A DUAL DIELECTRIC APPROACH FOR PERFORMANCE AWARE REDUCTION OF GATE LEAKAGE IN COMBINATIONAL CIRCUITS

Valmiki Mukherjee

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

May 2006

APPROVED:

Saraju Mohanty, Major Professor
Elias Kougianos, Committee Member
Murali Varanasi, Committee Member
Armin Mikler, Departmental Graduate Advisor
Krishna Kavi, Chair of the Department of Computer Science and Engineering
Oscar Garcia, Dean of the College of Engineering
Sandra L. Terrell, Dean of the Robert B. Toulouse School of Graduate Studies
Mukherjee, Valmiki. *A Dual Dielectric Approach for Performance Aware Reduction of Gate Leakage in Combinational Circuits*. Master of Science (Computer Engineering), May 2006, 68 pp., 3 tables, 30 illustrations, bibliography, 90 titles.

Design of systems in the low-end nanometer domain has introduced new dimensions in power consumption and dissipation in CMOS devices. With continued and aggressive scaling, using low thickness SiO$_2$ for the transistor gates, gate leakage due to gate oxide direct tunneling current has emerged as the major component of leakage in the CMOS circuits. Therefore, providing a solution to the issue of gate oxide leakage has become one of the key concerns in achieving low power and high performance CMOS VLSI circuits. In this thesis, a new approach is proposed involving dual dielectric of dual thicknesses (DKDT) for the reducing both ON and OFF state gate leakage. It is claimed that the simultaneous utilization of SiON and SiO$_2$ each with multiple thicknesses is a better approach for gate leakage reduction than the conventional usage of a single gate dielectric (SiO$_2$), possibly with multiple thicknesses. An algorithm is developed for DKDT assignment that minimizes the overall leakage for a circuit without compromising with the performance. Extensive experiments were carried out on ISCAS'85 benchmarks using 45nm technology which showed that the proposed approach can reduce the leakage, as much as 98% (in an average 89.5%), without degrading the performance.
ACKNOWLEDGMENTS

I would like to express gratitude to my major professor, Dr. Saraju P. Mohanty, for his guidance and advise throughout my masters degree program. I am highly indebted to Dr. Elias Kougnjos, my co-major professor for his timely help and advice at every stage of my work. I am obliged to Dr. Murali Varanasi for sparing some time out of his busy schedule for reviewing my thesis and approving it. My sincere thanks are due to Ramakrishna Velagapudi, my dear friend and co-author of my work for being of immense assistance to me. I also thank my fellow labmates Anil, Gayathri, Rahul, Atmaram and Yomi for their continued support and encouragement. I would thank my parents for instilling in me the virtues of diligence, discipline and dedication. My higher studies would not have been possible without the selfless cooperation from my younger brother Vidyapati who took care of my ailing parents while I was away in my job and studies. Finally I would like to thank my wife Gayatri for being a pillar of support during all the highs and lows.
CONTENTS

ACKNOWLEDGMENTS ii

LIST OF TABLES v

LIST OF FIGURES vi

CHAPTER 1. INTRODUCTION 1
  1.1. A Brief Background on CMOS VLSI Technology 2
  1.2. VLSI Design Flow and Electronic Design Automation 4
  1.3. An Introduction to Scaling of CMOS 6
  1.4. Short Channel Effects and Leakage Issues 11
  1.5. Why Low Power VLSI Design? 12
  1.6. Introduction to Leakage Reduction Techniques 14
  1.7. Contribution of This Work and Intended Impact 16
  1.8. Organization of the Thesis 18

CHAPTER 2. LOW POWER CMOS VLSI - CURRENT STATE OF THE ART 20
  2.1. Research on Technology Predictions and Scaling 20
  2.2. Research in Physical and Quantum Mechanical Modeling 21
  2.3. Research Based On Compact Modeling 23
  2.4. Research in Leakage Reduction and Analysis 23

CHAPTER 3. NEED FOR HIGH K GATE DIELECTRICS 25
  3.1. Fundamental Limits of Scaling of Gate Oxide 25
  3.2. Alternative Technologies in High-K for Nanoscale CMOS Devices 26
3.3. Alternatives to SiO₂ as Gate Dielectrics 26
3.4. Feasability for Alternative Dielectrics 27

CHAPTER 4. PROBLEM DEFINITION 28
4.1. Dual Dielectric - Reinforcing the Requirements 28
4.2. Motivation for Dual Dielectric Assignment Technique 29
4.3. Problem Definition for Dual Dielectric Assignment 29

CHAPTER 5. MODELING AND CHARACTERIZATION 31
5.1. Modeling and Calculation of the Gate Leakage 31
5.2. Study of Transient Response of a 2-Input NAND Gate 37
5.3. Results of Characterization 40

CHAPTER 6. THE DUAL DIELECTRIC ASSIGNMENT ALGORITHM 44
6.1. The Overall VLSI Design Flow 44
6.2. Dual Dielectric Assignment Algorithm 45
6.3. Illustration of the DKDT Assignment Algorithm 47
6.4. Algorithm Complexity 48

CHAPTER 7. THE EXPERIMENTAL SETUP AND RESULTS 53
7.1. The Modeling Test Bench 53
7.2. Customizing SIS - The Experimental Environment 53
7.3. The ISCAS Benchmark Circuits 54
7.4. Experimental Results 57
7.5. Observation From The Experiments Performed 59

CHAPTER 8. CONCLUSION AND FUTURE DIRECTIONS 61

BIBLIOGRAPHY 62
LIST OF TABLES

5.1 Characterization of gate leakage and delay for various logic gates 43

7.1 Experimental results showing reductions in gate leakage 58

7.2 Results using various approaches for gate leakage reduction 60
LIST OF FIGURES

1.1 The Y-chart depicting the design partition in VLSI over the three domains: behavioral, structural and physical or geometrical [87]. 5
1.2 Classical CMOS VLSI design flow [1]. 6
1.3 Levels of design abstractions [4]. 7
1.4 The classical Moore's Law of scaling [9]. 7
1.5 The increase in transistor count over the years in the Intel® microprocessors [18, 26]. 8
1.6 The increase in gate count in microprocessors over the years [49]. 9
1.7 With scaling of device dimension, the operating frequency also doubles every two years [18, 26]. 9
1.8 Practically CMOS scaling has beaten all the predictions and barriers. This figure shows the technology predictions and the actual scaling trends [18, 26, 77]. 10
1.9 With scaling of CMOS device the power density is increasing at an alarming rate [16, 18, 26]. 13
1.10 Break up of the components of power consumption [86]. 15
1.11 Normalized trends of static and dynamic power dissipation predicted by the ITRS [46]. 15

5.1 Labels of hierarchical abstraction used for characterization. 31
5.2 Test bench for fixing the rise time ($t_r$) for cell characterization. 32
5.3 Determination of delay and rise time. 34
5.4 Gate leakage paths in various switching states of a 2-input NAND gate.

5.5 Transient response of a 2-input NAND gate: output voltage, propagation delay and gate leakage components for fixed load.

5.6 Gate leakage for a 2-input NAND gate for different input switching states when dielectric constant \(K\), gate thickness \(T_{gate}\), and supply voltage \(V_{DD}\) is varied.

5.7 Propagation delay \(T_{pd}\) for a 2-input NAND gate with dielectric for different input switching states when dielectric constant \(K\), gate thickness \(T_{gate}\), and supply voltage \(V_{DD}\) is varied.

5.8 Average gate leakage \(\overline{I_{ox}}\) vs. gate dielectric electrical thickness \(T_{gate}\).

5.9 Average gate leakage \(\overline{I_{ox}}\) vs. gate material relative dielectric constant \(K_{gate}\).

5.10 Max. propagation delay Vs gate dielectric electrical thickness \(T_{gate}\). NAND2 is the best performing of all 2-input gates.

5.11 Max. propagation delay Vs gate material relative dielectric constant \(K_{gate}\). NAND2 is the best performing of all 2-input gates.

6.1 The overall VLSI design flow with two possible design paradigms: bottom-up and top-down [10].

6.2 Dual dielectric assignment algorithm for performance aware gate leakage reduction.

6.3 Logic design flow with the proposed dual-K-dual-T (DKDT) technique embedded.

6.4 The various states of the ISCAS C17 benchmark during DKDT assignment.

7.1 Test bench for modeling of gate leakage comprising of a 2-input NAND gate.

7.2 Flow of default SIS operations for technology mapping.

7.3 High level diagram of the SIS implementation of DKDT algorithm.

7.4 Gate leakage reduction for various benchmarks with DKDT assignments.
CHAPTER 1

INTRODUCTION

This chapter presents a broad introduction of the pervasiveness of digital electronics and its consequences and prepares a background for the overall work presented in the thesis. It has been overwhelmingly felt that digital electronics has been the driving force behind modern technological advances in the society. It has penetrated into every walk of life and has been embedded into all our activities. Whether it is the built-in feature of an alarm clock in our mobile phones that we wake up with in the morning or the sensor embedded into the microwave oven that heats our breakfast, the sleek, handheld personal digital assistant device or the laptop that keeps a busy executive on the move or the latest mp3 player that we take to our jog, everything has that “Digital” tinge in it. In short every accessory or gadget that we use in our daily life is becoming smaller, cheaper, efficient and more feature packed and this what a digital equipment has come to mean to a general user! This has led to skyrocketing of the demands for more and more features in every device. Consequently, in recent years there has been a phenomenal increase in the demand for low power and high performance digital devices. This market of digital electronics is driven by the prowess of immensely versatile, efficient and economical device called the complimentary metal oxide semiconductor (CMOS). CMOS forms the backbone of today’s circuits deploying very large scale integration (VLSI).

The transistor-feature sizes have dramatically shrunk with the technology scaling. With the accompanying shrinking of feature size there has been paradigm shift in the power profiles of the devices. There has been an exponential change in the leakage components of the fundamental CMOS device where each component of the total leakage has gained in relative importance. This has resulted in a drastic change in the leakage components of the device
both in the inactive as well as active modes of operation. As the CMOS device dimensions reach nanometer ranges, the dynamic power consumption remains almost unchanged, but the leakage power dissipation increases significantly and becomes a large portion of the total power dissipation as the technology changes. This has necessitated a critical analysis of the leakage components in the nanoscale CMOS VLSI devices and a concomitant need for the exploration of efficient reduction techniques of these leakage components, which continues to be dissipated even when a device is not doing any useful operation. With this objective at hand this thesis identifies the importance of the most significant of the leakage components at the nanometer levels, the gate tunneling leakage and presents an effective performance aware solution for the minimization of this critical mode of leakage.

