Impact of Technology Scaling on Leakage Reduction Techniques

by

Payam Ghafari

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2007

©Payam Ghafari 2007

AUTHORS DECLARATION FOR ELECTRONIC SUBMISSION OF THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

CMOS technology is scaling down to meet the performance, production cost, and power requirements of the microelectronics industry. The increase in the transistor leakage current is one of the most important negative side effects of technology scaling. Leakage affects not only the standby and active power consumption, but also the circuit reliability, since it is strongly correlated to the process variations. Leakage current influences circuit performance differently depending on: operating conditions (e.g., standby, active, burn in test), circuit family (e.g., logic or memory), and environmental conditions (e.g., temperature, supply voltage). Until the introduction of high-K gate dielectrics in the lower nanometer technology nodes, gate leakage will remain the dominant leakage component after subthreshold leakage. [1] Since the way designers control subthreshold and gate leakage can change from one technology to another, it is crucial for them to be aware of the impact of the total leakage on the operation of circuits and the techniques that mitigate it.

Consequently, techniques that reduce total leakage in circuits operating in the active mode at different temperature conditions are examined. Also, the implications of technology scaling on the choice of techniques to mitigate total leakage are investigated. This work resulted in guidelines for the design of low-leakage circuits in nanometer technologies. Logic gates in the 65nm, 45nm, and 32nm nodes are simulated and analyzed. The techniques that are adopted for comparison in this work affect both gate and subthreshold leakage, namely, stack forcing, pin reordering, reverse body biasing, and high threshold voltage transistors. Aside from leakage, our analysis also highlights the impact of these techniques on the circuit's performance and noise margins.

The reverse body biasing scheme tends to be less effective as the technology scales since this scheme increases the band to band tunneling current. Employing high threshold voltage transistors seems to be one of the most effective techniques for reducing leakage with minor performance degradation. Pin reordering and natural stacks are techniques that do not affect the performance of the device, yet they reduce leakage. However, it is demonstrated that they are not as effective in all types of logic since the input values might switch only between the highly leaky states.

Therefore, depending on the design requirements of the circuit, a combination, or hybrid

of techniques which can result in better performance and leakage savings, is chosen. Power sensitive technology mapping tools can use the guidelines found as a result of the research in the low power design flow to meet the required maximum leakage current in a circuit. These guidelines are presented in general terms so that they can be adopted for any application and process technology.

Acknowledgements

This research project would not have been possible without the support of many people. I would like to thank my research advisors Professor Mohamed I. Elmasry and Professor Mohab Anis for their guidance and encouragement through out this research. I would also like to thank my colleagues in the electrical and computer engineering department for their suggestions and assistance. Special thanks are due to my parents, sister and numerous friends who endured this long process with me, offering me endless support and love.

To my

dearest family

Contents

1 Introduction **1**

1.1 CMOS Scaling 1

1.2 Importance of Leakage 1

1.3 Research Approach 2

1.4 Contributions 3

1.5 Thesis Organization 3

2 Background **5**

2.1 Introduction 5

2.2 Sources of Power Consumption in CMOS 5

2.3 Impact of Technology Scaling 6

2.4 Understanding Leakage 7

2.5 MOS Leakage Mechanisms 8

2.5.1 Gate leakage 8

2.5.2 Gate Induced Drain Leakage (GIDL) 11

2.5.3 Band to Band Tunneling (BTBT) Leakage Current 12

2.5.4 Subthreshold leakage 12

2.5.5 Punch-through Current 14

2.6 Leakage Reduction with Device Techniques 15

2.6.1 Halo Doping 15

2.6.2 Source/Drain extensions (SDE) 16

2.6.3 Super Steep Retrograde Well 16

2.6.4 High-K Gate Dielectric 16

2.7 Leakage Reduction by Circuit Techniques 17
2.7.1 Reverse Body Biasing (RBB) 17
2.7.2 Dual V_{th} Transistors Assignment 20
2.7.3 Natural/Forced Stacking 20
2.7.4 Pin Reordering 22
2.8 Summary 23

3 Characterization and Simulation Setup 25
3.1 Introduction 25
3.2 Design Parameters 25
3.3 Characterization 27
3.3.1 Predictive Technology Model (PTM) 27
3.4 Summary 32

4 Logic Gates 34
4.1 Introduction 34
4.2 Design Metrics 34
4.3 Effect of Gate Leakage 36
4.4 Inverter 39
4.4.1 DC Characteristics 39
4.4.2 Change in Rise and Fall Times 41
4.4.3 Total Leakage versus Input 43
4.4.4 Effect of Temperature 50
4.5 Two Input NAND Gate 52
4.5.1 DC Characteristics 52
4.5.2 Change in Rise and Fall Times 53
4.5.3 Total Leakage versus Inputs 55
4.5.4 Effect of Temperature 59
4.6 Two Input NOR Gate 61
4.6.1 DC Characteristics 61
4.6.2 Change in Rise and Fall Times 62

4.6.3 Total Leakage versus Inputs 64
4.6.4 Effect of Temperature 69
4.7 Three Input NAND Gate 70
4.7.1 DC Characteristics 70
4.7.2 Change in Rise and Fall Times 71
4.7.3 Total Leakage versus Inputs 73
4.7.4 Effect of Temperature 79
4.8 Three Input NOR Gate 80
4.8.1 DC Characteristics 80
4.8.2 Change in Rise and Fall Times 81
4.8.3 Total Leakage versus Inputs 82
4.8.4 Effect of Temperature 88
4.9 Low Leakage Logic Gate Selection 89
4.10 Power Aware Technology Mapping Tools 94

5 Conclusions **98**

5.1 Future Work 100

List of Figures

2.1 Active power vs. leakage power consumption 6

2.2 Typical scaling of a MOSFET by a factor of S 7

2.3 Scaling of Vdd and Vth 8

2.4 Short channel MOS leakage mechanisms [2] 9

2.5 Gate direct tunneling current vs. gate voltage [3] 10

2.6 Components of gate leakage [2] 10

2.7 Inverted n+ region with a high negative gate-drain voltage [2] 11

2.8 NMOS drain current vs. gate voltage [2] 13

2.9 Bulk CMOS structure [2] 15

2.10 Self-adjusting threshold voltage scheme [4] 19

2.11 Altera Stratix III programmable power technology [5] 21

2.12 Forced NMOS stack effect in an inverter 22

2.13 Two input NAND gate 23

2.14 Pull-down network of a two input NAND gate for 10 and 00 inputs 24

3.1 Six techniques to reduce the total leakage 26

3.2 Subthreshold current measurement setup 27

3.3 Subthreshold current and V_{th} vs. different technology nodes 28

3.4 Subthreshold current vs. gate voltage for different technology nodes 29

3.5 Gate current measurement setup 29

3.6 Gate to channel leakage current for different technology nodes 30

3.7 V_{th} Values for an inverter using PTM and ST Microelectronics Models . . 31

3.8 Gate to channel leakage current for different technology nodes 32

4.1 Noise margin definition (DC simulation) 35
4.2 Delay, rise and fall time definitions (DC) 36
4.3 Sample of input mapping for the table above 37
4.4 Inverter characteristics 65nm @ 25 °C 40
4.5 Inverter noise margins @ 25 °C 41
4.6 Inverter switching voltage across technologies 25 °C 42
4.7 INV Leakage vs. input @ 25 °C 45
4.8 INV leakage vs. input for various leakage reduction techniques @ 25 °C . . 48
4.9 INV leakage vs. input for various leakage reduction techniques @ 25 °C . . 49
4.10 INV average total leakage @ 25 °C 50
4.11 INV % leakage savings @ 25 °C 51
4.12 INV % leakage savings @ 90 °C 51
4.13 NAND2 DC characteristics 65nm @ 25 °C 52
4.14 NAND2 noise margins @ 25 °C 53
4.15 NAND2 switching voltage across technology nodes @ 25 °C 54
4.16 NAND2 leakage vs. inputs @ 25 °C 55
4.17 2-input NAND gate leakages in steady state 56
4.18 NAND2 leakage vs. input for various leakage reduction techniques @ 25 °C 57
4.19 NAND2 leakage vs. input for various leakage reduction techniques @ 25 °C 58
4.20 NAND2 average total leakage @ 25 °C 59
4.21 NAND2 % leakage savings @ 25 °C 59
4.22 NAND2 % leakage savings @ 90 °C 60
4.23 NOR2 DC characteristics 65nm @ 25 °C 61
4.24 NOR2 noise margins @ 25 °C 62
4.25 NOR2 switching voltage across technologies @ 25 °C 63
4.26 NOR2 leakage vs. inputs @ 25 °C 64
4.27 NOR2 inputs and leakages in steady state 65
4.28 NOR2 leakage vs. input for various leakage reduction techniques @ 25 °C 66
4.29 NOR2 leakage vs. input for various leakage reduction techniques @ 25 °C 67
4.30 NOR2 average total leakage @ 25 °C 68

4.31 NOR2 % leakage savings @ 25 °C 68
4.32 NOR2 % leakage savings @ 90 °C 69
4.33 NAND3 DC characteristics 65nm @ 25 °C 70
4.34 NAND3 noise margins @ 25 °C 71
4.35 NAND3 switching voltage across technologies @ 25 °C 72
4.36 NAND3 leakage vs. inputs @ 25 °C 73
4.37 NAND3 inputs and leakages in steady state 74
4.38 NAND3 leakage vs. input for various leakage reduction techniques @ 25 °C 76
4.39 NAND3 leakage vs. input for various leakage reduction techniques @ 25 °C 77
4.40 NAND3 average total leakage @ 25 °C 78
4.41 NAND3 % leakage savings @ 25 °C 78
4.42 NAND3 % leakage savings @ 90 °C 79
4.43 NOR3 DC characteristics 65nm @ 25 °C 80
4.44 NOR3 noise margins @ 25 °C 81
4.45 NOR3 switching voltage across technology nodes @ 25 °C 82
4.46 NOR3 leakage vs. inputs @ 25 °C 83
4.47 NOR3 inputs and leakages in steady state 84
4.48 NOR3 leakage vs. input for various leakage reduction techniques @ 25 °C 86
4.49 NOR3 leakage vs. input for various leakage reduction techniques @ 25 °C 87
4.50 NOR3 average total leakage @ 25 °C 88
4.51 NOR3 % leakage savings @ 25 °C 88
4.52 NOR3 % leakage savings @ 90 °C 89
4.53 Different implantations of the function 91

List of Tables

2.1 Total leakage and threshold voltage of an inverter 22

2.2 Total leakage and threshold voltage of a NAND gate 23

3.1 Design parameters 26

4.1 Worst case input rise and fall time calculation 36

4.2 Inverters average total leakage @ 25 °C 38

4.3 Logic gates average gate leakage to the total leakage @ 25 °C 39

4.4 Inverter rise and fall times @ 25 °C 43

4.5 NAND2 rise and fall times @ 25 °C 54

4.6 NOR2 rise and fall times @ 25 °C 63

4.7 NAND3 rise and fall times @ 25 °C 72

4.8 NOR3 rise and fall times @ 25 °C 83

4.9 Comparison of the total leakage of two input NAND and NOR gates 90

4.10 Total leakages for the implementations in Figure 4.53 92

4.11 Comparison of the average total leakage of two input NAND and NOR gates 93

Chapter 1

Introduction

1.1 CMOS Scaling

CMOS technology has been scaling down to meet the performance, production cost and power requirements of the industry. With the rapid growth of portable electronic devices, low power design is crucial in the design of electronics. Therefore, the focus of the industry has changed from high performance designs to low power designs to meet the demands of the portable electronics. Since the battery lifetime is a key factor for a portable device, both the electronic industry and the battery industry have been attempting to address this need. However, power reduction is also an issue in high performance computers where heat dissipation has become a major bottle neck in keeping the processors at proper operating temperatures. This is due to the growing power per unit area as a result of the reduction in the minimum feature size of the process technology, which allows a denser integration of transistor with higher speeds in each chip.

1.2 Importance of Leakage

The tremendous increase in the transistor leakage current is the primary disadvantage of technology scaling. Leakage affects not only the standby and active power consumption, but also the design margins, since it is closely related to process variations. As a result, it is vital for a circuit designer to be aware of the impact of leakage on the operation of the

circuit and techniques to mitigate it.

Leakage current affects the circuit differently depending on the operating conditions (eg. standby, active and burn in test), circuit family (eg. logic or memory) and environmental conditions (eg. temperature and supply voltage). Consequently, specific solutions exist for the particular condition of the circuit that it is being applied to. Principally, the focus of this work is on techniques that mitigate leakage in circuits operating in the active mode with various temperatures and supply voltages.

Since there is no single technique that will deal with all sources of leakage and their impact simultaneously, this problem must be tackled at various levels. So there exist solutions that will address leakage current at the system, circuit and device level. The circuit level solution is chosen for investigation in this work.

The three main components of MOS transistor leakage are gate, subthreshold and junction tunneling leakage. Presently, subthreshold leakage is the dominant component of the total leakage in current manufacturing technology nodes and it will remain to be the dominant component even at the lower technology nodes at higher than room temperatures. There are several different techniques that tackle these leakage components in various angles. Some of these techniques only focus on component, whereas others address more than one component at the same time. Also, some techniques might reduce one component of the MOS leakage but they might increase the other components. Therefore it is advantageous to know the limitations of each of these techniques, and their effectiveness in the lower technology nodes.

1.3 Research Approach

To understand the limitations and effectiveness of each leakage reduction techniques different logic structures such as inverter, two or three input NAND and NOR gates are simulated for the 65, 45 and 32nm technology nodes by using Predictive Technology Model (PTM) [6]. Parallel or series devices in the pull-up and pull-down network feature the behavior of each applied leakage reduction technique in all the cases of static CMOS. The

techniques are compared with respect to their effect on noise margins, delay, and total leakage.

The leakage reduction techniques studied in this work consist of stack forcing, high V_T CMOS, pin reordering and reverse body biasing. The reason for choosing these techniques is because they incorporate reduction of both gate and subthreshold leakage, either implicitly or explicitly, and do not necessitate expensive fabrication processes. These techniques are applied to the aforementioned gates and simulated for different technology nodes, temperatures, leakage considerations (subthreshold leakage only v.s. subthreshold and gate leakage) and various simulations types (e.g. DC and transient). The number of these simulations add up to more than a thousand. Therefore, to facilitate extracting data and running simulations due to the high number of simulations, a framework is created by using various technologies such as Perl, SQL and VB Script to run simulations and extract data into a human readable format.

1.4 Contributions

Several contributions of this work can be summarized as follows:

- The analysis of six different leakage reduction techniques
- The comprehension of the impact of technology scaling on the reduction techniques' leakage, noise margin, and delay.
- Guidelines for designing low power digital circuit

1.5 Thesis Organization

Chapter 1 presents the impact of technology scaling and the growing importance of leakage currents. Leakage mechanisms and leakage reduction techniques are explored in Chapter 2 to understand the effectiveness of each leakage reduction technique for a given set of requirements and conditions. The technology models and simulation setup are discussed in Chapter 3.

In Chapter 4 logic gates, such as inverter, two and three input NAND and NOR gates are simulated and compared with respect to their noise margins, delay and leakage. Lastly, Chapter 5 provides conclusions and recommendations for leakage reduction techniques and how they are impacted with respect to technology scaling.

Chapter 2

Background

2.1 Introduction

Microelectronics have grown tremendously in the past three decades because of the consistent scaling of CMOS technology. This reduction in size has enabled very dense transistors chips that have improved speed, functionality, and power compared to their predecessors. To achieve an optimal design, trade offs exists between power and performance at each stage of the design. Therefore the designer must understand the sources of power consumption and make these tradeoffs.

After the effect of technology scaling on power and transistor characteristics is explored, each leakage current component in a MOSFET is inversitgated.

2.2 Sources of Power Consumption in CMOS

For digital circuits, the power can be examined at its peak and average. In this research the average power is used, as peak power, is more related to reliability and performance of the device. In CMOS circuits average power consumption can be separated in two categories: active or dynamic power and passive or leakage power. Dynamic power is the result of logic gates switching states. Switching, glitching, and short circuit power are all components of dynamic power. During this process the capacitance that is associated with the gate will charge or discharge which causes power dissipation. Static power is the power that is

dissipated when the circuit is idle, or not switching states. Static power consumption is caused by the leakage currents, while the gates are idle.

Leakage currents affect the circuit both during active and idle mode of operations. Figure 2.1 shows the relative power consumed in active mode of operation versus the subthreshold leakage. It is evident the leakage power will eventually exceed the active power if no leakage reduction scheme is used. Leakage power reduction during active mode of operation is investigated in this work.

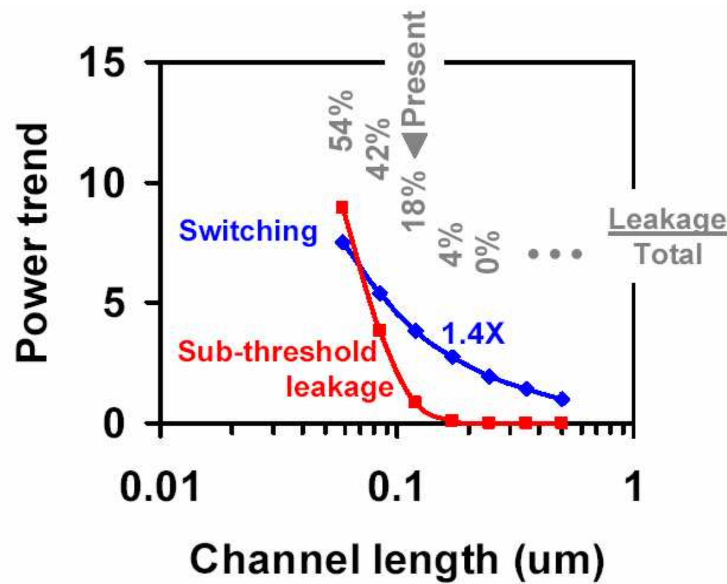


Figure 2.1: Active power vs. leakage power consumption

2.3 Impact of Technology Scaling

in 1975, Gordon Moore, predicted that the number of transistors that are intergraded on a single die will increase exponentially with time [7]. For the past last three decades, this prediction has held. In each new process technology most of the device dimensions scale to allow higher device integration. Figure 2.2 reflects the scaling of a typical MOSFET that has an immediate impact on the performance and power of the device. The primary effect of the scaling is to reduce the capacitances, which in turn, reduces power and delay.

There are two major types of scaling schemes for MOSFET devices. One is called Constant Field Scaling (CFS) and the other one is Constant Voltage Scaling (CVS). In CFS, all the dimensions of the transistors as well as the supply voltage are scaled down by a factor of S and the doping densities are increased by the same factor to preserve the internal electric field [8]. In CVS, the same scaling down occurs as CFS, however, the supply voltage remains unchanged in this case. To maintain the charge-field relationship, the doping densities are increased by a factor of S^2 [9].

