. The basic idea of all circuit analysis procedures is to ensure a high performance of the

circuit while minimizing the computational cost, design overheads like area, leakage power etc, and finally to reduce the time and effort put in by the designer. Keeping these objectives, in mind the methodology followed by designers to optimally design a circuit are broadly categorized as following.

2.1 NOMINAL/DETERMINISTIC DESIGN TECHNIQUES

The Nominal or deterministic approach refers to the circuit design technique wherein parametric variations are not taken as a factor of consideration, when the circuit is designed. The trade offs between functionality is analyzed for every design point. The authors in [6] highlight one such approach for a 6-T SRAM cell which is used to determine initial cell design. The futility of this approach can be considered by a simple observation that one of the circuit performance parameters like Subthreshold current has an exponential dependence on threshold voltage V_{th} variation. Hence a small change in process parameters will result in a huge change in leakage current value [7]. The deterministic approach is very optimistic and can lead to a significant yield loss in sub 100nm technologies where the variability plays a dominant role in determining circuit performance, thereby, rendering this method quite ineffective for sub 90 nm CMOS designs.

2.2 WORST CASE ANALYSIS

As the name suggests, worst case approach deals with analyzing the circuit performance at every worst case condition of its circuit performance parameters, individually. Traditionally, this has been the method of choice for designers for a number of years due to ease of design and minimal computational cost. In this method, designers consider the effect of process variations by assuming worst case device characteristics, usually 2-3 σ from the typical or nominal value and identify design problems arising as a result of parametric variations. The design is then verified with respect to the specifications and a redesign is done for the corner cases till all the specifications are met. However, the major problem with this method is that rarely does

design corners of circuit performance, occurs simultaneously. A case in point being the read and write margins of a SRAM cell. A design corner guaranteeing a good read stability can have extremely degraded write stability, as they are two contradictory performance criteria. The authors in [6] conclude that the worst case optimization approach for an SRAM memory cell, overestimates the underlying process variations, which leads to increased leakage power consumption. Furthermore, the settings of all threshold voltage, V_{th}, parameters to their worst case values (Nominal $\pm$ 3 σ) rarely occur in reality. Hence, this one-at-a-time corner analysis yields very pessimistic results, leads to over design and potentially increased area and leakage power over heads.

2.3 MONTE-CARLO ANALYSIS

Monte Carlo simulation is a method of simulation with unknown variables. In Monte Carlo simulation values for unknowns are randomly selected according to their statistical distribution. The process is repeated for a number of simulation runs, each with a new set of values for the unknowns. The distribution of the final results is taken to be representative of the behavior over the range of inputs. In circuit analysis, Monte Carlo simulation is a popular method of dealing with the large number of correlated and uncorrelated variables involved in circuit design. Process parameters can be characterized as a distribution of transistor behaviors giving the designer a large amount of data to deal with. Monte Carlo simulation allows all of these variables to be considered during simulation. Monte Carlo simulation is frequently used to give inputs to statistical timing software or to predict circuit yield and sensitivity. This method provides the most realistic estimate of true worst case performance of the circuit and serves as a benchmark against which all modeling and analysis techniques are tested for accuracy. In Monte-Carlo analysis the, the error in estimation reduces with the number of samples n as $O(n^{-1/2})$. Thus the number of simulations, n, needed to obtain a good accuracy is large and a Monte-Carlo based optimization method is too computationally demanding and time consuming [8].

2.4 STATISTICAL ANALYSIS

The presence of parametric variations demands a more comprehensive approach to circuit design, without the inaccuracies, design time involved and computational overheads, as mentioned in previous sections. Statistical analysis for optimization deals with developing failure probability and yield prediction models for circuit performance parameters in the presence of process variation. They are then used to analyze the combined failure probability of the circuit performance parameters as a function of design parameters and process parameters.

In the domain of statistical analysis, various methods for modeling the performance parameters exist. The can be broadly categorized as (a) *analytical* and (b) *empirical models*. The analytical models are generally first order approximations and lack in accuracy. However they can be used for initial analysis of the circuit. Empirical models, on the other hand can be any order. Higher the order, higher is the accuracy. The authors in [9] propose one such method to analyze the stability of SRAM cells in the presence of random fluctuations in the device parameters. According to their analysis, they develop linear models for the Read, write and access time models for circuit analysis. As an example the model for read noise margin *(RNM)* is shown in (1.1)

$$RNM = RNM_0 + \sum_{i=1}^{6} k_i \left(\Delta V_{th} \right)_i \quad (1.1)$$

However, the circuit performance parameters are seldom linear [10] and higher order models, along with the effects of other sources of random variations (like gate oxide thickness) needs to be analyzed, to optimally and accurately predict impact of parametric variations.

The authors in [6] propose a similar approach for a 6-T SRAM cell, wherein all the performance models are considered linear and an expression for the mean and variance values are extracted from them. Moreover, they only consider the effect of Subthreshold leakage as their objective function, completely ignoring the effect of gate leakage. The approach outlined by [6] can lead to inaccurate results on three accounts.

Firstly, on the basis of linear empirical models, secondly, on ignoring gate leakage from the objective function and thirdly, by not considering other sources of parametric variations, which can have a significant impact on SRAM performance and stability. The impact of the second point can be analyzed from the fact that gate leakage, for sub 65 nm technologies, contributes to nearly 45% of total device leakage [7].

Recently, the authors in [11] , [12] proposed an extremely elegant approach to SRAM circuit optimization, taking all the SRAM circuit performance parameters (read, write, access time and cell leakage) into account and also considering the effect of gate leakage component in the total cell leakage. However, their models donot consider other sources of variations as mentioned above. Furthermore, none of [6], [9], [11]-[14] consider the effect of varying the voltage source like cell voltage and wordline voltage, on performance of the SRAM memory cell. In effect the cell voltage can be taken as inputs along with the CMOS design parameters, in the optimization process to comprehensively analyze the SRAM memory cell performance.

3. RESEARCH OBJECTIVES

In view of the detrimental effects to technology scaling, it becomes increasingly difficult for circuit designers to design a reliable and robust SRAM memory cell, which has an optimum performance at every design point. The major driving point of this thesis is to develop a technology-aware design methodology which takes into account, the variation in process parameters at every design point. Thus a fast, yield-aware and robust flow would be formulated which optimizes the design of the SRAM memory cell to ensure a low failure probability of the cell, ensuring lower probability of read, write, hold and access time failures, while simultaneously optimizing the on-chip area and reducing leakage current for every design.

The performance parameters of a nano-scaled 6-T SRAM cell are modeled as an accurate, yield aware, empirical polynomial predictor, in the presence of intra-die process variations. The estimated empirical models are used in a constrained non-linear, robust optimization framework to design an SRAM cell, for a 45 nm CMOS technology,

having optimal performance, according to bounds specified for the circuit performance parameters, with the objective of minimizing on-chip area. This statistically aware technique provides a more realistic design methodology to study the trade off's between the performance parameters of the SRAM.

Furthermore, a dual optimization approach is followed by considering SRAM power supply and wordline voltages as additional input parameters, to simultaneously tune the design parameters, ensuring a high yield and considerable area reduction. In addition, the cell level optimization framework is extended to the system level optimization of caches, under both cell level and system level performance constraints.

In view of the above discussion an ideal circuit optimization method is one which has lower computational cost, less design time and low overheads, while ensuring a high circuit performance and low failure probability. In this research proposal, a robust technology aware statistical optimization flow is presented to optimize a 6-T SRAM memory cell with 3 broad objectives in mind:

a) To develop a yield aware optimization framework for a *6-T SRAM cell*, with the Widths of the MOS transistors being the design variables, under the constraints of the failure probability of SRAM circuit performance parameters $P(Y_{cell})$. The 5 SRAM performance parameters considered are Static Read Noise Margin, Static Write noise Margin, Read Access Time, Write access time and Leakage power dissipation.

b) To develop a yield aware optimization framework for a *tunable 6-T SRAM cell*, with the dimensions of the MOS transistors, the cell Voltage supply and the Wordline voltages being the design variables, under the constraints of the failure probability of SRAM circuit performance parameters $P(Y_{cell})$. The 5 SRAM performance parameters considered are Static Read Noise Margin, Static Write noise, Read Access Time, Write access time and Leakage power dissipation.

c) The cell level optimization framework in (a) is extended to the system level optimization of caches, given the failure probabilities of cell level, $P(Y_{cell})$, and system level performance constraints $P(Y_{system})$. This would be achieved by modifying the CACHE performance analysis tool CACTI 5.2, developed by HP labs. The modified circuit parameters involves constraint evaluation at the cell level using circuit simulators like HSPICE and for a memory cache deploying those cells, using CACTI 5.2.

II. DESIGN PARAMETERS FOR SRAM MEMORY CELLS

A memory unit in any electronic device should ensure extremely high reliability of operation and data storage. Thus, a memory cell must be designed very carefully for reliable operation while simultaneously ensuring a low overhead of on-chip area and power consumption. In a 6-T SRAM cell, there are three operation modes of which two are active (read and write operation) and one is passive (hold or the retention mode). In this section the design parameters for various operation modes of a 6-T SRAM memory cell are lucidly described.

1. STANDBY LEAKAGE POWER

Standby leakage power has become an important constraint in today's processor design. According to the figures posted by International Technology Roadmap for Semiconductors (ITRS) in [2], the leakage power is set to become a dominant source of power dissipation in sub 90 nm technologies. With the technology node fast approaching the 45 nm mark, the leakage power is expected to dominate the dynamic switching power and account for more than 50% of the total chip power. As a result, Standby leakage power becomes an important constraint to be considered and controlled while designing digital circuits.

With decreasing technology node, supply voltage is continually scaled to reduce the dynamic power dissipation. The continued voltage scaling offsets in lowered device speeds. To compensate this decrease, the threshold voltage, Vth, is reduced commensurately. The reduced V_{th} in turn has a large impact on standby leakage power of the device, thereby increasing the ratio of standby leakage power to the total leakage power. The major components of leakage power in current generation CMOS technologies are sub-threshold leakage and gate leakage. Sub-threshold leakage was typically the dominant component of leakage as I_{Sub} was a significant percentage of the total leakage current. As the gate length of MOSFET's continued to be scaled down in the sub 100nm regime, gate oxide thickness have value of less than 20 Å. Consequently,

it yields in significant gate leakage current by various tunneling mechanisms, which have undesirable effects on standby current and memory operation.

1.1 SUBTHRESHOLD LEAKAGE

Subthreshold leakage or subthreshold drain conduction refers to the current that flows between the source and drain of a MOSFET when the gate-to-source voltage is below the threshold voltage (V_{th}). This region of operation of a transistor is also referred to as the subthreshold region of operation. The transistor is essentially considered to be turned "off". The commonly used model of Subthreshold leakage current (I_{Sub}), through a transistor is [15]

$$I_{Sub} = I_{ds0} e^{(V_{gs} - V_{th})/\eta V_T} \left(1 - e^{-V_{ds}/V_T}\right) \tag{2.1}$$

$$I_{ds0} = \beta V_T^2 e^{1.8} \tag{2.2}$$

$$\beta = \mu_0 C_{ox} (W_{eff} / L_{eff}) \tag{2.3}$$

Where, V_{gs} and V_{ds} are the gate-to-source and drain-to-source bias voltages respectively, V_T is the Boltzmann constant, μ_0 is the zero bias electron mobility, C_{ox} is the gate oxide capacitance, W_{eff} and L_{eff} are the effective transistor width and length respectively.

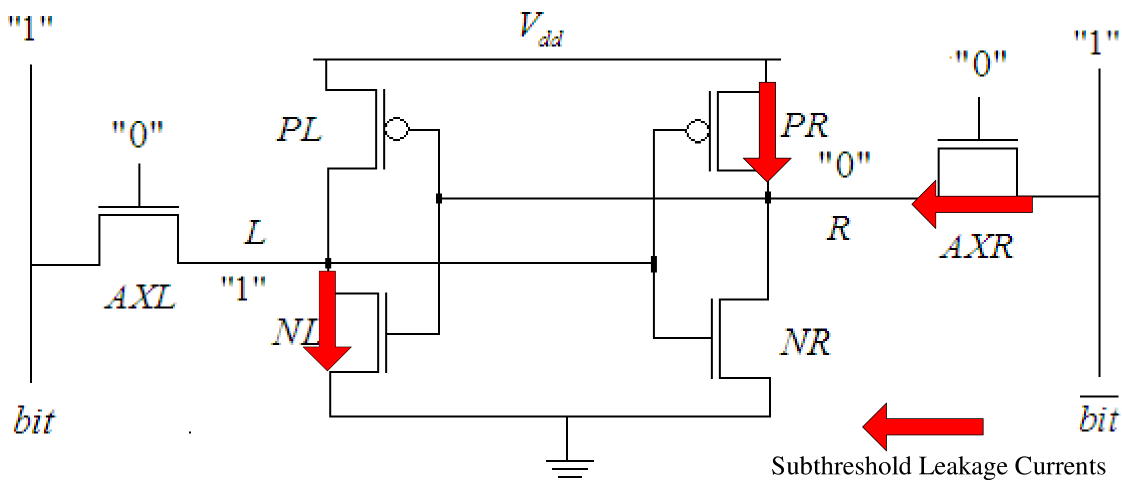

Fig. 2.1 Subthreshold leakage components in a 6-T SRAM cell in the retention mode.

In Fig. 2.1(a), the SRAM cell is operating in the retention mode, the node L stores a "1" and node R stores a "0". The Wordline is turned off and the bit and bit-bar lines are precharged high. The pull-up transistor PR, the pull down transistor NL and the access transistor AXR are turned "off" and contribute towards the Subthreshold leakage current of the circuit.

1.2 GATE LEAKAGE

Gate leakage is the other contributing factor towards the total standby leakage current of the memory cell, in the retention mode. Gate dielectric leakage current becomes a serious concern as sub-20Å gate oxide prevails in advanced CMOS processes.

When the oxide thickness of a device is reduced there is an increase in the amount of carriers that can tunnel through the gate oxide, from the bulk silicon to the gate. For NMOS transistor the majority carriers are electrons and for PMOS transistors it is holes. This offsets itself in gate leakage current. Gate tunneling current as shown in Fig. 2.2 consists of three components:

- Gate to source and gate to drain overlap current (I_{gso} and I_{gdo}).
- Gate to channel (I_{gc}), a part of which goes to the source (I_{gcs}) and rest to the drain (I_{gcd}).
- Gate to substrate current (Igsb)

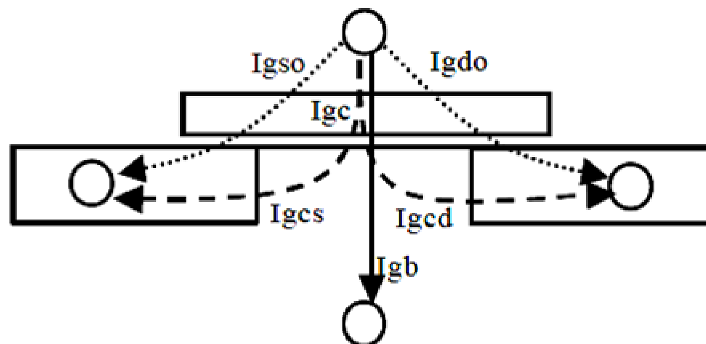

Fig. 2.2. Gate tunneling components in a MOSFET.

In bulk CMOS technology, the gate to substrate leakage current is several orders of magnitude lower than the overlap tunneling current and gate to channel current [16]. On the other hand, while the overlap tunneling current dominates the gate leakage in the "OFF" state, gate to channel tunneling dictates the gate current in the "ON" condition. Since the gate to source and gate to drain overlap regions are much smaller than the channel region, the gate tunneling current in the "OFF" state is much smaller than gate tunneling in the "ON" state [16]. The total gate leakage current I_{gate} is linearly dependent on the area of the device and has an exponential relationship with oxide thickness. Consequently, with technology scaling and voltage scaling with sub 100 nm CMOS devices, there is a subsequent reduction in oxide thickness, which manifests itself in increased leakage current.

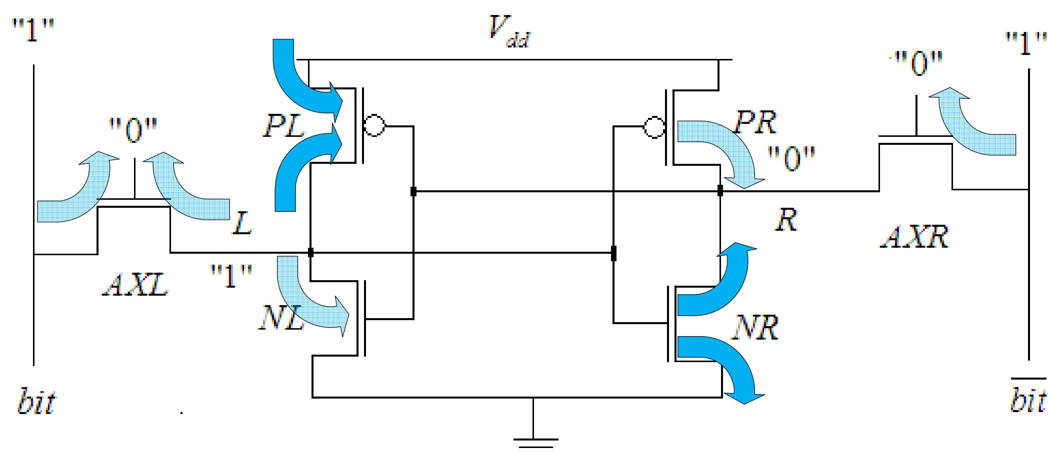

Fig. 2.3 Gate Leakage components in the retention mode of the SRAM.

The gate leakage current components of a 6-T SRAM memory cell are shown in Fig 2.3. As previously stated, the CMOS devices in the "ON" state would contribute more to gate leakage than in the "OFF" state. Also from fig 2, for the node L storing a "1", pull-up transistor PL and the pull-down transistor NR of the cross-coupled inverters are in the "ON" state. Therefore, the gate leakage current components are predominant through these two MOS transistors.

2. STATIC NOISE MARGIN

In Static Noise analysis, the inputs and outputs of the cross-coupled inverters in an SRAM or any other logic circuits are assumed to be DC signals, Fig 2.4. It is widely considered as a good measure to analyze the stability of a logic circuit [17], [18]. Static noise margin is defined as the unity gain points on the voltage transfer curve of a logic gate as illustrated in Fig 2.5.

Some of the reliable methods for reliable static noise margin estimation, widely used in literature are

- Unity small signal loop gain
- Maximum Square method
- Coincidence of the roots of the flip-flop equation
- Jacobian of the Kirchhoff equations is zero

For CMOS logic gates one of the most widely used criteria for accurately estimating static noise margin is the maximum square method [19]. In this method, the static noise margin is given by the sides of the largest square that can be inscribed in the normal and mirrored DC voltage transfer curves. This method will be used in this work, to estimate the Static noise margin for the read and write operation, of a 6-T SRAM cell.

2.1 STATIC READ NOISE MARGIN (RNM)

The *read static noise margin* is one of the most important criteria to analyze the performance during the read mode of the SRAM memory cell. The stability of a SRAM cell in read mode is characterized by RNM when the wordline is activated and the bit and bitbar lines are pre-charged high for the read mode. RNM is defined in [20] as the maximum DC noise voltage ($\pm V_n$) that can be tolerated at the cell storage nodes L and R as shown in Fig 2.4, without the cell flipping its state.

As shown in Fig 2.4, the two DC noise voltage sources are placed in series with the cross coupled inverters, with the worst polarity at the internal nodes. Voltage source $(V_n)_i$ is used to model the DC noise source. For a perfectly symmetric SRAM cell V_{n1} and V_{n2} are equal and are unequal for an asymmetric SRAM cell.

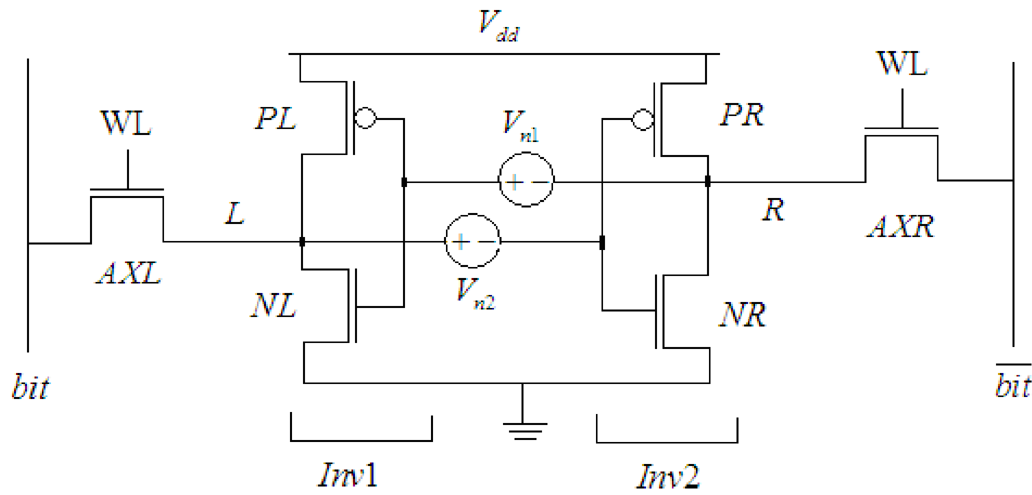

Fig. 2.4 Two cross-coupled inverters with static-noise voltage source V_{n1} and V_{n2} for a 6-T SRAM cell.

In the latter case, the RNM is given by

$$(RNM)_{asymmetric} = \min\{V_{n1},\ V_{n2}\} \tag{2.4}$$

RNM for a 6-T SRAM cell is calculated by measuring the sides of the largest square that can be inscribed in a butterfly curved, formed by drawing the voltage transfer curve (VTC) of Inv_1 and the mirrored VTC of Inv_2.