In the next section there is an overview of the CMOS VLSI technology. Then the concept of scaling is introduced. The concept of short channel effect (SCE) is then elaborated with emphasis to various leakage phenomenons in the short-channel CMOS device. Also the need for low power CMOS VLSI is explained with the requirements. A brief introduction to the various leakage reduction techniques is given. This is followed by a summary of the contribution of this work and its impact on the technology is given. Finally the organization of the chapters in this thesis is laid out in the last section of this chapter.

1.1. A Brief Background on CMOS VLSI Technology

The increasing need for integrating more and more complex functions into ever smaller systems led to the development of VLSI systems. However the goal of integrating continually larger and more complex circuits into monolithic packages and systems seemingly would not have been possible without the aid of a very fundamental device called the “Complementary Metal Oxide Semiconductor”, referred to as CMOS. CMOS is the building block of this vast array of integrated circuits available to us in multifarious forms.

CMOS finds its origin in the metal oxide field effect transistor (MOSFET) which was first proposed by J. Lilienfield in 1925 [87]. This was finally given a physical shape in 1960 with the
introduction of silicon planar processes. The predominant technology at this time was the p-type MOSFET. In 1971 the nMOS silicon-gate technology came into picture. However even before the advent of PMOS, another revolutionary form of transistor was introduced, which was called CMOS. Fairchild Semiconductor Research and Development patented the first concept of CMOS and three basic gates: the inverter, the NAND-gate and the NOR-gate.

CMOS was originally favored to the faster TTL transistor logic for its low-power feature and found applications where the battery life was crucial and held more significance than speed, especially in the watch industry and in similar fields. However with time, CMOS has become the predominant technology in digital integrated circuits. And this is for some of its fabulous features that till date is limited to CMOS, some of which are less area, high operating speed, tremendous energy efficiency and economical manufacturing costs. Another most significant feature of CMOS technology is the classical scaling of CMOS which has enabled packing more and more features in the same area of a chip with every new generation of semiconductor manufacturing processes. Moreover, the simplicity and comparatively low power dissipation of CMOS circuits have allowed for integration densities likely not possible with its analog counterparts like bipolar transistors.

As of today CMOS has become so popular a device in the digital world that it only refers to a transistor device but it has a lot of other meanings too. CMOS is a particular style of digital circuitry design and the family of processes used to implement that circuitry on integrated circuits (chips). It identifies a logic family which comprises of a collection of n-type MOSFETs arranged in a pull-down network between the output and the lower-voltage power supply rail ($V_{SS}$) or the ground. Instead of the load resistor of NMOS logic gates, CMOS logic gates have a collection of p-type MOSFETs in a pull-up network between the output and the higher-voltage rail ($V_{DD}$). The p-type transistor network is complementary to the n-type transistor network, so that when the n-type is off, the p-type is on and vice-versa. CMOS logic on a CMOS process dissipates less energy and is more dense than other
implementations of the same functionality. As this advantage has grown and become more important, CMOS processes and variants have come to dominate, so that as of 2006 the vast majority of integrated circuit manufacturing by dollar volume is on CMOS processes.

1.2. VLSI Design Flow and Electronic Design Automation

The VLSI design flow is evolutionary in nature. It consists of a number of levels of design and test to conform to the user or design specification. A typical VLSI design flow is best represented by the Gajski and Kuhn’s chart which first appeared in [79] and describes the design flow using design activities on three different axes (domains) which resemble the letter Y and hence called the Y-chart. The three major domains shown in the Y-chart in fig. 1.1 are:

- Behavioral domain
- Structural domain
- Physical domain

Behavioral domain allows design at higher levels of abstraction and allows the translation and optimization of a behavioral description, or high-level model, into an register transfer language (RTL) implementation. In behavioral domain, the focus is on what the design does, not on how it is built. The structural domain describes what the structure needed to achieve a particular behavior. It gives the interconnection between the various modules to achieve a particular function and the individual gates and transistors forming each module is specified. Physical domain describes how the particular design is physically constructed. The classical CMOS VLSI design flow is shown in fig. 1.2.

The work in this thesis considers design in the logical domain which is the process of mapping from the system-level design handoff (currently at the RT-level) to a gate-level representation that is suitable for input to physical design. This precedes the circuit design which addresses creation of device and interconnect topologies (standard cells, full-custom analog, etc.) that achieve prescribed electrical and physical properties while remaining feasible with
Figure 1.1. The Y-chart depicting the design partition in VLSI over the three domains: behavioral, structural and physical or geometrical [87].

respect to process and manufacturability-induced constraints. This subsequently contributes to the physical design which addresses aspects of chip implementation (floorplanning, placement, routing, extraction, performance analysis) related to the correct spatial embedding of devices and interconnects. The output of physical design is the handoff ("tapeout") to manufacturing (currently centered around a generalized data stream (GDS) II stream file), along with verifications of correctness (design rules, layout versus schematic, etc.) and constraints (timing, power, reliability, etc.). Together, logical, circuit and physical design comprise the implementation layer of design technology that supports system-level design.

The logic or gate level can also be shown in the perspective of a design abstraction level hierarchy which is shown in fig. 1.3.
1.3. An Introduction to Scaling of CMOS

Scaling of CMOS implies proportional reduction of the geometrical features as well as parametric features (device characteristics) of the CMOS transistor. The idea of scaling began with the classical Moore’s Law, which had a phenomenal impact as a driving force for the entire semiconductor industry. Gordon Moore, cofounder of Intel® in his article [63],
Figure 1.3. Levels of design abstractions [4].

Moore's Law

Figure 1.4. The classical Moore’s Law of scaling [9].
gave an empirical observation which was finally called as the Moore’s law. It states that at the current rate of technological development, the complexity of an integrated circuit, with respect to minimum component cost, will double in about 18 months. This was more of an economical observation. However, over 25 years this law has been the prime motivation for the entire semiconductor industry. The effect can be seen on case of lead microprocessors as Intel as shown in fig. 1.4 and fig. 1.5.

![Graph showing the increase in transistor count over the years in Intel microprocessors.](image)

**Figure 1.5.** The increase in transistor count over the years in the Intel® microprocessors [18, 26].

Scaling of CMOS has become very well defined practice in the semiconductor industry. The design goals for scaling have been well defined for long. According to Borkar [18], the main three typical goals of classical technology scaling are:

- to reduce gate delay by about 30% which results in an increase in operating frequency of about 43% (fig. 1.7)
- to double transistor density which results in packing ever more components in the same chip (following Moore’s Law) (fig. 1.6) and
Figure 1.6. The increase in gate count in microprocessors over the years [49].

Figure 1.7. With scaling of device dimension, the operating frequency also doubles every two years [18, 26].
Figure 1.8. Practically CMOS scaling has beaten all the predictions and barriers. This figure shows the technology predictions and the actual scaling trends [18, 26, 77].

- to reduce energy per transition by about 65% and thus saving 50% of power (at a 43% increase in frequency).

Most of the goals have been maintained over all the years to a great benefit and has fuelled the market demand for portable devices. Also the scaling of CMOS has always been beaten predictions and has advanced faster than predicted as shown in fig. 1.8. The benefits of scaling have been manifold:

- Decreasing device size
- Increasing chip density and component density
- Increasing performance and speed
- Decreasing cost
- Decrease in supply voltage
- Decrease in power requirement

However there are some obvious roadblocks that can be credited to scaling:
• Tremendous increase in power density in a chip
• Issues in reliability and robustness
• Ever increasing mask costs and fabrication issues
• Tremendous increase in complexity
• Decreasing design flexibility
• Random dopant fluctuations
• Increasing process variations at nano levels
• More likelihood of soft errors

1.4. Short Channel Effects and Leakage Issues

The channel length of a device represents the technology node that a particular process follows. The current industrial processes now are inching down to reach the nanometer level where the channel length is of the order of a fraction of a micron. The industry has already put 65nm node process into production. However at these nano levels the channel is much shorter than the conventional channel length and the device characteristics are very much affected by these shorter channel lengths.

A number of adverse effects peculiar to scaling which results in a significantly reduced channel length are identified as short channel effects or SCEs. These SCEs mostly pertain to the power dissipation or leakage characteristics of the nanoscaled CMOS device. There are several forms of leakage currents in short channel nanometer transistor [74], such as:

• Reverse biased diode leakage
• Subthreshold leakage
• Gate oxide tunneling leakage
• Hot carrier gate current
• Gate induced drain leakage (GIDL)
• Channel punch through current

In addition, there are several SCEs (Roy et al. [74], Agarwala et al. [13]) such as:
• Drain induced barrier lowering (DIBL)
• Large $V_{th}$ roll-off
• Diminishing $I_{ON}/I_{OFF}$
• Band-to-band tunneling (BTBT)

These SCE among others are gaining prominence with the phenomenon of scaling. The reverse biased diode leakage current and gate oxide tunneling current flow during both active and sleep mode of the circuit, whereas other currents flow during the sleep mode only.

To overcome these SCEs the International Technology Roadmap for Semiconductors (ITRS) roadmap envisages that high performance CMOS circuits will require ultra-low gate oxide thickness [6]. However, such devices will be susceptible to a more profound leakage mechanism due to tunneling through the gate oxide ($I_{ox}$)(Sultania et al. [80]). With this the gate oxide tunneling current is emerging as the major component of the static power consumption of a low-end nanoCMOS device. Thus, there is a critical need for analysis, explanation and characterization of the various tunneling mechanisms, targeted towards design for manufacturing (DFM) and process variation modeling.

1.5. Why Low Power VLSI Design?

As seen in the sections 1.3 and 1.4, power dissipation has become the most significant issue for consumers as well as designers in the nanoscale CMOS regime. As the demand for more and more portable, battery operated devices increase the issue of low power VLSI design become more critical. The obvious reasons why power becomes an issue can be summerized as below:

• With scaling new and ever more obstrusive forms of power leakage becomes prevalent.
• Static power consumption becomes as significant as dynamic power, implying the device is dissipating half the power without doing any useful work.
• High power dissipation increases the packaging and cooling costs, which are not economical even at higher technology nodes.
• The battery technology, like many other technologies, is not expected to keep pace with the scaling.
• Environmental effect of wasting precious power resources.
• Scaling has resulted in manifold complexity in today's CMOS circuits, this along with diverse and profound power leakage makes reliability and robustness of the device a prime concern.