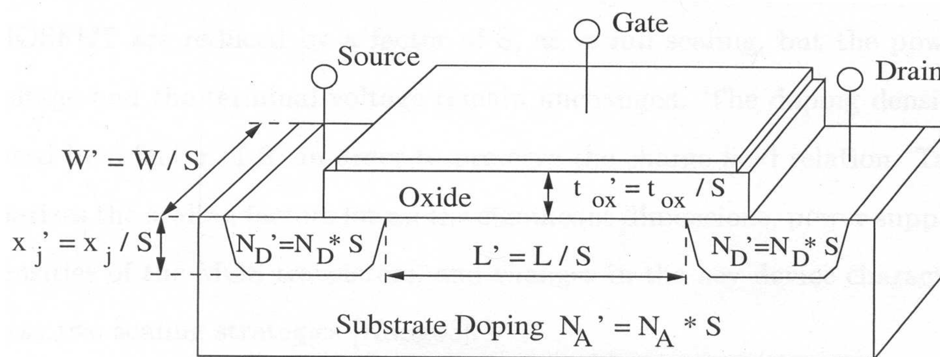


Figure 2.2: Typical scaling of a MOSFET by a factor of S

It can easily be observed that in case of CFS, the power consumption is reduced by a factor of S^2, and increased by a factor of S in the case of CVS. Even though CFS has this attractive power reduction feature, Intel had used CVS with 5V supply to maintain compatibility with supply voltage of the conventional system and to achieve better performance [10]. CFS has been used since the 0.5 μm generation, since the large increase in the drain current density can cause reliability problems such as electron migration, hot carrier degradation, and oxide breakdown. However, the negative impact of CFS is that it causes the subthreshold current to increase exponentially as the device scales, and turns it into a major component of the total power.

2.4 Understanding Leakage

The average switching power is proportional to the square of the supply voltage, which means a reduction of the supply voltage reduces power significantly. However, the trade off

here, is between performance and power. As a result, to maintain the device performance, the threshold voltage and the gate oxide thickness of the device must be scaled with the supply voltage. In Figure 2.3 the scaling of supply voltage and threshold voltage over time is plotted.

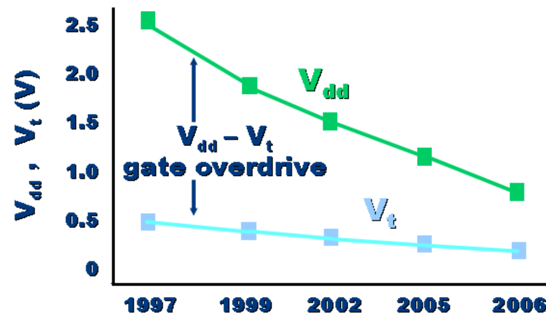


Figure 2.3: Scaling of Vdd and Vth

Such scaling has the disadvantage of increasing the device leakage exponentially, a major issue in nanometer process technologies. This phenomena is discussed in greater detail Section 2.5.4.

Now, each leakage mechanism of a MOSFET and its roll in the total leakage of the device is explored. Then, the techniques to mitigate the leakage through circuits and device structures are presented.

2.5 MOS Leakage Mechanisms

As illustrated in Figure 2.4 in there are five main short channel leakage mechanisms that have become the bottleneck of transistor scaling. Here each of these mechanisms of gate leakage, gate induced drain leakage, band to band tunneling leakage, subthreshold leakage and punch through leakage will be described.

2.5.1 Gate leakage

Gate leakage is becoming more visible and even a dominant component of leakage in the nanometer technologies. As silicon dioxide gate dielectric thickness decreases to keep up

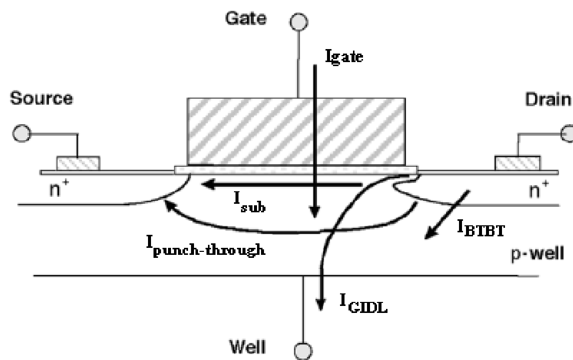


Figure 2.4: Short channel MOS leakage mechanisms [2]

with the technology scaling, the gate direct tunneling leakage current increases. The low oxide thickness, combined with the increased electric field across the oxide, results in significant electron tunneling from the substrate to the gate and vice versa.

Gate tunneling can be divided into two major mechanisms: Fowler-Nordheim (FN) tunneling and direct tunneling. In the first mechanism, tunneling occurs in the conduction band of the gate oxide, and in the latter case, electrons directly tunnel through the forbidden band gap of the oxide. Electron tunneling from Conduction Band (ECB), Electron tunneling from Valence Band (EVB), and Hole tunneling from Valance Band (HVB), are three means by which direct tunneling happens in MOS transistors. In the case of PMOS, the gate to channel leakage in inversion is regulated by HVB and gate to body tunneling is regulated by EVB and ECB in depletion-inversion and accumulation respectively. In the case of NMOS, the gate to channel current in inversion is regulated by ECB and the gate to body tunneling is regulated by EVB and ECB similar to PMOS. Due to the lower barrier height for ECB, compared to HVB the tunneling current associated with ECB is much higher than HVB. Therefore, NMOS experiences more gate leakage current than PMOS [11].

An increase in the voltage across the oxide or a decrease in the oxide thickness increases gate direct tunneling exponentially as reflected in Figure 2.5.

The current density of the gate tunneling is given by

$$J_{DT} = AE_{ox}^2 exp(-\frac{B[1-(1-\frac{V_{ox}}{/\phi_{ox}})^{3/2}]}{E_{ox}})[2] \tag{2.1}$$

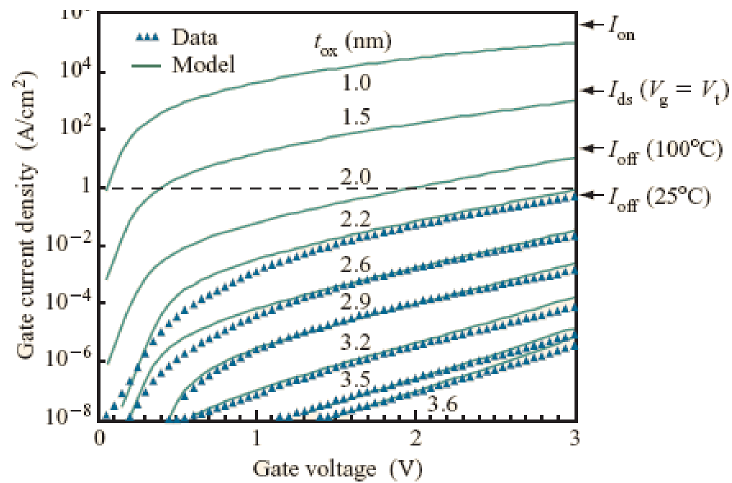


Figure 2.5: Gate direct tunneling current vs. gate voltage [3]

where E_{ox} is the field across the oxide,ϕ_{ox} is the barrier height for electrons in the conduction band, m^* is the effective mass of an electron, $A = q^3/16\pi^2 h\phi_{ox}$, and $B = 4\sqrt{2m^*}\phi_{ox}^{3/2}/3hq$

Five major components of the gate direct tunneling current are denoted in Figure 2.6. Gate to source (Igso) and gate to drain (Igdo) overlap region. Gate to channel (Igc) which is composed of gate to source (Igs) and gate to drain (Igd), and finally, the gate to substrate leakage (Igb).

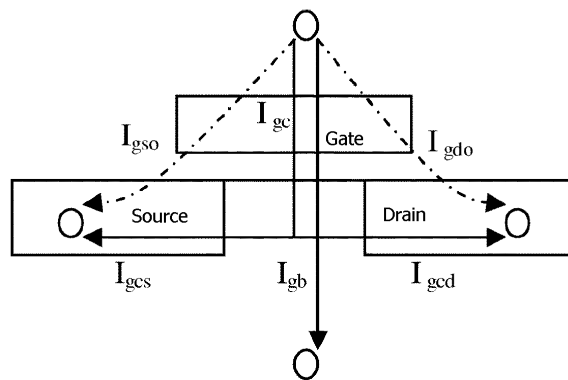


Figure 2.6: Components of gate leakage [2]

To overcome the limits of silicon dioxide scaling, several alternatives have been explored such as the use of metal gates, high permittivity (K) gate dielectric material, other transistor structures, and circuit techniques. The use of high-K material looks to be a promising

solution in nanometer technologies, because it results in a thicker oxide with the same capacitance as silicon dioxide gates. Therefore, the rise in oxide thickness, due to use of high-K material causes gate leakage to become insignificant compared to the other leakage components. However, high-K material has not been incorporated in any technology process for production yet, since there are some issues with their yield [12].

2.5.2 Gate Induced Drain Leakage (GIDL)

GIDL occurs when a negative potential is applied to drain and gate of a MOS transistor ($Vgd < 0$). The negative potential causes a depletion layer to form under the gate to drain over lap region. When the negative gate potential applied is large then the n+ drain region under the gate inverts or becomes depleted as shown in Figure 2.7. This causes a increase in high field effects such as band to band tunneling. In case of NMOS electrons will be collected by the drain and holes are collected by the bulk. This also can occur via near-surface trap assisted tunneling. However, BTBT has a higher dependency on electric field, and is also the dominant form of tunneling. [13]

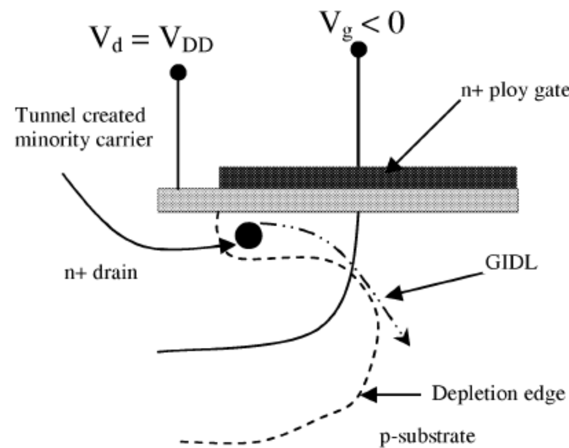


Figure 2.7: Inverted n+ region with a high negative gate-drain voltage [2]

GIDL varies exponentially with gate-drain voltage, and increases as the gate oxide thickness decreases (the electric field increases). Well doping heavily affects GIDL. At low drain doping there is no tunneling due to lower electric field but at very high doping the

depletion width is limited causing less tunneling. Therefore, the worse case is when the drain is moderately doped where the depletion width and electric field are significant. In general, to minimize GIDL very high and abrupt drain doping is required since it lowers the series resistance and increases drive current. [2]

2.5.3 Band to Band Tunneling (BTBT) Leakage Current

Typically in the normal mode of operation of MOS drain and source to well junctions are reverse biased. This causes a reverse biased pn junction leakage. In advanced nanometer transistor both n and p are heavily doped shallow junctions, and require halo doping to control short channel effects, BTBT dominates the pn junction leakage. Significant tunneling will occur from the valance band of the p region to the conduction band of the n region due to the high electric field across the reverse biased pn junction.[2] PMOS has considerably higher junction BTBT current than NMOS because effective density of state of hole is less than that of electron.

BTBT has a exponential dependence on body bias voltage as analyzed later, this causes reverse body biasing, a circuit technique, that is used to mitigate subthreshold leakage become less effective. [14]

2.5.4 Subthreshold leakage

Supply voltage must decrease as part of technology scaling, hence the threshold voltage of the device has to scale to maintain the gate delay reduction. This causes a significant increase in subthreshold current as the threshold voltage and subthreshold current are exponentially coupled.

Subthreshold current occurs between the source and drain of MOS when the gate voltage is below the threshold voltage. This current flow is due to diffusion of minority carriers at ($V_g < V_{th}$). In this region MOS transistor is behaving like a lateral bipolar transistor. Source and drain correspond to the emitter and collector respectively and substrate corresponds to the base. Based on bipolar modeling the current equation can be

approximated as follows [15]:

$$I_{sub} = I_s e^{\frac{q(V_{gs}-V_{th})}{nkT}}(1 - e^{\frac{-qV_{ds}}{kT}})(1 + \lambda V_{ds}) \tag{2.2}$$

where I_s and n are empirical parameters. To measure the quality of the device the slope factor S is defined as how much should the V_{gs} be reduced to get a drop factor of 10 in the drain current. From above equation [15]:

$$S = n(\frac{kT}{q})ln(10) \tag{2.3}$$

Where n is typically between 1 and 2, and S is expressed in mV/decade. To achieve the sharpest roll off, n should equal to 1, resulting S to equal 60mV/decade at room temperature. However, n is normally greater than one for typical bulk CMOS process, resulting in a slower rate of reduction.

In short channel MOS transistors, the deplition region around the drain increases as V_{ds} increases. The potential barrier between the source and the drain is reduced further or it is enhanced with the application of V_{ds}. This is referred to as Drain Induced Barrier Lowering (DIBL), which causes the the threshold voltage to decreases. There subthreshold current increases exponentially with high drain voltages. Figure 2.8 shows the effect of DIBL. Shorter channel lengths and higher drain voltages increase the effect of DIBL.

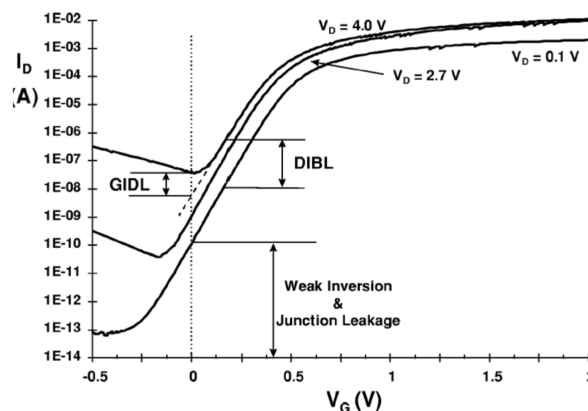


Figure 2.8: NMOS drain current vs. gate voltage [2]

2.5.5 Punch-through Current

In short channel MOS transistor the depletion regions at source-substrate and drain-substrate extend into to the channel due to scaling of the transistor geometry. The boundaries between the depletion region decreases as channel length is further reduced, and as the junction becomes highly reversed biased. At some point, this will lead to the boundaries merging, which then is called punch-through. This lowers the potential barrier for the electrons in case of an NMOS. Therefore, more of these carriers will enter into the substrate and some may flow to the drain as well. Punch-through current has a quadratic dependency on drain voltage, and increases the slope factor which reflects the drain leakage. [16]

2.6 Leakage Reduction with Device Techniques

Leakage current can be minimized by properly adjusting the device geometry and doping concentrations. Figure 2.9 shows the structure of a bulk CMOS with different well engineering techniques. To reduce the short channel effects several techniques are used in the doping profile of the transistor such as source/drain extensions(SDE) or lightly doped drain(LDD), halo doping and retrograde well which are discussed below.

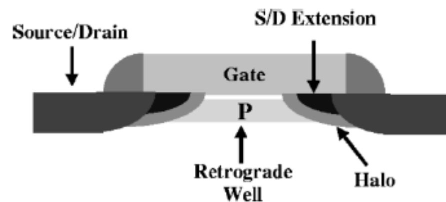


Figure 2.9: Bulk CMOS structure [2]

2.6.1 Halo Doping

A non-uniform implant, called halo, is used in short channel devices to mitigate the short channel effects [17]. This method allows controlling the dependence of threshold voltage on the channel length. A more highly p-type regions are added near the ends of the channel for a n-channel MOSFET, as it is shown in Figure 2.9. Charge-sharing from source and drain fields will be reduced as a result of the halo implants. This, in turn, will reduce the width of the depletion region, which reduces barrier lowering in the channel (DIBL reduction). The threshold voltage dependence on channel length variation is reduced due to reduction of charge-sharing effects. Therefore, the subthreshold leakage current becomes less sensitive to channel length variation.

Punch-through currents are reduced because the effective distance between source and drain depletion region is larger due to the high doping of channel edges. However the higher doping near the edges of the channel cause higher GIDL and BTBT currents which are the bottle neck to the halo doping level. [2]

2.6.2 Source/Drain extensions (SDE)

SDE are used to reduce the effect of wide electric fields in the drain and source regions. This effect reduces the DIBL and the V_{th} roll off which are the short channel effects (SCE). The depth of the SDE junction is tightly coupled with the SCE. The deeper the SDE, the higher the SCE, and spreading of the depletion region into the channel. However, the trade off between shallow SDE is an increase in the series resistance of the transistor [18]. Therefore, there exists an optimum point which reduces the series resistance and lowers the SCE. Moreover, if the junction is too lateral or too abrupt it will degrade the V_{th} roll-off. Therefore, there exists an optimum point for lateral abruptness. These parameters are used in combination, to yield the minimum SCE with a high drive current. [19]

2.6.3 Super Steep Retrograde Well

The gate-controlled depletion width and the oxide thickness have to be scaled in proportion to the channel length to reduce the SCE. Which translates to increasing the channel doping as the gate length is decreased, in order to maintain an acceptable subthreshold leakage. But this causes the threshold voltage to increase and reduce the device performance. As shown in Figure 2.9, retrograde well is a non-uniform vertical channel doping used to reduce the gate-controlled depletion width, hence maintaining the V_{th} reduction trend [2]. Retrograde well improves the SCEs and increases surface channel mobility, because the low surface concentration minimizes the channel impurity scattering. It also increases the linear drive current, resulting into performance improvement for logic gates [20] [21].

2.6.4 High-K Gate Dielectric

High permittivity (K) gate dielectric material are used to overcome the limits of silicon dioxide scaling, since the gate direct tunneling leakage current increases rapidly as the oxide thickness decreases. Using high-K material seems to be a promising solution in deep nanometer technologies as it results in a thicker oxide with the same capacitance as silicon dioxide gates. Therefore, this increase in oxide thickness, due to use of high-K material causes gate leakage to become insignificant compared to the other leakage components.

There are studies such as [22] that show that the transistor drive will be reduced when high-K dielectric is used and to overcome this problem a layer of low k dielectric between the substrate and high-K layer is used.

2.7 Leakage Reduction by Circuit Techniques

There are various circuit techniques that tackle different leakage components and are targeted for different types of operating modes of the circuits. Here, techniques that reduce leakage during active mode of the circuit will be explored, even though some might also reduce leakage during standby as a secondary effect. As discussed in section 2.5, gate and subthreshold leakage are the dominant components of leakage and only techniques that effect them will be investigated. More subthreshold leakage reduction techniques are discussed as it tends to be the dominant source of leakage, specially at high temperatures due to its exponential dependency with temperature.