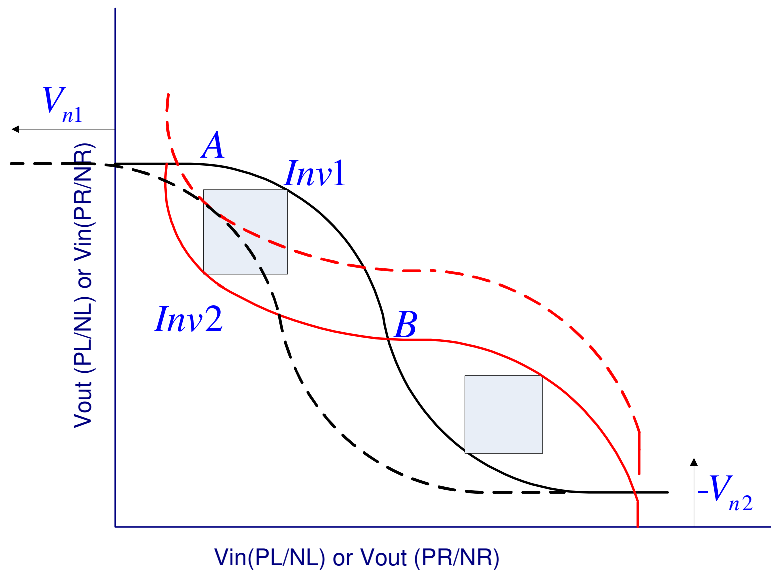

Fig 2.5 Read Static Noise Margin (RNM) for a symmetric 6-T SRAM cell.

The RNM value corresponds to the side of the smallest square of the two squares that can be inscribed. As stated above, the 2 square will have different dimensions for an asymmetric cell and dimensionally equal for a symmetric SRAM cell. When V_n, as shown in Fig 2.5, is equal to RNM the VTC's move horizontally and vertically until the stable point A and the meta-stable point B coincide.

2.2 STATIC WRITE NOISE MARGIN (WNM)

During the write operation, Fig 2.6, the value that has to be stored in the SRAM cell is reflected in the bit line voltage. If the node L is storing a value of "1" and a "0" has to be written to it during the write operation, in that case the bit line voltage is set to "0" before the wordline is pulsed high. If due to the switching activity of the cells in that column of the SRAM array, a voltage is induced in the bitline, due to capacitative coupling, such that it fail to write a "0" to the node L, in that scenario a write failure occurs. Thus write margin is defined as the minimum voltage difference between the bitlines necessary to flip the activated cell. Alternatively, it can be understood as the Minimum Square that can be inscribed in the VTC of the cross coupled inverters, during a write operation as shown in Fig 2.6 and Fig 2.7.

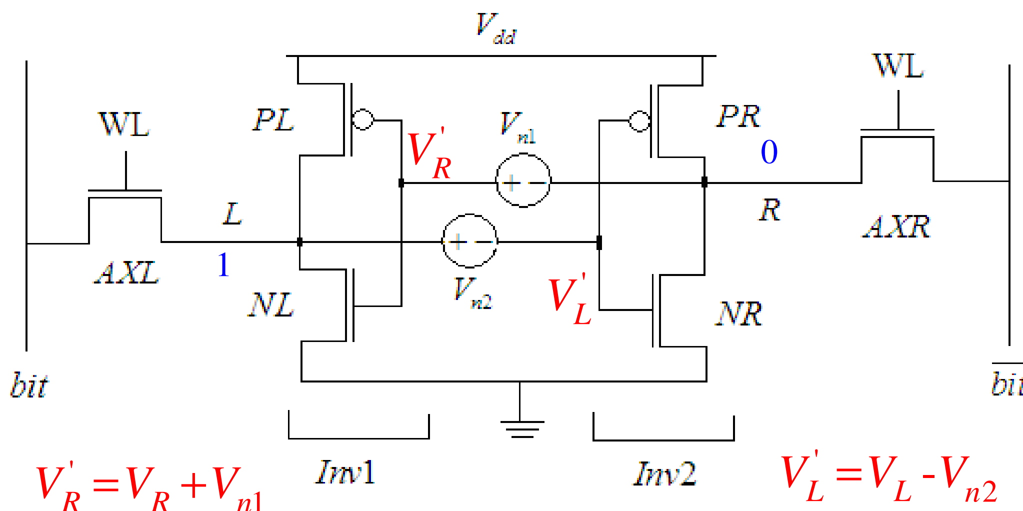

Fig. 2.6 SRAM during a write operation.

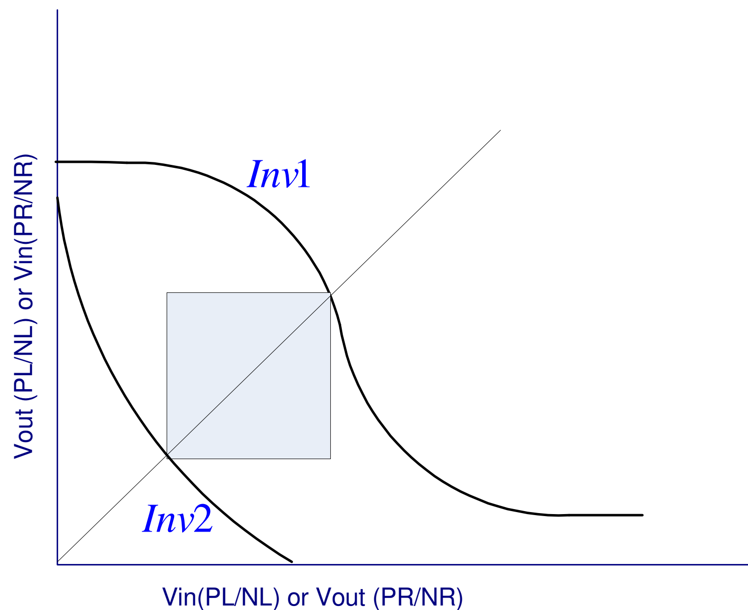

Fig. 2.7 Write Noise Margin. Sides of the minimum sized square that can be inscribed in the VTC of the cross coupled inverters.

3. READ ACCESS TIME

The read access time is a measure of the SRAM operation during its read mode, to access the stored data and is one of the main timing criteria in memory circuits. The read access time is defined as the time required to produce a pre specified voltage difference the two bit-lines. The prespecified bit-differential is a fixed quantity which is dependent on the voltage sensing sensitivity of the sense amplifier. For a successful read operation, the bit-lines should have a voltage difference within the time period, during which the wordline is high ($T_{wordline_read}$). The voltage differential is established in the bit-lines during the read operation, due to the cumulative pull down action of the access transistor and the pull down transistor of the SRAM for the node storing a 0.

To ensure a write failure does not occur

$$T_{read_access_time} \leq T_{wordline_read} \tag{2.5}$$

4. WRITE ACCESS TIME

In addition to read access time, write access time is another important metric in the timing criteria of memory circuits. The write access time is a measure of memory speed during the write mode of the SRAM circuit. The write access time is defined as the time required to write a specified value to the SRAM cell, during the write operation, from the moment the wordline is activated. Thus the window of opportunity to complete a successful write is the width of the high-pulsed wordline voltage ($T_{wordline_write}$).

To ensure a write failure does not occur

$$T_{write_access_time} \leq T_{wordline_write} \tag{2.6}$$

III. PROCESS VARIATIONS AND YIELD AWARE RESPONSE SURFACE MODELING

1. MODELING PROCESS VARIATIONS

As the transistor continues to be scaled to finer feature sizes, it becomes increasingly difficult to control the process variations. The increasing fluctuations in manufacturing process have introduced unavoidable and significant uncertainty in circuit performance. There are two variation sources in digital circuits which induce variability, namely *Inter-die variations* and *Inter-die process variations*. Inter-die variations refer to the die-to-die variations occurring during the photo-lithographic wafer manufacturing process. The inter-die variation in a parameter, say gate oxide thickness, T_{ox}, modifies the value of the parameter of all the transistors in a die in the same direction, which implies that the gate oxide thickness of all the transistors in a die either increase or decrease by the same value. This principally results in a spread of performance parameters such as access time delay and leakage but does not cause any mismatch between transistors.

Intra-die variations on the other hand refer to the within-die variation of all the process parameters. The intra die variations shift the process parameters of different transistors within the same die in different direction. For instance, in the case of the 6-T SRAM cell, the gate oxide thickness of some transistors might increase and of others, might decrease. Intra-die variations can be either systematic or random. If the source of intra die variation is systematic, then there is a distance dependent correlation between the process parameters of transistors. Thus, shift in the parameter of one transistor depends on the shift in the same parameter of a neighboring transistor. Contrary to this is observed in random intra die variations, wherein, there is no proximity dependent correlation between parameters and shift in process parameters of neighboring transistors can be completely independent of each other.

The sources of inter-die and intra-die variations in process parameters includes channel length (L_g), channel width (W), gate oxide thickness (T_{ox}), line edge roughness, threshold voltage (V_{th}), and random dopant fluctuation (RDF). These sources of variations result in a significant variation in the performance metrics of the SRAM. Hence, any constraint model developed to analyze the performance of a SRAM should consider the variability in process parameters and is more relevant as technology nodes begin to reach sub 90nm levels. Modeling and analyzing these random process variations to ensure manufacturability and improve yield has been identified as a top priority for today's IC design problems.

In general, parametric variation can be modeled as

$$\delta_{total} = \delta_{inter} + \delta_{intra} \tag{3.1}$$

where, δ_{inter} is the inter-die variation and δ_{intra} is the intra-die variation. Since the present analysis is based on a single 6-T SRAM cell, only the effect of intra-die variations will considered. The systematic intra-die variations does not result in large variations in transistors which are in close spatial proximity, especially so for the transistors in a 6-T SRAM cell. Thus the random effect of intra-die process variations would be a major contributing factor in the predictor model development, for the SRAM and would lead to mismatch in the transistors.

Amongst the random sources of intra-die process variation, the ones considered for model development are gate oxide thickness variation (T_{ox}) and variation in threshold voltage (V_{th}) due to random dopant fluctuation (RDF). According to ITRS [2, the total V_{th} 3σ variation, for the 45nm technology node is 42% and RDF accounts for 40% of it. Thus it is one of the major contributing factors for threshold voltage variation. Gate oxide thickness on the other hand is a relatively well controlled parameter; the T_{0x} variation for the 45nm technology node for a MOSFET is 4 %.

As mentioned before, the random effects of intra-die process variations are being considered in this work. Hence, the threshold voltage and gate oxide thickness of the MOSFET's of a 6-T SRAM cell are going to be independent parameters and bear no correlation with each other. This principally means that the models developed will have

6 distinct and independent V_{th} and T_{ox} values. To ensure a high yield, the range of distribution of the process variations is assumed to lie within the domain of values [-6σ, 6σ].

2. VARIATION AWARE RESPONSE SURFACE MODELING

In order to account for process variations, *response surface models* [21] are used to capture the circuit performance parameter variation, caused as a result of fluctuations inherent in die production process at the wafer fabrication labs. In statistics, response surface methodology (RSM) explores the relationships between several explanatory variables and one or more response variables. The thought behind this method is to perform a sequence of design experiments, by varying the input parameters, to obtain an optimal response. Thus, *response surface modeling* or RSM is a technique used to form the polynomial functions to predict the circuit performance in terms of design parameters, while simultaneously accounting for process variations.

The application of Response surface modeling (RSM) in developing empirical predictor models for circuit parameters is explained in this section. For a fixed circuit topology, the circuit performance parameters, like Static Noise margin, Access Time and Leakage, are a function of design parameters, like Transistor channel widths, W, as well process parameters, like gate oxide thickness, T_{ox}, and threshold-voltage, V_{th}. Mathematically, RSM can be represented as

$$Y = F_1(D) + F_2(P) \tag{3.2}$$

Where, Y is the Predictor model of the parameters to be determined, $F_1(D)$ is the function of static variables and $F_2(P)$ is the function statistically varying parameters

In the domain of digital\analog circuit the above equation; for a fixed topology circuit of *m* transistors; can also be more intuitively written in terms of dependent design variables and process variation as

$$Y_N = F_{1N}(W_1, W_2, ...,W_m) + F_2(\Delta V_{th1},...,\Delta V_{thm}; \Delta T_{ox1},...,\Delta T_{oxm}; \Delta L_{g1},...,\Delta L_{gm}) \tag{3.3}$$

Where, Y_N is the Nth circuit performance parameter to be determined, F_{1N} is a function of

widths of the *m* transistors in the circuit and F_{2N} is the function of random process variables of the *m* transistors, like V_{th}, T_{ox}, gate length L_g etc.

A common methodology is followed during the process of determining the predictor model, which is outlined in the steps below.

- *Step 1: Fixing the design parameters*

  During the optimization process, the design parameters are optimized and fixed during that iteration. The predictor models developed would only be a function of process variations/random variables, to account for any uncertain variation. Consequently, the Nth circuit performance parameter has the form

$$Y_N\big|_{W_i=fixed} = F_2(\Delta V_{th1},...,\Delta V_{thN};\Delta T_{ox1},...,\Delta T_{oxN};\Delta L_{g1},...,\Delta L_{gN}) \quad (3.4)$$

- *Step 2: Selecting the template and order for RSM*

  A fixed template needs to be selected for the performance parameter, determined by the accuracy of the model required. For Circuit performance parameters that are linearly dependent on random variables, a linear model suffices, for nonlinear models a second order or higher model can be used depending on the accuracy of the template in estimating the circuit performance parameter Y. Generally, in nano-scaled analog and digital circuits, the linear model is not sufficiently accurate. Hence, applying quadratic RSM or higher order RSM would improve the accuracy.

  Given a set of fixed design parameters, the Nth circuit performance can be approximated by [20]

$$\textit{Linear Model:} \quad Y(X) = B^T X + C$$

$$\textit{QuadraticModel:} \quad Y(X) = X^T A X + B^T X + C \quad (3.5)$$

where $X = [x_1, x_2,\ldots, x_m]^T$ represents the process variations and m is the total number of variation process parameters,

$C \in R$ is the constant term,

$B \in R^N$ represents the linear coefficients,

$A \in R^{NxN}$ denotes the quadratic coefficients.

Higher Order Polynomials:

$$Y(X) = a_0 + \sum_{i=0}^{m} a_i x_i + \sum_{i<j}^{m} a_{ij} x_i x_j + \sum_{i=0}^{m} a_{ii} x_i^2 + \ldots.. \quad (3.6)$$

- *Step 3: Design of Experiment (DOE): Sampling in the domain [-6σ, 6σ] of X_i , the process variations.*

  It is apparent by now that a statistical modeling methodology needs to be followed to account for the process variations. One problem that particularly arises is that the observed changes in a response variable, Y, may be correlated with, but not caused by, observed changes in individual *factors,* X_i (process variables). Simultaneous changes in multiple factors may produce interactions that are difficult to separate into individual effects. Observations may be dependent, while a model of the data considers them to be independent [22].

  Design of Experiments or *DOE* is the possible way out of the aforementioned predicament without loss in reliability as well as saving the computational overhead involved. DOE is a method of effectively collecting experimental data at the *effectively sampled* points of the individual factors, X_i, in their domain (*-6σ to +6σ* to ensure higher quality of the data points). In a designed experiment the data samples, at which the performance of the circuit is evaluated, is actively manipulated to improve the quality of the data and remove redundant information. A common goal of all designed experiments is to be parsimonious with respect to the number of sampled points while simultaneously ensuring sufficient information for high degree of accuracy in modeling of the circuit performance.

  Some common methods followed for deriving ideal sampling points for response Y are

  1. Full factorial designs
  2. Plackett-Burman designs
  3. Box-Wilson designs
  4. Box-Behnken designs

5. D-Optimal designs
6. Random sampling

The first 2 methods are ideal for *linear* RSM, while 3, 4 are ideal for *quadratic* RSM and 5 is best suited for models in which coefficients to be fitted are nonlinear, a case in point being a model for circuit performance estimation.

Random sampling, generally, on the other hand fails to capture correlation between process parameters. This however, is not objectionable in the template used for RSM, in this research exploration. As stated previously in section (A), the effect of random intra-die process variations is taken into account, and the parameters, X_i, are totally random, independent variables and bear no correlation with each other. Hence, if the RSM used is quadratic, in equation (15) the coefficients of the cross terms $x_i x_j$, a_{ij} would be zero.

- *Step 4: Solving the over-determined equation to find the coefficients, using suitable regression methods.*

  Once the ideal sampled points (X_i) are selected, HSPICE simulations are run to determine the respective circuit performance parameter, Y_N. Using the set of sampled data $\{X_i, \tilde{Y}_i\}$, also can be referred to as the training data, the coefficients of the over determined equation is found out by suitable regression methods. Since, all the models used for RSM for an SRAM are linear with respect to the coefficients, any linear regression method should suffice, which guarantees a good confidence interval for the estimated unknown variables.

The methodology of formulating an empirical model is summarized in the flowchart Fig 3.1

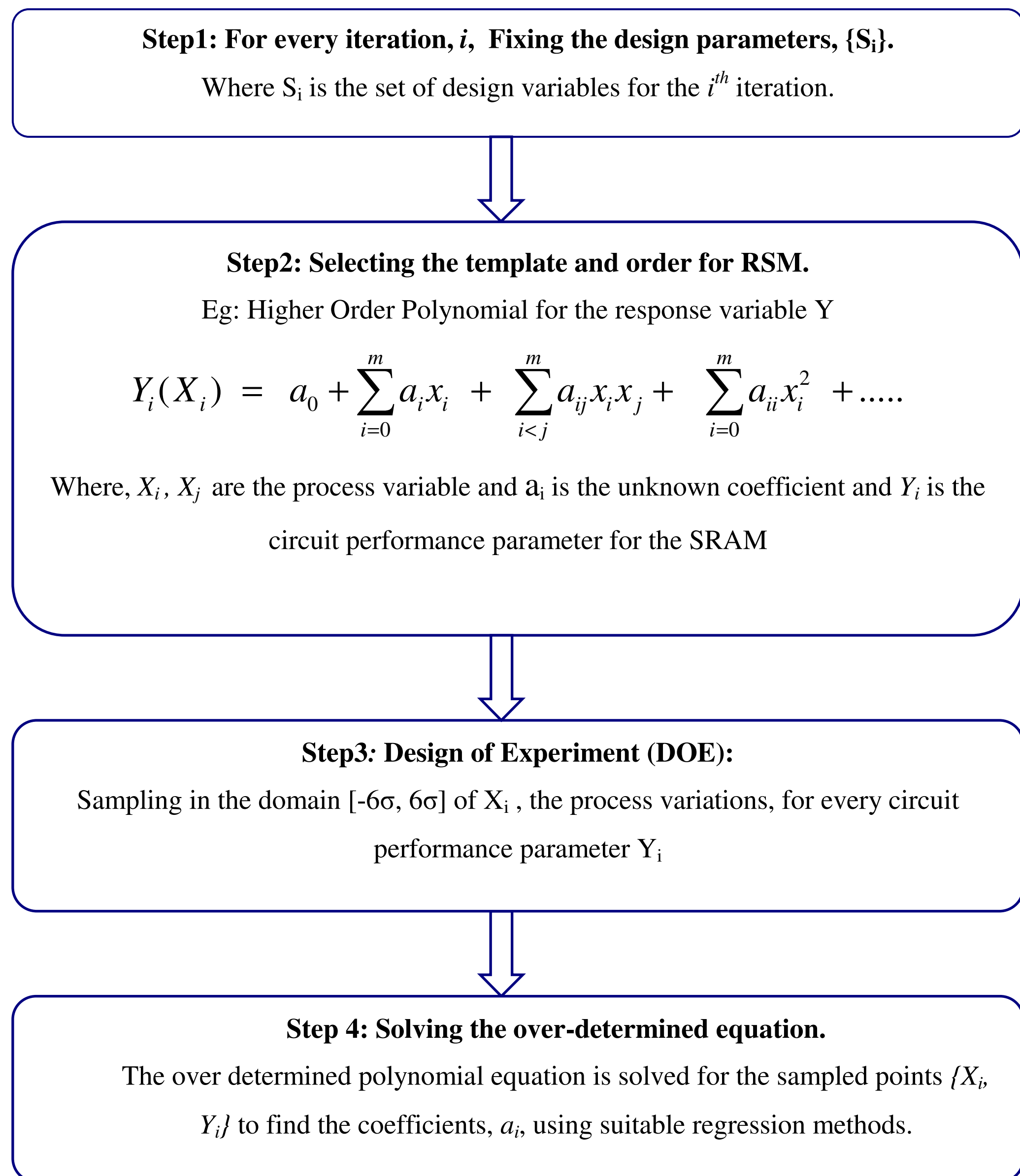

Fig 3.1 Flow chart depicting the response surface methodology flow used for circuit parameters.

IV. STATISTICAL MODELING OF SRAM PERFORMANCE PARAMETERS

This section describes the predictor models developed, for the performance metrics of a 6-T SRAM cell. The predictors are mathematical, empirical estimators formulated after the regression analysis of the over determined polynomial equation. These models are then used to determine the value of a performance metric of a circuit, without running further SPICE simulations, without loss in reliability and offering a SPICE like accuracy. The performance metrics for the 6-T SRAM cell described and estimated in this section are, Standby Leakage power, Static Noise margin (Read and Write), Dynamic Noise Margin (Read and Write) and Read access time. These models will form the main components in our optimization process.

The models are developed in the presence of process variation, which manifests itself in device mismatch. Hence, all the models developed are characterized for an *asymmetric* 6-T SRAM cell. Also the distribution of process variations is assumed to be *Gaussian.*

1. CELL LEVEL SRAM PERFORMANCE MODELING

In this sub-section, models for cell level performance parameters of the SRAM memory are developed. Furthermore, the SPICE level simulations and the model predicted values are compared and are shown to have a very accurate fit with minimal relative error.

1.1 STANDBY LEAKAGE CURRENT

The figure 4.1 represents the SRAM memory cell in the standby/data retention mode. The cell is storing a “1” which is reflected in the value of the node L. Bitlines are precharged high and since the cell is not being accessed, wordline is pulsed low. The case shown in the figure represents the worst case leakage scenario as the bitlines are pulsed high, which contributes to leakage power through both Subthreshold and gate leakage.