![Power Density vs. Year Graph](image)

Figure 1.9. With scaling of CMOS device the power density is increasing at an alarming rate [16, 18, 26].

In order to obviate the power issues, a specialized methodology of design known as low power VLSI design has gained prominence. Key principles in low power design involves:

• using the lowest possible supply voltage
• using the smallest geometry, highest frequency devices, but operating them at lowest possible frequency
• using parallelism and pipelining to lower required frequency of operation
• power management by disconnecting the power source when the system is idle

Low power VLSI may be considered in terms of average power reduction, peak-power reduction and power fluctuation reduction. The requirement for the various types power reduction can be summerized as below:

*Requirement for peak-power reduction*

• to maintain supply voltage at normal levels
• to increase reliability
• to use smaller heat sink
• to minimize packaging requirements

*Requirement for average power reduction*

• to increase battery life time
• to enhance noise margin
• to reduce energy costs
• to reduce use of natural resources
• to increase system reliability

*Requirement for power fluctuation reduction*

• to reduce power supply noise
• to enhance noise margin
• to reduce cross-talk and electromagnetic noise
• to increase battery efficiency

1.6. Introduction to Leakage Reduction Techniques

With the growing concern for leakage and power dissipation, a number of leakage reduction techniques have evolved. The trend for leakage reduction has started with the reduction of active and stand-by leakage, total leakage and finally sub-threshold leakage and then zeroed
Figure 1.10. Break up of the components of power consumption [86].

Figure 1.11. Normalized trends of static and dynamic power dissipation predicted by the ITRS [46].

onto the other components that are becoming more significant including gate leakage. The reduction techniques available now can be summerized as below.

- Voltage and process scaling
• Low K dielectric and copper interconnect
• Power-aware compiler and architecture design
• Power control and management techniques
• Dynamic voltage and frequency scaling based on workload
• Better cell library design and resizing methods
• Circuit design techniques
• Low power-driven bus encoding techniques
• Power-conscious synthesis and design tools

1.7. Contribution of This Work and Intended Impact

This work focuses on a novel technique for reduction of gate leakage in the nanometer scale CMOS devices. The CMOS process is based upon the usage of Silicon Dioxide (SiO₂) as the gate dielectric. The gate leakage of the CMOS device is dependent on the physical and electrical parameters that pertain to the gate.

Assuming that \( V_{DD} \) = supply voltage of the transistor, \( T_{ox} \) = gate SiO₂ thickness, \( W_{gate} \) = gate width, the gate oxide leakage current can be expressed as follows [46, 22]:

\[
I_{ox} = KW_{gate} \left( \frac{V_{gs}}{V_{th}} \right)^{2} \exp \left( -\alpha \frac{T_{ox}}{V_{th}} \right)
\]

where, \( K \) and \( \alpha \) are experimentally derived factors. This explains the fact that a small change in \( T_{ox} \) can have a tremendous impact on gate oxide current. Based on the above equation, the following possible options can be considered for reducing gate leakage power consumption:

• Decreasing supply voltage \( V_{DD} \),
• Increasing gate SiO₂ thickness \( T_{ox} \) and
• Decreasing gate width \( W_{gate} \).

The most popular (as well as necessary) option for decreasing the dynamic power consumption has been decreasing the power supply voltage \( (V_{DD}) \). There has been a consistent
decrease in the supply voltage and it will continue to play its role in the reduction of leakage power as well [60]. But, increasing the gate SiO2 thickness may not be an option due to technology scaling for which it should rather be decreased. Moreover, reduction of gate width may not be an attractive option as gate leakage current is only linearly dependent on it. Thus, it is concluded that replacing the SiO2 with insulators with high dielectric constants and high thicknesses is an attractive option. Intuition behind the conclusion is that employing multiple dielectric thicknesses (denoted as $T_{gate}$) and multiple dielectrics (denoted as $K_{gate}$), the leakage power (current, denoted as $l_{gate}$) may be reduced significantly while maintaining the performance.

With the limits being reached in the efficacy of gate oxide thickness optimization and assignment techniques, it has now become desirable to find a suitable alternatives to the SiO2 as the gate dielectric itself [45, 56, 83]. Recently, silicon oxynitride (SiON) has attracted attention as a replacement for SiO2, as it has been used in the silicon technology and is compatible with the established IC technology [30, 45, 56]. In this paper use of two dielectrics silicon dioxide (SiO2) and silicon oxynitride (SiON) each with two different thicknesses is focussed on. Here it is assumed that there are four different types of transistors available: a transistor can use a gate dielectric with relative permittivity $K_1$ or $K_2$ and of thickness $T_1$ or $T_2$. Assuming that all the transistors of a logic gate are made of the same $K_{gate}$ and $T_{gate}$, we consequently have four different types of logic gates for each cell in the library. Further, we develop an algorithm for the assignment of these four available alternative logic gates to a CMOS circuit logic network. The performance is preserved as in case of the conventional $K_1T_1$, while the algorithm ensures that the total gate tunneling current of the circuit is minimized.

The contributions of this thesis to the current state-of-art can be summarized as follows:

- The work introduces a new approach of dual dielectric (SiO2 and SiON) assignment for the tunneling current reduction.
• Logic level minimization of average-case gate oxide tunneling current dissipation of static CMOS circuits accounting for both on and off states of the device.

• Introduction of a heuristic algorithm that assigns dual dielectric material in combination with dual thickness and achieves the objective of minimization of gate leakage in CMOS circuits.

• Consideration of both resource and time constraints to provide the user the ability to explore the design space of tunneling current dissipation, circuit performance and cost of silicon.

• A characterization methodology is shown for logic gates using 45nm technology to calculate the average case tunneling. Also a dual gate dielectric and thickness component library that includes various logic gates like inverter, AND, OR, NAND, NOR and various other compound gates is developed.

This work is the first in terms of using alternative dielectric techniques for gate leakage reduction. The heuristic proposed is simplistic yet an efficient methods leading to significant reduction in the overall power dissipation minimization. The impact of this work has already been realized in the industry and some industry leaders in process technology are opting for multiple gate oxide approach [5].

1.8. Organization of the Thesis

The rest of this thesis is organized as follows. The various related works and the state of the art in CMOS VLSI design are summerized in chapter 2. In the chapter 3 the need of the hour for technologies in alternative gate dielectrics is presented and in chapter 4, the gate leakage reduction as an optimization problem is defined and contributions are summarized. Chapter 5 describes the modeling and calculation of the gate leakage for deriving the cell library. Chapter 6 describes the steps in the dual dielectric and dual thickness assignment algorithm. The experimental results are presented and discussed in chapter 7. Finally the
findings are stated and the thesis is concluded with a mention of the future directions of research in chapter 8.
CHAPTER 2

LOW POWER CMOS VLSI - CURRENT STATE OF THE ART

The state of art of the complimentary metal oxide semiconductor (CMOS) technology is a comprehensive effort of the past three or more decades of motivated and relentless research. In the current scene, the research community has been focused on looking into various mechanisms for reducing leakage in various forms in the nanometer devices. Most of the literatures that have been published however have been focusing on decreasing sub-threshold leakage. However, as we approach the lower end of nanometer technology, especially sub-65nm technology node, the component that predominates all others in contributing to leakage happens to be the gate leakage current and more specifically the direct gate tunneling current. There are very few work available in literature providing approaches for gate leakage reduction. This chapter begins with an outline of works in physical and quantum mechanical modeling followed by works in compact modeling. Various power reduction methods are summerized next. This is followed by logic level works in leakage reduction and behavioral works in leakage power minimization.

2.1. Research on Technology Predictions and Scaling

A number of works present the requirements that arise out of scaling of CMOS and the current technology trends. Silicon technology trends have been predicted by Ning in the work [7]. In [18] Borkar has outlined the design challenges in technology scaling. There have been predictions about the end of CMOS scaling with the advent of new materials and structural changes in the MOSFET design as presented by Skotnicki et al. [78]. Scaling issues with respect to a number of crucial effects and power constraints have been discussed by Frank [33]. A comprehensive analysis of the challenges and design choices in nanoscale
CMOS has been done by Narendra [68]. The benefits of scaling and ways to sustain them has been reviewed by Nowak [69] and Taur [82]. A number of low-voltage and low-power design techniques have been discussed in [85]. Nowak et al. [29] presented a discussion on new device alternatives to traditional CMOS devices such as double-gate and FINFETs. All these discussions and predictions of the effects of scaling indicate an urgent need for looking into alternative technologies and devices that can withstand the limits of scaling and sustain it in the future.

There has been extensive effort from the research as well as the industrial community to overcome the barriers to scaling of devices. This has extended from the domain of design and fabrication to the study of device physics and material science. While the design engineers have been trying to devise new methods and techniques to adapt to scaling, Physicists have been propounding new materials and evaluating their characteristics and suitability for integration into the current manufacturing techniques. The following sections brief about the works in the fields of physical - quantum modeling and compact modeling to aid in technology progression and scaling of devices.

2.2. Research in Physical and Quantum Mechanical Modeling

There has been variety of works in physical and quantum mechanical modeling of alternative gate dielectrics having higher relative permittivity than SiO2. Yang et al. [89] have presented modeling of gate leakage and demonstrated that direct tunneling through oxide/nitride can be lower than through only oxide. Gate leakage models through high-K dielectrics have been presented by Vogel et al. [83] demonstrating the effect of dielectric constant and barrier height. Guo et al. [88] provide a systematic study of effect of the oxygen/nitrogen content in the gate tunneling leakage current. Various works also are exploring other high-K dielectrics that can be used to replace SiO2 and their effects studied. The need for high-K gate dielectrics in the current scenario of ultra-thin gate oxide have been presented by Manchanda et al. [56]. Initially oxynitrides have been the alternatives of choice and have
been widely explored for various effects. Gusev et al. [28], have discussed various device characteristics such as gate leakage, flatband voltage shift, mobility for high-K gate dielectrics at high temperatures in their work. Hole-tunneling current study through both oxynitride and oxinitride/oxide stack is presented by Yu et al. [38]. Yang et al. [90] have presented the performance dependence of CMOS on Silicon Substrate orientation for ultrathin oxynitride and HfO$_2$ gate dielectrics.