Reverse body biasing, dual V_{th} transistors assignment, stacking and pin reordering are some of the circuit techniques that reduce leakage during active mode of operation and will be explored here. Pin reordering and stacking are techniques that will allow gate leakage reduction, as well as subthreshold leakage reduction.

2.7.1 Reverse Body Biasing (RBB)

RBB has been a technique that has been used since mid 1970s for different purposes depending on the type of the circuit and technology. In the early days, RBB was mostly used in memory chips to reduce the risk of latch up since there weren't enough substrate contacts for high density cell layout.[4] Now body biasing is being used to control subthreshold leakage, controlling V_{th} variation and reducing active power by reducing V_{th}.

RBB Scaling with Technology

In body biasing, V_{th} is controlled by utilizing the body effect. The formula for V_{th}, which incorporates body effect, is shown in the following equation [4]:

$$\begin{aligned} V_{th} &= V_{th0} + \gamma(\sqrt{2|\phi| - V_{BS}} - \sqrt{2|\phi|}) \\ \gamma &= \frac{t_{ox}}{\epsilon_{ox}}\sqrt{2\epsilon_{si} q N_A}, \quad \phi = \frac{kT}{q} ln(\frac{N_A}{N_i}) \end{aligned} \tag{2.4}$$

Where γ is the body effect coefficient, V_{BS} is the substrate potential, ϵ_{ox} is the dielectric constant of the SiO_2, V_{th0} is the value of V_{th} when there is no body bias, t_{ox} is the gate oxide thickness, ϵ_{si} is the permittivity of silicon, N_i is the carrier concentration in intrinsic silicon, N_A is the doping concentration density of the substrate, q is the electric charge, k is Boltzmann's constant, and T is the absolute temperature.

As it can be seen from equation 2.4 the threshold voltage is proportional to the square root of the body bias. Therefore, it is evident that as the technology scales down, the supply and threshold voltage are reduced. Obviously, body biasing cannot have the same impact on the threshold voltage, because it will not be able to create a large enough change in the threshold voltage. For example, if CMOS is scaled by a factor of k where k is greater than one in a constant field scaling scheme. The body effect coefficient scales by a factor of $\frac{1}{\sqrt{k}}$ since $\gamma \propto t_{ox}\sqrt{N_A}$, therefore, the body bias required to create a change in the threshold voltage needs to be scaled by a factor of $\frac{1}{\gamma^2} = k$. This is difficult to achieve, since the supply voltage is also scaling by a factor of 1/k. In addition, other parameters that should be taken into account are BTBT, and other short channel effects such as DIBL. As RBB is increased, subthreshold current is reduced but the BTBT is increased. Therefore, there exists an optimum point where the total leakage can be minimized by using RBB.

As per measurement results of [23] for using RBB as their leakage reduction technique, BTBT is the dominant junction leakage and the maximum achievable leakage power reduction diminishes by around four times per technology generation (assuming constant field technology scaling). RBB also causes the depletion layer of the drain-substrate to be extended, which further worsens SCE and V_{th} variations across a die. Also, the body effect coefficient is reduced in short channel devices since the channel potential is more

influenced by the drain due to DIBL than the substrate. [4]

Body Biasing Schemes

There are various methods of using body bias to our advantage. Each of these methods have their own limitations and advantages. The specific method used to achieve the body bias is beyond the scope of this work, and only the result of using body bias is evaluated for different circuits. To take advantage of the leakage savings of RBB scheme, the total leakage power saved must be more than the power of extra control block and circuitry needed to implement RBB. Also the performance loss must be taken into account so that the circuit still operates in the specified performance margins. An example of an implementation of body bias control is given below.

Self-adjusting threshold voltage scheme is a method that is used to compensate for variations in V_{th}. As shown in Figure 2.10, a leakage current monitor, monitors the subthreshold leakage and activates the substrate bias circuit when the leakage current is more than the reference current. Hence rasing the V_{th}, which in turn reduces the leakage current and then substrate bias circuit deactivates. However, the junction leakage raises the body bias voltage due to the impact ionization which reduces V_{th}. Therefore, subthreshold leakage increases again and the process reiterates. So, the variation in the V_{th} is mitigated by properly choosing a reference current signal. This scheme can be modified or extended to also reduce subthreshold leakage in the standby mode.

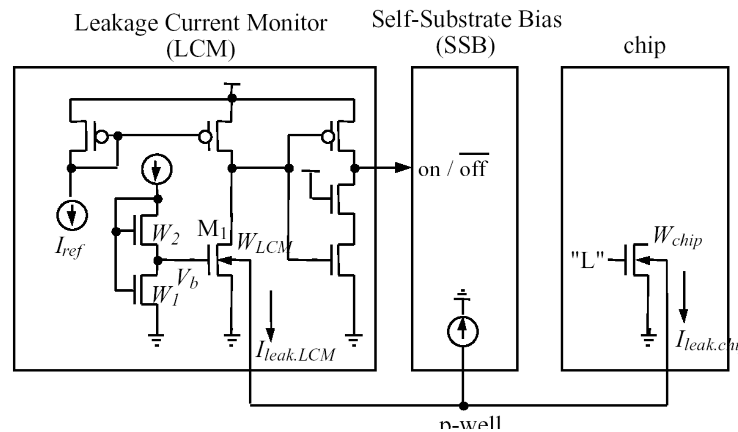


Figure 2.10: Self-adjusting threshold voltage scheme [4]

2.7.2 Dual V_{th} Transistors Assignment

Using dual threshold transistors, the designer is able to use both high and low V_{th} transistor to address leakage and performance simultaneously. High V_{th} transistors are assigned to some of the transistors that are not in the critical path and are highly leaky. Because if all or too many transistors are assigned as high V_{th} transistor this may alter the circuital path of the circuit. In contrast, low V_{th} transistors are assigned to transistors in the critical paths. In this manner the circuit will still maintain its performance while mitigating its leakage power.

Depending on the process technology there are different restrictions on how the dual V_{th} assignment can be done. For instance, the process might only allow you to use the same type of transistor in a stack, pull-up/down network, logic or just allow you to assign the type freely. These restrictions create cheaper fabrication process and more robust algorithms for Computer Aided Design (CAD) tools. However, these restrictions do not impact the leakage reduction significantly.

Field Programable Gate Array (FPGA) is a well-known application for using dual V_T transistors technique to reduce power. Since, in most designs, there is excess slack in the non critical paths they do not require high performance logic all the time. Therefore, the FPGA can contain a combination of logic elements that are high performance or low power, and the circuit is routed in a way to use the low power blocks for the non critical paths and the high performance blocks for the timing critical logic paths. Figure 2.11 shows Altera Inc.'s Stratix III programable power technology which uses a more sophisticated implementation of the concept above.

2.7.3 Natural/Forced Stacking

Subthreshold leakage is exponentially related to the threshold voltage of the device, and the threshold voltage changes due to body effect. From these two facts, one can reduce the subthreshold leakage in the device by stacking two or more transistors serially. The transistors above the lowest transistor will experience a higher threshold voltage due to the difference in the voltage between the source and body. Also, the V_{ds} of the higher transistor

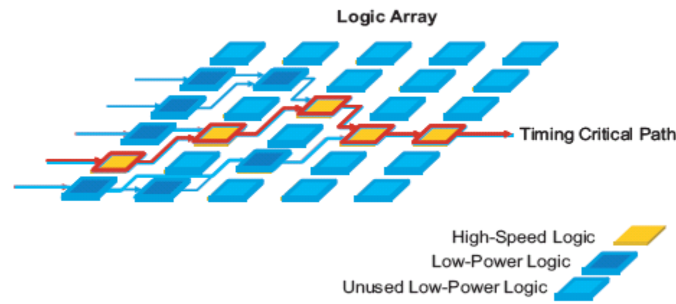


Figure 2.11: Altera Stratix III programmable power technology [5]

is decreased, since the intermediate node has a voltage above the ground. This results in reduction of DIBL effect hence better leakage savings. However, forced stack devices have a strong performance degradation that must be taken into account when applying the technique.

Figure 2.12 portrays an inverter with a forced NMOS stack. It is evident that the aforementioned effects are explained by looking at the threshold voltage and leakage values of this inverter in Table 2.7.3. From the table it can be seen that when the input of the inverter is zero, the leakage savings is large when the natural inverter and the forced NMOS case are compared. The new threshold voltages of the device verifies this fact. The node voltage at V_m does the following:

- increases N1's threshold voltage due body effect
- increases N1 and N2's threshold voltage due to lower V_{ds} (lower DIBL)
- puts N1 into strong off state since V_{gs} is negative

Another fact that is obvious in this table is that the threshold voltage of the short channel devices is strongly dependent on the drain-source voltage. The threshold voltage of the NMOS and PMOS varies about 40%, 70% respectively in 65nm technology. This change is contributed to the drain induced barrier lowering explained in Section 2.5.

From above, it was concluded that the subthreshold leakage strongly depends on the input applied to the circuit. This creates another method for reducing subthreshold leakage which is explained next.

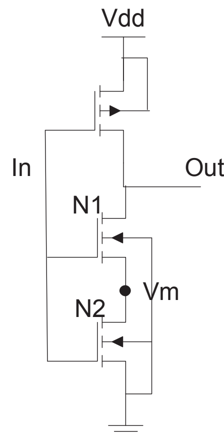


Figure 2.12: Forced NMOS stack effect in an inverter

Table 2.1: Total leakage and threshold voltage of an inverter using 65nm PTM Models @ 25 °C

Circuit	In	Total Leakage(A)	$V_m(V)$	$V_{thP}(V)$	$V_{thN1}(V)$	$V_{thN2}(V)$
Forced NMOS	1	3.84E-08	6.90E-05	0.2119	0.4245	0.4244
Forced NMOS	0	8.48E-10	9.01E-02	0.3651	0.3286	0.4142
Natural	1	3.86E-08	NA	0.2119	0.4244	NA
Natural	0	1.75E-08	NA	0.3651	0.2992	NA

2.7.4 Pin Reordering

By adopting the same concept, explained above, the input to gate can be ordered in a way that reduces leakage. For example, pin reordering is understood better by looking at a two input NAND gate in Figure 2.13. The total leakage is indicated in Table 2.7.4, and is the lowest when both inputs are zero, and the highest when both inputs are one. When the inputs are equal to zero, the pull-down network has two off NMOS transistors in series, achieving the lowest leakage. But, when the inputs are equal to one the pull-up network has two off PMOS transistors in parallel, creating less resistance than the series case, leading to a higher leakage. In the case of 10 or 01, there is only one off NMOS transistor, and due to the body effect and lower DIBL, the case of 10 (where 1 is applied to the NMOS closest to ground) is less leaky.

However, depending on the technology and oxide thickness, the gate leakage component

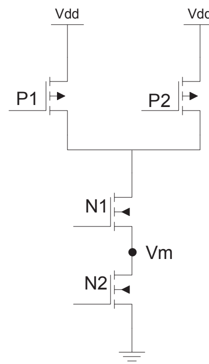


Figure 2.13: Two input NAND gate

Table 2.2: Total leakage and threshold voltage for a NAND gate - 65nm PTM @ 25 °C

Input	Total Leakage (nA)	$V_m(V)$
00	3.51	0.090
10	15.9	0.847
01	44	0.000
11	85.7	0.000

might shuffle these cases around. If a process is using a very thin gate oxide, then the gate leakage might be higher than the subthreshold at room temperature. For instance, the gate leakage is less in the case of 10 compared to 00. Since at 10, the intermediate node has a voltage of 0.85, resulting in a V_{gs} significantly less than VDD for N1 (lower I_{gc}), hence, reduced edge directed tunneling (EDT) current for N2, because the intermediate node is less than VDD. However, as shown in Figure 2.14, in the case of 00, N1 experiences high EDT current (I_{gdo}) and medium-low I_{gso} and N2 only experiences a medium-low I_{gdo}. Therefore, the gate leakage is higher in the case of 10 compared to 00.

2.8 Summary

It is shown that the primary reasons for technology scaling are to increase performance, and to reduce power and area to meet the industry needs. Constant field scaling which is used to achieve these requirements has created secondary issues such as short channel effects. This

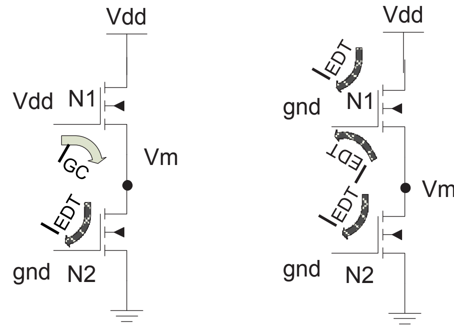


Figure 2.14: Pull-down network of a two input NAND gate for 10 and 00 inputs

type of scaling has made leakage currents an important source of power consumption with subthreshold leakage and gate leakage being the main components. Five main sources of short channel leakage mechanisms such as: gate leakage, gate induced drain leakage, band to band tunneling leakage, subthreshold leakage and punch through leakage are studied.

To suppress the leakage current in a circuit, a combination of device and circuit level techniques must be used. Halo doping, source/drain extensions, high-K gate dielectric, and super steep retrograde well are among the device level leakage reduction techniques that are discussed. Circuit techniques such as stack forcing, high V_T CMOS, pin reordering, and reverse body biasing are explored to mitigate gate and subthreshold leakage, the major components of these leakage currents. Gate and subthreshold leakage are further discussed in the following chapters, since the other components of leakage are orders of magnitude smaller, and can be neglected without impacting the results.

Chapter 3

Characterization and Simulation Setup

3.1 Introduction

In the previous chapter, the various methods of leakage reduction are described. The next step is to apply the techniques and analyze them. To further understand the effects of scaling through simulations and have a valid comparison, the validity of the process model used for simulation and the setup of each simulation is presented.

3.2 Design Parameters

An inverter is designed for equal rise and fall times in various technology nodes. It is discovered that the ratio of NMOS to PMOS should be about 2.7 to achieve this goal in PTM models for all technologies. The inverter is used as a reference to size the more complex logic gates. Table 3.1 illustrates how this inverter is setup with the sizing of each transistor. The minimum width used, is four times the channel length of the technology node. This reduces the more complex second order effects such as narrow width effect and reverse short channel effect. However, this also allows us to investigate other leakage reduction methods such as stack forcing. Since the width of the forced stack transistor should be halved, if a smaller minimum width is used then the technology model rules is

violated.

Table 3.1: Design parameters

Technology Node	NMOS (nm)	PMOS (nm)
65nm	W = 260, L = 65, Toxe = 18.5A, Vdd = 1.1V	W = 260*2.7, L = 65, Toxe = 18.5A, Vdd = 1.1V
45nm	W = 180, L = 45, Toxe = 16.5A, Vdd = 1V	W = 180*2.7, L = 45, Toxe = 16.5A, Vdd = 1V
32nm	W = 128, L = 32, Toxe = 15.5A, Vdd = 0.9V	W = 128*2.7, L = 32, Toxe = 15.5A, Vdd = 0.9V

Figure 3.1 depicts how the inverter is setup for the six different leakage reduction techniques that are used.

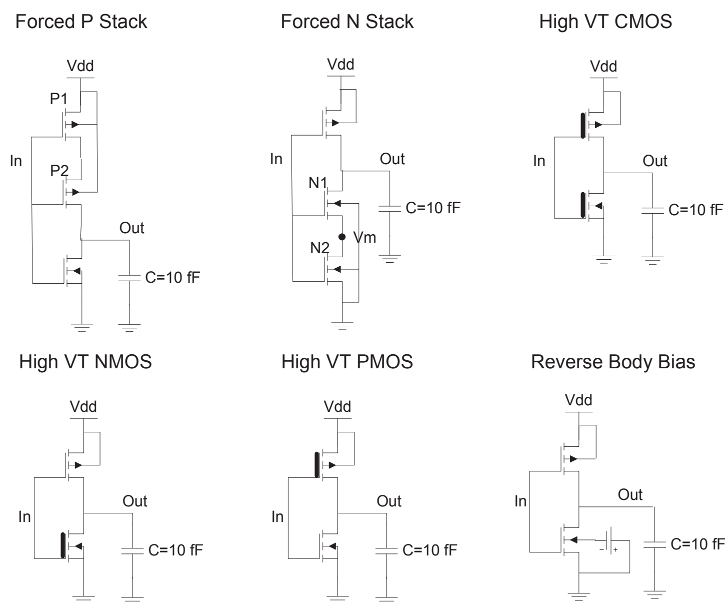


Figure 3.1: Six techniques to reduce the total leakage

Breaking down the high V_T transistors assignment technique into three flavors of usage in the pull-up network, pull-down network and a combination of both, helps to determine when and where this technique is best fit. Individual assignment of high V_T transistors faces fabrication limitation as only the transistors of the same type can be placed adjacent

to each other. Consequnelty, this technique is applied only to all the transistors in the pull-up network or pull-down network, or both.

3.3 Characterization

In this work various gates are simulated using Predictive Technology Models (PTM) [6] at three technology nodes: 65, 45 and 32nm. In addition, these gates are simulated after the techniques discussed in Chapter 2.7 are applied. The gates that are analyzed are: Inverter, 2 and 3 input NAND gates, 2 and 3 input NOR gates. All the logic gates analysis, is relative to the inverter's simulation results.

3.3.1 Predictive Technology Model (PTM)

PTM [6] is used to simulate logic circuits in nanometer technology nodes. Although these models incorporate gate leakage modeling (Berkeley Short-Channel Insulated-Gate Field-Effect Transistor Model, version 4) [11], they neglect the effect of BTBT. Since BTBT leakage is less than the gate leakage, neglecting BTBT leakage can be justified, if the gate leakage current is not suppressed by employing high-K material.

Here the characteristics of the PTM is explored. Subthreshold leakage, gate leakage and threshold voltages of PMOS and NMOS in the PTM is are examined. Figure 3.2 reflects the simulation setup for subthreshold and threshold voltage measurement.

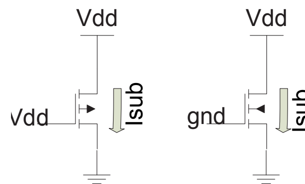


Figure 3.2: Subthreshold current measurement setup

Figure 3.3 illustrates how subthreshold current and threshold voltage vary with respect to technology scaling. The simulated transistors have a $\frac{W}{L} = 4$ and high V_{DS}. From the graph, it is obvious that the PMOS is subjected to lower subthreshold current, as expected,

due to the lower mobility of holes. Also the exponential dependency of subthreshold leakage and threshold voltage is depicted from the graph. The threshold voltage value on the graph is the final V_{th} , after taking into consideration the effect of DIBL and channel length modulation.