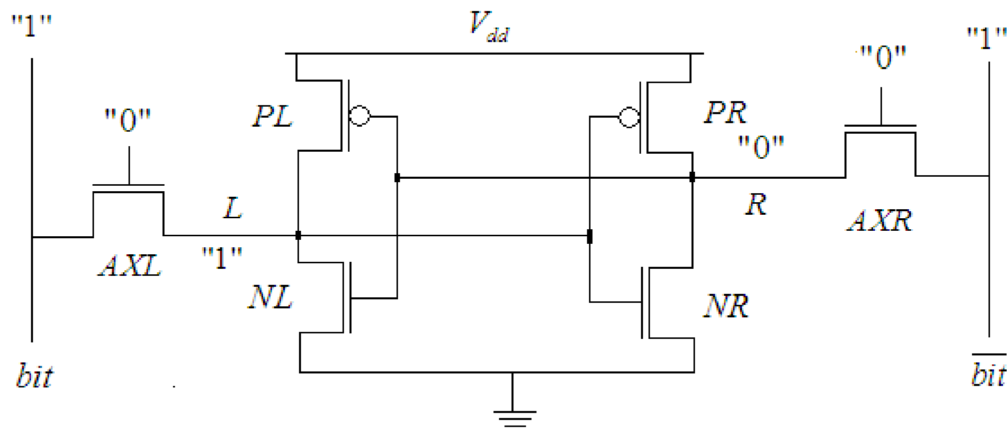

Fig 4.1 SRAM cell during the data retention mode. This represents the worst case standby leakage, when the bitlines are pulsed high.

Three dominant leakage paths can be identified in this scenario, whence the bitlines are set “high”

- Leakage from bitlines to wordline
- Leakage from bitlines to ground
- Leakage from Power supply line to ground

In the current section, each leakage component is analyzed in the presence of process variations and subsequently an empirical model is obtained, the accuracy of which is duly justified on the basis of actual HSPICE simulations.

1.1.1 Subthreshold Leakage

As stated previously, Subthreshold leakage or subthreshold drain conduction refers to the current that flows between the source and drain of a MOSFET when the gate-to-source voltage is below the threshold voltage (V_{th}). Thus “off” transistors are analyzed to compute the Subthreshold component of leakage.

From equations (4.1) (4.2) and (4.3), it is clear that I_{Sub} has an exponential dependence on threshold voltage (V_{th}) of the transistor. Furthermore, the transistor model used is a BSIM4 model, which describes V_{th} as a function of various process parameters such as effective gate length, L_{eff}, gate oxide thickness, T_{ox} and doping concentration of the channel, N_{sub} [23]. According to ITRS [2], the total 3σ variability of V_{th} is nearly 42% of the nominal value, of which random dopant fluctuation (RDF) in the channel of a transistor contributes to nearly 40% of the total V_{th} variability MOS device. Oxide thickness, T_{ox} variability at the 45nm node has a 3σ value of 4%, according to ITRS [2]. Furthermore, T_{ox} is a fairly well controlled parameter and more importantly, its effect on I_{sub} is not significant as compared to other factors hence it can be safely ignored, while modeling I_{sub}. Thus the empirical model for the Subthreshold leakage for an "off" transistor can be written as

$$(I_{Sub})_{gate} = (I_0)_{off}\, e^{F(\Delta V_{th})} \tag{4.1}$$

From circuit simulations, it was determined that a linear model for F(ΔV_{th_Nsub}) sufficiently models $I_{Sub.}$ The total subthreshold leakage current for the 6-T SRAM cell schematic shown in Fig 4.1 is given below.

$$\left(I_{sub}\right)_{total} = I_{nl0}.e^{-\alpha_1 \Delta Vth_nl} + I_{pr0}.e^{\alpha_2 \Delta Vth_pr} + I_{axr0}.e^{-\alpha_3 \Delta Vth_nr} \tag{4.2}$$

Where, I_{nl0}, I_{pr0} and I_{axr0} are the Subthreshold leakage current contributors in the SRAM cell, without any process variation and ΔV_{th_nl}, ΔV_{th_pr}, ΔV_{th_nr} are the change in threshold voltage values of the respective transistors as shown in Fig 4.1. It should be noted in the figure that the transistors NL, PR and AXR are in the "off" state. Consequently, only these three $(I_{ds})_{off}$ currents need to be considered while modeling the Subthreshold leakage of a SRAM cell shown in Fig 4.1. The model above is general and can be extended to include other effects due to random variations. From equation (4.1), if gate length variation has to be taken into account it can be modeled as

$$(I_{Sub})_{gate} = (I_0)_{off}\, e^{F_1(\Delta V_{th}) + F_2(\Delta L_g)} \tag{4.3}$$

Where $F_1(\Delta V_{th})$ and $F_2(\Delta L_g)$ are empirically determined models.

1.1.2 Gate Leakage

To develop an accurate model, which predicts gate leakage in the presence of process variation, the various leakage mechanisms and their impact in a MOSFET needs to be analyzed. The three major gate leakage components are Electro conduction-band (ECB) tunneling, Electron Valence-band tunneling (EVB) and Hole Valence band tunneling (HVB) [24]. Efforts on empirical gate leakage modeling as in [23] and [7] donot consider effect of random dopant fluctuation (RDF) and its effect on threshold voltage V_{th}, this as per analysis and HSPICE simulations, can lead to an incorrect model predictor. The authors in [18] derive an analytical intrinsic gate leakage model for a MOSFET with physical source/drain current partition which has been implemented in BSIM4 to a fair degree of accuracy. According to their models the leakage current density, due to the three major gate leakage current components, in addition to gate oxide thickness, T_{ox}, also has an exponential dependence to threshold voltage, V_{th}.
The gate tunneling current density as a function of x, the distance from source to drain in the channel, is [24]

$$J_g \approx A.E_{ox}^2 e^{-B/E_{ox}} \approx A.E_{oxs}^2 e^{-BT_{ox}/\left(V_{oxs}-V(x)\right)}$$

$$\equiv J_{g0}.e^{-B^* V(x)} \tag{4.4}$$

Where J_{g0} is the gate tunneling current density with $V_{ds} = 0$, $V_{oxs} \approx V_{gs}$ and $B^* = B.T_{ox}/V_{oxs}$.
V(x) is the voltage along the channel from source-to-drain and is approximately given by the expression

$$V(x) \approx \left(V_{gs} - V_{th} - V_{ds}/2\right).V_{ds}/\left(V_{gs} - V_{th}\right)x \tag{4.5}$$

Where V_{gs} is gate to source voltage, V_{ds} is the drain to source voltage and x is the distance from source to drain, in the channel.

From (4.4) and (4.5) it is clear that gate leakage current has an exponential dependency on both threshold voltage V_{th} and gate oxide thickness T_{ox}, vindicating our previous assumption and has shown to be true in our simulation results using HSPICE. Thus the gate leakage of a transistor can be modeled as

$$I_{Gate} = I_{gate0} \cdot e^{F_1(\Delta V_{th}) + F_2(\Delta T_{ox})} \tag{4.6}$$

Here I_{gate0} is the gate leakage current of a transistor with no process variation. From analysis, it was found that both functions F_1 and F_2 are linear in ΔV_{th} and ΔT_{ox} respectively. Thus the simplified model that can be used for an SRAM cell as shown in Fig 4.1 is

$$\left(I_{Gate}\right)_{total} = I_{nr0} . e^{-\beta_1 \Delta Vth_nr - \beta_2 \Delta Tox_nr} + I_{pl0} . e^{\beta_3 \Delta Vth_pl - \beta_4 \Delta Tox_pl} \tag{4.7}$$

Where, I_{nr0} and I_{pl0} are the gate leakage current values for an SRAM memory cell without taking process variations into consideration.

1.1.3 Total Leakage

The total leakage current of an SRAM cell can be modeled as a sum of the major leakage components.

$$\left(I_{leak}\right)_{total} = I_{sub} + I_{Gate} \tag{4.8}$$

Furthermore, as previously stated we are considering the V_{th} variation due to RDF, hence the V_{th} of different transistors in an SRAM cell are independent random variables [7]. Also according to [7] the T_{ox} of different transistors in an SRAM cell are random variables.

α_1, α_2, α_3, β_1, β_2, β_3, β_4 are the fitting parameters which can be determined using nonlinear regression methods in MATLAB one example, which was used in this work, is the Gauss-Newton algorithm with Levenberg-Marquardt modifications, to guarantee a global convergence. The accuracy of the above model can be judged by the confidence interval of the fitted parameters α_i, β_i.

In Table 1, the fitting of the coefficients are found to be very good thus vindicating our assumptions. Furthermore, the distribution of logarithm of leakage current of the SRAM cell is non central Fig 4.2, thus the single lognormal approximation of sum of lognormals, for the total leakage current [7] is also not accurate when the number of lognormals to be added is few.

TABLE I

FITTED PARAMETERS USING NONLINEAR REGRESSION

Coefficient	Value of coefficient	Confidence Interval	
β_1	6.5912	6.3017	6.8806
β_2	1.1901	1.0901	1.2901
β_3	7.9052	7.8465	7.9639
β_4	0.7844	0.1727	1.3961
α_1	26.9242	25.6304	28.2180
α_2	26.4226	26.3937	26.4515
α_3	28.0944	27.5719	28.6168

In statistics, a confidence interval (CI) is a particular kind of interval estimate of a population parameter. Instead of estimating the parameter by a single value, an interval likely to include the parameter is given. Thus, confidence intervals are used to indicate the reliability of an estimate. How likely the interval is to contain the parameter is determined by the confidence level or confidence coefficient. Increasing the desired confidence level will widen the confidence interval [25].

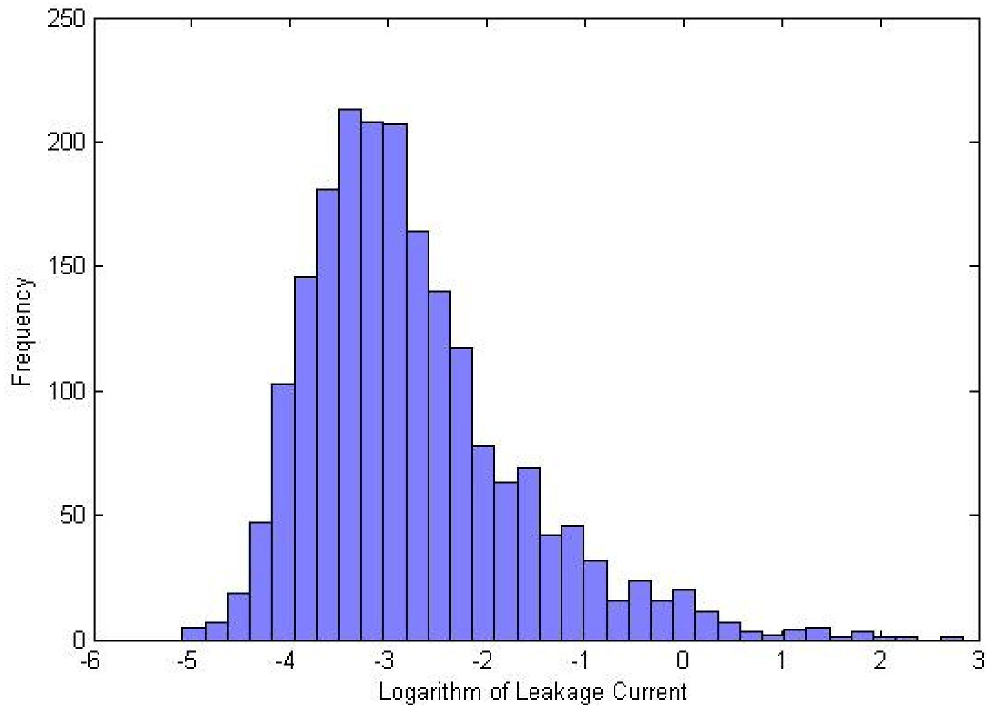

Fig. 4.2 Non-central distribution of the logarithm of leakage current by derived by varying process variables, in HSPICE simulation.

1.2 STATIC READ NOISE MARGIN (RNM)

As stated before, the RNM is defined as the maximum DC noise voltage ($\pm V_n$) that can be tolerated at the cell storage nodes without changing the stored bit. A positive value of RNM represents a stable read while a zero or negative value will result in a read failure. The effect of random variations on RNM value is characterized by varying individual Vth and Tox value of the MOS transistor constituting the SRAM cell between its ($\mu_{mean} \pm 6\sigma$) values. The HSPICE based Monte-Carlo simulations, Fig 4.3, are based on a 45 nm CMOS technology BSIM4 model. A second degree polynomial trend line is made to fit the plotted sensitivities. The sensitivity based characterization is an important step in determining the order of the polynomial, which has to be chosen for response surface modeling, to accurately fit the circuit parameter. A second order trend line

accurately fits the characterized curve, thus for read noise margin, a second degree polynomial would model it to a fair degree of accuracy Fig 4.4

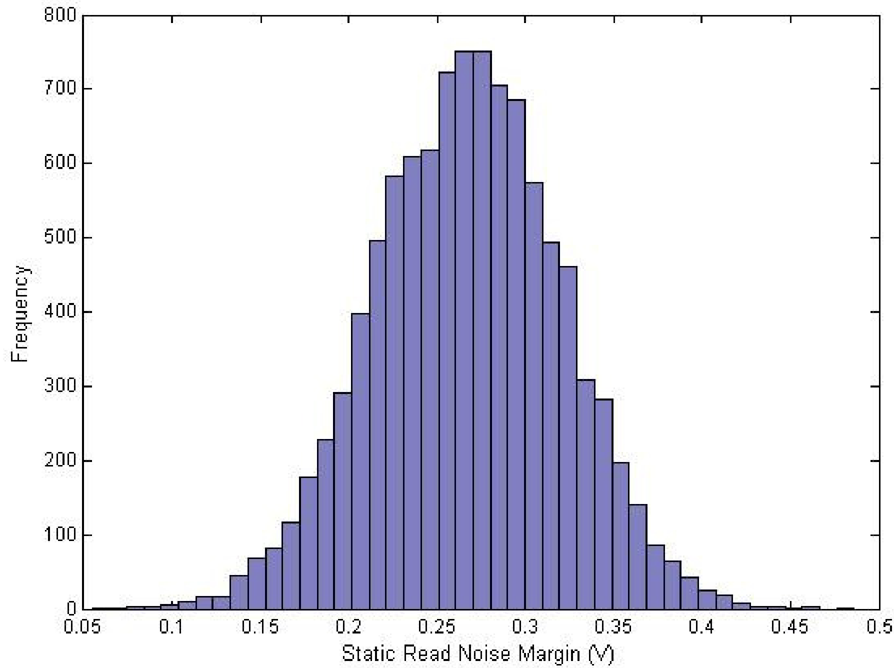

Fig. 4.3 Distribution of RNM on varying process variables.

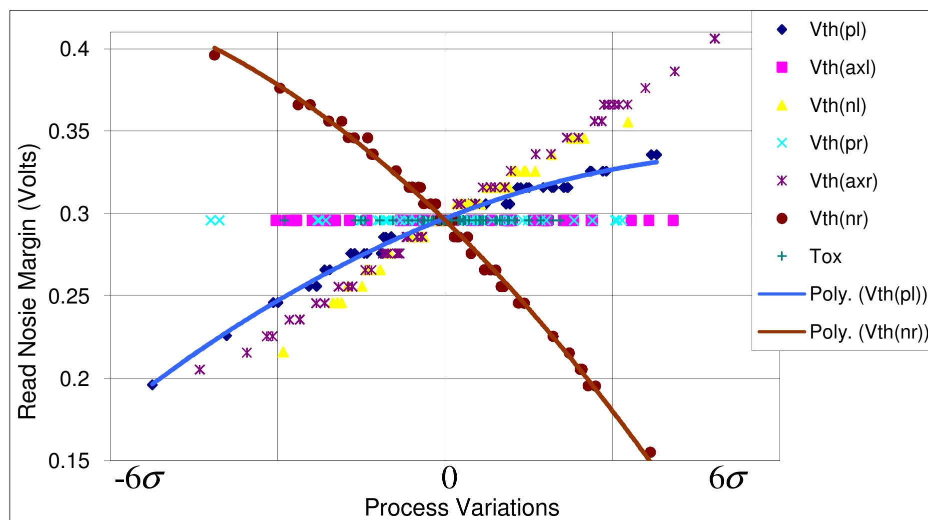

Fig 4.4 Characterization of RNM on varying V_{th} and T_{ox} .

The Gaussian nature of the distribution of RNM Fig 4.3 and the second order nature of the characterization curve indicates that RNM for an SRAM can be modeled as a second order polynomial, which is a function of process variations. The model is represented below

$$RNM(X) = X^T AX + B^T X + C \tag{4.9}$$

Where,

$X = \left[\Delta V_{th1},\ \ldots\ ,\Delta V_{th6,} \Delta T_{ox1,}\ \ldots\ ,\Delta T_{ox6}\ \right]^T$ is the column vector of random variations

$$A = \begin{bmatrix} a_{11} & a_{12} & .. & .. & .. & a_{1\,12} \\ . & a_{22} & & & & . \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ a_{12\,1} & .. & .. & .. & .. & a_{12\,12} \end{bmatrix}$$ is a 12x12 Matrix

$$B^T = [b_1, b_2,\ \ldots.\ ,\ b_{11}, b_{12}] \tag{4.10}$$

$C = \mu_{RNM}$, is the mean value of read noise margin for fixed Widths and lengths values of the MOS transistor and in the absence of process variations.

Matrices A and B are the coefficients of second order and first order terms respectively, in the polynomial model predictor. They constitute the unknowns to be determined after response surface modeling and regression analysis.

Since the random effects of intra-die process variations are independent in nature, the numbers of unknowns to be determined are reduced. Thus, the process variations in matrix X are independent and bear no correlation with each other. Consequently, the cross terms, a_{ij} , in matrix *A* are zero and matrix *A* is a *diagonal matrix*. This greatly reduces computation overhead and results in a simple, accurate, pure quadratic model for read noise margin.

Fig. 4.4 shows that RNM is most sensitive to V_{th} variation in the access transistor AXR, pull down transistor NR and to a lesser extent, to pull up transistor PL. To recapitulate, L is the node storing a “1” and R is the node storing a “0” in the SRAM cell Fig 4.1. This

observation is quite expected as these "ON" transistors (PL and AXL), form a pull down path for the bit bar line (Fig 4.1). In order for the sense amplifier to detect the contents of the cell, a differential voltage needs to be created between the two bitlines. The voltage division between these two transistors determines the magnitude of noise injected at the internal storage node during the read operation.

A stronger NR and a weaker AXR results in a better RNM due to smaller read disturbance at node R. The drive strength of these transistors is directly proportional to (V_{gs}-V_{th}). Thus if V_{th} for NR decreases and V_{th} of AXR increases, the drive strength of NR increases and AXR decreases, resulting in better read stability. Fig 4.4 verifies experimentally the aforementioned theoretical analysis.

1.3 STATIC WRITE NOISE MARGIN (WNM)

The write noise margin is characterized in a similar way to read noise margin. The characterized curve is plotted by varying the Vth and Tox value of each transistor of the SRAM cell, independently. The Gaussian nature of WNM, Fig 4.5, and the second order nature of the characterized sensitivities, Fig 4.6, indicates that WNM can be modeled as a second order polynomial and a degree twelve function of random variations.

The sensitivity curve based on HSPICE based simulations on a 45 nm CMOS process, shows that a second order trend line fits the write noise variation with the varying process parameters. Thus, a second order polynomial empirical estimator should suffice, without loss in accuracy. This is furthermore verified in fig 4.7 which plots the actual HSPICE simulation determined value and the model predicted value, for a set of points not used as training data in the original response surface modeling procedure.
Mathematically it can be represented as second order function in the Euclidian space:

$$WNM(X) = X^T A X + B^T X + C \quad (4.11)$$

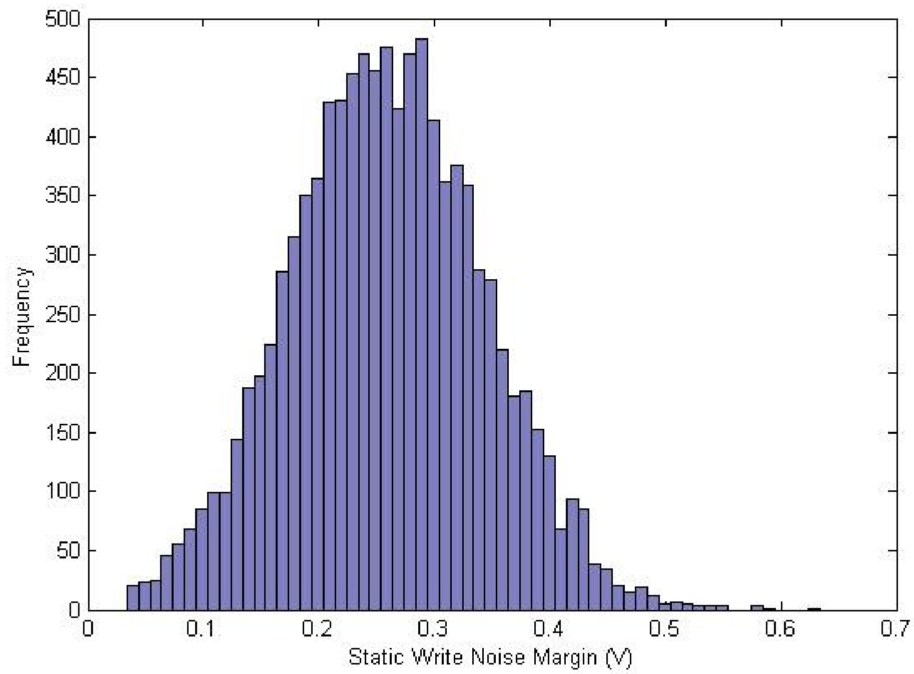

Fig. 4.5 Monte-Carlo sampling of WNM on varying process variables.