There have also been various works on different emerging CMOS device structures and scaling basically focussed on methods of reducing gate tunneling leakage. However most of the works have concentrated on the use of gates having SiO$_2$ as the gate dielectric and hardly any work has concentrated on the aspect of integrating gate dielectrics with higher K as an alternative solution. Mukhopadhyay et al. [66] study the effect of gate tunneling current in ultra-thin gate oxide MOS devices of effective length of 25 nm and 90 nm using device simulation and model leakage in a stack of transistors. Agarwala et al. [13] have presented a high-level leakage power analysis and reduction algorithm using device-level models for leakage to precharacterize a given register-transfer level module library. An approach for modeling and analysis of gate-to-channel leakage in different DGFET structures is presented by Mukhopadhyay et al. [65]. Mukhopadhyay et al [67] have developed analytical models to estimate the mean and the standard deviation in variation of different leakage components in scaled CMOS devices considering variation of some crucial device parameters. An analysis of loading effect on leakage and a method of estimation of the total leakage in a logic circuit is proposed by Mukhopadhyay et al. [64]. Guindi [36] presents a methodology for estimating and minimizing the total amount of gate-tunneling current in CMOS combinatorial circuits taking advantage of the state-dependency exhibited by the gate-leakage. Roychoudhury et al. [73] have presented a methods for modeling and estimation of leakage in sub-90nm devices. Agarwala et al. [11, 12] have presented a comprehensive review of models, estimation and tools for analysis and reduction of leakage power.
2.3. Research Based On Compact Modeling

There has been hardly any compact model based analysis and characterization research work for nanoscale high-$K$ logic gates addressing power-performance issues. Hu [42] presented the limits of gate oxide scaling and gave projections for various related parameters. Hirose et al. [31] has explored the fundamental limit of gate oxide thickness scaling. Rao [72] presents mechanisms for statistical estimation of leakage current due to process variations. Rao [71] has presented a method for steady-state gate leakage estimation based on state characterization. Lee et al. [52] presented a heuristic method for simultaneous analysis of $I_{\text{gate}}$ and $I_{\text{sub}}$ and the complicity involved in the interaction. Guindi [37] has elaborated compared logically equivalent but structurally independent CMOS circuits and have formulated "state-dependent" gate leakage tables and provided guidelines for designs focusing on gate tunneling leakage current. Various characterization and modelling issues of ultra thin oxide MOS devices are discussed by Ghibaudo et al. [34].

Lo et al. [53] provide quantum mechanical models for direct tunneling current through the ultrathin oxide of a MOSFET. Choi et al. discuss a surface potential model and compact quantum mechanical models to predict direct tunneling through thin oxide [24]. Hou [41] provides a direct tunneling current calculation for holes using a modified Wentzel-Kramers-Brillouin (WKB) approximation. [23] provides simulation based studies for gate direct tunneling including edge-direct tunneling (EDT). Maitra [55] proposes analytical schemes to combine the channel component and the edge component of gate oxide direct tunneling current. Kougianos et al. [48] presented a methodology for calculating the gate tunneling capacitance.

2.4. Research in Leakage Reduction and Analysis

Techniques like BGMOS devices using dual $T_{ox}$ and dual $V_{Th}$ and PMOS dominated circuits have been proposed by Inukai et al. [44]. Rao et al. [70] have presented a sleep state assignment techniques and applied this to MTCMOS circuits for reduction of both gate and
subthreshold leakage. Lee et al. [50] develop techniques to estimate total leakage power of a circuit employing sub-threshold leakage and gate tunneling current. They suggested pin reordering to minimize gate leakage during standby positions of NOR and NAND gates. Lee et al. [51] discuss the effect of gate tunneling with scaling as well as present pin-reordering as a solution for minimization of gate leakage and resolving state dependencies. Sultania et al. [80] have proposed a heuristic for dual $T_{ox}$ assignment and the consequent tradeoff with the delay in the circuit. Also, Sultania et al. [81], suggest an approach to minimize gate tunneling utilizing dual $T_{ox}$ along with pin-reordering to reduce the total leakage current. Bhunia [17] presented a technique for reduction of active leakage power in 70nm node. Mohanty et al. in [59] presented a new analytical approach for calculating gate leakage at the behavioral level. Mohanty et al. have [62, 75, 57, 61, 58] presented a number of behavioral level and gate level leakage reduction works.

As can be seen from above that various works have focused on development of methods that use oxide of different thicknesses for gate tunneling reduction. We see that they do not handle emerging dielectrics that will replace SiO$_2$ to handle the tunneling current for low end nanometer technology. The research work proposed in this thesis presents a maiden and unified performance aware heuristic that determines the assignment of both dielectric material and the gate dielectric thickness. This research also considers the gate tunneling current for both ON and OFF state of the device which is not handled by any of the previous works.
CHAPTER 3

NEED FOR HIGH K GATE DIELECTRICS

This chapter begins with a recapitulation of the fundamental limits of scaling of gate oxide in complementary metal oxide semiconductor (CMOS) devices. Then some alternative gate dielectrics for CMOS is considered. Finally alternative technologies for high-K dielectrics in the CMOS devices in the nano regime is presented.

3.1. Fundamental Limits of Scaling of Gate Oxide

The scaling of CMOS device has had the most significant effect on the gate oxide, which serves as the dielectric of the gate. The physical dimension has reached practical and physical limits. As shown by Hirose in [31], the ultrathin gate oxide leads to significant fluctuations transconductance and leads to critical limits when the gate oxide thickness is scaled down to about 0.8 nm. For 45nm and below the gate oxide thickness is going to be far less, of the order of 14-17Åwhich will be susceptible to worse scaling effects. Frank et al. in their work [27] also suggested that, at these fundamental limits simple scaling rules do not apply and different parameters such as supply voltage or doping profile, need to be scaled at different proportion. With SiO₂ as the gate oxide there are a number of side effects that need to be encountered. Solutions need to be found for scaling power supply and threshold voltage scaling. Moreover the processes should take care of the resulting short channel effect, especially high field effects and the effects of random dopant distribution. However lithographic techniques will be a significant challenge even if techniques of minimization of gate leakage can be found within the premises of scaling.
3.2. Alternative Technologies in High-K for Nanoscale CMOS Devices

SiO$_2$ has hit the fundamental limit in its role as the gate dielectric of choice [45]. This is basically due to the fact that decrease in the gate oxide thickness is associated with a concomitant and significant increase gate leakage due to direct tunneling through the gate oxide. This inevitable drawback and the impending increase in power dissipation has called for finding alternatives candidates for replacement of SiO$_2$. A suitable candidate needs to have a higher dielectric constant than SiO$_2$ [56]. This would allow for scaling down the gate thickness as well as maintaining the effective dielectric barrier height to prevent gate tunneling current. The use of high K-gate dielectrics would also result in maintaining the equivalent oxide thickness (EOT) [32, 8] of at most 0.1nm and maintaining the mobility comparable to SiO$_2$. There are various metrics that define the performance of a high-K gate dielectric such as $D_{it}$, low frequency CV hysteresis and frequency dispersion, fixed charge density, minimal dielectric charging, interface degradation or stress induced leakage current(SICL) as result of voltage stress and surface mobility. All these parameters should be comparable to that of silicon to almost 95%. Manchanda et al. in [56] suggest a prominent figure of merit for comparing various high-K materials, this is given by $(m^*)^{1/2}\phi^{1.5}(T_{phys})$, where $m^*$ is the electron effective mass. $\phi$ is the band-offset (recommended > $V_{DD}$ for controlled schottky, thermionic emission as well as controlled Fowler-Nordheim (FN) leakage currents). and $T_{phys}$ is the film’s physical thickness, taken to be sufficiently thick to correspond to the FN tunneling regime. The corresponding gate leakage current density goal for high performance microprocessor units (MPUs) has often been utilized as < 1A/cm$^2$.

3.3. Alternatives to SiO$_2$ as Gate Dielectrics

Recently a host of materials have been investigated for use with SiO$_2$ in the CMOS devices such as ZrO$_2$, TiO$_2$, BST, HfO$_2$ [90], Al$_2$O$_3$ [56] and a host of Silicon Nitride compounds. It is definitely a challenging task in itself to integrate these materials into the conventional process and is a topic of continued research [88]. There has been a lot of progress in
the development of various technologies for high-K gate dielectric deposition [30]. This includes the extension of chemical vapor deposition (CVD), single wafer methodologies such as rapid thermal chemical vapor deposition (RTCVD), rapid plasma-enhanced chemical vapor deposition (RPECVD) and liquid source misted chemical vapor deposition (LSCVD). Other techniques such as physical vapor deposition (PVD)[84], jet vapor deposition (JVD) [54], oxidation of metallic films [14], molecular beam epitaxy (MBE)[35, 47]. All these processes are aimed at controlling the film structure and therefore its properties such as gate stack EOT, gate leakage current and transistors characteristics.

3.4. Feasability for Alternative Dielectrics

It is an intuitive conclusion that along with the efforts in introducing high-K gate dielectrics, futuristic synthesis methodologies should be developed in order to incorporate them into the existing design flows. This leads to the proposal of the idea of using dual gate dielectric in CMOS devices along with dual thickness which promises to be an efficient and effective alternative to the conventional method of single dielectric, predominantly composed of SiO₂.
CHAPTER 4

PROBLEM DEFINITION

In this chapter the requirement for dual dielectric assignment technique is reinforced and the problem of dual dielectric assignment is formally defined. It builds the basis for a practical implementation of the dual dielectric assignment. A mathematical representation of the logical problem helps in choosing a proper algorithm for its implementation. Particularly in case of huge networks of nodes, as is prevalent in the logic circuits, the choice of a network traversal algorithm has a lot of impact on the runtime as well as the turn-around time of the actual design automation technique. Since the algorithm that implements the technique is the most crucial element of the solution it needs a thorough emphasis. In the following sections, the assignment of dual $K$ and dual $T$ is considered in the light of both, leakage reduction and preserving performance of the device.