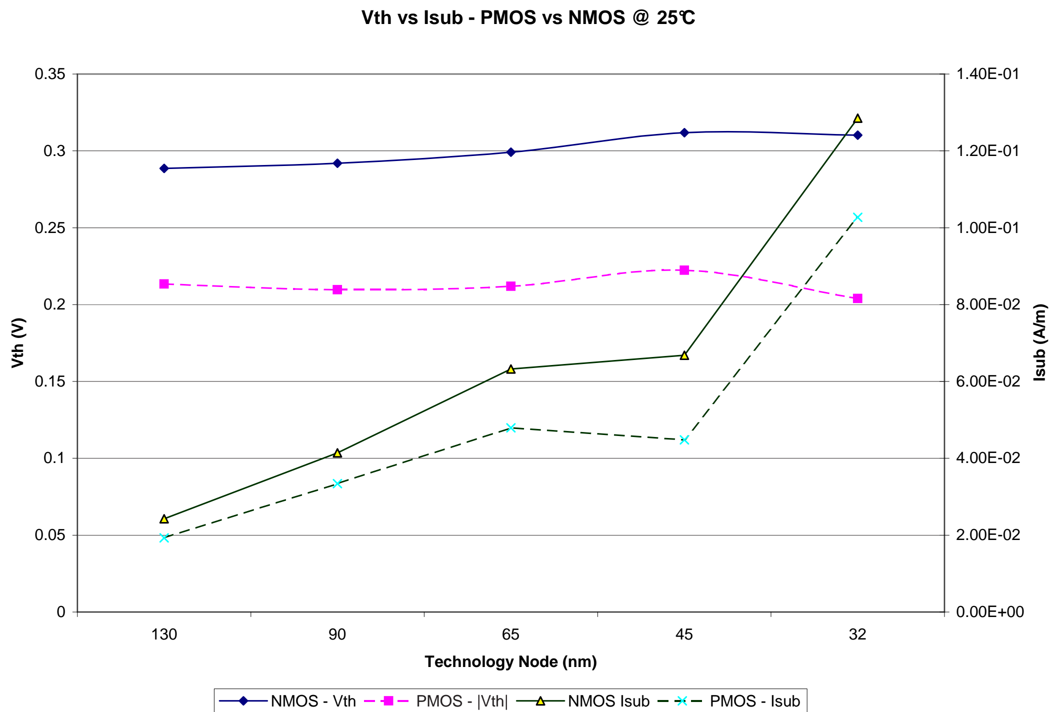


Figure 3.3: Subthreshold current and V_{th} vs. different technology nodes

Figure 3.4 signifies the dependency of subthreshold leakage with the gate to source voltage. Also the effect of using RBB on the NMOS is illustrated as well. The effect of RBB on leakage seems to decrease as technology scales, which is true due to reduction of supply voltage and body effect coefficient, and the increase in BTBT. As a result, an optimum point exists that reduces subthreshold leakage, and at the same time does not increase BTBT to overcome this reduction. However, since BTBT is not modeled in PTM the optimum value of RBB bias can not be found and the values are scaled from the optimum value for 130nm. The relationship between BTBT and RBB have been explained in Section 2.7.

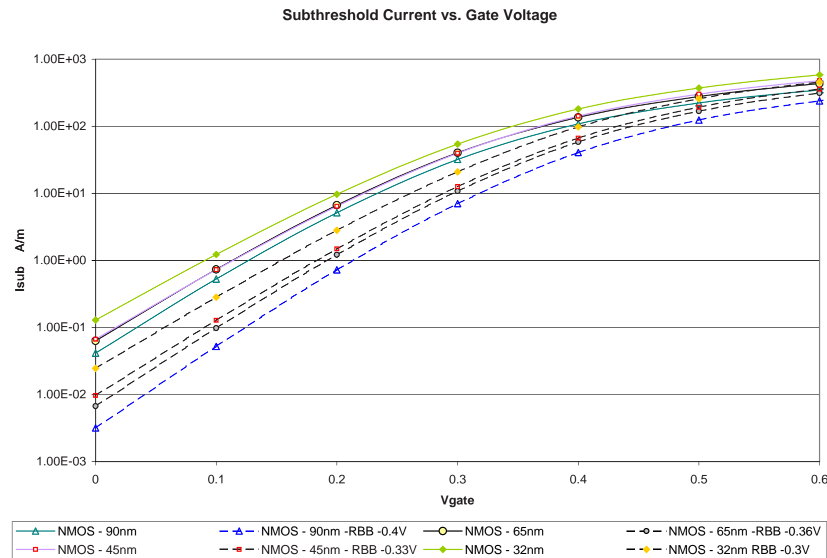


Figure 3.4: Subthreshold current vs. gate voltage for different technology nodes

Figure 3.5 represents the simulation setup for gate to channel leakage measurement. Using this setup the gate leakage current is plotted in Figure 3.6.

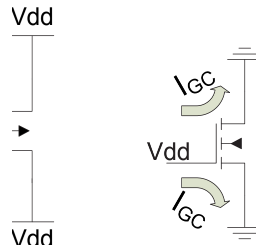


Figure 3.5: Gate current measurement setup

I_{EDT} is neglected in this comparison since it is an order of magnitude less than I_{gc} [14]. Also, the I_{gc} for a PMOS with a ratio of 2.7 times the NMOS is simulated, since in practice, most of the time PMOS is sized up to match the NMOS performance. Obviously, I_{gc} is directly proportional with the width of the transistor. It is also noteworthy that I_{gc} of PMOS is about an order of magnitude lower than that of the NMOS.

The effects of the leakage mitigation techniques on the V_{th} are compared and validated against the 90nm technology model from ST Microelectronics. In Figure 3.7 the V_{th} values

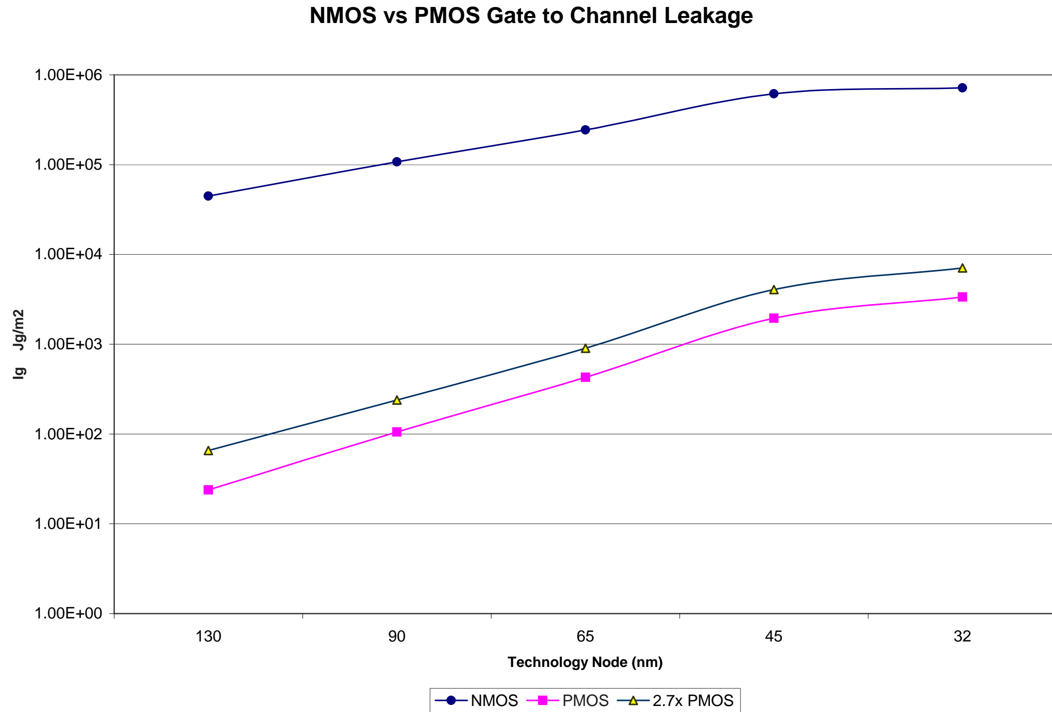


Figure 3.6: Gate to channel leakage current for different technology nodes

of all the transistors in an inverter with all the applied techniques are contrasted. Where V_{thx} is the threshold voltage of the NMOS (closer to the ground) or PMOS (closer to the output) that is used to create a forced stack pull-up or pull-down network. The details of the setup of this simulation are explored in the next chapter, and the values will be contrasted against each other. For example, by comparing the natural and high V_{tN} inverters it is seen that both ST and PTM models display higher threshold voltage for the NMOS as expected. It is clear that using the PTM, results in a lower threshold voltage and exaggeration of DIBL and stack effect in general. However, the change in the V_{th} values appears to be relatively the same as the ST Microelectronics model. Hence the relative comparison of simulations at various technology nodes with PTM, are appropriate and yield to correct conclusions.

Since nanometer technology nodes were not available to the university from the industry the PTM for these technologies is used to analyze the effect technology scaling. Using PTM for technology nodes of 65, 45 and 32nm the effect of DIBL on the threshold voltage is

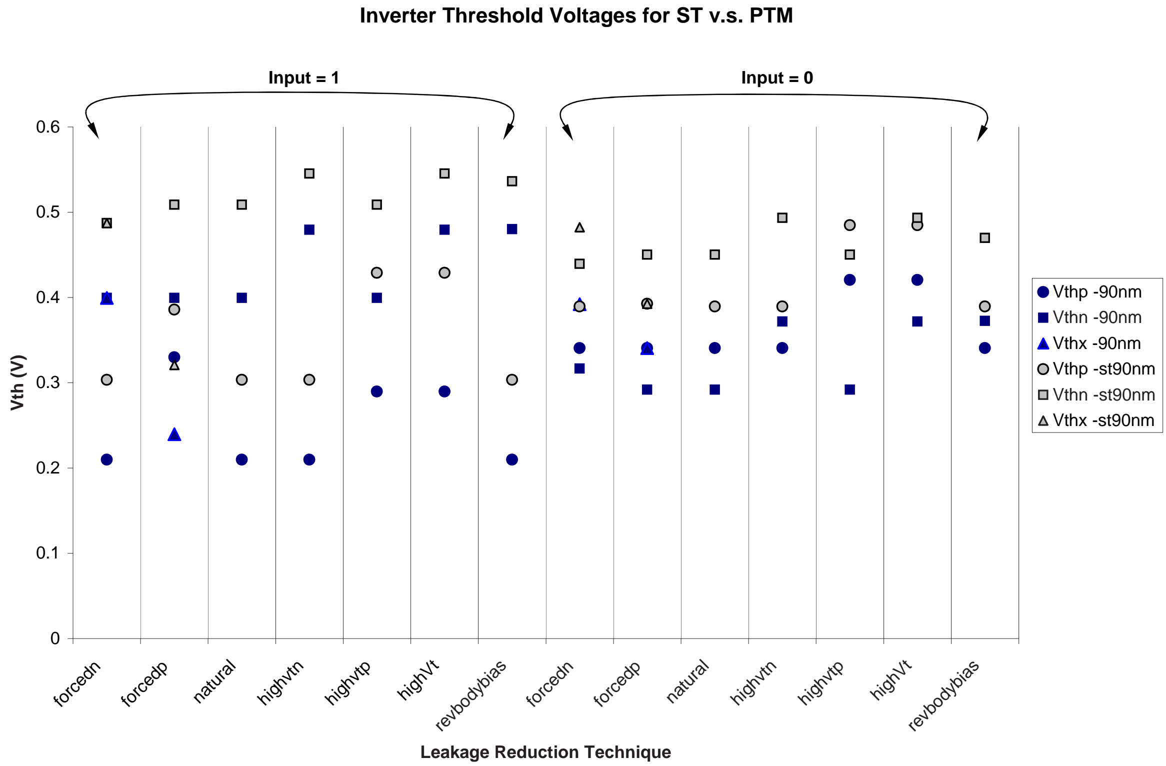


Figure 3.7: V_{th} Values for an inverter using PTM and ST Microelectronics Models

exhibited in Figure 3.8, where V_{thx} is the threshold voltage of the NMOS (closer to the ground) or the PMOS (closer to the output) that is used to create a forced stack pull-up or pull-down network. When the inverter's input is equal to one the PMOS transistor experiences DIBL effect and channel length modulation. Therefore, PMOS's threshold voltage is reduced, and the opposite is seen when the input is equal to zero. Also the effect of each technique on the threshold voltage can be contrasted with that of stack forcing, which has the largest effect on V_{th} . All three technologies exhibit similar behavior with respect to the threshold voltage, which again validates the consistency of the PTM simulation results.

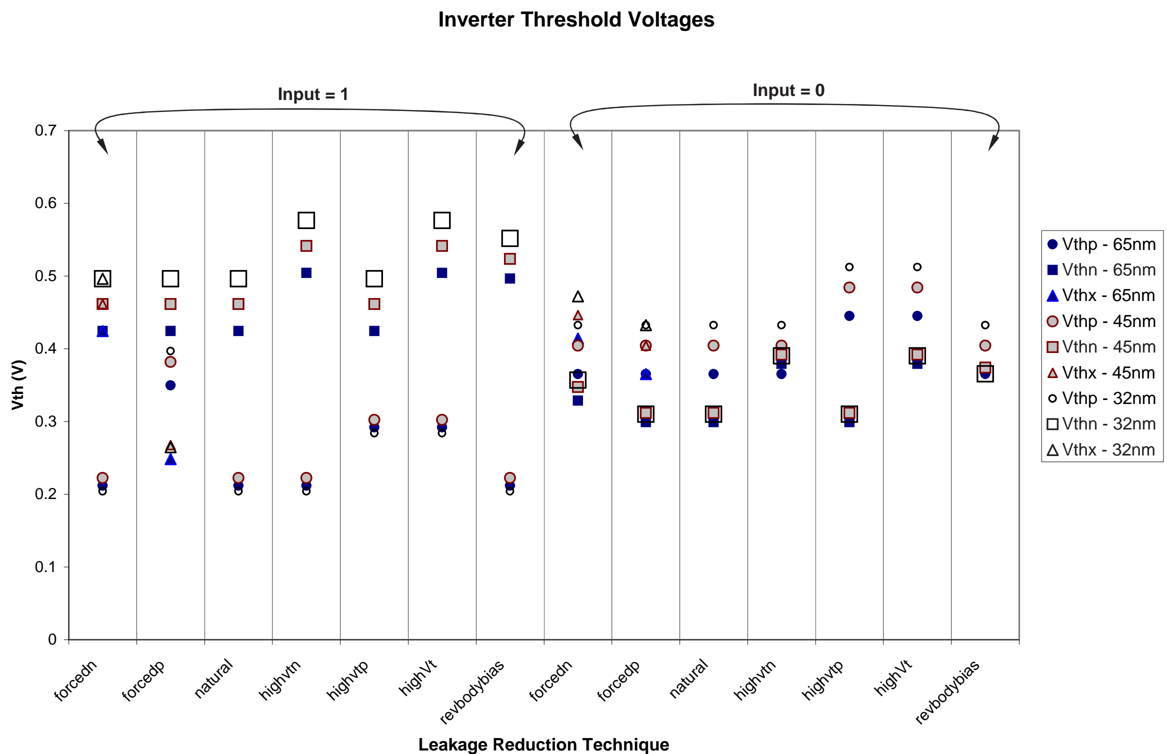


Figure 3.8: Gate to channel leakage current for different technology nodes

3.4 Summary

To understand the effects of scaling through simulations and to establish a valid comparison, the validity of the process model used for simulation and the setup of each simulation is discussed here. Since nanometer technology nodes are not available to the university from the industry, PTM is used for these technologies to analyze the effect of technology scaling. Threshold voltage, gate and subthreshold leakage are simulated for different technology nodes and their trends are explored, since the result of simulation of each logic gate is highly dependent on these trends. The 90nm model is also compared with the ST-Microelectronic's model, and even though the simulation results differ, the relative changes are in good agreement with each other. The characterization of PTM showed that the models are consistent in different technology nodes and exhibit similar results that are similar to those of industrial models.

The setup to apply six leakage reduction techniques are shown for an inverter. Similar setups will be used to simulate two and three input NAND and NOR gates which are discussed in detail in the following chapter.

Chapter 4

Logic Gates

4.1 Introduction

To realize the impact of technology scaling on logic gates, three key aspects of power, performance and reliability of an inverter and two and three input NAND and NOR gates are analyzed. Noise margin, rise and fall times and leakage current are used as a measure of reliability, performance, and power consumption respectively. These gates are chosen because the parallel or series devices in the pull-up and pull-down network display the behavior of each leakage reduction technique in static CMOS. Therefore, NAND and NOR gates cover these aspects and the inverter is used as the most basic logic block, which is a point of reference for the design and comparison of different logic gates.

4.2 Design Metrics

Leakage

Total leakage is measured from the simulation as means of leakage comparison between different logic gates. Simulation and measurement of leakage is discussed in Section 3.

Noise Margin

The unwanted variations in currents or voltages is referred to as noise. Noise margin for a logic gate, is defined as the boundary at which the input noise can be attenuated as it

travels to the output. On the voltage transfer characteristics curve, the points where the gain is equal to one, is where the slope is equal to negative one, as shown in Figure 4.1. The unity gain points, defined here with the output high and low values, are used to define the following equation:

$$NM_H = |V_{OH} - V_{IH}|, NM_L = |V_{IL} - V_{OL}| \tag{4.1}$$

The logic gate operates as expected, as long as the inputs are within the noise margin ranges.

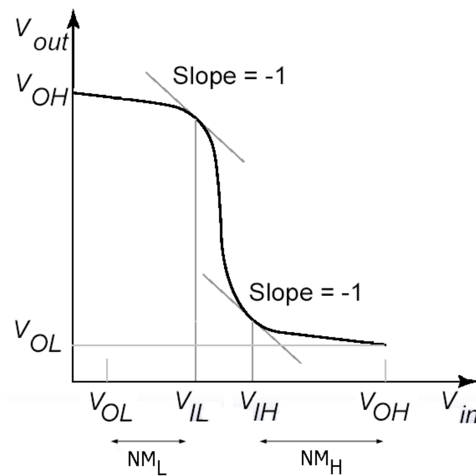
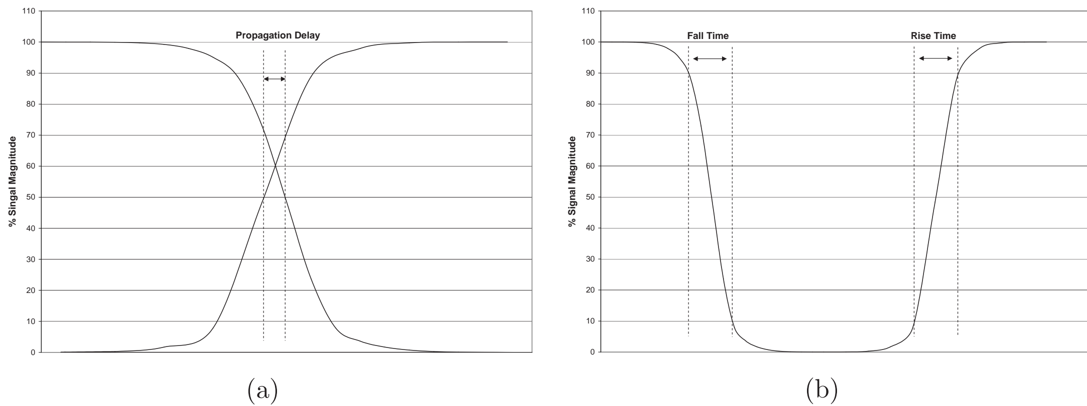


Figure 4.1: Noise margin definition (DC simulation)

Delay, Rise and Fall Times

Propagation delay, rise and fall times are defined as a measure to compare logic gates performance. Propagation delay is defined as the time that is required for a stable output, from the point when the input is stable. Here, propagation delay is defined as the time from when the input changes to 50% of its final value to the time when the output reaches 50% of its final value as denoted in Figure 4.2(a).