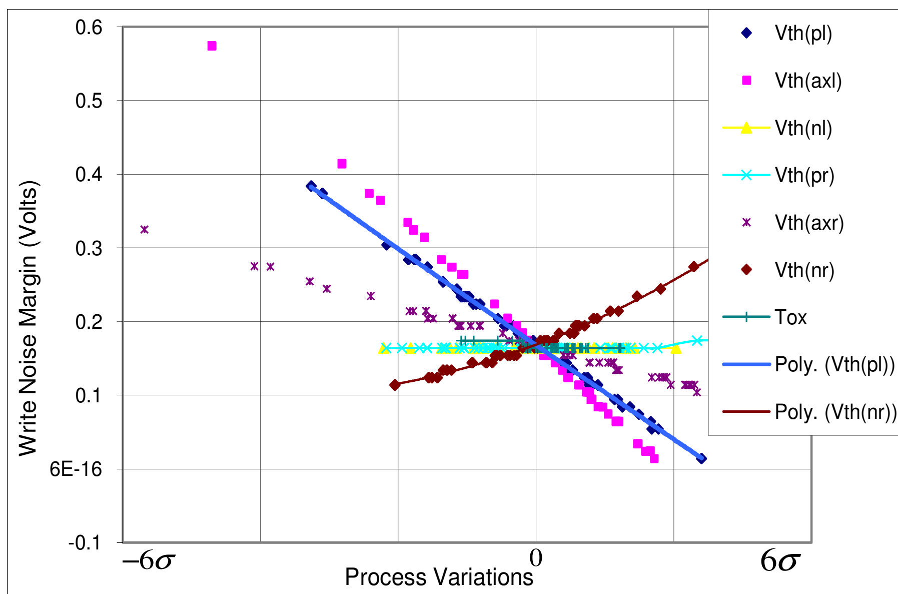

Fig 4.6 Characterization of WNM on varying V_{th} and T_{ox}.

Where X is a set of process variations, for every transistor in the SRAM memory cell, represented in the matrix form as

$$X = \left[\Delta V_{th1},\ \ldots\ ,\Delta V_{th6,}\Delta T_{ox1,}\ \ldots\ ,\Delta T_{ox6}\ \right]^T$$

$C = \mu_{WNM}$, is the mean value of write noise margin for fixed Widths and lengths values of the MOS transistor and in the absence of process variations.

$$A = \begin{bmatrix} a_{11} & a_{12} & .. & .. & .. & a_{1\ 12} \\ . & a_{22} & & & & . \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ a_{12\ 1} & .. & .. & .. & .. & a_{12\ 12} \end{bmatrix}$$ is a 12x12 Matrix

$$B^T = [b_1, b_2,\ \ldots.\ ,\ b_{11}, b_{12}]$$

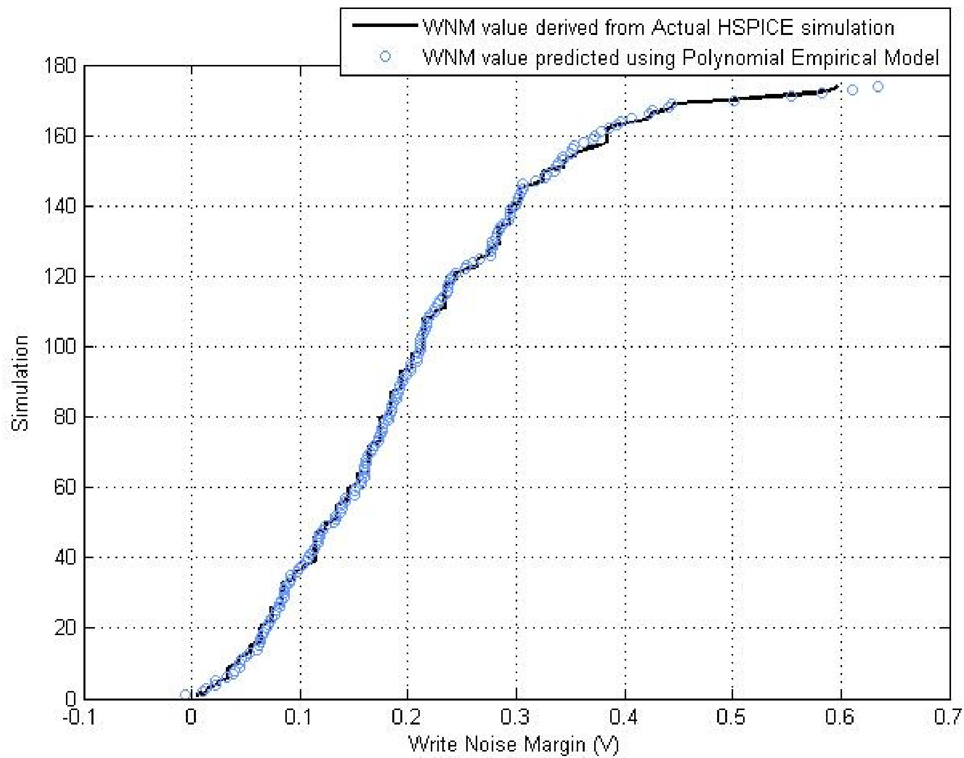

Fig 4.7 Comparison of HPSICE predicted and polynomial predicted values using sample test points.

Since the random variations are independent of one another and cross terms, a_{ij}, in matrix A are zero, the empirical model of write noise margin can be represented in a relatively simpler notation as in (4.12)

$$WNM(X) = \mu_{WNM} + \sum_{i=1}^{6}\left(a_{ii}\Delta V_{thi}^{2} + b_i \Delta V_{thi}\right) + \sum_{i=6}^{12}\left(a_{ii}\Delta T_{oxi}^{2} + b_i \Delta T_{oxi}\right) \quad (4.12)$$

Where,

a_{ii} are the diagonal elements of matrix A

b_i are the elements of the column vector B

1.4 READ ACCESS TIME

During the read operation, the word line is activated for a pre specified time. If a read operation does not occur during that time, a read failure is said to have occurred. The HSPICE based Monte Carlo analysis of read access time, results in a non-central distribution. However, from Fig. 4.8 it is quite evident that the distribution of inverse of read access time is Gaussian. Hence, the inverse can be modeled as an nth order polynomial function of Gaussian variables. These Gaussian variables are process variations of a 6-T SRAM cell. The order n of the polynomial can be gauged by the characterization of Read access time as a function of process variations. The random variables are Gaussian and they are sampled in the $[-6\sigma, 6\sigma]$ range. The resultant variation in read access time is characterized in Fig. 4.9, by individually varying the random variables.

A second order polynomial has a good fit with the threshold voltage variation data Fig 4.10, with a very small relative error ($\pm 2\%$), indicating a good fit. Consequently, a second order polynomial can be used as a good and a reliable template to accurately statistically model the variations in inverse read access time.

It can be mathematically represented as following

$$1/T_{read} = X^T A X + B^T X + C \quad (4.13)$$

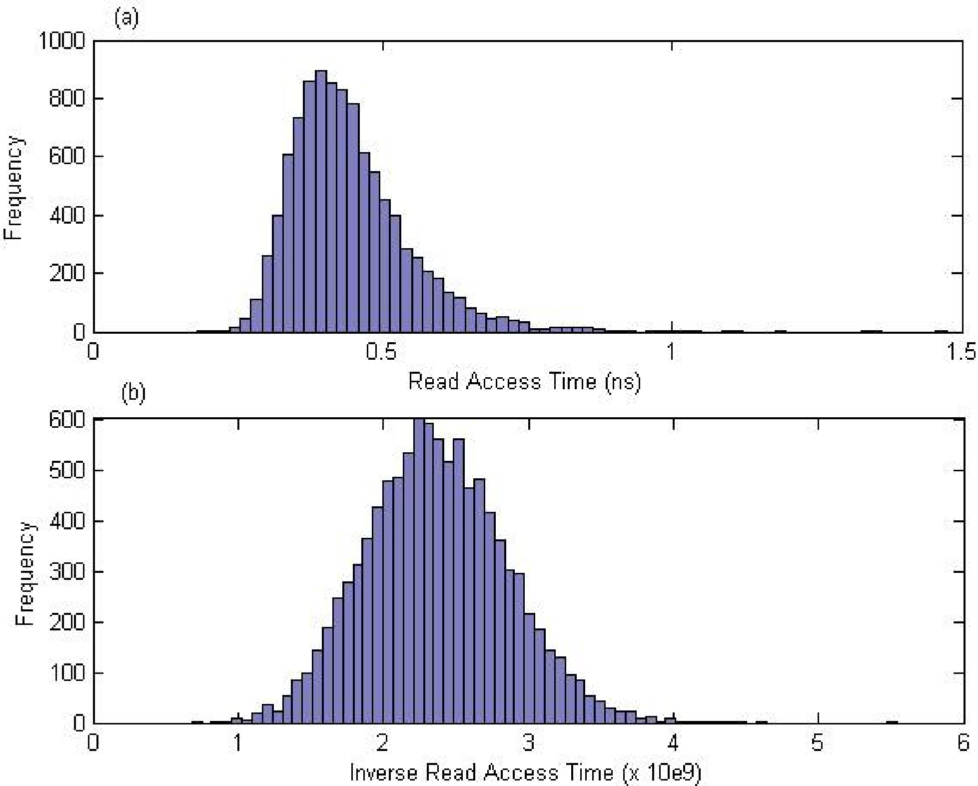

Fig 4.8 (a) Non-central distribution of read access time and (b) Gaussian distribution of inverse read access time.

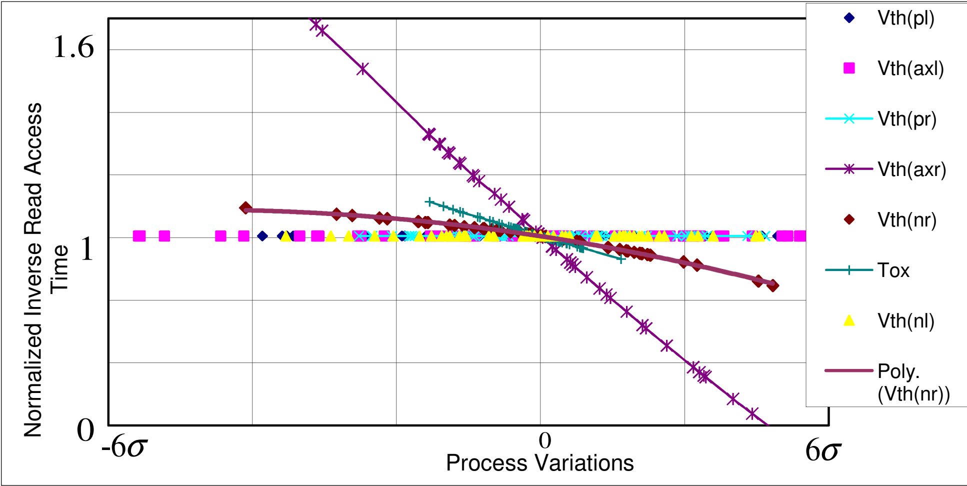

Fig 4.9 Characterization of read access time on varying V_{th} and T_{ox} of all the 6 transistors.

Where,

Matrices A and B are the coefficients of the second order and first order terms respectively (specify equation number of A &B).

C = $1/\mu_{read}$, is the inverse mean value of read access time in the absence of process variations.

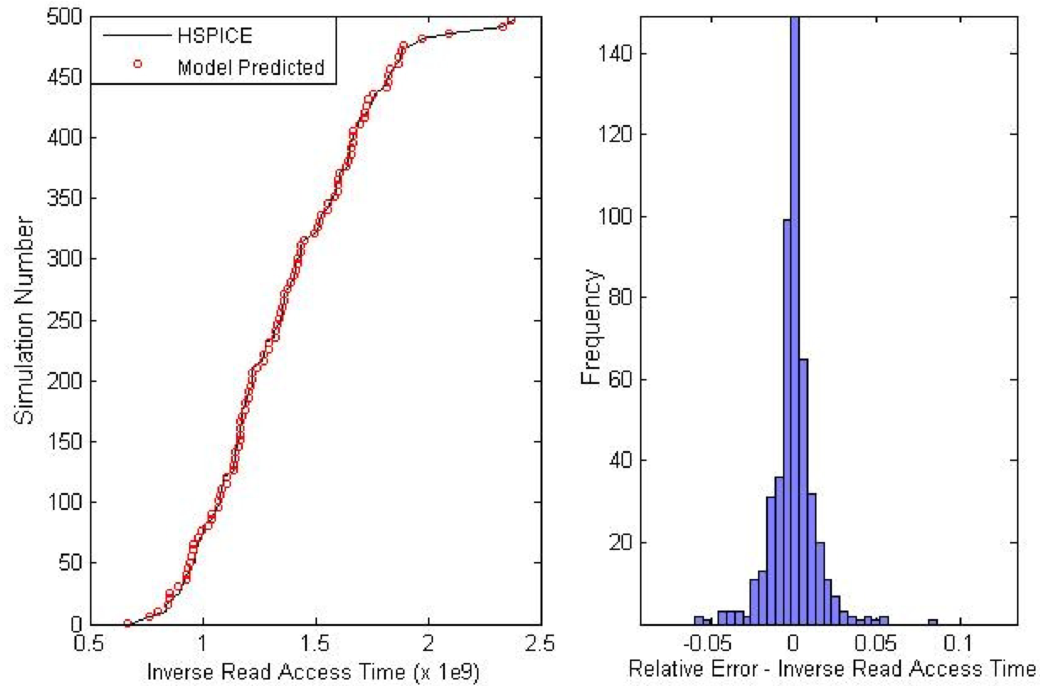

Fig 4.10 Comparison of HPSICE predicted and polynomial predicted values using sample test points and the distribution of relative error.

1.5 WRITE ACCESS TIME

For the write access time a linear model was chosen based on the model developed by the authors in [9]. The model is modified to include the contribution of oxide thickness variations to the total variability. The template of the model is given below

$$1/T_{Write} = B^T X \tag{4.14}$$

Where,

$$B = \left[b_1, b_2, \; .... \; , \; b_{11}, b_{12} \right]^T$$

$$X = \left[\Delta V_{th1}, \; ... \; , \Delta V_{th6,} \Delta T_{ox1,} \; ... \; , \Delta T_{ox6} \right]^T$$

Fig 4.11 depicts the fit between the model and HSPICE simulations indicating a good fit between the model and actual value.

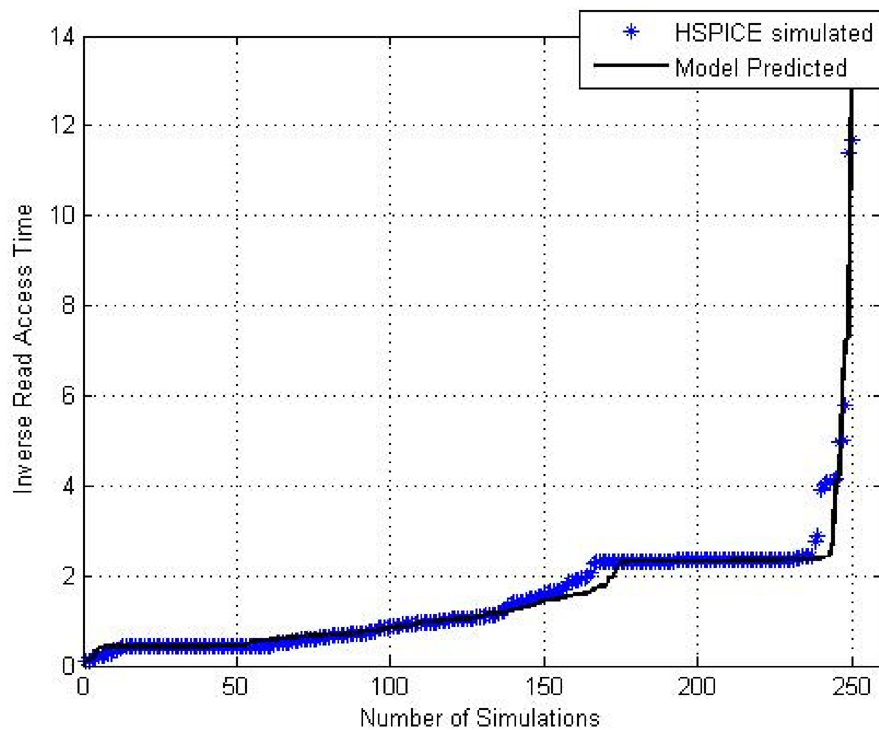

Fig 4.11 Model Vs HSPICE fit for write access time.

2. STATISTICAL MODELING OF PERFORMANCE PARAMETERS FOR CACHES

In the previous sub section, various circuit performance parameters were developed for the SRAM memory cell. The models take into account the variability in

threshold voltage and gate oxide thickness for the 45 nm CMOS technology. From various Monte-Carlo simulations, it was quite evident that the variability has a huge impact on the performance of the memory cell, for a fixed set of design parameters. The SRAM memory cells constitute the building blocks of on-chip L-1 and L-2 cache memory. Thus is becomes imperative to analyze the circuit performance of the on-ship cache memory, in the presence of process variations which might have a huge impact on its performance.

In this sub section, polynomial empirical models are developed for Standby leakage power and Access time of the cache in the presence of variability. For the purpose of modeling them, *a modified* CACTI 5.2 [26], Cache performance tool, developed by HP labs is used. CACTI is an integrated cache and memory access time, cycle time, area, leakage, and dynamic power model. By integrating all these models together, users can have confidence that tradeoffs between time, power, and area are all based on the same assumptions and, hence, are mutually consistent. The tool is based on circuit assumptions in sync with modern design practices.

However it assumes a fixed transistor, memory cell area and memory cell aspect ratio, for a particular technology node. Moreover, it does not consider the effect of process variations on the performance of caches. These inherent flaws in CACTI deem it unfit for analyzing realistic performance of the cache, and the circuit performance values reported by it quite meaningless, from the circuit designer's perspective. This observation is quite pertinent for sub 90 nm CMOS technologies, where variability plays a major contributing factor in determining circuit performance.

To adequately integrate these two important factors for the cache performance values at the system level, a flow is developed which integrates CACTI with a circuit simulator for determining actual performance values. The empirical models are then developed in MATLAB, based on the training data and the sampled values in the variability space. The models developed here form the basic building blocks of the nonlinear probability constraints, formulated during the optimization step.

2.1 CACHE LEAKAGE POWER

The leakage power estimation of CACTI is quite simplistic. It considers the drain-to-source Subthreshold leakage current for all the transistors that are "OFF" with V_{dd} applied across their drain and source. This however does not take into account the gate leakage component of the total leakage dissipation or the inherent process variations.

To get a picture of the modeling methodology used in this thesis, a brief overview of the cache total leakage power equation [26] used in CACTI is given below

$$P_{leak}=P_{leak_request_network} + P_{leak_reply_network} + P_{leak_mats} \quad (4.15)$$

Where,

P_{leak_mats} is the total leakage current associated with the predecoding/decoding logic of all the mats (which are shared by all the subarrays of a single mat) and the total leakage power of the cells in the data array

$P_{leak_request_network}$ and $P_{leak_request_network}$ is the leakage in the routing network from the Bank to the subarray of the cache

In the modeling methodology used to model the cell level variability, the process variations are considered for the cells in the SRAM array. Thus the total leakage power of the cache can be simplistically represented as

$$P_{leak} = P_{leak_cells} + P_{interconnects_periphery} \quad (4.16)$$

Where,

$P_{interconnects_periphery}$ represents the leakage current value in the in the routing network from bank to the subarray and the peripheral logic unit of every subarray.

In the presence of the process variations, the total standby leakage equation can be considered as

$$P_{leak}(X) = P_{leak_cells}(X) + P_{interconnects_periphery}$$

Where, $X = \left[\Delta V_{th1}, \ ... \ , \Delta V_{th6,} \Delta T_{ox1,} \ ... \ , \Delta T_{ox6} \ \right]$ represents the variation space for the random variables of a 6-T SRAM cell. Furthermore, the total leakage can be represented as a function of the total leakage current per cell.

$$P_{leak_cells} \ (X) = N_{subarrys_per_data_array} . P_{leak_mem_cell}(X) \tag{4.17}$$

$$P_{leak_mem_cell}(X) = N_{cells_per_subarray} . P_{mem_cell}(X) \tag{4.18}$$

$$P_{mem_cell}(X) = V_{dd} . I_{leak_per_cell}(X) \tag{4.19}$$

$$N_{cells_per_subarray} = N_{subarr_rows} . N_{subarr_cols} \tag{4.20}$$

$$N_{subarrys_per_data_array} = N_{banks} . N_{subbanks} . N_{mat_in_subbanks} \tag{4.21}$$

Here $N_{subarrys_per_data_array}$ and $N_{cells_per_subarray}$ are the number of subarrays per data array nd the number of SRAM cells per subarray. These values are evaluated on the fly by CACTI to achieve an optimum delay-power product [27].

The value of $I_{leak_per_cell}$ consists of total leakage current components due to both gate tunneling component and Subthreshold leakage component and can be developed using the appropriate lognormal models described in the previous sub-section. The $I_{leak_per_cell}$ is developed as a lognormal approximation of the total leakage current component of every transistor in that cell and the process variability parameters (V_{th} and T_{ox}) associated with every transistor. Thus it is a lognormal leakage approximation of the total leakage current of a single cell. The same is represented below

$$I_{leak_per_cell}(X) = E^T . U \tag{4.22}$$

$$E = \left[e^{\alpha_1 \Delta Vth_pl} . e^{\alpha_6 \Delta Tox_pl}, e^{\alpha_2 \Delta Vth_nl}, e^{\alpha_3 \Delta Vth_pr}, e^{\alpha_4 \Delta Vth_axr}, e^{\alpha_5 \Delta Vth_nr} . e^{\alpha_7 \Delta Tox_nr} \right]^T \tag{4.23}$$

$$U = [\zeta_1, \ \zeta_2, \ \zeta_3, \ \zeta_4, \ \zeta_5]^T \tag{4.24}$$

Where,

ζ_i and $\alpha_{\rm i}$ are parameters that have to be fitted using suitable non linear regression methods.

It should be noted from the discussion on leakage modeling in the previous section that gate leakage is computed for the "ON" transistors and the Subthreshold leakage component of current has negligible dependence on gate oxide thickness, T_{ox}.

The Monte-Carlo distribution of the logarithm of the total leakage power computed from the modified CACTI, for a 45 nm CMOS technology, is shown in Fig 4.12, which predictable has a wide degree of variation and again highlights the fact that variation leakage plays a significant role in cache performance.