4.1. Dual Dielectric - Reinforcing the Requirements

In the chapter 3 the need for an alternative dielectric for the gate of a complementary metal oxide semiconductor (CMOS) device was upheld. Here the dual dielectric assignment, combined with dual thickness assignment is proposed as a formal assignment problem. It is expected that with the assignment of a dielectric of higher dielectric constant the gate leakage decreases significantly. However if this is done indiscriminately, there will be a tremendous penalty that has to be paid in case of performance of the device. However the objective of this work is to present a performance aware solution along with achieving the objective of minimizing gate leakage. So the strategy is to utilize a dual dielectric approach which allows for assigning dielectrics of higher $K$ to components which do not have a critical effect on the performance of the circuit, in other words that do not lie in the critical path(s) of the circuit.
So with a combination of components made of high and low \( K \) in combination with high and low dielectric thickness can give a feasible solution for a performance aware dual dielectric approach for gate leakage reduction.

4.2. Motivation for Dual Dielectric Assignment Technique

In this section the state-of-the art and the need of the future as discussed previously are summarized and based on this the problem definition is formulated. subsequently the contribution of this work to address these needs is provided. As per the discussion in section 3.2 and the current works as cited in chapter 2 the need for alternative futuristic methodologies using dual dielectric can be summarized as below.

- various logic synthesis works available at present (chapter 2) focus predominantly on the subthreshold leakage. This calls for a new logic synthesis approach considering tunneling current for low-end nanotechnology domain.
- The logic synthesis works (chapter 2) focus on solutions along the traditional lines of using \( \text{SiO}_2 \) only. Also various techniques in tunneling current reduction that exist offer methods that are limited to design and logic methodologies without any consideration to the material perspective.

Thus, it is observed that simultaneous assignment from a combination of dual dielectric and dual thickness to various transistors in a logic gate will prove to be an effective and futuristic technique aimed to minimize the gate leakage of a CMOS device while keeping its performance under control. While desired performance tradeoff can be incorporated in the form of a delay tradeoff factor, the cost of silicon is taken into account as resource constraints.

4.3. Problem Definition for Dual Dielectric Assignment

A combinational circuit can be modeled as an weighted directed acyclic graph \( G(V, E) \). The nodes \( v, \epsilon V \) are comprised of the primary inputs (PIs), the primary outputs (POs), and the
combinational processing elements (PEs). The edges $e_{i,j} \in E$ represent the interconnections between node $v_i$ and $v_j$. A PI is a node that has no fanins (incoming edges) and a PO is a node that has no fanouts (outgoing edges). Using this interpretation let us introduce the formulation of the dual dielectric assignment problem. The weight on the nodes can be associated with the delay in the processing elements. Also, the dual dielectric assignment occurs at the technology mapping phase and as the exact layout information is not available in this stage the interconnecting edges can be modeled as a constant delay.

*Given an weighted directed acyclic graph (WDAG) $G(V, E)$ it is required to find the best possible assignment of thickness and gate dielectric such that the total tunneling current is minimized and latency constraints (circuit performance) are satisfied.*

This can be viewed as an optimization problem as follows. Let $V$ be the set of all vertices and $V_{CP}$ be the set all vertices in the critical path from the PIs to POs. The power and performance driven two dimensional problem can thus be formulated as follows.

$$
(2) \quad \text{Minimize } \sum_{v \in V} I_{ox}(v_i)
$$

Where, $I_{ox}(v_i)$ is the tunneling current consumed per sample node $v_i$ of the DAG, such that the following latency constraint is satisfied.

$$
(3) \quad \sum_{v_i \in V_{CP}} D_i(v_i) \leq D_{CP} \quad \forall \ v_i \in V_{CP}
$$

The constraints in Eqn. (3) ensure that the summation of all delays $D_i(v_i)$ in a given path is less than the critical path delay $D_{CP}$. 
CHAPTER 5

MODELING AND CHARACTERIZATION

This chapter presents the methodology followed for the characterization of the devices at the logic level in vast detail. A complete study of the gate leakage with respect to process variation in case of a 2-input NAND gate is presented and the results are extended to other basic logic gates. Depending on the results of characterization, the fundamental unit of characterization was decided as a 2-input NAND gate and all the other library elements are derived in terms of it. The Cadence® ICFB tool was used throughout the characterization methodology.

5.1. Modeling and Calculation of the Gate Leakage

A complete transistor level characterization of a number of logic gates (NOT, NAND, NOR, AND and OR) was performed with respect to tunneling leakage and input-output delay using Cadence® Design Systems’ Spectre® analog circuit simulator [43]. Even though leakage and timing were concentrated upon in this analysis, the test benches are fully parameterized to account for load and power supply variations as well as the physical dimensions of the devices. A hierarchical model of the logic gates is followed as shown in Fig. 5.1.

![Hierarchical Abstraction Diagram]

Figure 5.1. Labels of hierarchical abstraction used for characterization.
Figure 5.2. Test bench for fixing the rise time \( t_r \) for cell characterization.

One of the obstacles in performing such technology characterization, particularly in the sub-65\( nm \) region, is the lack of commercially available processes for which data have been published. The Berkeley predictive technology model (BPTM) [21] was chosen to be used for characterization since it is well established and is able to predict the general trend of effects such as gate leakage and delay. *Since the algorithms described in this work depend on trends rather than exact values, it is believed that this approach has its merit as opposed to waiting for commercial semiconductor fabs to publish data.* Specifically, the gate leakage calculations are expected to be highly accurate as their dependence on short channel effects is not strong.

The BPTM BSIM 4.4 decks generated represent a hypothetical 45\( nm \) CMOS process with oxide thickness \( T_{ox} = 1.4nm \), threshold voltage \( V_{th} = 0.22V \) for the NMOS and \( V_{th} = -0.22V \) for the PMOS. The nominal power supply is \( V_{DD} = 0.7V \). These decks are also
scalable with respect to $T_{ox}$ and channel length. The effect of varying oxide thickness ($T_{ox}$) was incorporated by varying $TOXE$ in the spice model deck directly while the effect of varying dielectric material was modeled by first calculating an equivalent oxide thickness ($T_{ox}^*$) according to the formula:

$$T_{ox}^* = \frac{K_{gate}}{K_{SiO_2}} \times T_{ox}$$

Here, $K_{gate}$ is the dielectric constant of the gate dielectric material other than SiO$_2$, while $K_{SiO_2}$ is the dielectric constant of SiO$_2$ (=3.9). It may be noted that the length of the device is proportionately changed to minimize the impact of higher dielectric thickness on device performance [80]. Moreover, the length and width of the transistors are chosen to maintain a ($W/L$) ratio of 4 : 1 for NMOS and 8 : 1 for PMOS to ensure proper flow of current through the devices.

The first step in the characterization was the selection of an appropriate capacitive load. A value of 10 times the total gate capacitance $C_{gg}$ of the PMOS device was used [39]. This value depends strongly on the condition of the channel and has been calculated from the BPTM model for each case and operating condition.

The effect of the switching pulse rise time ($t_r$) was initially examined on the delay characteristics of the various gates. Following standard approaches [15] the delay is defined as the time difference between the 50% level of the input and the output waveforms. For worst-case scenarios in the development of the algorithm, we chose the maximum delay time regardless of whether this was due to a low-to-high or a high-to-low transition. In order to eliminate an explicit dependence of the algorithm results on $t_r$, we chose a value that is realistic yet does not affect the delay significantly. For $t_r = 10\, ps$ the dependence of the delay on $t_r$ is minimal for all gates.

After holding $t_r$ fixed at the selected value, we characterized first the maximum delay time for each gate and subsequently the gate leakage by evaluating all tunneling components for each PMOS and NMOS device in the logic gate. There are also several components of
the gate leakage within each device, such as $l_{gs}$ and $l_{gd}$ (components due to the overlap of gate and diffusions), $l_{gcs}$ and $l_{gcd}$ (components due to tunneling from the gate to the diffusions via the channel) and $l_{gb}$, the component due to tunneling from the gate to the bulk via the channel. The total gate leakage for each device was then calculated by summing all components:

\begin{equation}
\frac{\partial}{\partial x} = l_{gs}[i] + l_{gd}[i] + l_{gcs}[i] + l_{gcd}[i] + l_{gb}[i],
\end{equation}

where the index $i$ identifies the device within a gate. A total gate current for the gate ($I_{ox}$) was then calculated by summing the gate currents over all the devices in the gate:

\begin{equation}
I_{ox} = \sum_i I_{ox}[i].
\end{equation}
Figure 5.4. Gate leakage paths in various switching states of a 2-Input NAND gate.

During its various states of operation, each gate presents different dominant leakage paths, depending on the combination of inputs. For the 2-input gates we considered in this
Figure 5.5. Transient response of a 2-input NAND gate: output voltage, propagation delay and gate leakage components for fixed load.

work, the characterization is straightforward as all states can be simulated thus resulting in a complete characterization. For example, in Fig. 5.4 identify the various paths of the leakage in a NAND gate as a function of the inputs.

For each of the four possible states (00, 01, 10 and 11), the overall gate leakage ($I_{00}, I_{01}, I_{10}, and I_{11}$, respectively) was calculated from eqs. 5 and 6. Since all states are assumed to occur with equal probability, an average gate leakage ($\overline{I_{ox}}$) was assigned to each gate:

$$\overline{I_{ox}} = \frac{I_{00} + I_{01} + I_{10} + I_{11}}{4}. \tag{7}$$

The only exception is the NOT gate which has only 2 possible inputs (0 or 1) and the average gate leakage was then calculated by:

$$\overline{I_{ox}} = \frac{I_{0} + I_{1}}{2}. \tag{8}$$

Following standard approaches [15] the delay is defined as the time difference between the 50% level of the input and output waveforms. For worst-case scenario, the maximum
delay time is chosen regardless of whether this was due to a low-to-high or a high-to-low transition.

5.2. Study of Transient Response of a 2-Input NAND Gate

The gate leakage involved in the transient response of a 2-input NAND was first evaluated and the characteristics of the various voltage and gate leakage components in the device was identified. Various process and design parameters were varied and their effect was observed. This provided a wide variety of data for implementing the dual-K-dual-T (DKDT) algorithm and characterizing a wide variety of logic gates.

5.2.0.1. Effect of Parameter Variation on Gate Leakage. Here the effects of process and design variation on $I_{gate}$ is presented. A 2-input NAND gate was studied with respect to the variations in process parameters (relative permittivity and gate dielectric thickness) and also the effect of variation of a crucial design parameter (supply voltage). As can be observed from Fig. 6(a) all the components of the gate leakage current show a uniform trend when the gate dielectric ($K$) is varied assuming a fixed dielectric thickness $T_{gate}$, which is assumed as the default minimal thickness from the BSIM 4.4 model and a supply voltage of 0.7V. It can be seen that with an increase in the value of $K$ there is a steady decrease in the value of gate leakage, $I_{gate}$. This trend continues up to a knee region of $K = 12$, after which all the components eventually become almost constant and there is not much of a decrease in the value of $I_{gate}$ with an increase in $K$. This indicates the presence of a saturation zone in the gate leakage versus $K$ characteristics which implies that there is not much benefit in deploying gate dielectrics of very high-$K$, at least for the gate leakage.