Rise time is defined as the time it takes for the output to increase from 10% of the logic high to 90% of the logic high as signified in Figure 4.2(b). Fall time is defined as the time it takes for the output to go from 90% of the logic high to 10% of the logic high as shown



(a) (b)

Figure 4.2: Delay, rise and fall time definitions (DC)

in Figure 4.2(b).

Rise and fall time for the more complex gates are calculated for the worse case scenario. Table 4.2 summarizes these input transitions, and the mapping is found in Figure 4.3.

Table 4.1: Worst case input rise and fall time calculation

Gate	Rise Time	Fall Time
Two input NAND	11→10	10→11
Three input NAND	111→110	110→111
Two input NOR	01→00	00→11
Three input NOR	011→000	000→011

4.3 Effect of Gate Leakage

In the simulations for the two and three input NAND/NOR gates and the inverter, the total leakage is considered to determine the lowest leakage input combinations or average leakages. Here the impact of gate leakage on performance, total leakage and noise margins is studied. In order to understand gate leakage's impact, the gate leakage must be extracted and analyzed separately, or the percentage of subthreshold leakage, with respect to the total leakage, can be used a measure of the gate leakage. In the following simulations with the PTM gate leakage is ignored, therefore, the total leakage is approximately equal to the subthreshold leakage. This approximation is valid since the oxide thickness used in these

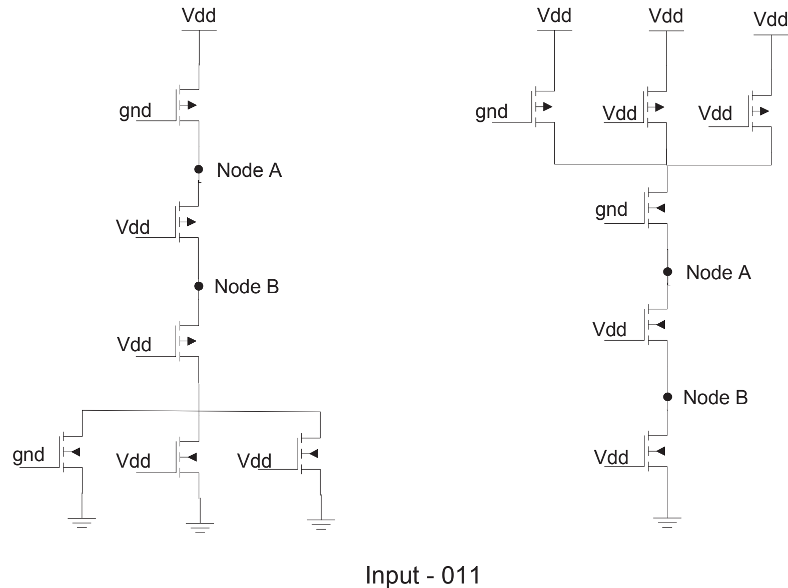


Figure 4.3: Sample of input mapping for the table above

technologies is at the higher side of the range with respect to International Technology Roadmap for Semiconductors (ITRS)[24]. As a result, the amount of gate leakage that the device experiences does not impact the node voltages enough to influence the subthreshold leakage.

Table 4.3 lists the inverters average total leakage for both cases of accounting for gate to channel current and gate to body current and for the case of ignoring these currents. The table also displays the percentage of leakage savings with respect to the natural form, standard deviation, and maximum to minimum leakage current. However, the standard deviation and maximum to minimum ratio convey more information on how the amount of leakage varies for different inputs which can be used to estimate the dependency of a logic gate's leakage on the order of its inputs.

The table shows that the amount of leakage savings is more when only subthreshold current is considered. This is understood as the absolute value of the total leakage has decreased by ignoring gate leakage. Therefore, a higher ratio of leakage savings from subthreshold reduction is achieved. Table 4.3 lists all the logic gates ratio of average gate

Table 4.2: Inverters average total leakage @ 25 °C

Tech	Method	Leakage (nA)	$\frac{Max}{Min}$	std (nA)	% Savings	Leakage (nA)	$\frac{Max}{Min}$	std (nA)	% Savings
		Igc and Igb included				Ignoring Igc and Igb			
32	forcedn	20.3	64.4	19.7	29.6	18.7	300.8	18.6	30.5
	forcedp	10.2	5.7	7.2	64.5	8.3	155.7	8.2	69.3
	natural	28.9	2.3	11.4	0.0	26.9	2.3	10.5	0.0
	highvtn	21.6	11.7	18.2	25.1	19.9	15.5	17.5	26.1
	highvtp	12.7	2.2	4.7	55.9	10.8	3.2	5.7	60.0
	highVt	5.5	2.2	2.1	80.9	3.7	2.1	1.3	86.1
	RBB	22.2	9.7	18.1	23.0	20.3	11.9	17.1	24.7
45	forcedn	14.1	30.5	13.2	31.4	11.3	146.5	11.2	34.4
	forcedp	9.3	2.6	4.2	54.7	6.1	95.9	6.0	64.9
	natural	20.5	2.0	7.0	0.0	17.3	1.9	5.3	0.0
	highvtn	14.9	8.7	11.9	27.3	12.1	14.1	10.5	30.2
	highvtp	10.6	1.7	2.9	48.2	7.4	4.4	4.6	57.3
	highVt	5.0	2.3	2.0	75.5	2.2	1.7	0.6	87.4
	RBB	15.4	8.6	12.2	25.0	12.1	12.9	10.4	29.7
65	forcedn	19.6	45.3	18.8	30.0	17.4	108.1	17.1	31.7
	forcedp	11.0	3.9	6.5	60.9	8.4	52.1	8.1	67.1
	natural	28.0	2.2	10.5	0.0	25.5	2.1	9.0	0.0
	highvtn	20.5	12.0	17.3	26.9	18.3	16.4	16.2	28.2
	highvtp	12.8	2.2	4.7	54.4	10.2	4.2	6.2	59.9
	highVt	5.3	2.3	2.1	81.3	3.0	1.9	0.9	88.1
	RBB	20.7	13.7	17.9	26.2	18.1	19.7	16.4	28.9

leakage to average total leakage. This tables declares how important the gate leakage is, at each technology node for a particular technique. It is observed that the percentage of gate leakage, as part of the total leakage, increases as different circuit techniques are used to mitigate subthreshold leakage. This amount increases by different values, depending on the effectiveness of the subthreshold reduction technique. For instance, in the case of an inverter employing high V_T transistors, gate leakages' share of the total leakage, moves from 16% to 57%.

The observation above indicates that gate leakage is about 10%-20% of the total leakage in nanometer devices at room temperature. Furthermore, this percentage increases, as subthreshold leakage reduction circuit techniques are applied. These values can increase even more, depending on the exact manufacturing process parameters such as gate oxide thickness. Therefore, to further reduce total leakage with less impact on the circuit performance, gate leakage must be considered. Using high-K material and/or pin-reordering appear to be the promising solution to reduce gate leakage.

However the circuit performance and noise margins are not affected by ignoring the gate leakage. This can be due to the fact that the relative change in leakage of pull-down and pull-up networks was insignificant for the worse case rise and fall time calculations.

Table 4.3: Logic gates average gate leakage to the total leakage @ 25 °C

Tech	Method	% average (I gate leakage)/(total leakage)				
		Inverter	**NAND2**	**NAND3**	**NOR2**	**NOR3**
32	forcedn	8.0	NA	NA	11.9	18.8
	forcedp	19.3	33.7	49.3	NA	NA
	natural	6.9	16.1	29.3	10.3	16.5
	highvt	32.0	54.8	72.2	42.2	55.5
	highvtn	8.1	19.9	36.2	11.5	17.7
	highvtp	15.2	29.5	45.2	24.0	36.5
	RBB	8.7	21.7	38.7	12.8	19.7
45	forcedn	19.8	NA	NA	25.9	36.0
	forcedp	34.8	51.6	65.8	NA	NA
	natural	15.8	31.9	49.3	22.0	31.6
	highvt	56.6	76.3	86.8	66.1	76.2
	highvtn	18.9	39.6	59.7	24.5	33.6
	highvtp	30.5	48.2	63.3	42.9	57.6
	RBB	21.5	42.9	62.9	27.2	37.1
65	forcedn	11.4	NA	NA	14.9	21.2
	forcedp	23.6	37.0	50.5	NA	NA
	natural	9.1	20.1	33.8	12.6	18.6
	highvt	42.3	64.0	78.1	51.0	62.2
	highvtn	10.8	25.2	43.2	13.5	18.9
	highvtp	20.3	34.0	48.0	30.0	43.3
	RBB	12.5	28.7	47.7	15.7	21.9

The DC characteristics of a logic gate do not depend on the amount of leakage which is confirming the fact that no change is observed.

4.4 Inverter

4.4.1 DC Characteristics

The inverter characteristics are examined with respect to six different leakage reduction techniques explained in the previous chapter.

Figure 4.4 denotes the DC simulation results of an inverter at 65nm technology node and the noise margins low and high are found in Figure 4.5(a,b). From the graph it can be seen that forcing a PMOS stack moves the characteristic curve to the left, which implies an increase in NM_H and a decrease in NM_L according to equation 4.1.

This behavior can be explained by the fact that stacking two PMOSs on top of each other increases the absolute value of the threshold voltage of the lower transistor. The pivotal factor is that the PMOS current drive is reduced significantly, since the width of the transistors is half of the original width. As expected, forcing a NMOS stack has the opposite effect, because the width of the transistors is half of the original NMOS, and

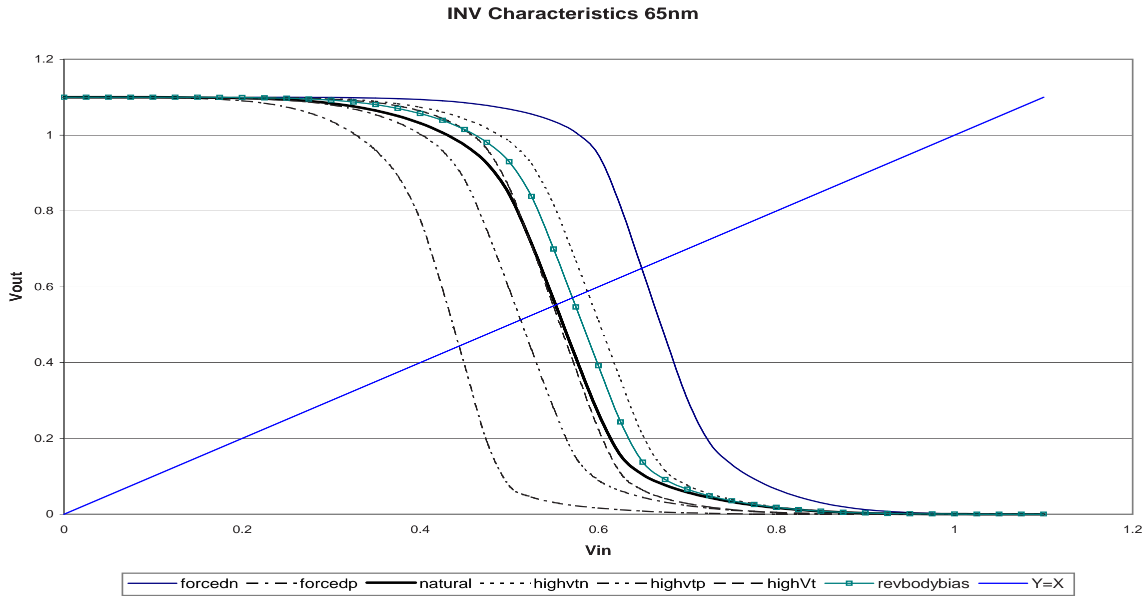
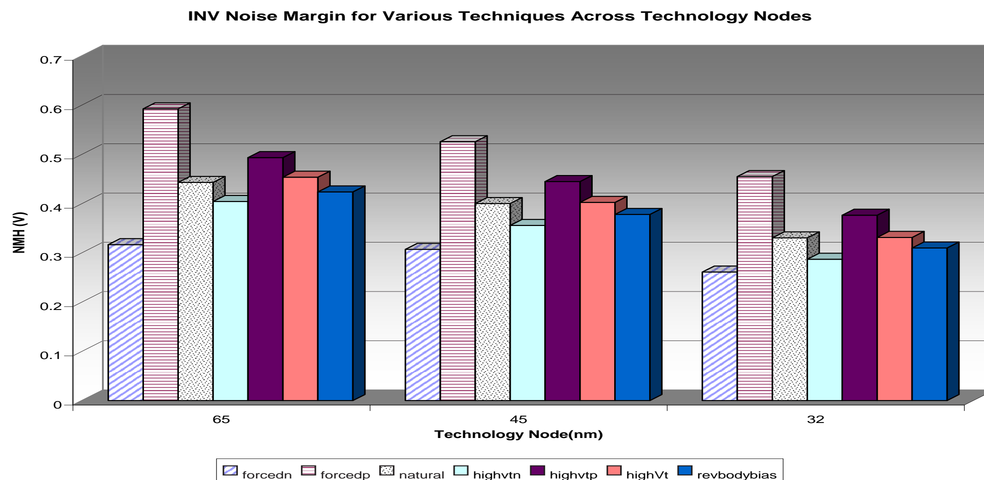


Figure 4.4: Inverter characteristics 65nm @ 25 °C

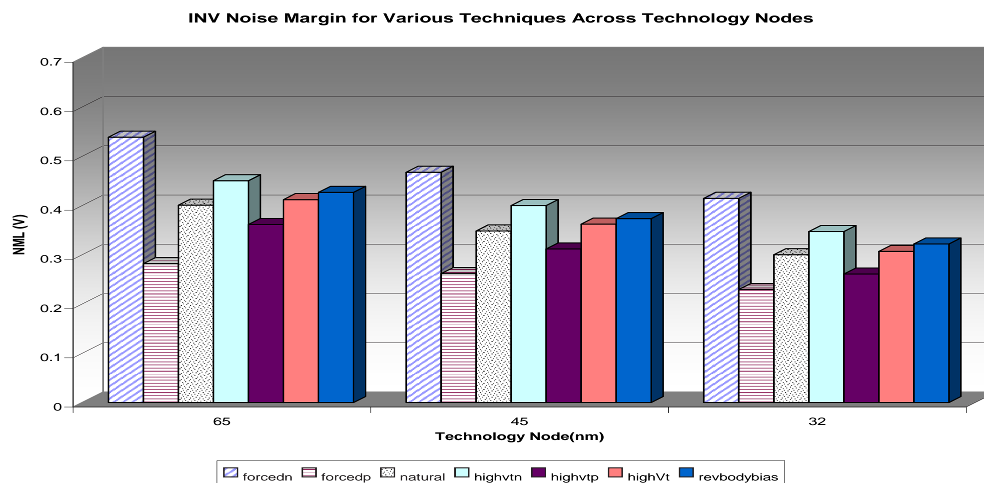
the current drive is reduced significantly. Also stacking two NMOSs on top of each other increases the threshold voltage of the higher transistor, further reducing the drive.

By applying the high V_T technique or the RBB technique, the inverter experiences a similar effect as that of forcing a NMOS stack. It is observed that this change is higher for the high V_T technique than the RBB. This is due to the fact that the high V_T NMOS transistor has a higher threshold voltage than the threshold voltage of the NMOS with applied RBB. Also, employing a high V_T PMOS, due to its proper size, has a higher current drive compared with the forced stack PMOS. As a result, it alters the noise margin less than the forced stack PMOS. The same conclusions are depicted for the 45 and 32nm technology nodes shown in Figure 4.5.

The intersection of the DC curve with $y = x$ is called the switching voltage and as shown in the graph, this point changes according to the applied leakage reduction technique. Figure 4.6 reflects the switching voltage for each technique at different technology nodes. It is obivous that the change in the switching voltage is proportional to the change in the noise margins for the particular technique and technology node. This change also seems to be equal in each different technology node. The other shift in total reduction of the



(a) NMH



(b) NML

Figure 4.5: Inverter noise margins @ 25 °C

switching voltage from one technology node to another is due to the reduction of the power supply voltage.

4.4.2 Change in Rise and Fall Times

Table 4.4 shows the rise time and fall time of the inverter for each applied technique for all three technology nodes. By looking at the percentage change in the rise and fall times

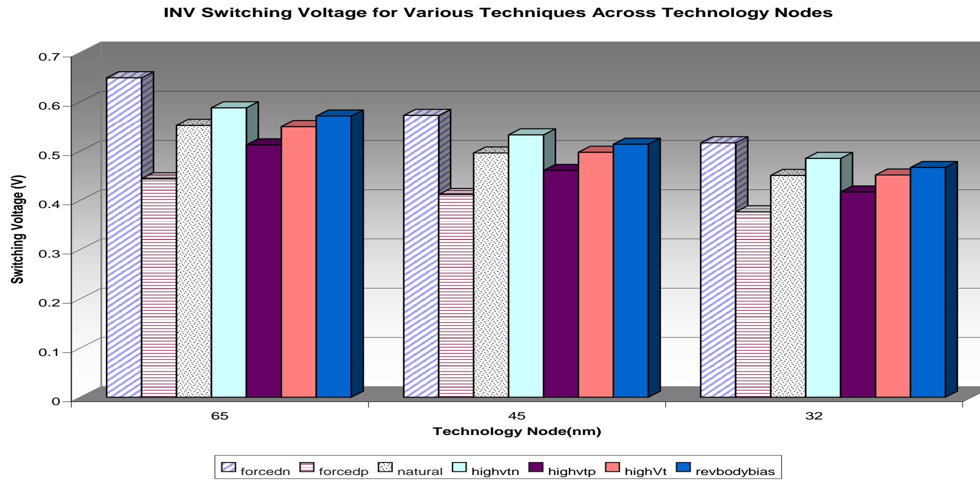


Figure 4.6: Inverter switching voltage across technologies 25 °C

of each technique compared to the original inverter it is deduced which technique has the most or the least impact on the performance of the gate.