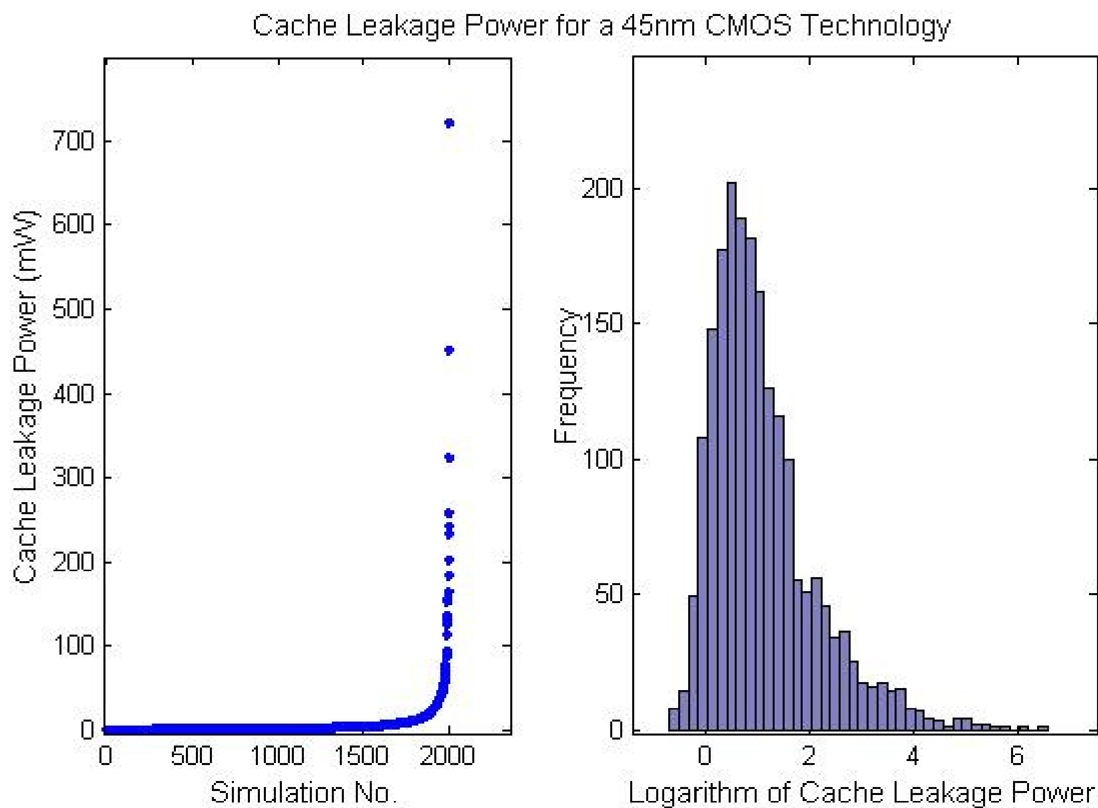

Fig 4.12 Total cache leakage Power as evaluated by a modified CACTI 5.2 under the influence of process variations.

The variation in leakage power for a L2- Cache simulated in CACTI 5.2, in the presence of process variation, shows a variation factor of 1400X from the minimum value to the maximum value and a factor of nearly 100X from the nominal value, as shown in table II. Thus the effect of variation is an important criteria and it is

incorporated in the modified version of CACTI by integrating it with a circuit simulator, while developing the framework for the performance analysis and optimization of memory at the system level. The figure 4.13 shows the fit between the model predicted and the one obtained by actual HSPICE simulations.

Table II

LEAKAGE POWER OF A 45nm L-2 CACHE

Cache Leakage Power (mW)		
Min	Nominal (No variation)	Max
0.5078 mW	7.5 mW	721.3505 mW

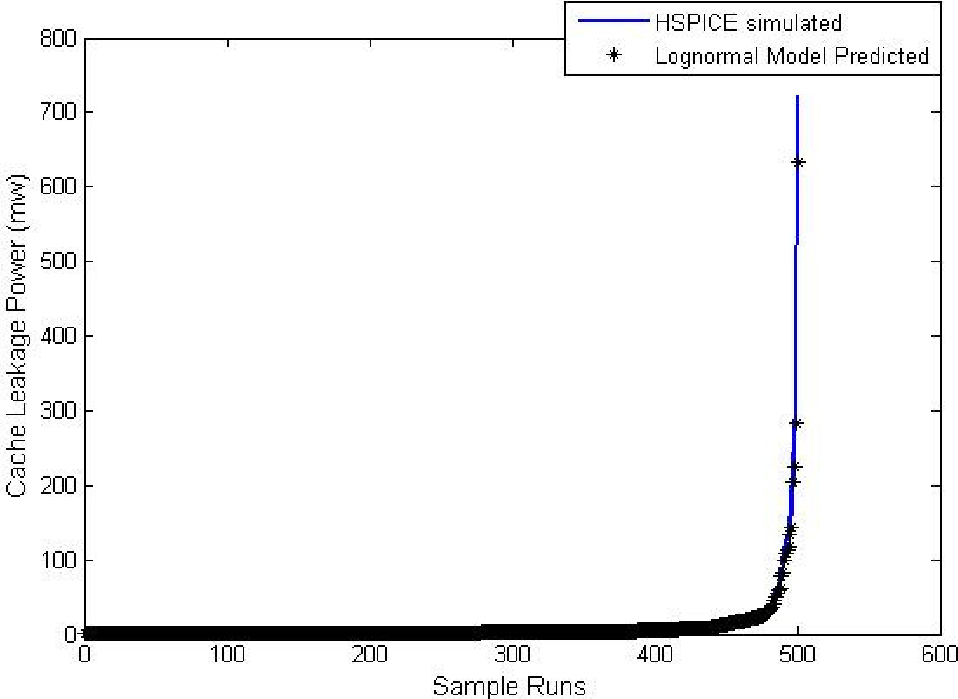

Fig 4.13 Total cache leakage power, a comparison between model evaluated and modified HSPICE-CACTI setup.

2.2 CACHE ACCESS TIME

Internally in CACTI, the total cache access time is evaluated by modeling the interconnects as pi models and replacing the on transistors by their parasitics, which are present in a look up table form. The basic equation used by cacti is shown below

$$T_{access} = T_{request_newtwork} + T_{mat} + T_{reply_network} \tag{4.25}$$

Where,

request_network is the interconnect centric H-tree network to route the address

reply_network is used to route the data value accessed

From the access time equations in [26], the T_{mat} is the delay caused due to the row predecoder, row decoder and the time taken by the cell to pull down the bitline. Since the performance modeling of the SRAM cell is considered, the variability in the SRAM cell would percolate to the total access time equations.

Thus for the variability aware modeling framework, cache access time can be considered to constitute the following

$$T_{access}(X) = \mathrm{T}_{\text{inerconnects+mat_decoding_logic}} + T_{bitline}(X) \tag{4.26}$$

Where, $X = \left[\Delta V_{th1},\ \ldots\ ,\Delta V_{th6,}\Delta T_{ox1,}\ \ldots\ ,\Delta T_{ox6}\ \right]$ represents the variation space for the random variables of a 6-T SRAM cell. CACTI evaluates the total bitline delay by considering the wordline rise time [26]. The approach used in this section also considers the wordline rise time by evaluating the total capacitive loading of the wordline driver. The parasitics for the bitline peripheral circuit, namely Sense amplifier, Bitline and sense amplifier Muxes, and the isolation transistors for the sense amplifier, is derived from CACTI. These are then used as inputs for the response surface modeling methodology described previously, by integrating it with HSPICE. The template for the empirical polynomial estimator is the same as (mention equation number of 1/Tread) and is represented below.

$$\frac{1}{T_{bitline}(X)} = X^T AX + B^T X + C \tag{4.27}$$

Where,

Matrices A and B are the coefficients of the second order and first order terms respectively, equations (4.9) and (4.10)

C = $1/\mu_{bitline}$, is the inverse nominal value of the bitline delay calculated by CACTI in the absence of process variations

Fig 4.14 below shows a comparison between HSPICE determined value and model predicted value. The error distribution shows a good fit validating the accuracy of the model derived in this section

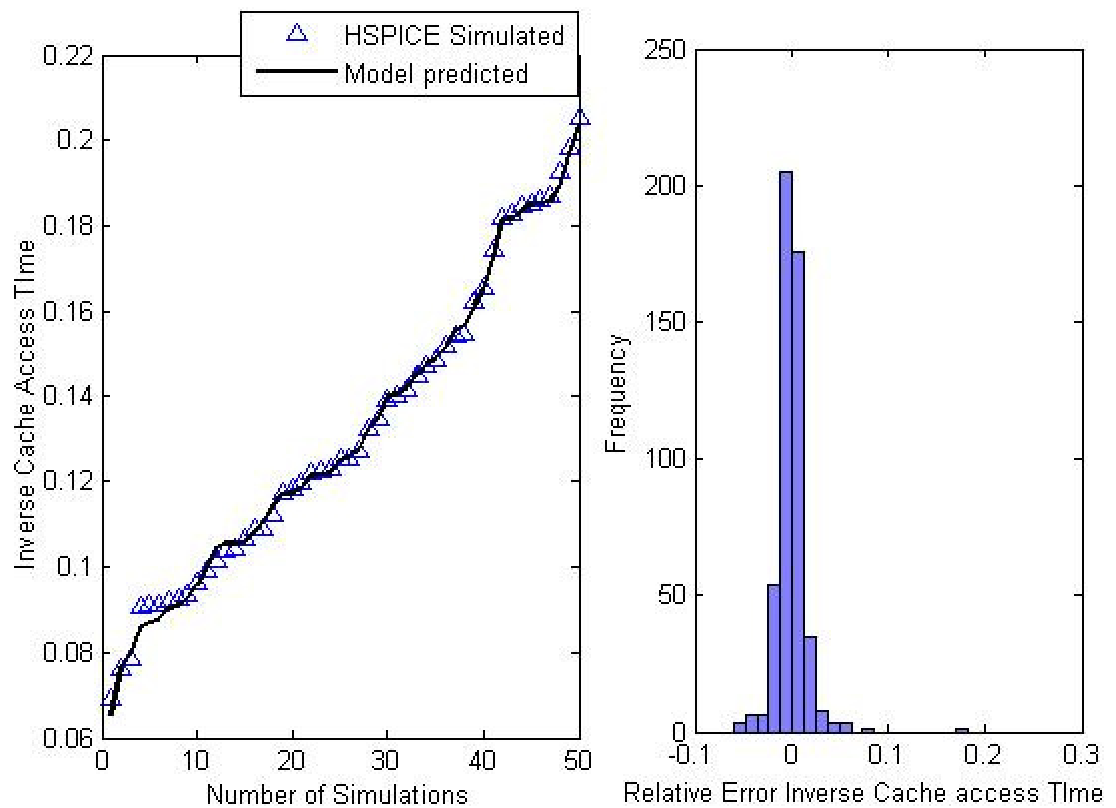

Fig 4.14 Fit between RSM predicted model and CACTI-5.2 for Cache access time

V. IMPACT OF DESIGN PARAMETERS ON SRAM PERFORMANCE

This section provides a basic intuition behind the optimization process and the effect of varying design parameters have on SRAM performance that avoids failure in operation. Another way to analyze performance is to look at various failure mechanisms in an SRAM cell and the impact deign parameters have on avoiding failure. Varying design parameters not only directly affect the circuit performance but also indirectly affect it due to the dependence of process variations on design parameters, specifically transistor widths. In this section the impact of transistor widths, bias voltage and wordline voltage is analyzed for a 6-T SRAM cell during normal modes of operation.

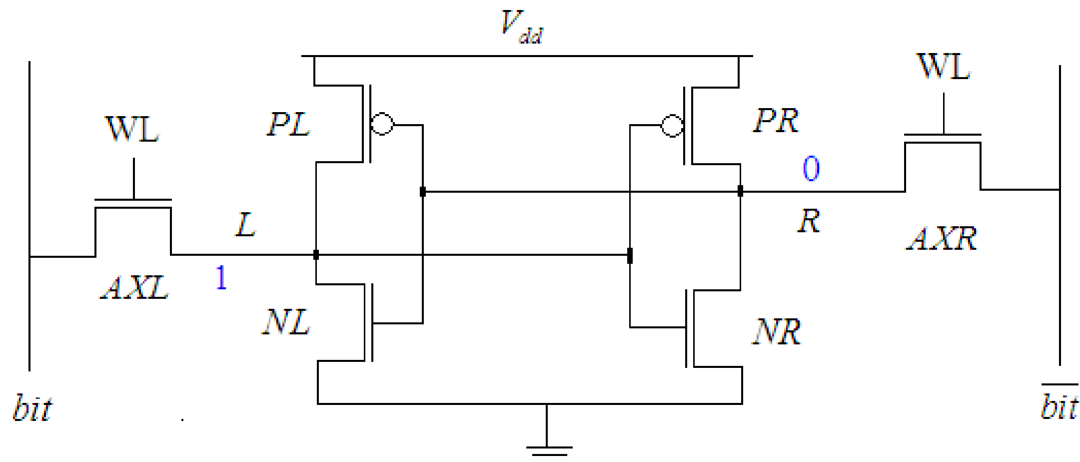

Fig 5.1 A 6-T SRAM cell storing a "1".

During the read operation of the circuit shown in Fig 5.1, the bitlines bit and bit' are set high and the cell is storing a "1" shown by the value at node "L". While during a write bit is set to "0" and bit' is set to "1'.

1. SIZING THE PULL-UP TRANSISTOR (WP)

In the figure the pull up transistor Wpl is switched "ON" and pull up transistor Wpr is switched "OFF". During a read operation Wpl helps in maintaining a "1" value at

the node L. A larger value of Wpl provides a low resistance from bias voltage to the node L thus effectively shorting V_{dd} and node L, leading to a stable read operation, thus increasing the Wp value results in read stability.

However, during a write operation i.e. writing a "0" to node L, sizing up the pull up, Wpl, leads to a degraded write operation as it resists the change in value at node "L". This is due to the relative low resistance between node L and supply voltage $V_{dd.}$ The voltage division between Wpl and Waxl during a write operation dictates that a smaller voltage drop should be between bit and node L and higher voltage drop between node L and power line. As transsitor "ON" resistance is inversely proportional to transistor width, the pull up Wpl should be weaker than access transistor Waxl to increase write ability.

Pull up transistor sizing has no impact on the read access time value while for write access time; a stronger pull up degrades the write access time. The effective capacitance on node "L" during a write has a discharge path through access transistor Waxl and charge path through pull up transistor Wpl. For a shorter write access time, the discharge rate should be faster than the charging rate, hence the negative impact of increasing Wp on write access Time.

2. SIZING THE ACCESS TRANSISTOR (WAX)

During the read operation the discharge path is through access transistors Waxr and Pull down Wnr. Due to the voltage divider operation between them a stronger Waxr implies a smaller "ON" resistance for it, hence, a larger voltage drop between node R and bitline. For read stability, the voltage at node R should not rise above the trip voltage for inverter PL-NL to cause a flip in state, hence the resistance of access transistor Waxr should be much less than pull down Wnr, implying a weaker access transistor compared to the pull down.

For read access time an increase in access transistor width is going to reduce the read delay due to the smaller value of "ON" drain-to-source resistance for it. This leads

to a decrease in the time constant value for the charging current resulting in reduced read access time.

To ensure write ability the access transistor should be stronger than the pull up because of the voltage division action due to Wpl and Waxl, at node L as outlined in the previous subsection. However, a stronger access transistor increases the discharge time for a write access resulting in increased write access time.

3. SIZING THE PULL-DOWN TRANSISTOR (WN)

The read stability of an SRAM can be ensured by making the Pull down transistor sufficiently stronger than the access transistor. This can be viewed from the above discussion or from Fig. 5.1. A stronger pull down Wnr ensures that the voltage division action in favor of node "R" storing a "0" by offering a low resistance path to ground. Thus, sizing up the pull down, not only offers improved read stability but also improves the read access time by decreasing the value of time constant for the discharge current.

4. SUPPLY VOLTAGE V_{DD}

The supply voltage has a significant impact on the SRAM performance both during the read as well as the write operation. Increasing the value of V_{dd} during the read operation results in improved read stability as a larger voltage swing between the access transsitor and node L, is required to flip the state. The increased V_{dd} increase the charging current at the node L, hence a large value of discharge current is required to decreased the potential at node L. This however results in reduced write margin during the write operation due to the above stated fact and a larger discharge current is required to drain the capacitance at node L.

5. WORDLINE VOLTAGE V_{WL}

Reducing the wordline voltage is effectively decreasing the drive current value of the access transistors of the SRAM as the reduced gate voltage results in reduced gate to

source voltage. Hence a reduced V_{wl} during a read operation improves read stability, however during a write it degrades write margin value. Thus circuit designers employ a dual supply techniques to lower the wordline voltage only during the read to improve the read stability while lowering supply voltage to curtail power requirements and improved write margins.

VI. PROBLEM FORMULATION AND DESIGN PROCEDURE

In the previous section, an in-depth analysis was done on the circuit performance parameters of the SRAM memory cell. The modeling procedure of the performance parameters forms the backbone of the optimization flow and would be again revisited later in the section. The SRAM cell is a difficult circuit module to optimize due to the inherent analog nature of the circuit itself. The nonlinearities in the circuit performance parameters and their dependence on statistically varying random variables, further complicates the design process and is a challenging objective for the design procedure. The optimization flow methodology presented in this section serves to optimize the SRAM cell to maximize performance of the circuit in the presence of process variations. To fully understand the impact of process variations, a technology node needs to be selected where the process variations play a significant role. For this purpose a 45 nm CMOS technology, predictive technology model (PTM) for low-power applications (PTM LP), incorporating high-k/metal gate and stress effect [28], was selected. This section serves the purpose of describing the basic problem formulation to be utilized for the optimization of the SRAM cell, which would then be leveraged to the three research objectives mentioned.

1. PROBLEM DESCRIPTION

In this problem formulation the threshold voltage variation ΔV_{th} and gate oxide thickness, ΔT_{ox}, for every transistor in the memory cell are taken into consideration, thus the intra-die random variations are comprehensively dealt with, in the design flow culminating in a comprehensive problem formulation.

The distributions of the process variations are assumed to be Gaussian for the SRAM cell shown in Fig. 6.1. The simulations are run in HSPICE on a 45nm PTM model for metal gate/high-k CMOS, for low-power applications. The proposed design problem is formulated as a minimization problem.

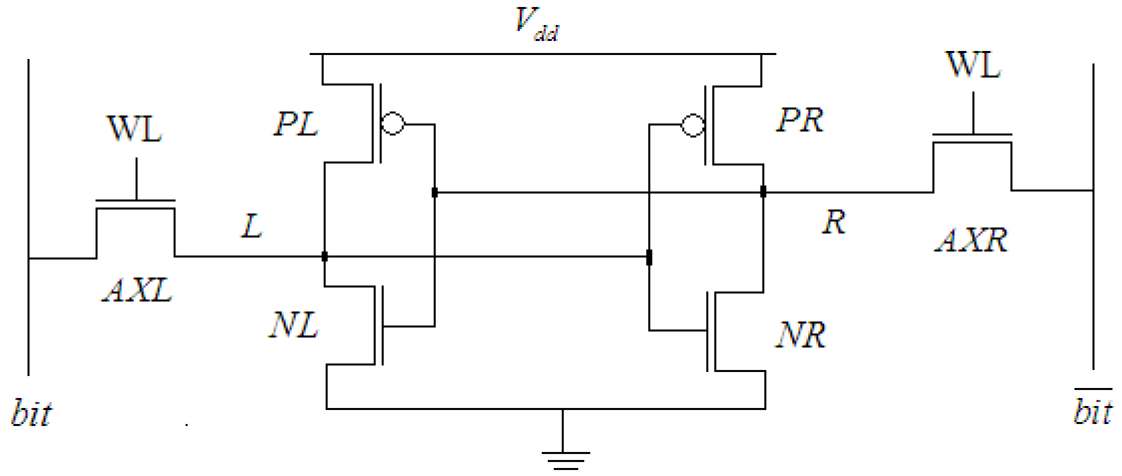

Fig 6.1 A schematic of a 6-T SRAM memory cell.

Minimize: $$f(S) = Area_{SRAM}$$

Subject to:

$$Yield(S,X_i)_{RNM} \geq \left(Yield(S,X_i)_{RNM}\right)_{\min} \qquad (6.1)$$

$$Yield(S,X_i)_{WNM} \geq \left(Yield(S,X_i)_{WNM}\right)_{\min} \qquad (6.2)$$

$$Yield(S,X_i)_{write_access_time} \geq \left(Yield(S,X_i)_{read_access_time}\right)_{\min} \qquad (6.3)$$

$$Yield(S,X_i)_{write_access_time} \geq \left(Yield(S,X_i)_{write_access_time}\right)_{\min} \qquad (6.4)$$

$$Yield(S,X_i)_{leakage_power} \geq \left(Yield(S,X_i)_{leakage_power}\right)_{\min} \qquad (6.5)$$

The design parameters in the design space S are bound quantities represented by: $(S_j)_{\min} \le S_j \le (S_j)_{\max}$ (6.6)

S_j is the jth element in the set {S}

Where,

S is the set of design variables of the 6-T SRAM cell (Fig 6.1)

$S = \left\{ W_{pl},\ W_{axl},\ W_{nl},\ W_{pr},\ W_{axr},\ W_{nr} \right\}$

$X_i = \left[\Delta V_{th1},\ \dots\ , \Delta V_{th6,} \Delta T_{ox1,}\ \dots\ , \Delta T_{ox6} \right]^T$ is the set of process variations for the ith iteration,

f(S) is a differential real function and is the objective or the cost function of the minimization problem.

For the dual optimization problem, where the Wordline and the V_{dd} voltages are taken as additional design parameters, along with the widths of the transistor, the set S can be modified as $S = \left\{ W_{pl},\ W_{axl},\ W_{nl},\ W_{pr},\ W_{axr},\ W_{nr}, V_{WL}, V_{dd} \right\}$

The above stated problem is essentially a non-linear optimization problem with nonlinear constraints. The yield calculations pertains to the probability that the circuit performance parameter (γ), is greater than its minimum threshold (say, $\gamma_{\min}$).

Alternatively, it can also be defined as the probability that the circuit performance parameter (φ), is less than its maximum threshold (say, $\varphi_{\max}$), set by the circuit designer.

The two definitions of yield evaluation are shown below.

$$Yield = P(\gamma \ge \gamma_{\min}) \quad (6.7)$$

$$Yield = P(\varphi \le \varphi_{\max}) \quad (6.8)$$

Thus, this framework provides a methodology to analyze the tradeoff between circuit performance parameters at the cell and system level, in the presence of intra-die process variations with the additional option to analyze the circuit performance in presence of voltage tunability. A similar problem formulation is used for the optimization

framework, the only difference is in the evaluation of the constraints and the objective function, but the underlining concept remains the same.