The effect of a variation in the process parameter gate thickness, $T_{gate}$, is shown in the fig. 6(b). As can be seen there is a uniform decrease among all the components of the gate leakage with an increase in the gate thickness. This is not beneficial to the design of low-end nanoscale devices as increasing the gate thickness is in conflict with the trend of scaling of the devices. However, this effect can be utilized in some design styles as in dual-dielectric or
Figure 5.6. Gate leakage for a 2-input NAND gate for different input switching states when dielectric constant $(K)$, gate thickness $(T_{gate})$, and supply voltage $(V_{DD})$ is varied.

dual-thickness approaches used by researchers as an effective means of minimizing the impact of gate leakage.

Fig. 6(c) shows the variation of $V_{DD}$ and its corresponding effect on $I_{gate}$. As can be seen an increase in the supply voltage leads to a corresponding increase in the level of the gate leakage. This trend supports the scaling down of the supply voltage along with the scaling of other device features.
(a) Propagation delay Vs dielectric constant ($K$).
(b) Propagation delay Vs gate thickness ($T_{gate}$).
(c) Propagation delay Vs supply voltage ($V_{DD}$).

Figure 5.7. Propagation delay ($T_{pd}$) for a 2-input NAND gate with dielectric for different input switching states when dielectric constant ($K$), gate thickness ($T_{gate}$), and supply voltage ($V_{DD}$ is varied).

5.2.0.2. Effect of Parameter Variation on Propagation Delay. Here the effect of various process and design parameters on the propagation delay ($T_{pd}$) of a 2-input NAND gate was observed. The variation of $T_{pd}$ is shown in fig. 7(a), from which it can be seen that there is a sharp increase in the value of $T_{pd}$ with an increase in the gate dielectric strength; this increase continues until a value of around $K = 6$ after which the slope is much lower. This shows that there is almost no impact on the performance of the device at higher dielectric strengths.
The effect of variation of propagation delay with respect to the process parameter gate thickness is shown in fig. 7(b). As can be seen from the figure, there is a slight increase in the propagation delay of the gate as the gate thickness increases. Therefore, when combined with the benefit gained out of decrease in gate leakage of the gate with an increase in the gate thickness, this can used in the design of low leakage gates with hardly any compromise on performance.

The supply voltage was also varied, keeping the dielectric and the gate thickness fixed among other parameters and the results are shown in fig. 7(c). It can be seen that the propagation delay shows a decreasing trend with an increase in the value of the supply voltage. This is due to the increase in the drive current resulting from the increase in the supply voltage. However, as scaling continues, the trend is to scale down the supply along with other device features and that is also compatible with the objective of decreasing the tunneling leakage current in the nanometer regime.

5.3. Results of Characterization

The consolidated results of characterization of the basic cells viz. inverter and 2-input NAND, NOR, OR and NOR gates, as determined in the characterization methodology laid out above are shown in the Table 5.1. The results are shown in Figs. 5.8 - 5.11 for all the five basic gates. Fig. 5.8 shows the gate leakage variation with dielectric thickness when the dielectric is SiO₂. When the gate dielectric constant \( K_{gate} \) is varied while keeping the gate thickness constant the gate leakage changes as shown in fig. 5.9 assuming a fixed dielectric thickness \( T_{gate} \), which is assumed as the default minimal thickness from the BSIM 4.4.0 model (1.4nm in this case). The results for the propagation delay shows that delay increases both in case of an increase of gate thickness as shown in fig. 5.10 and in case of gate dielectric constant, as shown in fig. 5.11 for all the gates.
Figure 5.8. Average gate leakage ($\overline{I}_{ox}$) vs. gate dielectric electrical thickness $T_{gate}$.

Figure 5.9. Average gate leakage ($\overline{I}_{ox}$) vs. gate material relative dielectric constant $K_{gate}$.
Figure 5.10. Max. propagation delay vs gate dielectric electrical thickness $T_{gate}$. NAND2 is the best performing of all 2-input gates.

Figure 5.11. Max. propagation delay vs gate material relative dielectric constant $K_{gate}$. NAND2 is the best performing of all 2-input gates.
Table 5.1. Characterization of gate leakage and delay for various logic gates

<table>
<thead>
<tr>
<th>Dual Dielectric and Thickness combinations</th>
<th>Logic Gate</th>
<th>Gate Leakage Current $i_{\text{gate}}$ in A/μm For different switching states</th>
<th>Propagation Delay $T_{\text{DMax}}$ in ps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INVERTER</td>
<td>100.4n 252.0n - -</td>
<td>129.6</td>
</tr>
<tr>
<td>$K_1 = 3.9$ $T_1 = 1.4\text{nm}$</td>
<td>NAND2</td>
<td>55.8n 172n 35.8n 247.6n</td>
<td>256.9</td>
</tr>
<tr>
<td></td>
<td>NOR2</td>
<td>102.1n 128.5n 121.3n 246.6n</td>
<td>378.2</td>
</tr>
<tr>
<td></td>
<td>AND2</td>
<td>179.6n 295.7n 160n 298.5n</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td>OR2</td>
<td>225.4n 179.6n 171.8n 297.7n</td>
<td>340.3</td>
</tr>
<tr>
<td></td>
<td>INVERTER</td>
<td>6.0n 15.9n - -</td>
<td>210.2</td>
</tr>
<tr>
<td>$K_1 = 3.9$ $T_2 = 1.7\text{nm}$</td>
<td>NAND2</td>
<td>3.1n 10.7n 1.0n 15.6n</td>
<td>423.7</td>
</tr>
<tr>
<td></td>
<td>NOR2</td>
<td>6n 8n 7.7n 15.6n</td>
<td>586.2</td>
</tr>
<tr>
<td></td>
<td>AND2</td>
<td>11n 18.6n 8.9n 18.7n</td>
<td>611.3</td>
</tr>
<tr>
<td></td>
<td>OR2</td>
<td>13.9n 11n 11n 19n</td>
<td>573</td>
</tr>
<tr>
<td></td>
<td>INVERTER</td>
<td>0.262n 0.770n - -</td>
<td>241.3</td>
</tr>
<tr>
<td>$K_2 = 5.7$ $T_1 = 1.4\text{nm}$</td>
<td>NAND2</td>
<td>134p 506p 29p 754p</td>
<td>495.3</td>
</tr>
<tr>
<td></td>
<td>NOR2</td>
<td>260p 382p 375p 755p</td>
<td>680.3</td>
</tr>
<tr>
<td></td>
<td>AND2</td>
<td>11n 18.6n 8.9n 18.7n</td>
<td>611.3</td>
</tr>
<tr>
<td></td>
<td>OR2</td>
<td>640p 513p 501p 885p</td>
<td>697</td>
</tr>
<tr>
<td></td>
<td>INVERTER</td>
<td>0.005n 0.018n - -</td>
<td>258.4</td>
</tr>
<tr>
<td>$K_2 = 5.7$ $T_2 = 1.7\text{nm}$</td>
<td>NAND2</td>
<td>2.8p 12.1p 0.4p 18.5p</td>
<td>519.2</td>
</tr>
<tr>
<td></td>
<td>NOR2</td>
<td>5.5p 9.4p 9.2p 18.6p</td>
<td>724</td>
</tr>
<tr>
<td></td>
<td>AND2</td>
<td>12p 21p 10p 21p</td>
<td>802.8</td>
</tr>
<tr>
<td></td>
<td>OR2</td>
<td>15p 12p 12p 21p</td>
<td>772</td>
</tr>
</tbody>
</table>
CHAPTER 6

THE DUAL DIELECTRIC ASSIGNMENT ALGORITHM

This chapter presents the principal contribution of the work - the dual dielectric dual thickness or DKDT assignment algorithm. It presents the overall design flow and the position where the algorithm fits in. The implementation of the algorithm is illustrated with the help of an example benchmark circuit - ISCAS C17 benchmark. Finally the complexity of the assignment algorithm is calculated.

6.1. The Overall VLSI Design Flow

VLSI design involves a number of steps at which a well defined set of task leading to the final circuit design is performed. However there are many choices of the methodologies for design and it depends on the tools and techniques accessible to the designer. A detailed VLSI design flow is given in the figure 6.1 [10]. It shows the two design methodologies:

- Bottom-up or the full-custom design: The bottom-up design methodology primarily involves manual design of circuit building blocks at the transistor and mask-layout level, these are then combined together into a complete design hierarchy. It is very convenient and flexible at lower levels of design this enables mitigating issues such as transistor size optimization and parasitics minimization. The bottom-up methodology is better suited for the design of very dense, high-performance digital modules as well as analog and mixed-signal integrated circuits. It allows for faster and more efficient circuits, however it is very involved.
- Top-down or the standard cell design: The top-down design methodology is primarily based on automatic (computer-aided) synthesis of the gate-level netlist, using
a behavioral or high level description of the system. The synthesized logic is subsequently mapped onto a standard-cell library or an field programmable gate array (FPGA). This design methodology serves well in case of purely digital designs with relatively short turn-around times and moderate area-performance requirements.

The DKDT assignment algorithm fits into the top-down design methodology as it uses a cell library and each component used in the design are pre-built or pre-characterized following a particular modeling technique or characterization policy. The generalized steps give enough room for the technique to be plugged into the design synthesis framework during the logic synthesis phase of the top-down design hierarchy. Further details of the assignment algorithm is given in the next and subsequent sections.

6.2. Dual Dielectric Assignment Algorithm

The new VLSI design flow with the dual dielectric approach embedded is presented in Fig. 6.1. In the new design flow, the dual dielectric assignment and optimization are performed before the placement and routing to obtain the gate leakage optimized netlist. This netlist is subsequently processed for the placement legalization and engineering change orders (ECO) routing before generating the final layout [19, 25].