For the case of 65nm technology node, about a 244% increase in the fall time is seen a forced NMOS stack. This is in line with the previous observation of the shifted characteristic curve, with the same reasoning mentioned above. The NMOS in the pull-down network has become weak thus, such increase in fall time is anticipated. The opposite holds true for the case of a forced PMOS stack in the pull-up network of the inverter. There is approximately 284% increase in the rise time of the inverter due to the weakening of the pull-up network. Using a high V_T transistor in the pull-down network has a much smaller effect than forcing a NMOS stack, affecting the fall time by about 11%. The same can be said for using a high V_T PMOS, which only changes the rise time by only 17% compared with the 284% of that of forced PMOS. It is evident that the effect of RBB on the fall time is about 6%, confirming that the threshold voltage does not change as much as it does in the case of high V_T technique.

The same trend is observed for the 45 and 32nm technology nodes. However, a look at the percentage change of rise and fall time values vertically across the technology nodes, conveys that the percentage changes is increasing for all the techniques. This can be explained by the fact that there is more leakage. The other effect that is interesting to

Table 4.4: Inverter rise and fall times @ 25 °C

Tech(nm)	Method	tRise(ps)	%change	tFall(ps)	%change
65	natural	40.8	0.0	42.5	0.0
	forcedn	40.6	-0.3	146.3	244.3
	forcedp	156.5	284.0	40.6	-4.4
	highvtn	40.8	0.0	47.3	11.3
	highvtp	47.7	17.0	42.5	0.0
	highVt	47.7	17.0	47.3	11.3
	revbodybias	40.7	0.0	45.0	5.9
45	natural	62.1	0.0	56.3	0.0
	forcedn	62.0	-0.1	202.7	259.9
	forcedp	252.4	306.3	54.6	-3.1
	highvtn	62.1	0.0	64.6	14.7
	highvtp	75.4	21.4	56.3	0.0
	highVt	75.4	21.4	64.6	14.7
	revbodybias	62.1	0.0	60.2	6.9
32	natural	82.0	0.0	72.2	0.0
	forcedn	82.0	0.0	277.2	284.1
	forcedp	356.7	335.1	70.6	-2.1
	highvtn	82.0	0.0	86.1	19.3
	highvtp	103.7	26.5	72.2	0.0
	highVt	103.7	26.5	86.1	19.3
	revbodybias	82.0	0.0	78.2	8.4

note is that in some cases, using one technique even though it causes a performance penalty in rise(fall) time, helps the fall(rise) time performance. For example, from Table 4.4, the forced PMOS technique in 65nm increases the rise time by 284% but reduces the fall time by 4.4%. This can be attributed to the amount of leakage reduction that is achieved in the pull-up (pull-down) network. Now, the transistors in the pull-up (pull-down) network has to battle less with the leakage in the pull-down (pull-up) network in order to charge up (discharge) the capacitance at the node, therefore, a little boost can be seen in the rise time (fall time).

On the same note, the performance gain achieved on the other side when using a leakage reduction technique is diminishing as the technology scales. This shows that the techniques are becoming less effective as they scale.

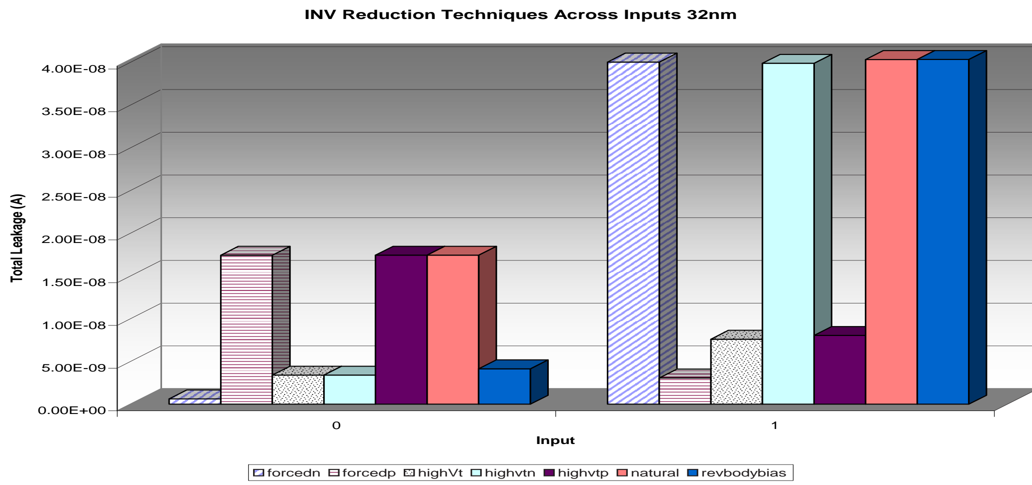
4.4.3 Total Leakage versus Input

Here, the impact of each of these techniques on the total leakage with all the possible inputs across all the technology nodes is examined. After the inverter is examined in its natural form with no techniques applied, then other techniques are compared with respect to the natural form.

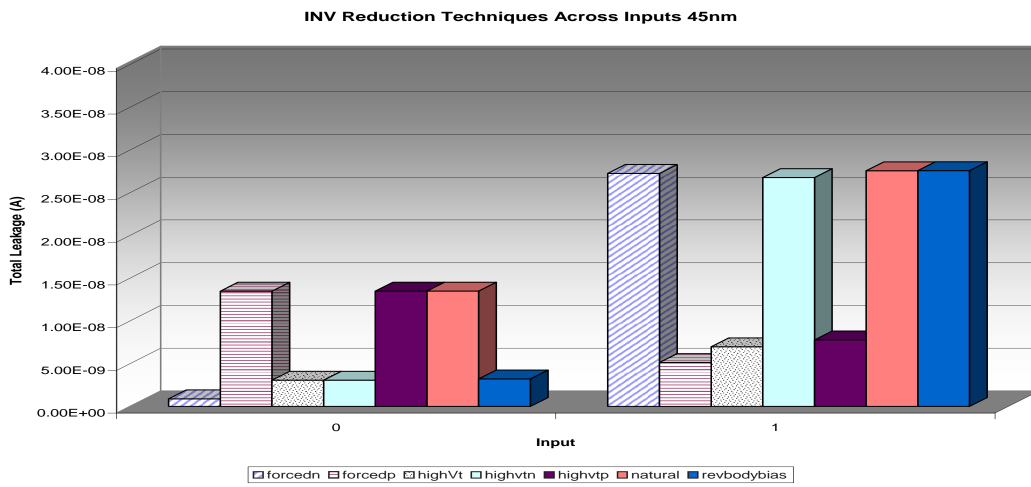
Figure 4.7, illustrates the total leakage of an inverter for each leakage reduction technique for all the different inputs at various technology nodes. Only for the inverter, the total leakage graph is included for all three technology nodes, and due to the similarities the other technology node graphs are omitted for the other logic gates. When the natural form of invertor is compared with the forced NMOS stack technique, as expected, the forced NMOS stack inverter has a significant leakage improvement when the input of the inverter is zero, and practically no change in the case that the input is one. This is because this technique only effects the leakage in the NMOS or the pull-down network, which is leaking when the input is zero, and there is not any change in the pull-up network's leakage.

In the same figure, similar results are seen comparing the total leakage of the inverter when a high V_T NMOS is used in the pull-down network, with when a RBB technique is used. By the same reasoning, as in the case of forced NMOS stack, the total leakage is reduced when the input is zero, and no change in when the input is one. Also, the amount of reduction is in line with the effect of each of these techniques on the DC characteristics of the inverter explained in Section 4.4.1. This indicates that the RBB technique has a lower leakage savings than that of high V_T NMOS, and the high V_T NMOS has a lower leakage savings compared with that of the forced NMOS stack.

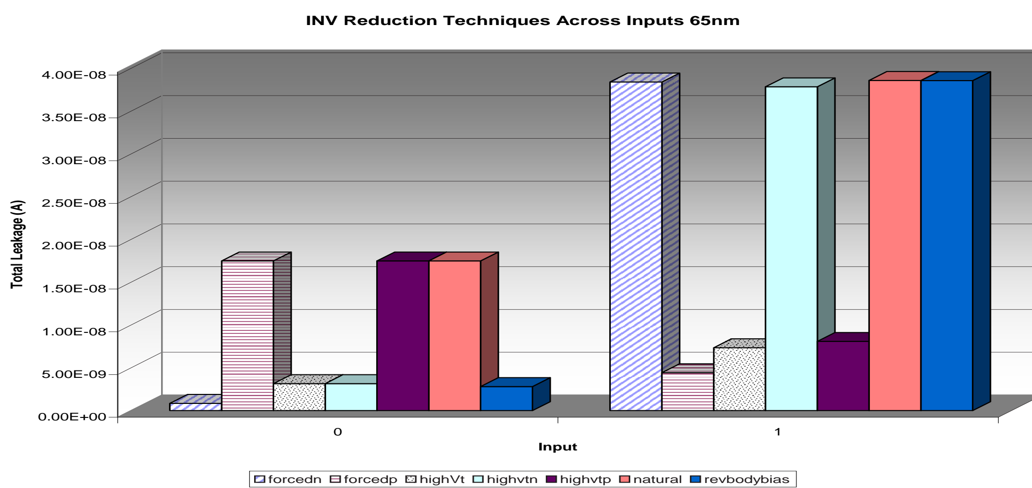
In the same figure, the total leakage of these two methods are contrasted: high V_T PMOS used in the pull-up network and a forced PMOS stack applied to the pull-up network. Comparing the two techniques with the inverter in its natural form, as expected the forced PMOS stack inverter is less leaky than the high V_T PMOS inverter. Both exhibit a significant leakage improvement when the input of the inverter is one, and practically no change when the input is zero. This occurs because these techniques effect the leakage only in the PMOS or the pull-up network, which is leaking when the input is one with no change in the pull-down networks leakage. Also, the amount of the reduction that is depicted, is in line with the effect of each of these techniques on the DC characteristics of the inverter, as explained in Section 4.4.1. Meaning that the high V_T PMOS has a lower leakage savings compared to the forced PMOS stack. When high V_T NMOS and PMOS are used in the inverters structure there are leakage savings for both of the possible inputs. This structure has very little effect on the DC characteristics of the inverter as explained in



(a) 32nm



(b) 45nm



(c) 65nm

Figure 4.7: INV Leakage vs. input @ 25 °C

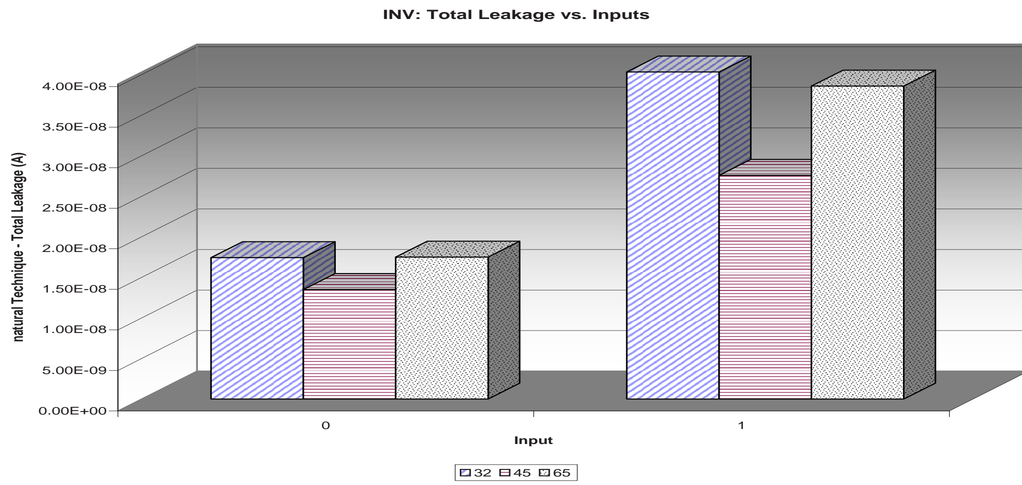
Section 4.4.1. However, one of the major down sides of this technique is that a performance penalty would be experienced in both rise and fall times. For the other technology nodes in Figures 4.7 (a,b,c), the same trend is seen for all the technologies, for each technique.

Figure 4.8 (a) shows the inverter's total leakage in its natural form across technologies. The figure shows a leakage reduction in the 45nm technology node for both inputs of zero and one. This behavior can be explained by the fact that in the PTM, the threshold voltage of the NMOS and PMOS device is a little higher than the 65nm and 32nm technology models as explained in Section 3.3.1.

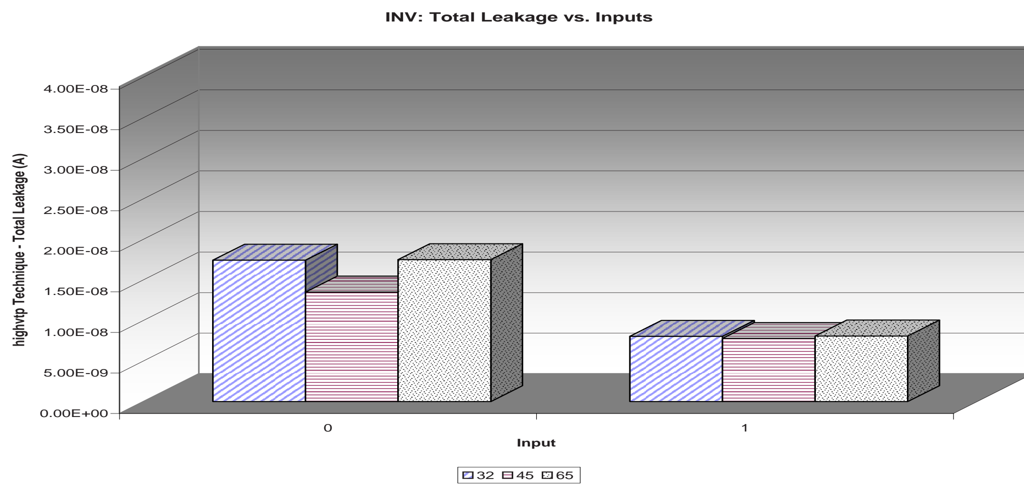
Figure 4.8 (b) shows the inverter's total leakage, when high V_T PMOS is used in its structure. The effectiveness of this technique in leakage reduction appears to be consistent across all the technology nodes.

Figure 4.8 (c) shows the inverters' total leakage, when a forced PMOS stack is applied. This figure portrays a little more savings at the 32nm technology node.

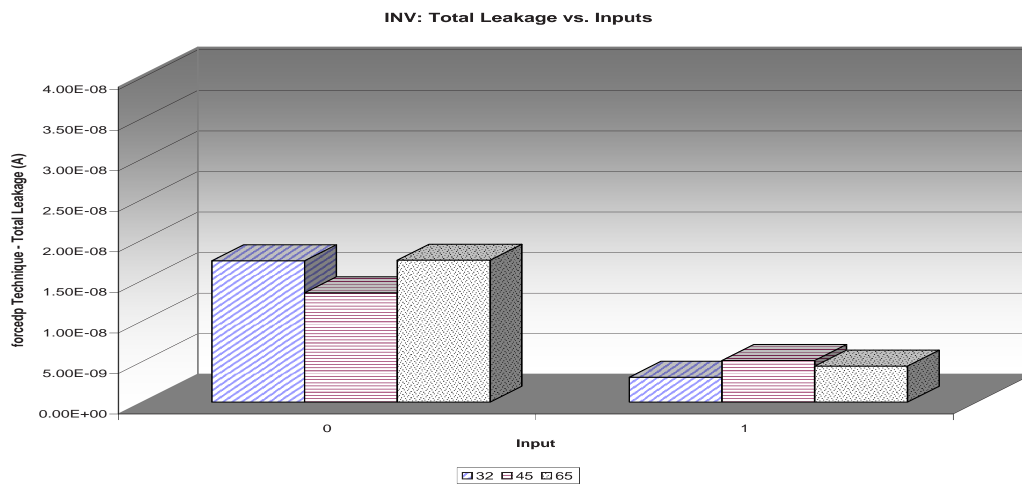
Figure 4.9 (a,b,c) show the inverter's total leakage when using a forced NMOS stack, high V_T NMOS and RBB technique in the inverters structure. All exhibit the same behavior in all technology nodes with their expected leakage savings, as analyzed above.



(a) Natural

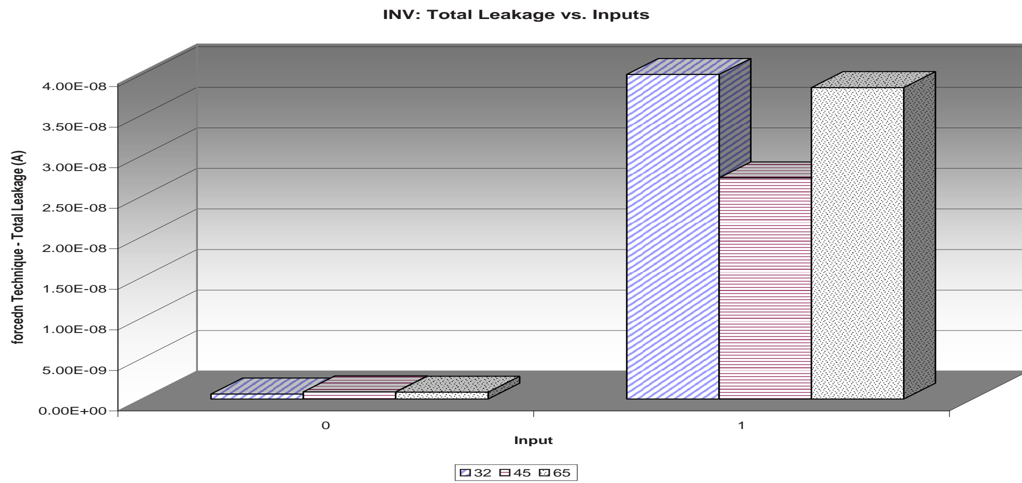


(b) High V_T PMOS

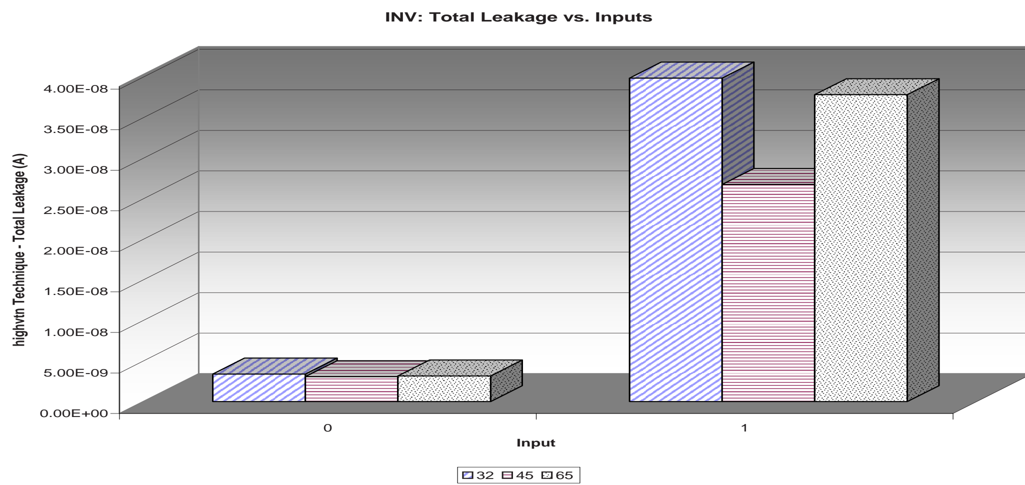


(c) Forced PMOS stack

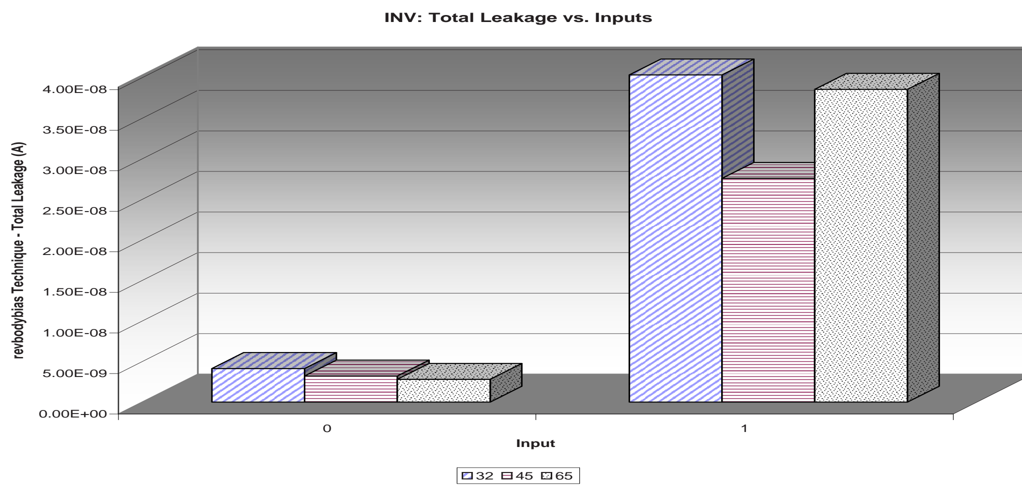
Figure 4.8: INV leakage vs. input for various leakage reduction techniques @ 25 °C