2. STOCHASTIC DESIGN OPTIMIZATION

The previous section describes the basic problem formulation which is used across in the different design optimization flows, to achieve the three research objectives mentioned previously. Here in this sub-section, the stochastic design flow would be developed in detail for the three different SRAM optimization problems, which form the research objectives, namely, 6-T SRAM memory cell, SRAM memory cell with dual voltage optimization and extension to system level CACHE performance optimization sub-problems. The main purpose is to include the information on device variability in a circuit optimization problem, for a mismatched asymmetric SRAM memory element. Here the empirical polynomial models of circuit performance parameters, described in the previous section would be used to calculate the yield values. The yield values, as mentioned in the problem formulations constitute the nonlinear constraints of the optimization problem.

In Section 2.1 the optimization flow is presented for a 6-T SRAM cell, in section 2.2 this flow is extended for a tunable circuit where the wordline voltage and V_{dd} are not constants and form a part of the design space along the widths of the transistor. In section 2.3 the cell level optimization flow methodology is extended to the system level optimization, where in the constraints of Cache performance are used to minimize its on chip area.

2.1 CELL LEVEL OPTIMIZATION FOR A 6-T SRAM MEMORY CELL

The problem formulation describes a non-linear constrained optimization problem. The objective function is continuous in the domain of S, the design space. The design space is bound by the constraints specified by the designer. A suitable cost

function is used for the minimization problem, using the layout of the SRAM cell [3], the total cell area is computed to be

$$Area(\overline{W}) = X_{cell}.Y_{cell} \tag{6.9}$$

$$X_{cell} = 5\lambda + 2\max(3\lambda, W_{ax}) + 2\max(L_p, L_n) \tag{6.10}$$

$$Y_{cell} = 9\lambda + \max(3\lambda, W_p) + \max(3\lambda, W_n) + L_{ax} \tag{6.11}$$

Where

λ , is the minimum feature size of a technology

[W_p, W_{ax}, W_n] and [L_p, L_{ax}, L_n] are the widths and lengths of the pull up, access and pull down transistors, respectively of the SRAM cell shown in Fig. 6.1

The problem formulation described in the previous section is of the form

$$\min_{s} f(\overline{s}), \quad s.t.\ b(\overline{s}) > 0,\ \mathrm{c}(\overline{s}) = 0 \qquad \forall\ s\ > 0 \tag{6.12}$$

This is a standard description of a Non linear programming (NLP) problem. Thus a non linear programming (NLP) optimization technique would form the working engine of the flow and a constrained algorithm is used to find a cell design which meets all the constraints on the cell performance parameters. A suitable NLP technique should be chosen, which is robust and guarantees an optimal or near optimal solution. Since the objective function is continuous as shown in equations (6.9)-(6.11), sequential quadratic programming or SQP would be an apt choice.

SQP is one of the most popular and robust algorithms for nonlinear continuous optimization. The process can be seen as finding search direction toward optimum sequentially through minimizing the quadratic approximation of the Lagrangian function with the linear approximation of the constraints, as known as quadratic programming. This is the reason why the approach called sequential quadratic programming [29].

The corresponding Quadratic programming (QP) problem form of the NLP problem (equation 5.12) for the kth iteration is

$$Q(\overline{d_k}) = \nabla f^T d_k + \frac{1}{2} d_k^T (\nabla L) \tag{6.13}$$

$$L = f(s) + \lambda^T c(s) + \mu^T b(s) \tag{6.14}$$

Where,

d_k is the search direction obtained by solving the quadratic problem in 6.13.

$Q(\overline{d_k})$ is the QP problem at the kth iteration of the constrained algorithm.

L is the Lagrangian augment form, in which the NLP problem is converted to.

The benefit of converting the NLP problem to a SQP problem is the linearization of the constraints and the optimization problem reduces to a one dimensional search problem. For the purpose of solving the SQP problem, any SQP solver can be utilized, which are easily available from various universities or in commercial software's like MATLAB.

The inputs to the constrained NLP algorithm are the widths of the Pull-up, Access and Pull down devices of the SRAM cell. The set of widths is chosen so that any combination of widths still results in an acceptable area-overhead for the memory design which constitutes the objective function for the constrained algorithm. The different steps of the algorithm are the following:

a) Specify a suitable starting value for the parameters in the design space, i.e. choose an initial nominal value for the widths of the SRAM cell,

   $S_0 = \{W_{pl0},\ W_{axl0},\ W_{nl0},\ W_{pr0},\ W_{axr0},\ W_{nr0}\}$.

   In addition to the design variables, the standard deviation values *(sigma value-σV_{th0} and σT_{ox0});* of the random variables (process variations), of the technology node in consideration; are taken as input parameters in the optimization framework. These values can be extracted from the ITRS roadmap and are reported in [2] for a minimum sized transistor of that technology node.

b) Specify suitable values of the performance parameters of the SRAM cell, which forms the minimum threshold values, beyond which the circuit performance is unacceptable. The probability of no failure at these values would form the lower bound for the constrained nonlinear problem.

c) Calculate the σV_{th} *and* σT_{ox} for the six transistors of the SRAM cell. From Pelgrom's law [30], it is a well know fact that the mismatch parameters or the sigma values, scales with device area and is inversely proportional to the area. The sigma value of the threshold voltage for the ith transistor in the circuit and k^{th} iteration is given by

$$\sigma V_{thi}{}^{k} = \frac{\sigma V_{thi0}.A_{\min}}{\sqrt{W_i^k.L_i^k}} \tag{6.15}$$

d) Evaluate the performance parameters and estimate the probability of no failure for the set of input design parameters, using empirical polynomial estimators. The method to formulate them was described in the previous section.

e) Calculate the Jacobian, J, of the objective function, f, at the current value of widths, defined as:

$$J = \left[\frac{\partial f}{\partial w_p}, \frac{\partial f}{\partial w_{ax}}, \frac{\partial f}{\partial w_n} \right]^T \tag{6.16}$$

$$f : E_3 \rightarrow E_1 \quad \text{and } w_p, w_{ax}, w_n \in S$$

Where E_n defines the Euclidean space in which the problem is defined and S is the set of design parameters.

The elements of the Jacobian matrix constitute the widths sensitivities of the objective function (equation 6.14) and are the inputs to the appropriate NLP method. They factored in evaluating the search direction (equation 6.13) to achieve the minimum value for the cost function (6.9).

f) Calculate the Jacobian/Width sensitivities of the non linear constraints, required by the SQP solver as shown in equation (6.14), to evaluate the value of the search direction vector d_k,, which further gives information as to which parameter in the design space has to be changed to minimize the objective function and satisfy the non linear constraints.

g) If the constraints are not satisfied the solver evaluates the design parameters for the next iteration, S^{k+1}, based on the search direction value at the kth iteration and value of the design parameters at the present iteration (S^k). Repeat steps (c) to (g).

h) If constraints are satisfied and the minimum value of the objective function is reached, based on the termination criteria of the solver, S^k is the value of the design parameters for which the area is minimum and the non-linear constraints are satisfied.

2.2 DUAL-OPTIMIZATION OF A 6-T SRAM MEMORY CELL WITH VOLTAGE TUNABILITY

The voltage tunability is based on multiple voltage levels The optimization flow for the tunable circuit is similar in implementation to section 2.1. The major difference is the expansion of the design space to include the cell voltage (V_{dd}) and wordline voltage (V_{wl}) of the SRAM cell. The cost function is the same and the SQP optimization methodology described in the previous section is used. The different steps in the constrained algorithm used are described below.

a) The design parameters S_0 and mismatch parameters (σV_{th0}, σT_{ox0}) are taken as inputs to the optimization flow.

$$S_0 = \{W_{pl0},\ W_{axl0},\ W_{nl0},\ W_{pr0},\ W_{axr0},\ W_{nr0},\ V_{dd0}, V_{wl0}\}$$

b) Specify desired values of circuit performance parameters of the Tunable SRAM cell.

c) Calculate the σV_{th} *and* σT_{ox} for the 6 transistors of the SRAM cell using 6.15.

d) Evaluate the yield values of the circuit performance parameters, the input to the yield evaluation routines are the widths and voltage values at that iteration.

e) Calculate the widths sensitivities of the objective function.

f) Evaluate the normalized sensitivities of the probability values to the parameters in the design space. The normalized sensitivity for the probability P with respect to the design parameter s_i is given by 6.17

$$\left.\frac{\partial P}{\partial s_i}\right|_{S_i=S} = \frac{\Delta P / P}{\Delta s_i / s} \qquad (6.17)$$

g) If the constraints are not satisfied or optimum value of cost function is not achieved, update the value of the design parameters and go back to step (c).

h) If constraints are satisfied and optimum value is reached, exit out of the iteration and the current value of the design parameters guarantee optimum circuit performance with minimized on chip area.

The entire flow is summarized in Fig 6.2.

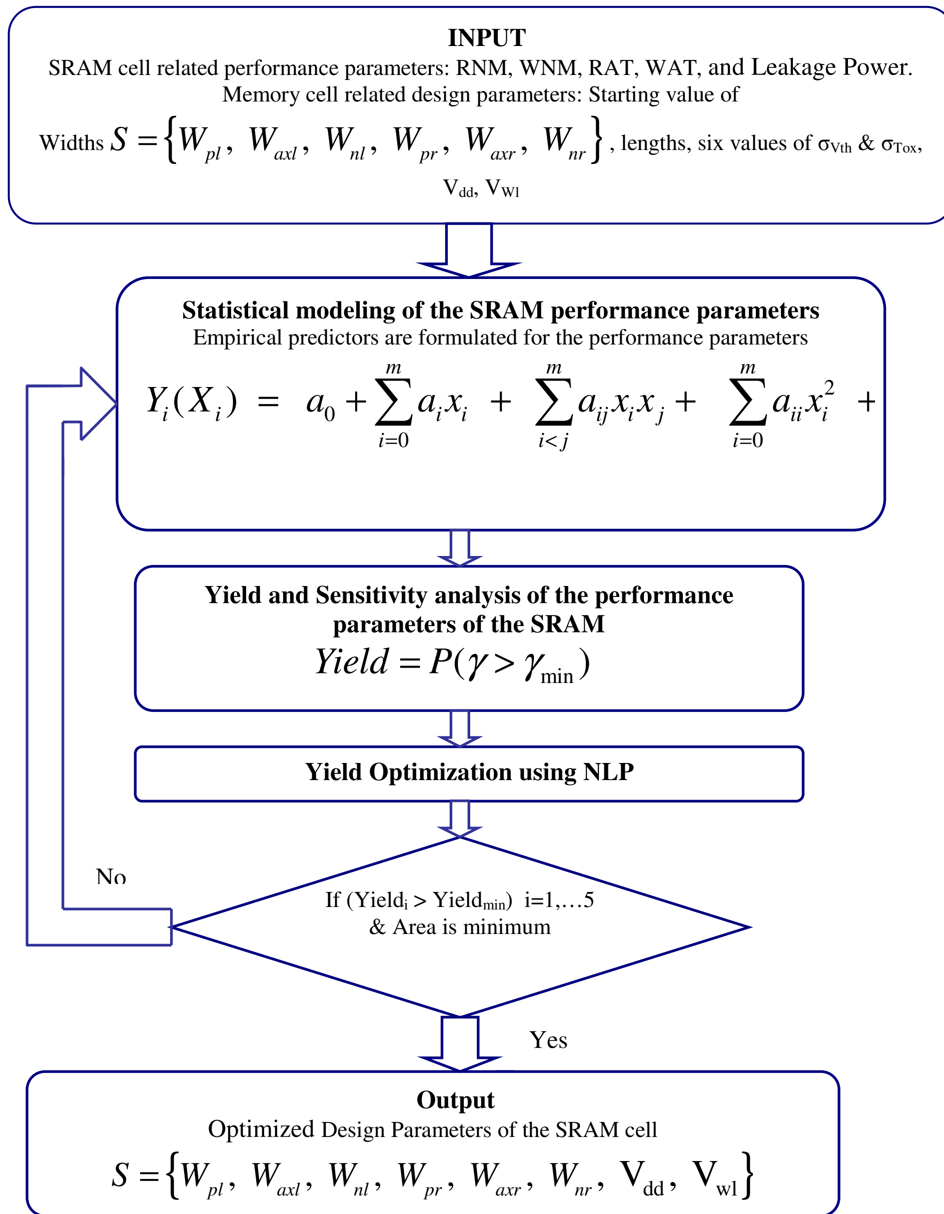

Fig. 6.2 Top level representation of the optimization loop of the NLP problem formulated.

2.3 SYSTEM LEVEL OPTIMIZATION OF CACHES

As stated in the section on modeling of system level performance parameters of Caches, the variability can cause the circuit performance to vary widely, rendering nominal design based performance analysis useless. For that purpose CACTI is modified

and integrated with a circuit simulator to update CACTI, on the fly, based on variability information supplied and cell performance simulated in HSPICE.

2.3.1 SRAM System Level and Cell Level Performance Parameter Interactions

To develop a CACTI based system level optimization framework for optimal design and reliable functioning of caches, the performance metrics for SRAM based memory elements are divided into two categories, cell level performance and system level performance parameters. Under the purview of reliable operation of the SRAM cell and hence the caches are Noise margin criterion. Noise margin, for read and write, of the SRAM cell determines the reliable performance of the cache without being cognizant of the overall design of the Cache. Simply put, if a cell flips due to noise injection, the whole column of cache is rendered inactive and nonuse able. In ECC memories a soft fault can be removed using efficient error correcting codes, but if it is a hard fault then the whole column has to be swapped, hence the idea of introducing redundancy in cache designs. Lower the failure probability lower in the redundancy. This directly implies a smaller on-chip area for the cache. Thus cell level parameters are of paramount importance while developing a system level optimization methodology.

The system level performance parameters encapsulate the standby leakage power and Read access time of the cache. Needless to say, both are very important performance parameters. With the faster processor designs being doled out by companies and SRAM based L-1 caches performing with the speed of the core, the read access time is a carefully controlled parameter and ultimately decided the overall system performance. The system level parameters are also directly or indirectly affected by cell level parameters. For instance the cache leakage power is a function of individual cell leakage current and the cache access time depends on the bitline delay in determining the memory cell access time.

Thus a system level optimization framework having cell level deign parameters is but an obvious spillover of the optimization infrastructure built for the cell level, as

the overall performance of the cache is a direct consequence of system level and cell level memory performance parameters and the inherent interactions between them.

2.3.2 System Level Optimization Methodology

The objective function is again minimization of on chip area, a parameter which is evaluated based on equations (6.9)-(6.11). The problem formulation remains the same as described in equations (6.1)-(6.8). The cell level performances are totally independent of the cache organization and system level design parameters like Associativity, Bank size, interconnect parasitics etc. and are based solely on the micro modeling of the SRAM cells in the cache. The cache design parameters can be easily integrated into the flow as they have a fixed combination of possible values, hence a lookup table methodology would ideally account for them. The optimization flow is explained below

a) Input the design parameters S_0, σV_{th0} and σT_{ox0} to the optimizer setup
b) Specify the optimum system level performance parameters for Standby leakage power, access time and cell level parameters for Static Noise margins (Read and Write)
c) Evaluate the σV_{th} *and* σT_{ox} for the 6 transistors of the SRAM cell using Pelgrom's law.
d) Input the design space and the sampled variation space to modified CACTI-HSPICE setup to perform Design of experiments for performance extraction.
e) Evaluate the Yield values (Y) for the empirical models developed and the sub-gradient of the Yield values to the design parameters.
f) Input the design space to CACTI to evaluate the objective function at a particular value in the 3 dimensional design space.
g) Use finite differencing to evaluate the sensitivities of the objective function to the widths.
h) Input Cell area, its width sensitivity and sub gradient of Yield values to the Nonlinear programming (NLP) solver.

i) Check for optimality condition and check if constraints are satisfied, exit. If not satisfied update the design parameters based on the search direction used for the Sequential programming solver (SQP solver).

VII. EXPERIMENTAL RESULTS

The statistical optimization was carried out on a 6-T SRAM memory cell design using a 45 nm CMOS technology based on a predictive technology model (PTM) for low-power applications (PTM LP), incorporating high-k/metal gate and stress effect [28]. The estimation of V_{th} and T_{ox} mismatch is carried out using the pelgroms's model described in the previous section. For solving the Non-Linear problem, as stated in the problem statement, a sequential quadratic programming (SQP) solver, DONLP2 [31] is used. DONLP2, the SQP solver developed by Dr. Spellucci's is a mixed SQP/ECQP-method for general continuous nonlinear programming problems. This version allows a choice between exact and numerical gradients. Since the sub gradients were internally evaluated in the routines developed for yield evaluation, this solver was optimal for the present optimization flow methodology.

In this section the impact of every design parameter on the failure probability of the 6-T SRAM memory performance is analyzed. This builds the foundation for understanding and developing an intuition for the statistical optimization results at the cell and system level. The general experimental setup is represented in Fig 7.1.

1. SENSITIVITY ANALYSIS OF SRAM PERFORMANCE YIELDS

A) Impact of Transistor Widths

The transistor widths of the cell (W_p, Wax, W_n) impacts the yield values in two significant ways. First, the nominal values of performance parameters of SRAMS (Static Noise Margin, Access Time and Leakage power) are a function of the transistor sizing and any change in widths is going to shift the nominal value. Second, according to pelgrom's law the V_{th} variation is a function of the transistor area; hence any change in transistor widths is going to vary the probability distribution of process variables. Since, the SRAM circuit parameters are a function of Threshold voltage values, the widths impact the mean and variance of the statistical nature of the circuit parameters. In Fig

7.2 (a), (b) and (c) the sensitivity analysis of SRAM performance with varying transistor widths is shown.

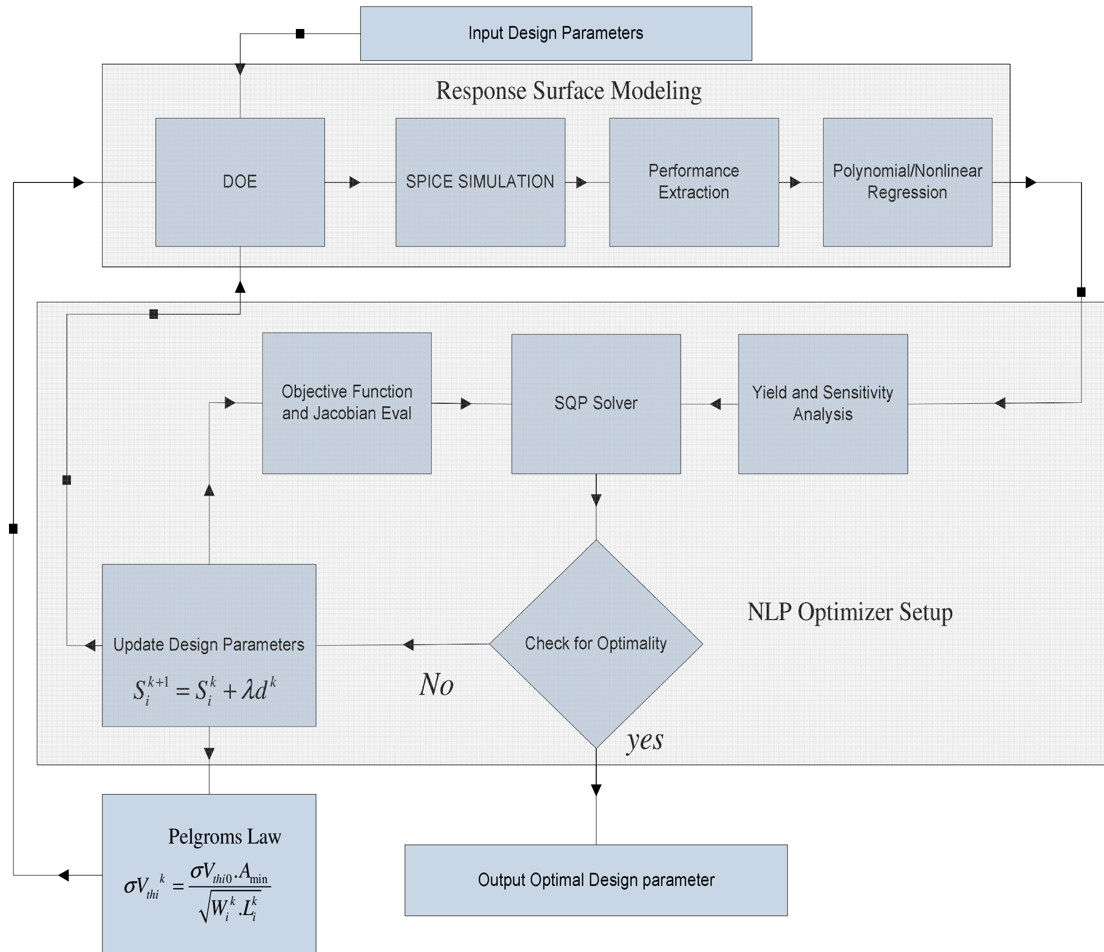

Fig 7.1 Experimental setup of the generalized optimization flow methodology.