The dual dielectric assignment algorithm is a performance aware gate thickness and dielectric assignment algorithm. It aims at minimizing the gate leakage without compromising the desired performance. In this technique, it is aimed at assigning from a combination of two dielectrics and two thicknesses that are assumed to be available as characterized cells. Let it be assumed that $K_1$ is the dielectric constant for SiO$_2$, $K_2$ is the dielectric constant for SiON, where $K_1 < K_2$, and thickness $T_1 < T_2$. It is assumed that there are four different types of transistors available, such as $K_1T_1$, $K_1T_2$, $K_2T_1$, $K_2T_2$. In other words, a transistor can use dielectric of relative permittivity $K_1$ or $K_2$ and of thickness $T_1$ or $T_2$. Assuming that all the transistors of a logic gate are made of same $K_{gate}$ and equal $T_{gate}$, consequently there are four different types of logic gates. It is anticipated that the tunneling leakage current
Figure 6.1. The overall VLSI design flow with two possible design paradigms: bottom-up and top-down [10].
of logic gates increases and the propagation delay decreases in the order $K_2 T_2$, $K_2 T_1$, $K_1 T_2$, and $K_1 T_1$ (this is also shown employing the cell characterization, which is explained in section 5.1). This has served as the basis of the heuristic algorithm proposed in this section, where a logic gate under consideration is assigned a higher order $K$ and $T$ to reduce leakage whenever corresponding increase in delay of the path does not violate the target delay.

In fig. 6.2 the proposed heuristic algorithm is outlined. The network is represented as a weighted direct acyclic graph $G (V, E)$. The algorithm traverses the above graph to identify the critical path for the circuit. Subsequently, the critical delay $D_{CP}$ for the current assignment is calculated, which is compared against a target delay $D_T$, for any further assignment. This ensures that there is no compromise with the desired performance of the network. And, this allows for a performance aware direct tunneling leakage reduction. While starting the assignment, each node is greedily considered as a candidate for assignment of $K_2 T_2$ which is the ideal case. In each iteration a node is considered and seen if the current assignment complies with the target delay limit $D_T$. Whenever an assignment is done, its effect on the overall critical delay is considered. For each assignment the slack $\Delta_D$, which is defined as the difference between target delay $D_T$ and critical delay $D_{CP}$ is calculated. If $\Delta_D < 0$ the $K_2 T_2$ assignment violates the target delay, then the next best assignment i.e. $K_2 T_1$ is considered and again the value of $\Delta_D$ verified for meeting delay constraint. In case, this assignment still violates the target delay, $K_1 T_2$ is assigned to the gate. If none of the above assignments meet the target delay requirement then $K_1 T_1$ is reassigned to the node under consideration.

6.3. Illustration of the DKDT Assignment Algorithm

The working of the assignment algorithm can be illustrated with the help of the smallest of ISCAS’85 benchmarks, the C17 [20, 3]. The complete steps for assignment of dual dielectric and dual thickness resources is shown in the figure 6.4. The initial directed acyclic graph (DAG) representation of the C17 benchmark circuit is as shown in fig. 4(a). After logic minimization and NAND realization, the DAG for C17 reduces to the one shown in 4(b), the
principal inputs and outputs are marked as PI and PO respectively. In this step the critical path is identified and the critical delay $D_{CP}$ is calculated. The initial critical delay $D_{CP}$ is fixed as the target delay $D_T$. This serves as a performance constraint for the assignment algorithm. In all cases the delay in each path from each PI to PO may not exceed the target delay $D_T$. This ensures that the performance of the circuit is not compromised while assigning nodes of higher dielectric constant thickness. To begin with the circuit is initialized with $K_1T_1$ assignment 4(c). The delay at each node in the DAG is marked on the top. The initial assignment starts from the very first node and default assignment of $K_2T_2$ is done. The delay in each path is calculated and compared with $D_T$ and the slack $\Delta_D$ in each path is found out. As there is no infringement of the $D_T$, the rest of the assignments are made accordingly. Upon assignment of $K_2T_2$ to other resources there is a violation of the target delay $D_T$ in the paths containing node 2, so, the node is reassigned with $K_2T_1$. However this assignment does not meet target delay so node 2 is reassigned with $K_1T_2$ and the delays are recalculated. It is seen that this assignment again does not meet the performance constraint placed on the circuit so the assignment is reverted back to $K_1T_1$. Likewise, every node is the graph is targeted with an attempt to assign $K_2T_2$, $K_2T_1$ and $K_1T_2$ in the respective order before reverting to the initial assignment of $K_2T_2$ and the slack is checked to be positive. If the assignment met the target delay it is maintained and the next node is targeted. The final assignment is shown in Fig. 4(j) and the gate leakage and critical delay of the circuit is calculated and compared.

6.4. Algorithm Complexity

Assuming that there are $n$ number of gates in the original network representing any circuit we now provide the time complexity of the proposed algorithm. The statements from line (01) to (03) take $\Theta(n^2)$ time in worst case, as they involve accessing an array of nodes and arranging them into an adjacency matrix of size $n \times n$. The run time for statements from line (04) to line (07) are of $\Theta(n^2)$ complexity. Finally, the loop from line (08) to line (24) which
assigns dual dielectric takes $\Theta(n^3)$. Thus, the overall worst case time complexity of the dual dielectric assignment algorithm is $\Theta(n^3)$. 
(01) Represent the network as a directed acyclic graph \( G(V,E) \);
(02) Initialize each vertex \( v \in G(V,E) \) with the values of
leakage and delay corresponding to \( K_1 T_1 \) assignment;
(03) Find the set of all paths \( P(\Pi_{in}) \forall v \in \Pi_{in} \), the set of
primary inputs, leading to primary outputs \( \Pi_{out} \);
(04) Compute the delay \( D_p \) for each path \( p \in P(\Pi_{in}) \);
(05) Find the critical path delay \( D_{CP} \) for \( K_1 T_1 \) assignment;
(06) Mark the critical path(s) \( P_{CP} \), where \( P_{CP} \subset P(\Pi_{in}) \);
(07) Assign target delay \( D_T = D_{CP} \);
(08) FOR each vertex \( v \in G(V,E) \)
(09) {
(10) Determine all paths \( P_v \) to which node \( v \) belongs;
(11) Assign \( K_2 T_2 \) to \( v \);
(12) Calculate new critical delay \( D_{CP} \);
(13) Calculate slack in delay as \( \Delta_D = D_T - D_{CP} \);
(14) IF \( (\Delta_D < 0) \)
(15) {
(16) Assign \( K_2 T_2 \) to \( v \); Calculate \( D_{CP} \) and \( \Delta_D \);
(17) IF \( (\Delta_D < 0) \)
(18) {
(19) Assign \( K_1 T_1 \) to \( v \); Calculate \( D_{CP} \) and \( \Delta_D \);
(20) IF \( (\Delta_D < 0) \)
(21) reassign \( K_1 T_1 \) to \( v \);
(22) } // end IF
(23) } // end IF
(24) } // end FOR

Figure 6.2. Dual dielectric assignment algorithm for performance aware gate
leakage reduction.
Figure 6.3. Logic design flow with the proposed dual-K-dual-T (DKDT) technique embedded.
Figure 6.4. The various states of the ISCAS C17 benchmark during DKDT assignment.
CHAPTER 7

THE EXPERIMENTAL SETUP AND RESULTS

The experiments for the research presented in this thesis were meticulously carried out in a universally acclaimed experimental setup, this chapter presents a detailed overview of the entire experiment. The chapter begins with a description of the experimental set-up of the modeling test bench. Then the DKDT implementation in the SIS logic synthesis framework is explained. This is followed by a brief overview of the benchmark circuits used. Next the experimental results are stated and finally the observations from the experiment are presented.

7.1. The Modeling Test Bench

The modeling was carried out using the Cadence® ICFB suite of tools. Model schematics were drawn for the library using the Cadence® Virtuoso® tool and the Cadence® Analog Environment with Spectre® simulator were used for a detailed simulation. The test bench comprising of a 2-input NAND gate is shown in Fig. 7.1. The characterization was done following the Berkeley predictive technology model using the BSIM4.4.0 model library for 45nm NMOS and PMOS. A number of basic logic gates including an inverter, a 2-input NAND and a 2-input NOR gate, and 2-input AND and OR gates were characterized as per the characterization methodology presented in the chapter 5. A complete 45nm cell library was developed for implementing the DKDT assignment algorithm.

7.2. Customizing SIS - The Experimental Environment

The framework of Berkeley sequential interactive synthesis tool [76](more popular as SIS) was used for implementing the DKDT assignment algorithm. The SIS framework provided APIs for developing and integrating custom interfaces and modules. The DKDT algorithm was developed as a separate module written in C language and was integrated into the main
Figure 7.1. Test bench for modeling of gate leakage comprising of a 2-input NAND gate.

SIS framework. An overall flow of the SIS operations carrying out the DKDT assignment is shown in Fig. 7.2. The whole experiment was written out to a script that took the name of high level benchmark circuits as an input and gave out the assignment of each node and finally the gate leakage and delay as the output. A detailed flowchart representing the high-level implementation of the algorithm in the SIS framework is shown in Fig. 7.3.

7.3. The ISCAS Benchmark Circuits

The ISCAS’85 benchmarks represent a research standard for logic synthesis [20, 40, 2]. The gate-level models form a set of hierarchically benchmark circuits and have been a useful research tool in several areas of digital design, including test generation, timing analysis, and technology mapping. A number of high-level benchmarks are available in various formats for the use in logic synthesis frameworks. In this work a set nine of ISCAS’85 benchmarks in the
Berkeley logic interchange format (blif) format were used. The benchmarks constituted of diverse set of circuit as follows:

- c432 : 27-channel interrupt controller
- c499/c1355 : 32-bit SEC circuit
- c880 : 8-bit ALU
- c1908 : 16-bit SEC/DED circuit
- c2670 : 12-bit ALU and controller
- c3540 : 8-bit ALU
One regular SIS application script that does all the nanotech independent optimizations

Run the script.init

Read Cell Characterization details

Map current network to an integer mapped node matrix

Initialize all the nodes with K1T1

Apply weighted network traversal algorithm to the matrix and populate corresponding field in node

Get the current critical path and the critical delay for the network

Get the DTF (Delay Tradeoff Factor) from the user and calculate the target delay Td or the network.