(a) Forced NMOS stack



(b) High V_T NMOS



(c) Reverse Body Bias

Figure 4.9: INV leakage vs. input for various leakage reduction techniques @ 25 °C

Figure 4.10 shows the average total leakage experienced by the inverter for a equiprobable input. Here again, the leakage in the 45nm node is reduced, due to the slightly higher threshold voltage in the model. The savings achieved relative to the natural form of the inverter are shown in Figure 4.11, indicating that the best method to reduce leakage current is to use high V_T NMOS and PMOS. It is noteworthy that the leakage savings is more when a high V_T PMOS is used, rather than a high V_T NMOS for equiprobable inputs. This is attributed to the size of the PMOS, which is about 2.7 times the NMOS. Even though the subthreshold leakage per length of NMOS is higher than PMOS the relative size of the PMOS to NMOS to achieve equal rise and fall time makes the PMOS the more leaky one in a logic gate.

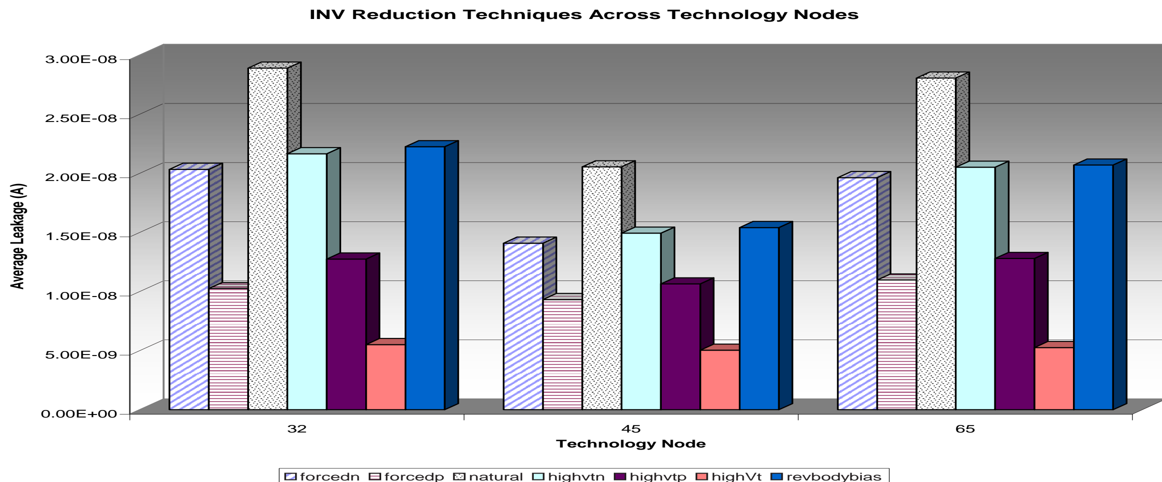


Figure 4.10: INV average total leakage @ 25 °C

4.4.4 Effect of Temperature

Figure 4.12 shows the percentage leakage savings for the different techniques relative to the natural form of the inverter at 90 °C. Compared with Figure 4.11, the effectiveness of these reduction techniques is revealed with respect to temperature. As explained earlier subthreshold leakage is exponentially related to temperature whereas the gate leakage is not tied with change in temperature. From the graph, it is evident that the techniques effecting the pull-down network become more effective at higher temperatures. Also, the

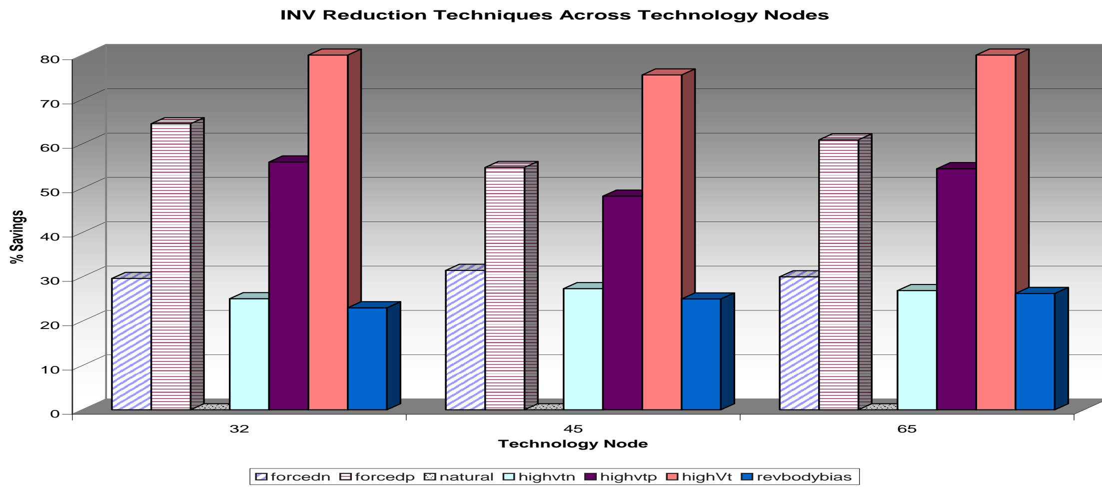


Figure 4.11: INV % leakage savings @ 25 °C

techniques effecting the pull-up network are more effective at lower temperature. This is due to the the decrease in the portion of the PMOS leakage with respect to the total leakage at higher temperatures. The doping levels of PMOS and NMOS directly effecting the change in threshold voltage with respect to temperature.

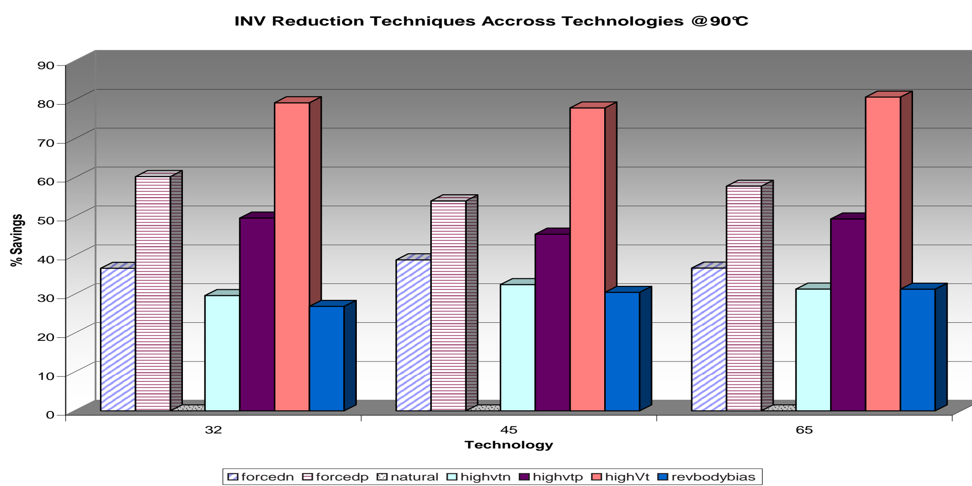


Figure 4.12: INV % leakage savings @ 90 °C

4.5 Two Input NAND Gate

4.5.1 DC Characteristics

Here, a two input NAND gate is simulated, sized relative to the inverter, as explained in Section 3.2. Figure 4.13 shows the DC simulation results for the 65nm technology node and the noise margins low and high can be found in Figure 4.14(a,b). When the graph is compared with that of the inverter, in Figure 4.4, very similar results are observed. Therefore, the same argument as that for the inverter can be used to explain the graph.

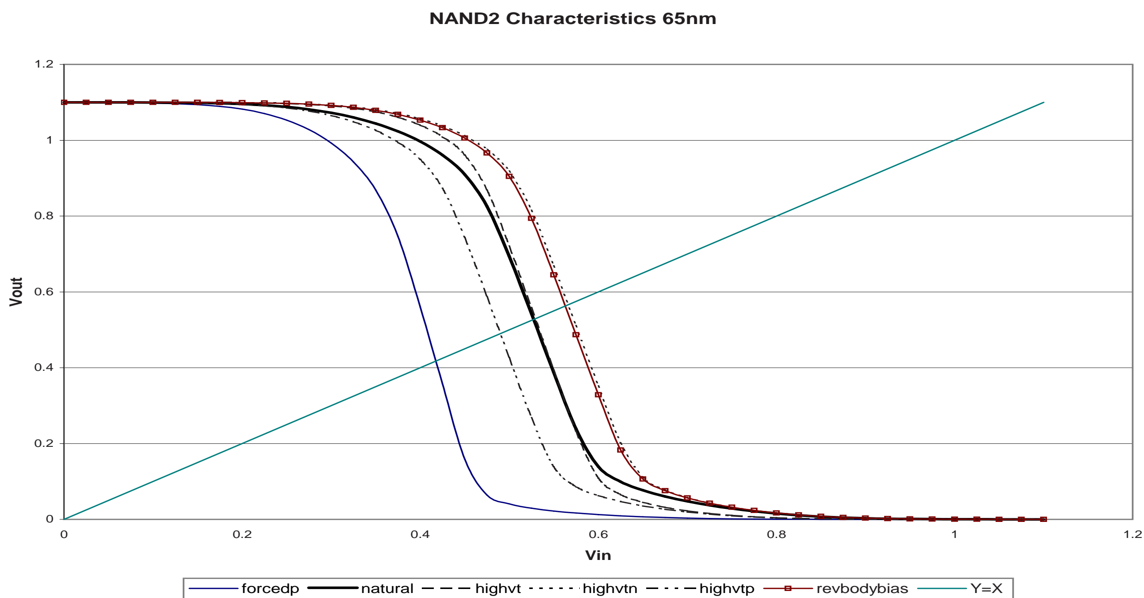
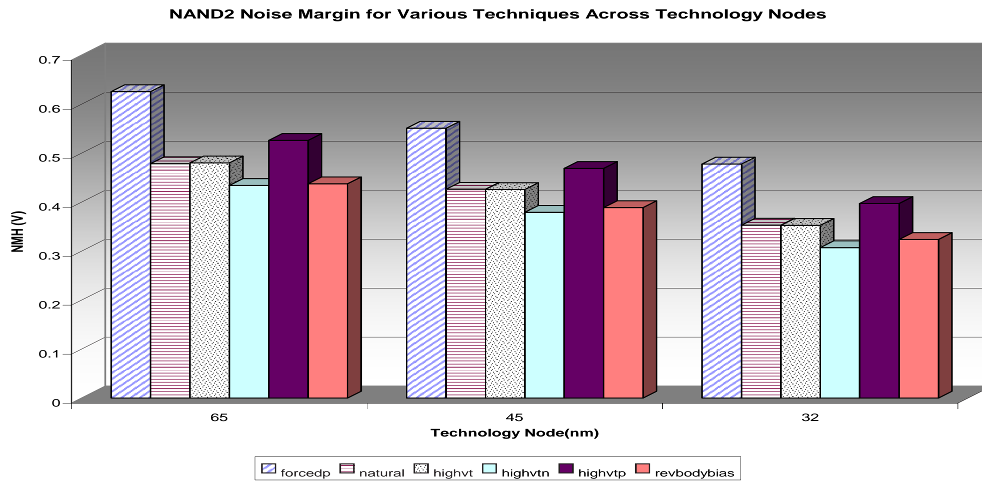
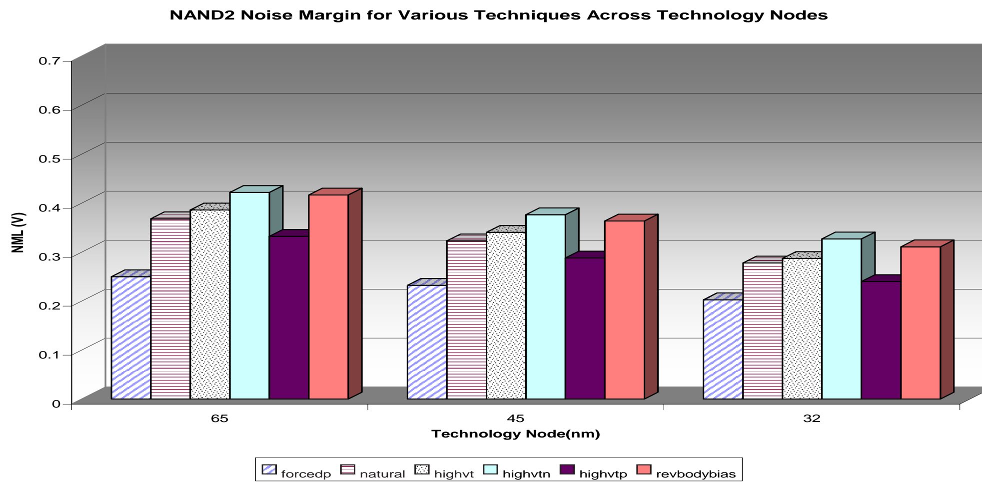


Figure 4.13: NAND2 DC characteristics 65nm @ 25 °C

Figure 4.15 portrays the switching voltage for each technique at the different technology nodes. The change in the switching voltage is proportional to the change in the noise margins for the particular technique and technology node. This change appears to be equal for each technology node. The other shift in the total reduction of the switching voltage from one technology node to the next is due to the reduction of the power supply voltage.



(a) NMH



(b) NML

Figure 4.14: NAND2 noise margins @ 25 °C

4.5.2 Change in Rise and Fall Times

Table 4.5 shows the rise time and fall time of the two-input NAND gate for each applied technique for all three technology nodes. By looking at the percentage change in rise and fall times of each technique compared to the original NAND gate the technique that has the most or the least impact on performance of the gate can be deduced.

As expected, the results shown in the table are in line with what was observed for the

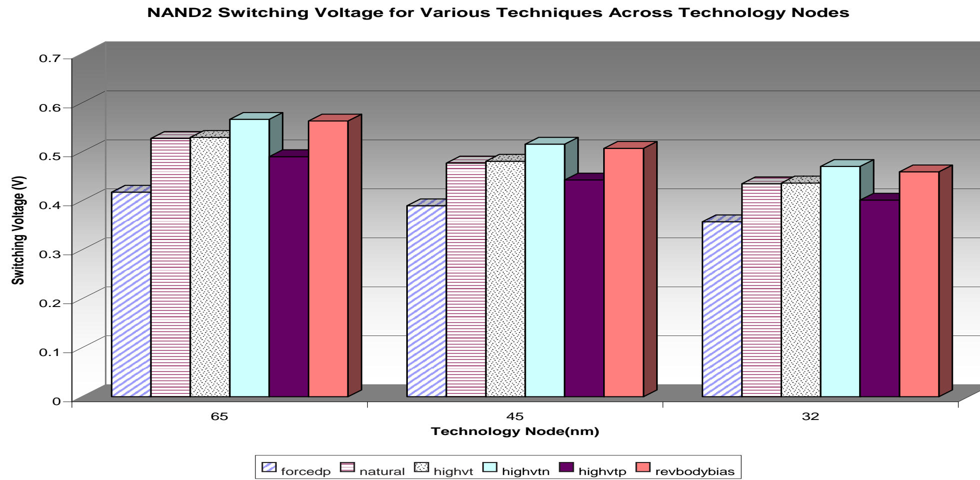


Figure 4.15: NAND2 switching voltage across technology nodes @ 25 °C

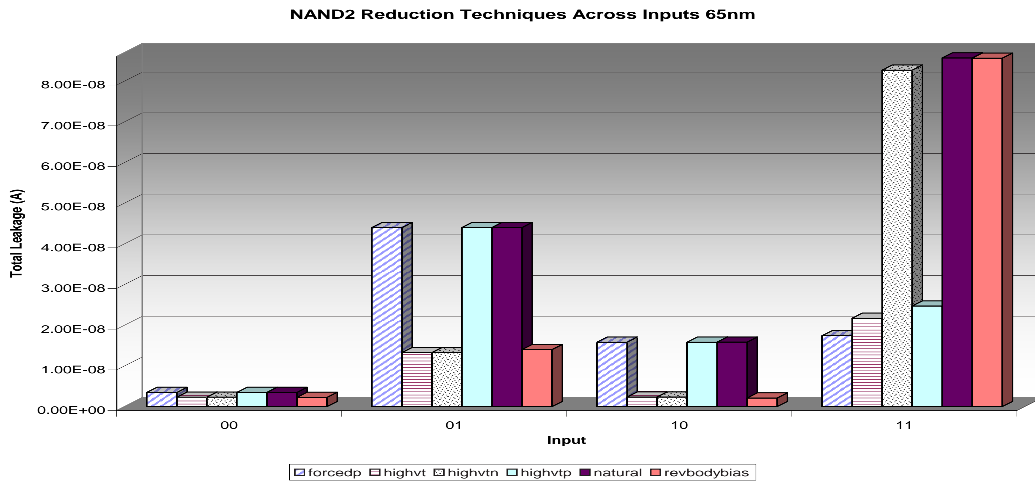
inverter. Meaning that reducing leakage in the pull-up network will impact the rise time, and reducing leakage in the pull-down network will impact the fall time. The most impact is observed from stack forcing as it reduces the drive of the transistor. The other thing to note is the performance gain on the opposite network is more sensible than the case of the inverter. This can be due to the fact that there are two more transistors leaking in the NAND gate therefore when the leakage is reduced the pull-down (pull-up) network has to fight less leakage from the extra NMOS (PMOS) in order to discharge the output node.