Fig. 7.2(a) shows that Static Noise Margin read (RNM) increase, WNM (Write Noise Margin) decreases and access time are not affected much by increasing the strength of the Pull up transistor W_p. The increase in the RNM value is due to the increase in the drive current value of the pull up transistor I_{dspl} or the decrease in the R_{ds_on} value of the

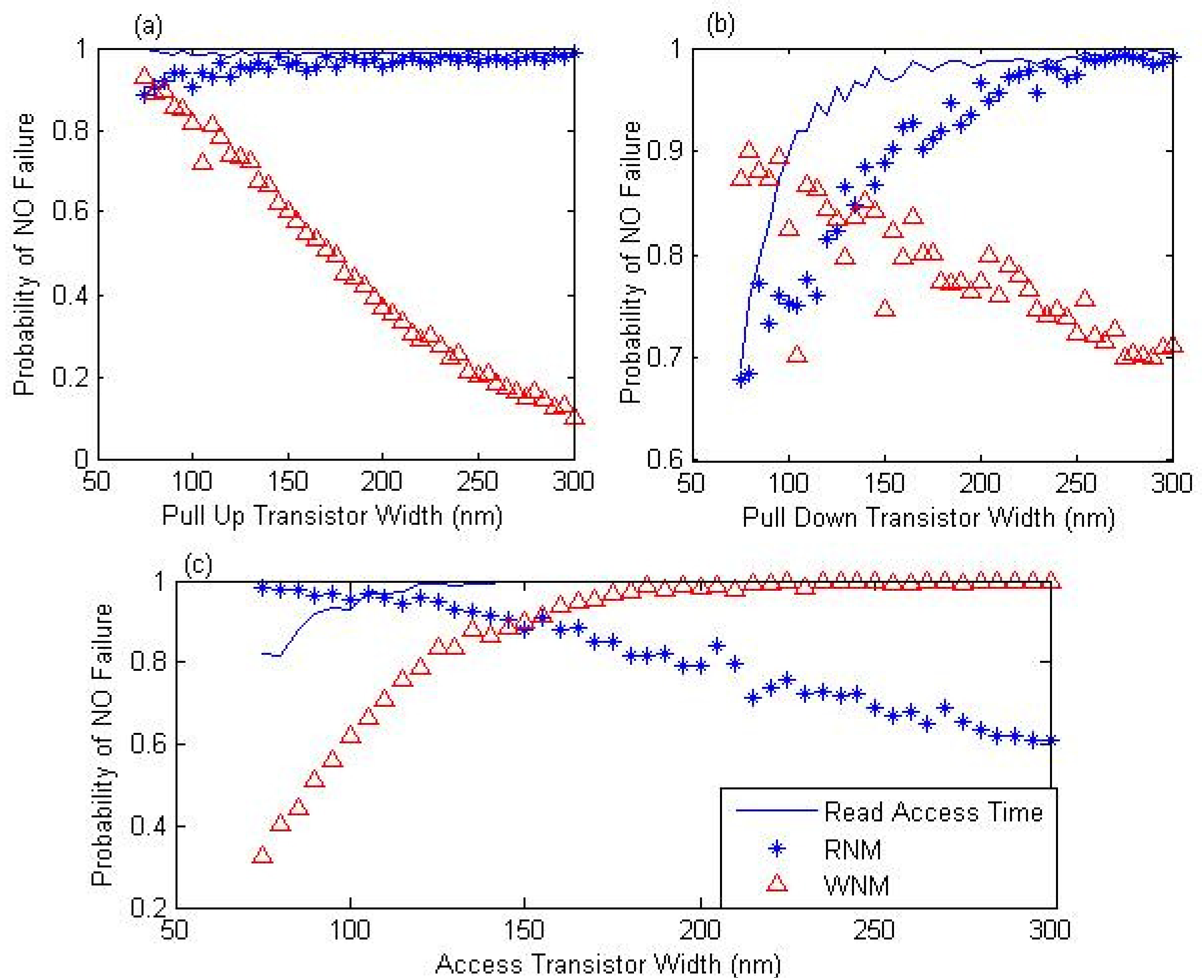

Fig 7.2 Effect of transistor widths on yield of SRAM performance.

pull-up transistor, thus the voltage at node storing a "1" i.e. node L, closely follows that of V_{dd} the bias voltage of the cross-coupled inverters. Hence, it increases the noise that can be tolerated at that node, increasing RNM. Increasing the pull-up strength, however, has the exact opposite effect on WNM. To write a "0" at a node storing a "1" (node L), the discharge current is the difference of the current through the access transistor AXL and the Pull-up, PL, thus a stronger pull up decreases the discharge current. If noise is induced at the bitlines, it decreases the value of V_{ds} of the access transistor, AXL, which further decreases the discharge current. Thus, with increasing pull up transistor strength less noise can be tolerated at the bitlines.

Increasing the strength of the access transistor increase its drive strength during both read and write operations hence decreasing the access time during read, furthermore, due to the increase in discharge current value of the node storing a "1" more noise can be tolerated at the bitlines. This leads to an increase in WNM with increasing width. The RNM, however, decreases due to the voltage division action due to the access and pull down resistances. This makes the node storing a "0" closely coupled with that of the bitline which has a value of "1". This can lead to the flip in value of the cell, if the voltage at this node rises above a tolerance value.

The increasing strength of the pull up transistor increase the drive strength of the pull down, hence the bitlines are discharged faster for the sense amplifier to notice a differential voltage between the bitlines, resulting in a faster access time. Also the increase in pull down transistor reduces the resistance of the pull down comparatively to the access transistor, thus node storing "0" is closely coupled with the ground voltage, improving RNM.

With the background formulated above for the sensitivity of transistor dimensions to the Yield values of SRAM performance, the optimization results are presented in the following subsections.

2. OPTIMIZATION RESULTS

The simulations were carried out on a 45 nm CMOS technology, using PTM models. The input parameters to the optimizer were the widths and the sigma values of V_{th} and T_{ox}. To have a comparative study between the three designs, the input values and the optimization targets have been kept the same in all the initial experimental runs, wherever a comparison is made.

2.1 CELL LEVEL OPTIMIZATION

The target values for the performance parameters used to evaluate the probability of NO failure are given below.

a) Leakage Current : 1nA

b) Read Noise Margin: 0.19 V
c) Write Noise Margin: 0.18 V
d) Read Access Time: 0.9 ns
e) Write Access Time: 0.5 ns

The total number of iterations in the Sequential quadratic programming (SQP), method to evaluate the optimized design parameters was 19 and the total run time was 77.760 $x10^3$ seconds.

After a run of the optimization flow the Yield values for every performance parameter is compared with the optimized value in Table III. To evaluate the tradeoffs of the optimization process, the initial and optimized values of design parameters are presented in table IV

Table III

PERFORMANCE RESULTS OF CELL LEVEL OPTIMIZATION

Performance Parameter	**Yield at Starting Point**	**Yield for Statistically designed Cell**
Leakage Current	99.66 %	99.21 %
Read Noise Margin	93.97 %	97.7 %
Write Noise Margin	52.53 %	91.84 %
Read Access Time	96.59 %	98 .43 %
Write Access Time	99.94 %	100 %

Table IV

OPTIMIZED DESIGN PARAMETERS OF A SRAM CELL

Design parameter	**Initial Value**	**Optimized Value**
Pull Up Transistor Width	140 nm	75 nm
Access Transistor Width	110 nm	117.9032 nm
Pull Down Transistor Width	210 nm	218.4346 nm
Cell Area	0.47590 μm^2	0.47791 μm^2

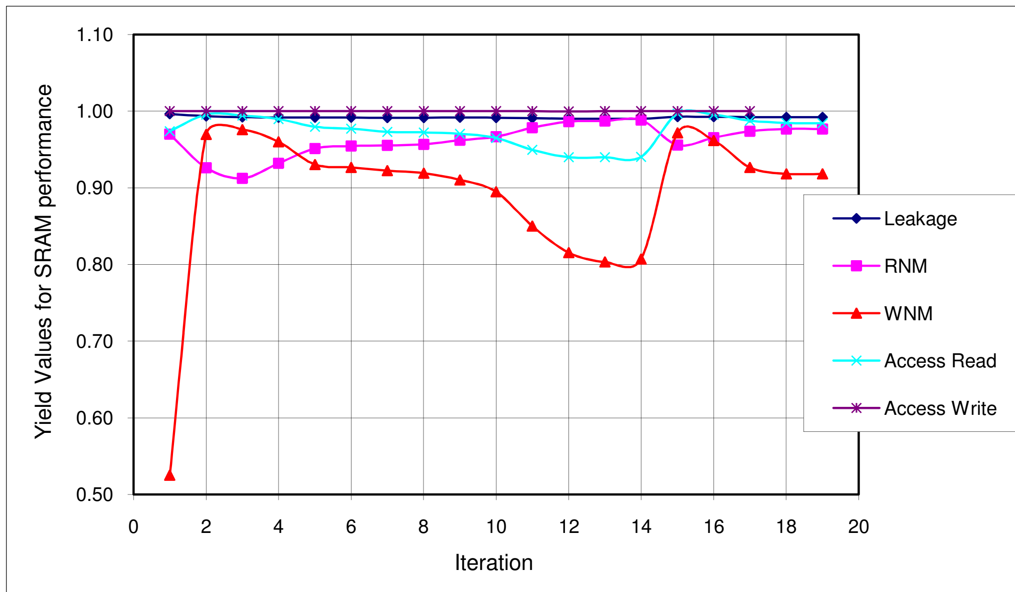

Fig 7.3 Yield Values for every iteration of the optimization process.

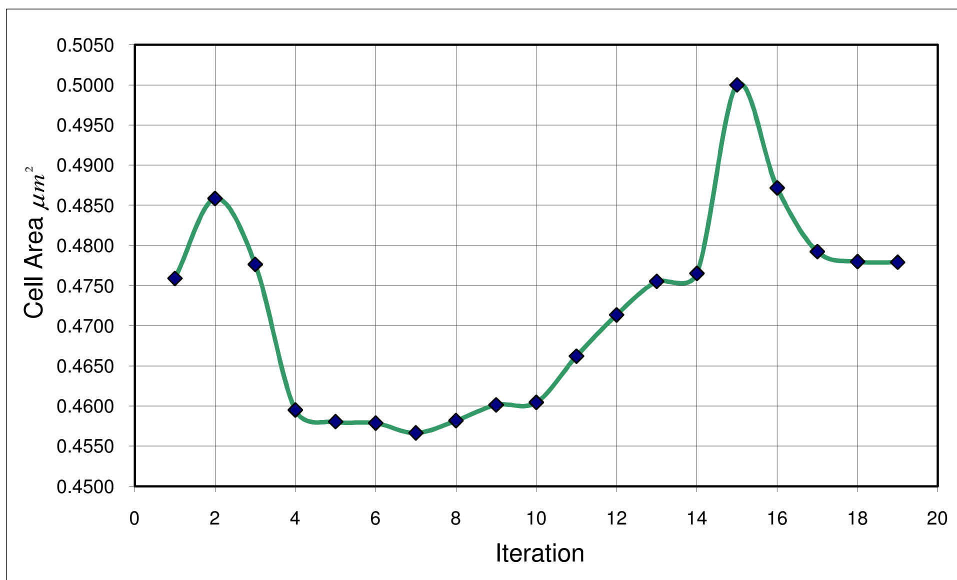

Fig 7.4 Cell area for every iteration of the optimizer.

The yield values and the cell area at every step of the optimization process are shown in Fig 7.3 and 7.4 respectively. Fig 7.5 gives a distribution of the transistor widths for each iteration. It is quite evident that there is a huge improvement in the reliability of the SRAM cell with nearly 40 % increase in the Yield of write noise margin and a minimal area penalty of 0.4 %. It can be observed from Fig 6.3 that the write failure probability is quite high at the beginning mainly due to the stronger pull up as compared to the access transistor. At the end of the optimization process, there is a significant decrease in the W_p value and this is quite evident from Fig 6.2 as the yield values of WNM improve with sizing down the pull down transistor. The increase in variability due to decreased transistor dimensions can offset the gains achieved in area. Further improvement in cell reliability can be achieved but at the cost of relaxing the constraints given. However, the flow is robust enough to provide good values with every run.

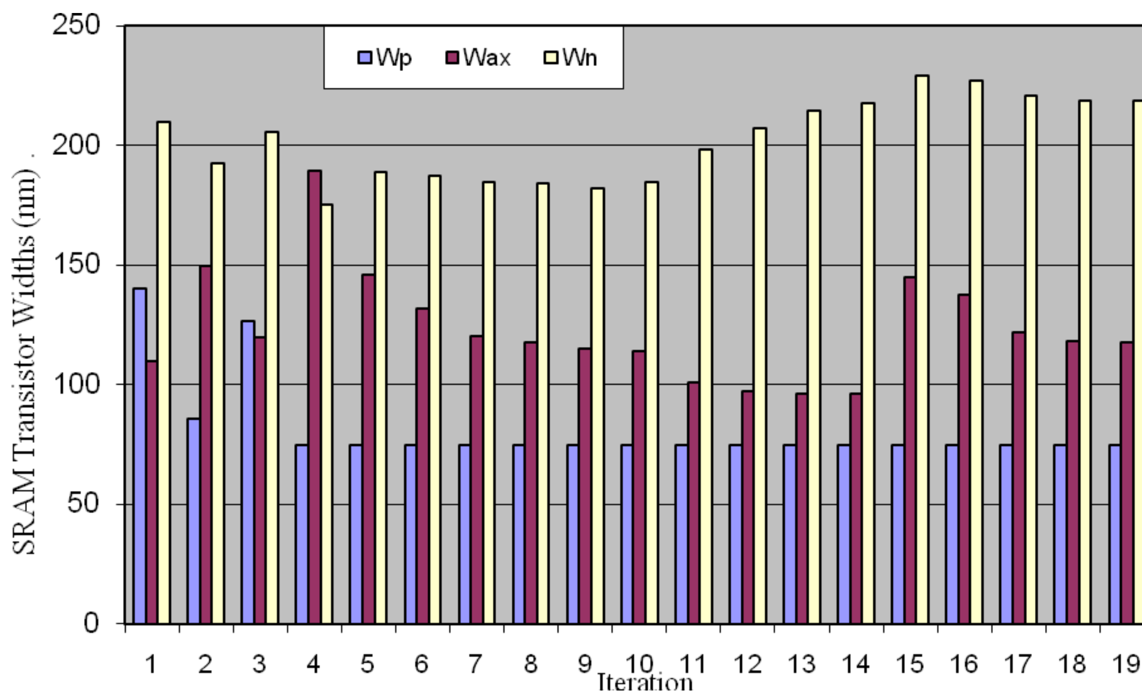

Fig. 7.5 SRAM transistor widths: pull-up (Wp), access (Wax) and pull-down (Wn) at every iteration of the optimizer run.

2.2 OPTIMIZATION OF A VOLTAGE TUNABLE SRAM CELL

In a tunable 6-T SRAM cell, the voltages have been taken as tunable parameters and are additional design parameters in the optimization flow and the circuit setup is as in [32, 33]. To have a comparison basis with the optimization of an SRAM cell the target values and the initial widths have been kept the same as in the previous case. The optimization took 10 iterations to converge to a solution and the time consumed was 39.03×10^3 seconds. The results are shown in tables V and VI.

The tables below show that a reliable SRAM design can be achieved for the same target values as compared to the previous cell optimization. The additional tunable parameter i.e. voltage helps in designing a reliable cell with the yield of WNM increasing by as much as 48 % from the initial value to nearly 98 %. The only drawback is a 27 % increase in the cell area, from the initial value. But for circuits requiring a reliable and robust operation criterion, like on-chip memory in micro-processors, the area is a trade off for improved performance. The joint sizing and voltage tuning is a feature that enables the designer to achieve extremely good levels of noise reliability in addition to reduction in leakage power and dynamic power dissipation.

Thus this dual optimization flow provides flexibility in design and highlights the tradeoff between reliability and area. Moreover, it designs a far more reliable SRAM cell than the one which considers only widths as design parameters in the optimization flow. However, this methodology does not always have an area penalty; the values are dependent on the target values set.

Table V

PERFORMANCE RESULTS OF A VOLTAGE TUNABLE SRAM CELL

Performance Parameter	**Yield at Starting Point**	**Yield-Statistically designed Cell**
Leakage Current	99.66 %	99.9 %
Read Noise Margin	93.97 %	97.83 %
Write Noise Margin	52.53 %	97.69 %
Read Access Time	96.59 %	98 .45%
Write Access Time	99.94 %	100 %

Table VI

OPTIMIZED DESIGN PARAMETERS OF A VOLTAGE TUNABLE SRAM CELL

Design parameter	**Initial Value**	**Optimized Value**
Pull Up Transistor Width	140 nm	92.41 nm
Access Transistor Width	110 nm	177.88 nm
Pull Down Transistor Width	210 nm	300 nm
Bias Voltage (V_{dd})	1.1 Volts	1.0762 Volts
Wordline Volatge (V_{wl})	1.1 Volts	0.9719 Volts
Cell Area	0.47590 μm^2	0.6027 μm^2

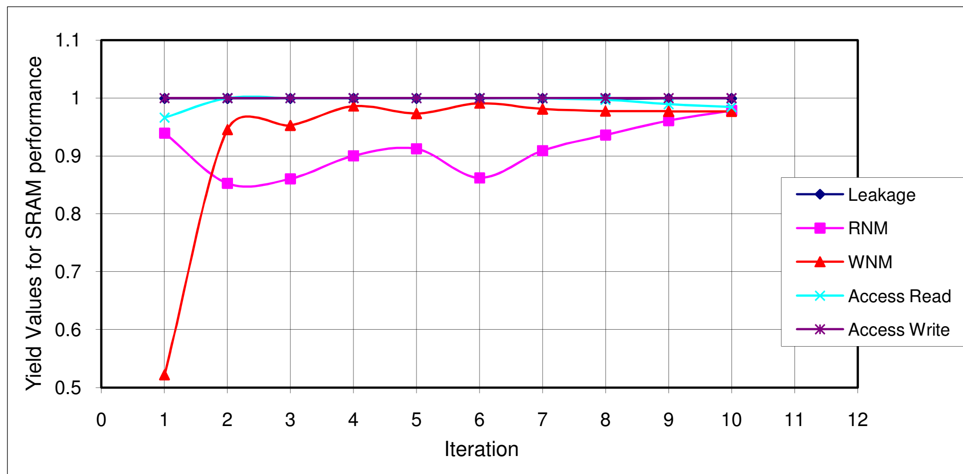

Fig 7.6 Yield values for every iteration of the optimization process of a tunable SRAM cell.

For some values a reduction in area is also possible. The yield values and the area for the above optimization are shown in fig 7.6 and 7.7. The transistor widths at every step of the optimization process are shown in Fig. 7.8. Fig. 7.9 depicts the variation in bias voltage and wordline voltage at every iteration.

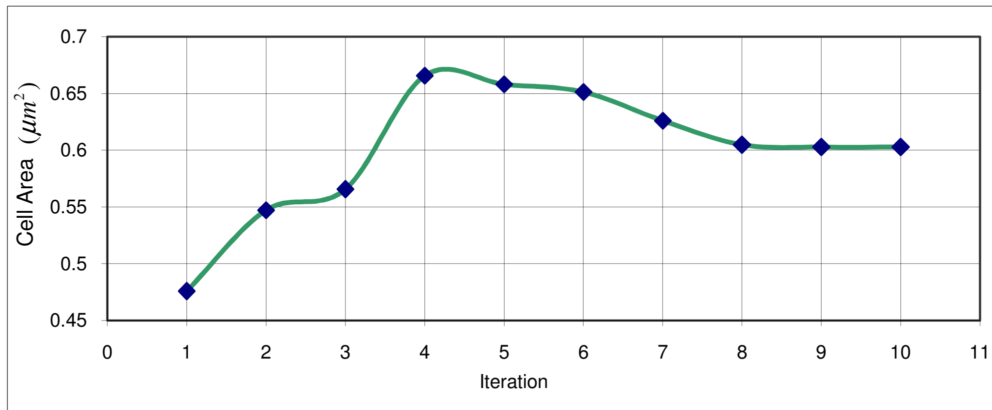

Fig 7.7 Cell area for every iteration of the optimizer for a tunable SRAM cell.

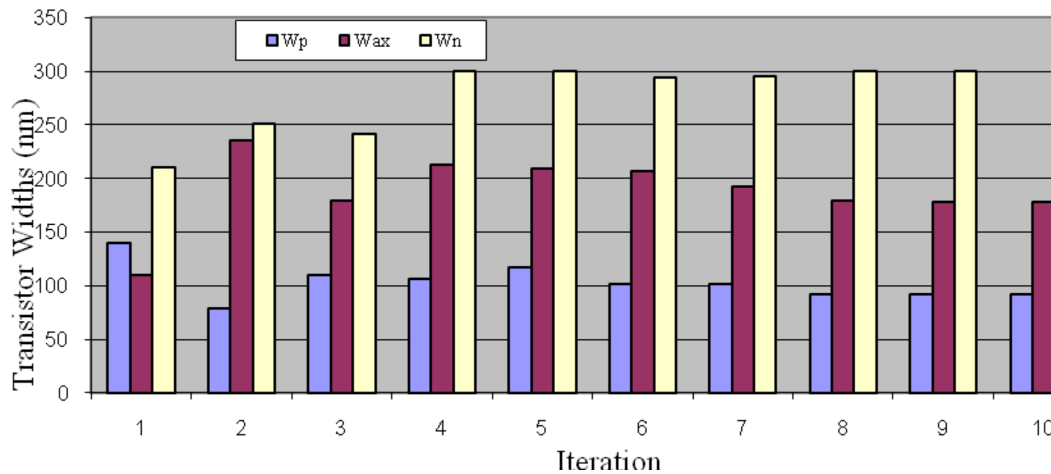

Fig. 7.8 SRAM transistor widths: pull-up (Wp), access (Wax) and pull-down (Wn) at every iteration of the optimizer run for a tunable SRAM cell.

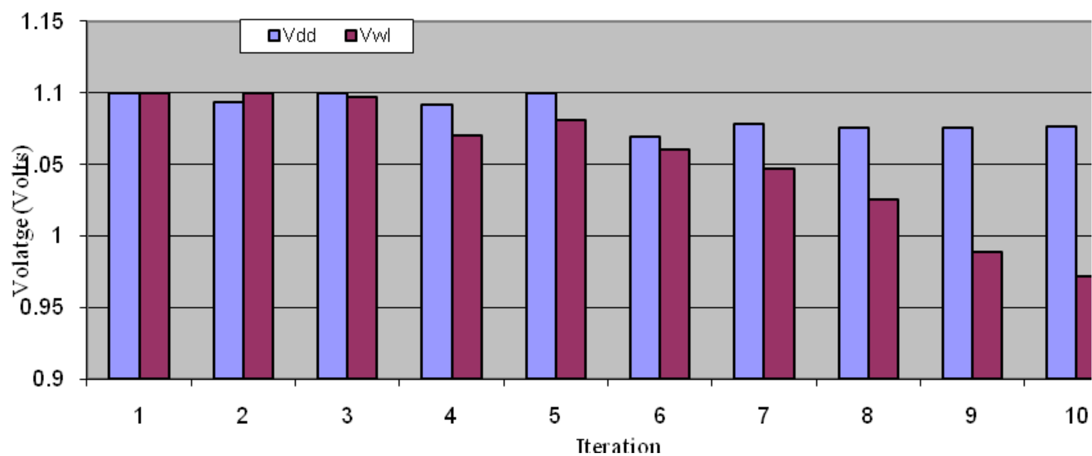

Fig. 7.9 SRAM voltages: supply voltage (Vdd) and wordline voltage (Vwl).