Start with node 1, i=0

n = n - 1

No

Populate the data structure of the node with the delay and the leakage values corresponding to KnTn, n=n_max

Check for delay constraint upon change, if CP flag is true, also check for delay in critical path as well.

n! = 1

No

Target delay Constraint is met

Yes

Increase the nth node type count by 1 and decrease (n−1) th node type count by 1

Reset KnTn values

Next node = node i + 1

No

Max node reached

Yes

Stop

This method is adopted to make implementation more efficient and scalable. The overheads of searching and sorting networks using Integer Mapped Node Matrix is much less than traveling original networks

This identifies the longest available path from a particular node to the output

This allows the user to determine the tradeoff from the very beginning and obtain a performance aware leakage optimization

Start with the default cell characterization values and then scale down to the most appropriate one

CP flag indicates if the particular node is in the critical path

This will depend on how many characterized cells are available for each type of basic logic cell

Figure 7.3. High level diagram of the SIS implementation of DKDT algorithm.
• c5315 : 9-bit ALU
• c6288 : 16x16 multiplier
• c7552 : 32-bit adder/comparator

These benchmarks were read using the sis: read.blif command and processed using the script.rugged available with SIS. Finally the DKDT assignment algorithm was run and the gate leakage and delay for each circuit was found out.

7.4. Experimental Results

The performance aware dual dielectric assignment algorithm proposed in the 6.2 was implemented in the framework of the SIS logic synthesis tool [76]. A dual dielectric and thickness library was characterized for 45nm technology using Cadence® Spectre® tool as described in section 5.1. The library included various logic gates like Inverter, AND, OR, NAND, NOR and various other compound gates with a combination of dual dielectric and dual thickness. The values of $K$ and $T$ chosen were $K_1 = 3.9$ (for SiO$_2$), $K_2 = 5.7$ (for SiON [83]), $T_1 = 1.4\,nm$, and $T_2 = 1.7\,nm$ to perform the experiments. The value of $T_1$ is chosen as the default value from the BSIM4.4.0 model card and value of $T_2$ is intuitively chosen based on the characterization process in the previous section. The dual dielectric approach was tested on all the major ISCAS’85 logic benchmarks.

The experimental results are presented in Table 7.1. The gate leakages reported correspond to the average case with contributions from both ON and OFF devices. It shows the values of gate leakage for $K_1T_1$ assignment (the base case), the gate leakage after dual dielectric assignment is done using the proposed algorithm, and percentage reduction for various values of $K$ and $T$ assignments. The results prove a considerable decrease in the gate leakage in case of all benchmarks. The data also shows that without much tradeoff in delay, there is tremendous reduction in leakage. This allows for a better performance as well as reduction in leakage.
Table 7.1. Experimental results showing reductions in gate leakage

<table>
<thead>
<tr>
<th>Benchmark Circuits</th>
<th>No. of Gates</th>
<th>No. of Nodes in NAND N/W</th>
<th>Critical Path Delay ((\text{ps}))</th>
<th>(I_{\text{Gate}}) for SiO(_2) with (K_1T_1) ((nA))</th>
<th>(I_{\text{Gate}}) with Dual dielectric Assignment ((nA))</th>
<th>Percentage Reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C432</td>
<td>160</td>
<td>398</td>
<td>3.848</td>
<td>3949.452</td>
<td>253.260</td>
<td>93.58</td>
</tr>
<tr>
<td>C499</td>
<td>202</td>
<td>503</td>
<td>2.054</td>
<td>5708.547</td>
<td>590.454</td>
<td>89.66</td>
</tr>
<tr>
<td>C880</td>
<td>383</td>
<td>576</td>
<td>6.162</td>
<td>6537.024</td>
<td>337.842</td>
<td>94.83</td>
</tr>
<tr>
<td>C1355</td>
<td>546</td>
<td>503</td>
<td>2.054</td>
<td>5708.547</td>
<td>274.644</td>
<td>95.19</td>
</tr>
<tr>
<td>C1908</td>
<td>880</td>
<td>856</td>
<td>6.675</td>
<td>9714.744</td>
<td>287.721</td>
<td>97.04</td>
</tr>
<tr>
<td>C2670</td>
<td>1193</td>
<td>1764</td>
<td>24.643</td>
<td>17863.326</td>
<td>1560.672</td>
<td>91.27</td>
</tr>
<tr>
<td>C3540</td>
<td>1669</td>
<td>1792</td>
<td>18.227</td>
<td>34637.148</td>
<td>2215.737</td>
<td>93.60</td>
</tr>
<tr>
<td>C5315</td>
<td>2406</td>
<td>2481</td>
<td>23.103</td>
<td>28156.869</td>
<td>1098.801</td>
<td>96.10</td>
</tr>
<tr>
<td>C6288</td>
<td>2406</td>
<td>4864</td>
<td>24.897</td>
<td>28474.641</td>
<td>372.564</td>
<td>98.69</td>
</tr>
<tr>
<td>C7552</td>
<td>3512</td>
<td>2987</td>
<td>26.438</td>
<td>33899.463</td>
<td>625.842</td>
<td>98.15</td>
</tr>
</tbody>
</table>

From the Fig. 7.4 and the supporting table (Table 7.1) it can be seen that the proposed technique gives a significant reduction in gate leakage values. The highest reduction is observed in case of the benchmark C6288 which is 98.59% while the lowest reduction is 49.96% in case of the benchmark C17 which is the smallest benchmark. Overall, the dual dielectric approach achieves an average a reduction of 89.5%. We also noted that for replacement of each node from \(K_1T_1\) to \(K_2T_2\) results in a reduction of almost 99% in gate leakage per node while there is an increase in the delay of about 67%. However, this increase in delay does not increase the critical path delay of the overall circuit, as the nodes belonging to the critical path are not assigned with \(K_2\) or \(T_2\). Consequently, the results vary according to the number of nodes in the critical path and the network as a whole. The more there is parallelism in the network with almost identical path lengths and all tending to the number of nodes in the critical nodes (assuming a uniformly decomposed circuit comprised of only say 2
input NAND gates) there are chances that assignment of $K_2T_2$ will be sparse. However, for bigger circuits with larger delays along critical path, the chances of $K_2T_2$ assignment increases and so does the drastic reduction in gate leakage as a consequence of this assignment. Fig. 7.4 shows this kind of a trend with almost all the benchmarks attaining a reduction in tunneling leakage more than 90%.

7.5. Observation From The Experiments Performed

We provide a overall view of the proposed work with other works available in literature handling gate leakage minimization in Table 7.2. The NA in table says that results are not provided in the corresponding paper and “0” in table indicate that there is no penalty. It is evident from the table that the DKDT approach makes a consistent improvement in almost all benchmarks. While Sultania, et. al. [80] work has resulted in appreciable reduction of 62.7% – 92.6% there is an average penalty of 25%. Similarly, the work Lee, et. al. [51] reduces leakage in the range of 27.67%—59.52% for 100nm technology node. The entire set
of results verify the claim that the assignment of dual dielectric to the MOSFET gates can be done to attain considerable reduction in leakage current without any loss of performance.

<table>
<thead>
<tr>
<th>Bench. Circuits</th>
<th>Sultania [81] (100nm) % Reduction</th>
<th>% Penalty</th>
<th>Lee <a href="100nm">51</a> % Reduction</th>
<th>% Penalty</th>
<th>Dual dielectric (45nm) % Reduction</th>
<th>% Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>C432</td>
<td>83.8</td>
<td>24.6</td>
<td>59.52</td>
<td>NA</td>
<td>93.58</td>
<td>0</td>
</tr>
<tr>
<td>C499</td>
<td>70.2</td>
<td>25.4</td>
<td>28.25</td>
<td>NA</td>
<td>89.66</td>
<td>0</td>
</tr>
<tr>
<td>C880</td>
<td>92.3</td>
<td>25.5</td>
<td>45.43</td>
<td>NA</td>
<td>94.83</td>
<td>0</td>
</tr>
<tr>
<td>C1355</td>
<td>67.9</td>
<td>25.0</td>
<td>33.50</td>
<td>NA</td>
<td>95.19</td>
<td>0</td>
</tr>
<tr>
<td>C1908</td>
<td>82.3</td>
<td>25.2</td>
<td>27.76</td>
<td>NA</td>
<td>97.04</td>
<td>0</td>
</tr>
<tr>
<td>C2670</td>
<td>92.6</td>
<td>25.3</td>
<td>33.80</td>
<td>NA</td>
<td>91.27</td>
<td>0</td>
</tr>
<tr>
<td>C3540</td>
<td>91.4</td>
<td>25.1</td>
<td>36.40</td>
<td>NA</td>
<td>93.60</td>
<td>0</td>
</tr>
<tr>
<td>C5315</td>
<td>91.7</td>
<td>26.1</td>
<td>34.34</td>
<td>NA</td>
<td>96.10</td>
<td>0</td>
</tr>
<tr>
<td>C6288</td>
<td>62.7</td>
<td>25.7</td>
<td>45.86</td>
<td>NA</td>
<td>98.69</td>
<td>0</td>
</tr>
<tr>
<td>C7552</td>
<td>91.6</td>
<td>25.3</td>
<td>28.10</td>
<td>NA</td>
<td>98.15</td>
<td>0</td>
</tr>
</tbody>
</table>
CHAPTER 8

CONCLUSION AND FUTURE DIRECTIONS

In this paper a new approach for gate leakage reduction in CMOS devices for both active and sleep state using dual gate dielectric of dual thickness was proposed. A heuristic algorithm was developed that could carry out such assignment for combinational circuits in reasonable amount of time. The experiments yielded significant reductions in gate leakage without making any compromise with the performance of the circuit. The assignment technique was effective in reducing the gate leakage by as much as 98% and on an average by 89.5%. It may be noted that there are process integration problems due to use of dual thickness and dual dielectric during fabrication process. Issues such as chemical reactivity, with existing or new gate dielectric materials to ensure optimal CMOS compatibility as appropriate and thermal budgeting are to be resolved. It may also need use of more number of masks for the lithographic process of circuit fabrication. But, it is believed that such cost would be compensated by the reduction of energy or power costs. However, the research on these materials by material science and engineering as well as electrical engineering community is on full swing and it is expected to see new process technology in future addressing these issues.

This work has focussed on heuristic based algorithms, but more optimal algorithms such as integer linear programming (ILP) are under development. Also future research is aimed at implementing DKDT assignment techniques for the more involved sequential circuits. In case of characterization and modeling future works might be attempted at transistor level DKDT assignment which is a rather ambitious goal. The objective is not only to achieve faster solutions, but also optimal solutions in future both with respect to circuit speed and power loss.
BIBLIOGRAPHY