2.3 SYSTEM LEVEL OPTIMIZATION OF CACHES

The system level optimization flow has a set of cell level constraints and system level constraints. Cache leakage power and cache access time constitute the system level constraints, while, RNM and WNM constitute the cell level constraints. The area estimation and the evaluation of the system level constraints is carried out by a modified CACTI 5.2 on a 45 nm technology node LSTP (low standby leakage power) device, for a single port L-2 cache RAM operating at a fixed temperature.

The system level design parameters initialized at the beginning of the simulation are

- Number of Banks : 4
- Set associatively : 2 way
- RAM size : 4 MB

Table VII

PERFORMANCE RESULTS OF A L-2 CACHE RAM

Performance Parameter For the L-2 Cache	**Yield at Starting Point**	**Yield for Statistically Cache**
Leakage Power	98.45 %	99.12 %
Read Noise Margin	93.97 %	96.01 %
Write Noise Margin	52.53 %	94.39 %
Read Access Time	97.22 %	99 .12%

Table VIII

OPTIMIZED DESIGN PARAMETERS OF A L-2 CACHE

Design parameter	**Initial Value**	**Optimized Value**
Pull Up Transistor Width	140 nm	79.1955 nm
Access Transistor Width	110 nm	129.7791 nm
Pull Down Transistor Width	210 nm	207.4264 nm
Cell Area	34.282566 mm^2	34.027384 mm^2

The results above in tables VII and VIII, show a 0.8 % reduction in on chip area of an L-2 Cache RAM while simultaneously improving the RAM performance. The Yield value due to WNM increases by 42 %, due to RNM improves by 3 % and Leakage power and Cache access time improve by 1 %. Thus the optimization process leads to a far more reliable Cache design the initial design. The robust optimization procedure estimates the reliability in the presence of process variations at 45 nm.

It is quite evident that a reduction in Pull up width and an increase in access transistor is the most prominent difference in design, as it leads to a better WNM, contributing in reducing the RAM failure probability.

VIII. FUTURE WORK

Any SRAM cell that is designed should be stable to noise induced at the cell storage nodes, hence noise stability much be insured for the cell-level design. Existing optimization approaches achieves this goal by considering traditional static noise margins (SNMs) [6], [11]-[14]. With shrinking access cycle times and the advent of advanced read/write assist circuits, SRAM operations become increasingly dynamical in nature. As a result, SNMs are increasingly inappropriate for specifying cell stability. Hence, it becomes absolutely imperative that any SRAM optimizer/design should consider new dynamic noise margins (DNMs) [34, 35] to accurately predict the circuit performance in the presence of important nonlinear dynamics. The use of conventional SNMs can have a negative impact on overall cell failure probability with a significant area and leakage overhead.

1. DYNAMIC VS STATIC

In Static Noise analysis, the inputs and outputs of the cross-coupled inverters in an SRAM or any other logic circuits are assumed to be DC signals. It was widely considered as a good measure to analyze the stability of a logic circuit [17]-[18]. However, noise in digital circuits is seldom DC in nature. Various sources of noise in digital circuits include both external noise sources, pertinently, single event upsets (SEU's) and on-chip noise sources, some of which include, power and ground network noises, capacitive coupling and noise injected by substrates. The aforementioned sources of noise are transient and a DC noise model does not account for the time dependence of the noise signal. Traditionally, static stability analysis considers only the maximum amplitude of the voltage deviation an input node can tolerate. This myopic viewpoint, ignores a poignant fact that not all transient noise affecting a sensitive node of a logic gate, or in our case, an SRAM cell, will cause the state to flip. A range of noise signals can be identified which cause only a temporary disturbance of the internal node voltage without flipping the state. In order to decide, whether a transient noise is detrimental and affects a flip in state of a noise sensitive node depends not only on the

amplitude of the signal, but also its time duration.

Input noise, even though its amplitude is greater than the static noise margin of the gate specified, might not cause a significant change in the voltage at the gate output [36]. As a result, static read noise analysis tends to be pessimistic and might lead to overdesign while the counterpart for write tends to be optimistic and might lead to write failures Hence, it becomes imperative to consider not only the amplitude of the injected noise, but its duration too. This transient nature of noise in logic circuits is defined as Dynamic noise analysis. Furthermore, static noise can be considered a subset of dynamic noise. In this alternative viewpoint, static noise can be equivalently described as a *dynamic noise of infinite duration and constant amplitude,* analogous to a unit step function. For SRAM based memory arrays, stability constraint is an essential design constraint. In sub-90nm technologies, the power minimization is achieved in lieu of reduced supply voltage. The supply voltage scaling reduces the design options available especially at sub-45nm technology nodes. Static noise Margin (SNM) based design furthermore limits the design options at the cost of reliability and possibly, at the cost of increased on chip area. SRAM stability using Dynamic noise margin (DNM) requires a complex transient stability analysis, but simultaneously, offers more flexibility to the design enabling, without compromising on reliability and accuracy.

To develop the notion of dynamic stability of a 6-T SRAM cell, an important characteristic to be considered is the stability boundary also defined as a *separatrix* [34, 35]. During the normal modes of operation of a SRAM cell during a read, if a transient noise perturbs the stable state across this boundary, the cell flips in state. While for a write, the transient noise affects the time to produce a state flip for a SRAM cell, thereby resulting in a write failure. For a perfectly symmetric SRAM cell, without any process variation this separatrix or stability boundary is a 45˚ line. Process variations introduce asymmetry into the SRAM cell due to device mismatch. Hence, the separatrix no longer remains a 45˚ line. In [34, 35], the authors come up with a nonlinear system theory to compute the separatrix for an asymmetric SRAM cell. The approach leads to a high level of SPICE like accuracy. In this paper their methodology to define our metrics for the

dynamic noise margins for the read and write operation of a 6-T SRAM cell. The empirical models for DNM are estimated by finding out the time duration for the SRAM cell to cross over the non linear separatrix, during the read and write operation.

2. DYNAMIC NOISE MARGIN

In section II we define the separatrix, a boundary between the stable and the unstable regions of SRAM operation in the presence of a transient noise source. The dynamic noise margin definition used by [34]-[36] estimate the dynamic noise margin in the time domain. In the presence of process variations the separatrix is not a 45^0 line, hence a good estimation needs to be done for noise analysis to guarantee accuracy. Here we estimate the transient nature of noise in the read and write operation modes of the SRAM by carrying out transient HSPICE simulation. We estimate the time it takes to cross the separatrix during the read and write operation [34, 35]. For a flip to take place the transient disturbance should be sustained longer than a minimum critical duration it takes for the operating point to cross the stability boundary. Based on the above definition derived from [35] we can model the dynamic noise margin for the read and write operation by estimating the time duration it takes to cross the metastable point on the separatrix.

2.1 READ DYNAMIC NOISE MARGIN

The read DNM is defined for a given wordline pulse width T_R. Let T_{across} be the time required for the transient state trajectory of the cell to reach the separatrix, or the stability boundary, of the hold mode, Fig 8.1 (a). The read DNM is mathematically defined as [33]

$$T_{DNM,R} = T_{across} - T_R \tag{8.1}$$

For a stable dynamic read operation $T_{DNM,R} \geq 0$, i.e. the transient noise pulse does not cause a state flip.

2.2 WRITE DYNAMIC NOISE MARGIN

The write DNM is defined for a given wordline pulse width T_W. Let T_{across} be the time required for the transient state trajectory of the cell to cross the separatrix, to cause a state flip, Fig 8.1 (b). The read DNM is mathematically defined as [35]

$$T_{DNM,W} = T_W - T_{across} \tag{8.2}$$

For a stable dynamic write operation $T_{DNM,W} \geq 0$, i.e. the time it takes to cause a state flip is less than the wordline pulse width.

We model the DNM for both read and write as second order polynomials, as a function of process variation, for response surface modeling and use the experimental setup in [35] for estimating the DNM value for both read and write.

$$DNM(X) = X^T AX + B^T X + C \tag{8.3}$$

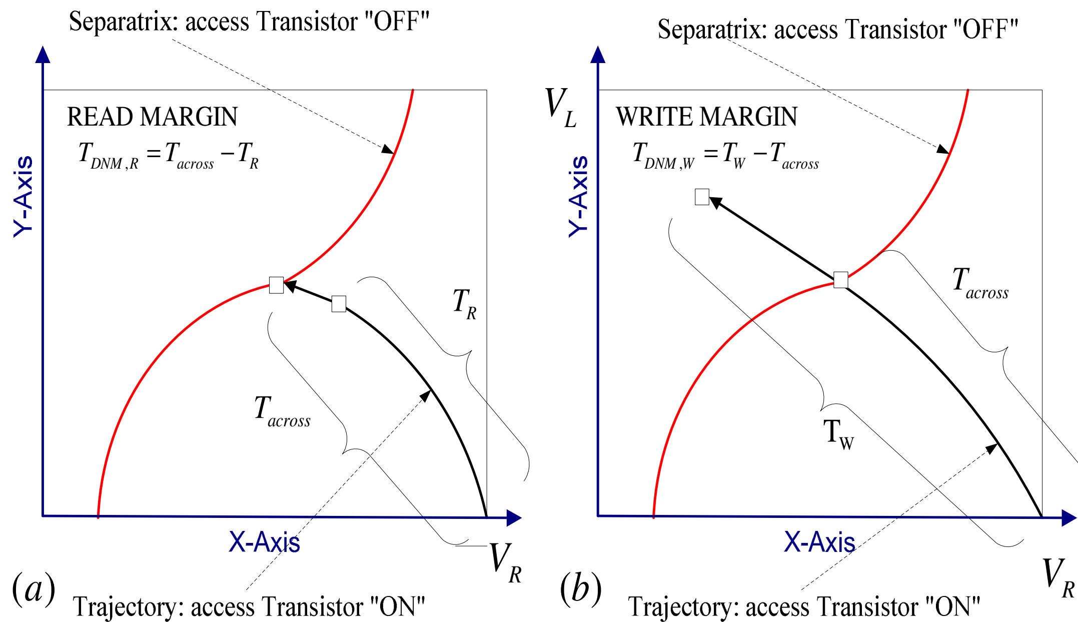

Fig 8.1 (a) Read Dynamic Noise Margin, (b) Write Dynamic Noise Margin.

3. PRELIMINARY OPTIMIZATION RESULTS

The same setup as that for static noise margin (SNM) based optimization is used for the dynamic case for a 6-T SRAM cell. Again the 45 nm CMOS PTM model is used for running HSPICE simulations. Here instead of considering the Static constraints, the dynamic constraints are considered. The target values for the SRAM performance are shown in Table IX

Table IX

PERFORMANCE THRESHOLD OF A 6-T SRAM CELL

Leakage Current	Dynamic Noise margin (Read)	Dynamic Noise Margin (Write)	Access Time (Read)	Access Time (Write)
0.8 nA	200ps	20 ps	700 ps	500ps

3.1 SRAM CELL LEVEL OPTIMIZATION : DYNAMIC PERSPECTIVE

This instance 3 of the optimization flow considers the dynamic noise constraints instead of the static noise constraints and the widths of the SRAM transistors are taken as the input design parameters in the current setup. The optimization results are highlighted in Table X.

Table X

OPTIMIZATION RESULTS FOR DYNAMIC CONSTRAINTS

Performance Parameter	Yield at Starting Point	Yield for Statistically designed Cell
Leakage Current	99.02 %	99.22 %
Dynamic Noise Margin (Read)	100 %	99.97%
Dynamic Noise Margin (Write)	2.46 %	90.02%
Read Access Time	96.59 %	99 .99 %
Write Access Time	99.94 %	100 %

Initial Area=0.47590 μm^2 , Optimized Area= 0.55334 μm^2

We observe a significant improvement in the dynamic stability of the SRAM cell at the optimized point with the dynamic noise margin for the read operation increasing by as much as 87.5%. Thus our optimized cell has an area overhead but ensures a reliable operation with respect to dynamic constraints.

IX. CONCLUSIONS

In this thesis, the performance parameters of a nano-scaled 6-T SRAM cell are modeled as an accurate, yield aware, empirical polynomial predictor, in the presence of intra-die process variations. The estimated empirical models are used in a constrained non-linear, robust optimization framework to design an SRAM cell, for a 45 nm CMOS technology, having optimal performance, according to bounds specified for the circuit performance parameters, with the objective of minimizing on-chip area. Leakage Performance parameters are approximated as sum of lognormals for accurately estimating the gate tunneling and Subthreshold component of leakage current.

For a cell level optimization a significant improvement in the SRAM cell performance was observed with an insignificant area penalty. The reliability improved by as much as 39% from the nominal starting value.

Furthermore, a dual optimization approach is followed by considering SRAM power supply and wordline voltages as additional input parameters, to simultaneously tune the design parameters, ensuring a high yield and considerable area reduction. This dynamic methodology provides extra degrees of freedom that can lead to a superior SRAM cell design with respect to reliability. The results achieved from this optimization prove the applicability of this method, with an improved yield for the noise margin levels by as much as 48 %. Area penalty is significant in some cases, while a relaxed set of target values can reduce the area penalty or even lead to a decrease in area.

In addition, the cell level optimization framework is extended to the system level optimization of caches, under both cell level and system level performance constraints. The cell level performance parameters have a significant impact on the total cache performance due to the inherent interaction between the cell level and system level parameters. Hence the optimization methodology uses a dynamic and robust design technique to consider both system and cell level parameters that governs optimal cache performance. The optimization process resulted in a reduction in area from the nominal value and a 42 % improved reliability in cache performance.

The statistically aware technique developed in this thesis provides a robust and a realistic design methodology to study the tradeoff between performance parameters of the SRAM with analysis for both cell level design and optimization of caches.

REFERENCES

[1] G.Moore, "Moore's Law: Made Real by Intel® Innovation", Feb. 7, 2008. [Online]. Available: http://www.intel.com/technology/mooreslaw/. [Accessed: July, 2009]

[2] ITRS (International Technology Roadmap for Semiconductors), [Online]. Available: http://www.itrs.net/Links/2007ITRS/2007_Chapters/2007_Design.pdf [Accessed: June, 2009]

[3] S. Borkar, "Design Challenges of Technology Scaling," *IEEE Micro,* vol. 19, no. 4, pp.23-29, July/August 1999.

[4] "Static Random Access Memory", [Online]. Available: http://en.wikipedia.org/wiki/Static_random_access_memory. [Accessed: July, 2009]

[5] N. Yoshinobu, H. Masahi, K. Takayuki and K. Itoh; "Review and future prospects of low-voltage RAM circuits," *IBM Journal of Research and Development*, vol. 47, No. 5/6,pp.525-552, 2003

[6] E. Grossar., M. Stucchi, K. Maex and W. Dehaene, "Statistically aware SRAM memory array design," in *Proc. IEEE ISQED*, San Jose, CA, 2006.

[7] H. Chang, S. S. Sapatnekar "Prediction of leakage power under process uncertainties". *ACM Transactions on Design Automation of Electronic Systems (TODAES)*. 12, 2 (Apr. 2007), 12.

[8] A. Srivastava, D. Sylvester, and D. Blaauw, *Statistical Analysis and Optimization for VLSI: Timing and Power*, Kluwer Academic Publishers, June 2005, pp14.

[9] K. Agarwal , S. Nassif, "The impact of random device variation on SRAM cell stability in sub-90-nm CMOS technologies", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol.16 no.1, pp.86-97, January 2008

[10] X.Li, J.Le, L.T. Pileggi, A. Strojwas, "Projection-based performance modeling for inter/intra-die variations", in *Proc. IEEE/ACM, ICCAD*, San Jose, CA, pp: 721 - 727 , 2005

[11] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, “Modeling and estimation of failure probability due to parameter variations in nanoscale srams for yield enhancement,” in *Proc. VLSI Circuits Symposium*, pp. 64–67, 2004.

[12] E. Grossar,”A yield-aware modeling methodology for nano-scaled SRAM designs”, in *Proc. ICICDT* ’05, p. 33-36, 2005.

[13] A. Srivastava, D. Sylvester, and D. Blaauw, ”Statistical optimization of leakage power considering process variations using dual-vth and sizing”, in *Proc. DAC ’04*, June 7-11, San Diego, CA, USA.

[14] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, ”Statistical design and optimization of SRAM cell for yield enhancement”, in *Proc. IEEE/ACM ICCAD ’04*, 7-11 Nov. 2004, pp. 10 -13

[15] N.H.E. Weste, D. Harris and A. Banerjee, *CMOS VLSI Design, A Circuit and Systems Perspective*. 3rd Edition, New Delhi, India, Pearson Education 2008, pp. 55–56.

[16] D. Lee, D. Blaauw, and D. Sylvester, “Gate oxide leakage current analysis and reduction for VLSI circuits,” *IEEE Trans. on Very Large Scale Integration Systems*, vol. 12, no. 2, Feb.2004, pp. 155-166.

[17] C. F. Hill, “Noise margin and noise immunity in logic circuits,” *IEEE Microelectronics*, vol. 1, pp. 16-21, 1968.

[18] J. Lohstroh, “Worst-case static noise margin criteria for logic circuits and their mathematical equivalence,” *IEEE Journal of Solid-State Circuits*, vol. SC-18, no. 6, pp. 803-806, 1983.

[19] J. R. Hauser, “Noise margin criteria for digital logic circuits, ”*IEEE Transactions on Education*, vol. 36, no. 4, pp. 363-368,

[20] E. Seevinck, F. List, and J. Lohstroh, "Static-NoiseMargin Analysis of MOS SRAM Cells," *IEEE Journal of Solid State Electronics*, Vol. SC-22, no. 5, pp 748-754, Oct. 1987.

[21] Design of Experiments, ”Engineering Statistics”. [Online]. Available: http://www.itl.nist.gov/div898/handbook/pri/section5/pri5.htm [Accessed: Aug 2008]

[22] Statistical Toolbox, The Mathworks^{TM}. [Online]. Available: http://www.mathworks.com/. [Accessed: March 2008]

[23] R. Rao, A. Devgan, D. Blaauw, D. Sylvester, “Parametric Yield Estimation Considering Leakage Variability” in *Proc. ACM/IEEE DAC*, pp.442-447, 2004

[24] W.-C. Lee and C. Hu, “Modeling gate and substrate currents due to conduction- and valence-band electron and hole tunneling,” in *Proc. IEEE Int. Symp. VLSI Techno,* pp.198, 2000.

[25] Confidence Interval. [Online]. Available: http://en.wikipedia.org/wiki/Confidence_interval. [Accessed: July, 2009]

[26] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi, “CACTI 5.1: An integrated cache and memory access time, cycle time, area, leakage, and dynamic power model” HP Laboratories, Palo Alto, CA, HPL-2008-20 April 2, 2008.

[27] M. Mamidipaka and N. Dutt, “eCACTI: An enhanced power estimation model for on-chip caches,” Center for Embedded Computer Systems (CECS), University of California, Irvine, CA, Technical Report TR-04-28, Sept. 2004

[28] Predictive Technology Models, 45 nm CMOS PTM models. [Online]. Available: http://www.eas.asu.edu/~ptm/. [Accessed: Feb, 2008]

[29] Sequential Quadratic Programming, SQP DD-WIKI. [Online]. Available: http://ddl.me.cmu.edu/ddwiki/index.php/Sequential_quadratic_programming. [Accessed: March, 2009]

[30] M. Pelgrom, A. Duinmaijer, A. Welbers “Matching properties of MOS transistors” *IEEE Journal of Solid-State Circuits,* vol. 24, no. 5, 1989, pp 1433–1439

[31] Dr. Peter Spellucci, T.U. Darmstadt, DONLP2. [Online]. Available: www.mathematik.tudarmstadt.de:8080/ags/ag8/Mitglieder/spellucci_de. [Accessed: March, 2009].

[32] M. Khellah, Y. Ye, N. S. Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb and V. De, “Wordline & bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65nm CMOS designs,” in *Proc. Symp. on VLSI Circuits*, June 2006, pp. 9-10.

[33] H. Pilo, J. Barwin, G. Braceras, C. Browning, S. Burns, et al. “An SRAM design in 65nm and 45nm technology nodes featuring read and write-assist circuits to expand operating voltage,” in *Proc. Symp. on VLSI Circuits*, June 2006, pp 30-35.

[34] G. M. Huang, W. Dong, Y. Ho, and P. Li. “Tracing SRAM separatrix for dynamic noise margin analysis under device mismatch”. In *Proc. IEEE Int. Behavioral Modeling and Simulation Conf.*, 2007, pp 6-10.

[35] W. Dong, P. Li, G. M. Huang, "SRAM dynamic stability: Theory, variability and analysis," in *Proc. IEEE/ACM International Conference on Computer-Aided Design*, 2008, pp. 378-385.

[36] B. Zhang, A. Arapostathis, S. Nassif, M. Orshansky, "Analytical modeling of SRAM dynamic stability," in *Proc. IEEE/ACM International Conference on Computer Aided Design,* 2006, pp 315-322.

VITA

Akshit Dayal received his Bachelor of Engineering degree in Electrical Engineering from Delhi College of Engineering, University of Delhi in 2007. He entered the Master of Science program at Texas A&M University in August 2007 and will receive his Master of Science degree in December 2009. His research interests include SRAM's, mathematical modeling of variation aware circuit performance and statistical circuit design and optimization.

In spring 2009, he did his internship at Texas Instruments (TI), Stafford, Houston, from January – May 2009. His role during the internship was that of a Design Verification Intern in the Low power design group, part of the ASP group at TI. There he carried out the Pre-silicon Functional Verification of a fixed point TI-DSP (TMS320-C5505) and was responsible for the verification of various peripherals/IP's at the SoC level, like: I2S, GPIO, DMA, MMC/SD, RTC, Timers, Watchdog timers and Interrupt Aggregation logic. Additionally he also carried out an Assertion based verification of the I/O bidirectional buffers at the reset/default state.

Mr. Akshit Dayal may be reached at 53/40 Ramjas Road, Karol Bagh, New Delhi-110005, India. His email id is akshitdayal "at" yahoo "dot" com.