Table 4.5: NAND2 rise and fall times @ 25 °C

Tech(nm)	Method	tRise(ps)	%change	tFall(ps)	%change
65	natural	15.9	0.0	10.6	0.0
	forcedp	45.6	187.6	7.7	-27.2
	highvt	17.5	10.3	12.1	13.9
	highvtn	15.6	-1.8	12.1	13.7
	highvtp	17.9	12.7	10.6	-0.3
	revbodybias	15.4	-2.7	11.8	11.0
45	natural	16.6	0.0	10.2	0.0
	forcedp	49.7	199.8	7.1	-30.4
	highvt	19.1	15.2	12.1	18.6
	highvtn	16.3	-2.0	12.1	18.7
	highvtp	19.5	17.7	10.2	-0.2
	revbodybias	16.2	-2.6	11.2	9.9
32	natural	15.9	0.0	9.9	0.0
	forcedp	49.3	210.6	7.0	-29.3
	highvt	19.0	19.5	12.3	25.1
	highvtn	15.5	-2.2	12.3	25.1
	highvtp	19.4	22.0	9.8	-0.1
	revbodybias	15.4	-2.8	10.9	10.4

4.5.3 Total Leakage versus Inputs

Figure 4.16 depicts the two-input NAND gate's total leakage for all the inputs for 65nm technology node. It can be observed that the inputs 00, 10, 01 and 11 are sorted in the order of lowest leakage to highest.



(a) 65nm

Figure 4.16: NAND2 leakage vs. inputs @ 25 °C

In Figure 4.17 the gate and subthreshold leakages are shown for the different inputs. An input of 00 has the least leakage, since N1 experiences a negative V_{gs}, body effect, and a reduced V_{ds}. Hence, the NAND gate exhibits very little subthreshold leakage, and the gate leakage through PMOSs is not significant. Given an input of 11 the most leakage occurs, since two PMOSs contribute to subthreshold leakage and all four transistors contribute to the gate leakage. Input of 10 results in less leakage than input 01, due to the intermediate node voltage which is approximately a V_T drop from the supply voltage.

As explained above it can be seen on Figure 4.16 that the main contributors to the leakage are the pull-down subthreshold for inputs of 10, 01, and 00, since only using high V_T NMOS or RBB techniques are the effective techniques for these inputs. Also, for input 11, the techniques that effect the pull-up network are effective, since the PMOSs are the main sources of the subthreshold current.

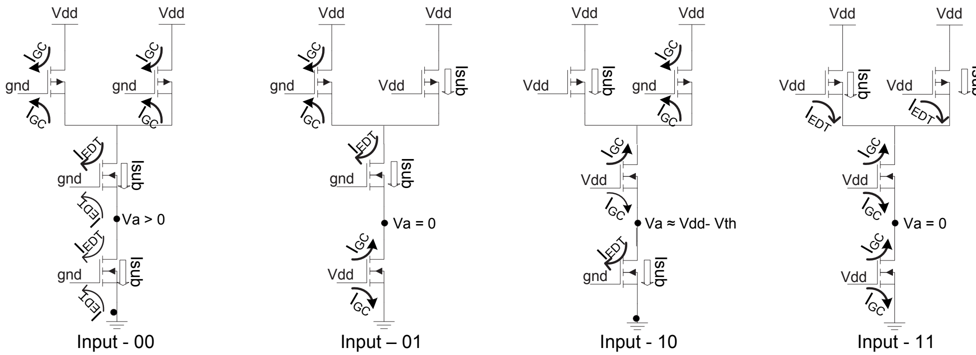


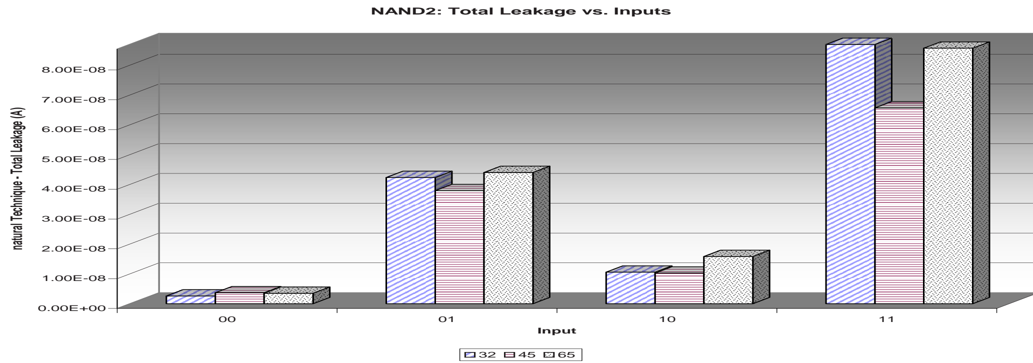
Figure 4.17: 2-input NAND gate leakages in steady state

Figure 4.18 (a) shows the two input NAND gate's total leakage in its natural form across technologies. The figure shows a leakage reduction in the 45nm technology node for inputs 11 and 01. This behavior for the inputs of 11 and 01 can be explained by the fact that in the PTM, the threshold voltage of the NMOS and PMOS device is a little higher at 45nm than that of the 65nm and 32nm technology nodes as explained in Section 3.3.1.

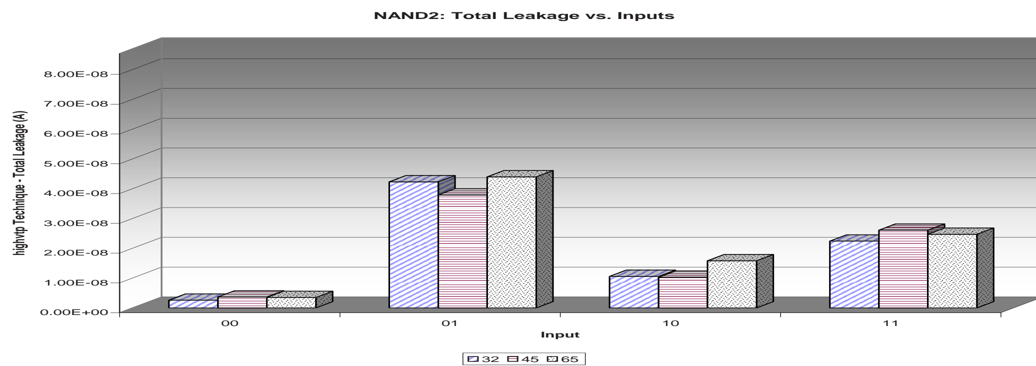
Figure 4.18 (b) shows the two input NAND gate's total leakage when high V_T PMOS is used in the gate's structure. The technique's effectiveness in leakage reduction seems to be consistent across all the technology nodes.

Figure 4.18 (c) displays the two input NAND gate's total leakage, when a forced PMOS stack is applied. This technique effects the leakage mainly for the 11 input and the leakage reduction is to be consistent across all the technology nodes.

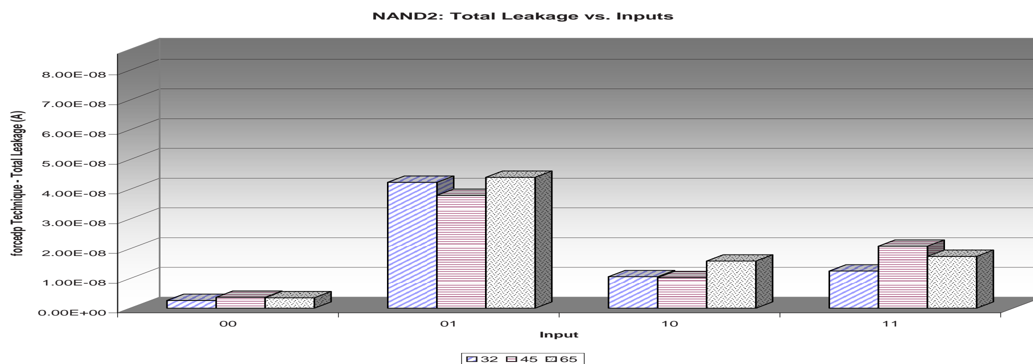
Figure 4.19 (a,b,c) reflect the two input NAND gate's total leakage, when using high V_T PMOS and NMOS, a high V_T NMOS and RBB technique in the NAND gate's structure. They all exhibit the same behavior in all the technology nodes with the previously analyzed, expected leakage savings.



(a) Natural



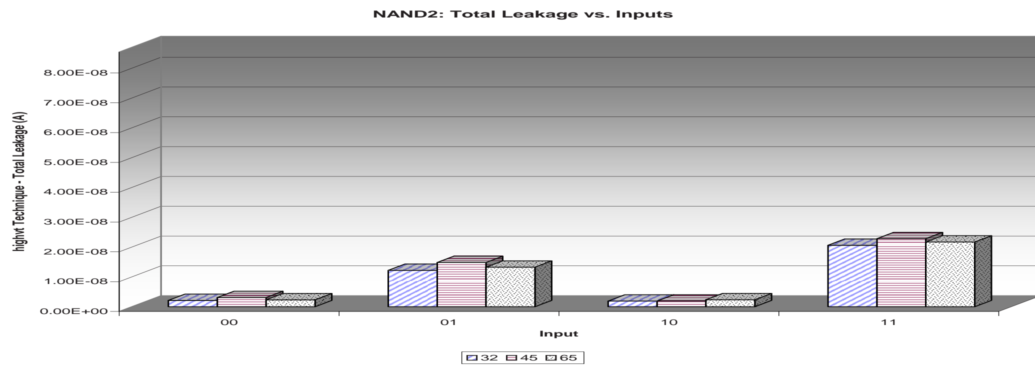
(b) High V_T PMOS



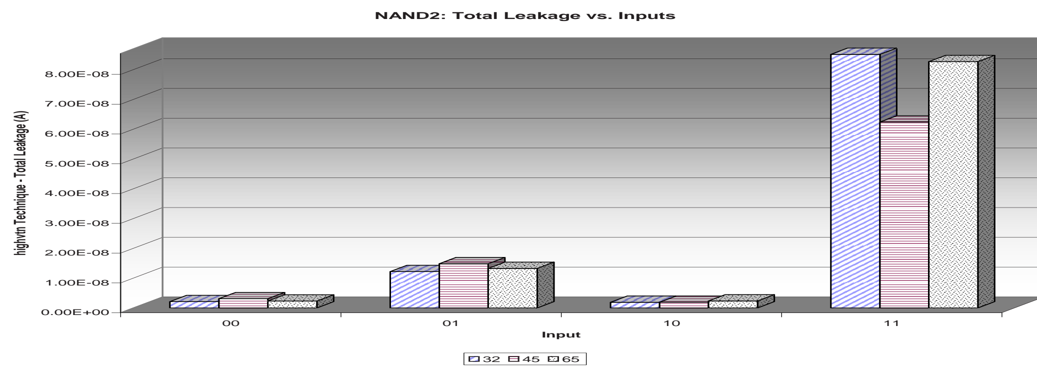
(c) Forced PMOS stack

Figure 4.18: NAND2 leakage vs. input for various leakage reduction techniques @ 25 °C

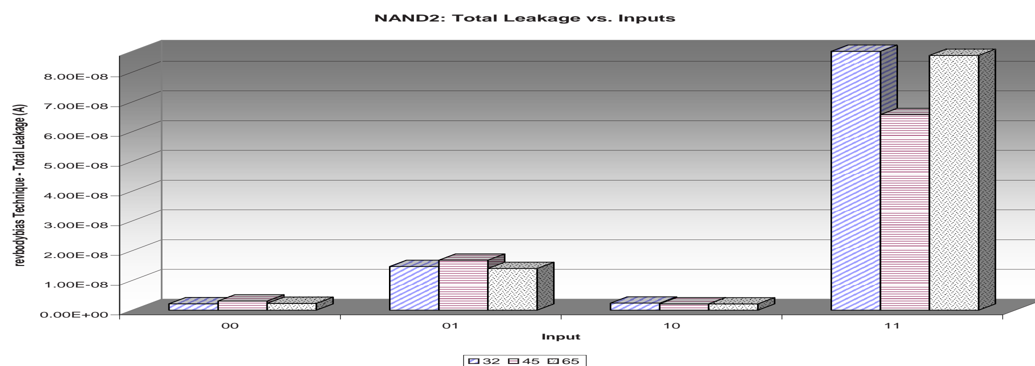
Figure 4.20 shows the average total leakage experienced by the two input NAND for an equiprobable input. Here again, a reduced leakage in the 45nm node is seen due to



(a) High V_T CMOS



(b) High V_T NMOS



(c) Reverse Body Bias

Figure 4.19: NAND2 leakage vs. input for various leakage reduction techniques @ 25 °C

the slightly higher threshold voltage in the model. The savings achieved relative to the natural form of the two input NAND is shown in Figure 4.21. It is evident that the best

method to reduce leakage current is to use high V_T NMOS and PMOS. It is interesting that there are more leakage savings when a high V_T PMOS is used, than a high V_T NMOS for equiprobable inputs. This is explained by the fact that the two PMOSs have more leakage than the NMOSs due to their size.

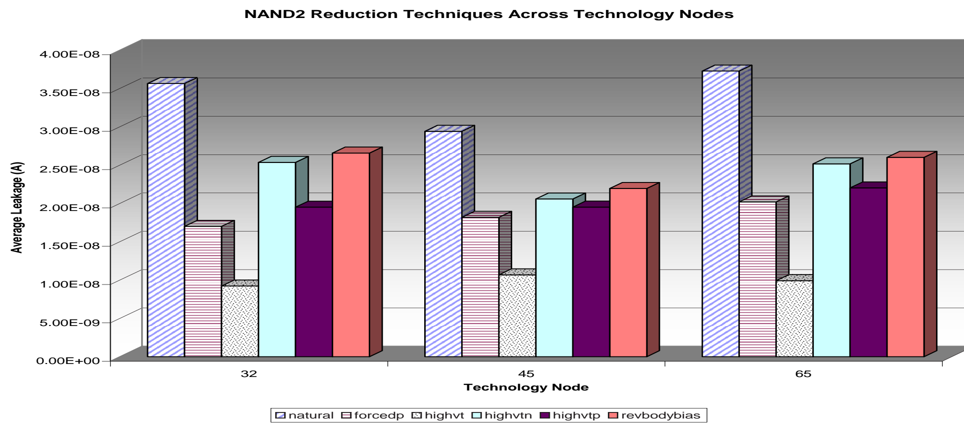


Figure 4.20: NAND2 average total leakage @ 25 °C

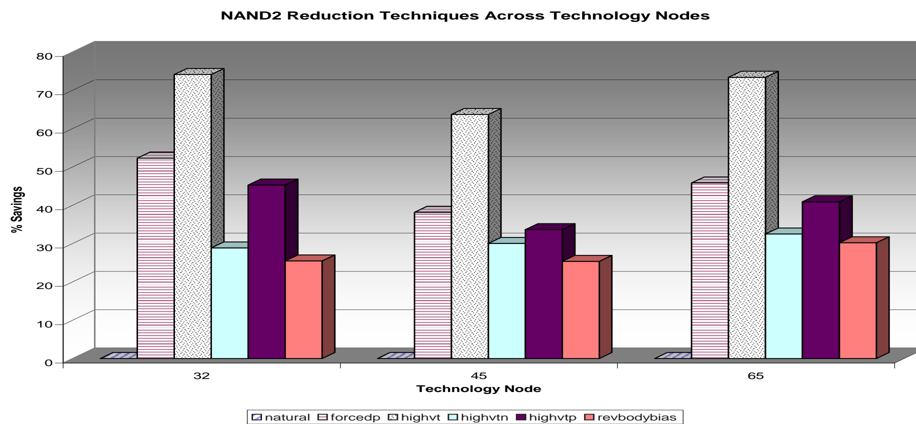


Figure 4.21: NAND2 % leakage savings @ 25 °C

4.5.4 Effect of Temperature

Figure 4.22 shows the percentage leakage savings using different techniques relative, to the natural form of the NAND gate at 90 °C. When this figure is compared with Figure 4.21,

effectiveness of these reduction techniques are observed with respect to temperature. As explained previously, subthreshold leakage is exponentially related to temperature, whereas the gate leakage is not tied with change in temperature. From the graph it is observed that the techniques effecting the pull-down network become as effective as the pull-up techniques. This is due to the the decrease in the portion of the PMOS leakage with respect to the total leakage at higher temperatures.

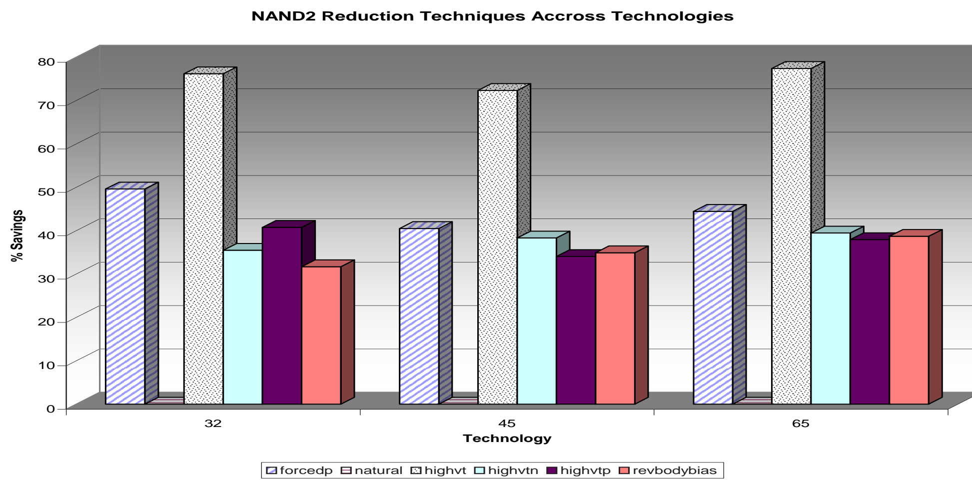


Figure 4.22: NAND2 % leakage savings @ 90 °C

4.6 Two Input NOR Gate

4.6.1 DC Characteristics

Here a two input NOR gate is simulated which is sized relative to the inverter in Section 3.2. Figure 4.23(a) signifies the DC simulation results for the 65nm technology node and the noise margins can be found in Figure 4.24(a,b). By comparing the graph with the one of the inverter in Figure 4.4 it is obivous that the results are very similar. Therefore, the same argument as the inverter can be used to explain this graph.

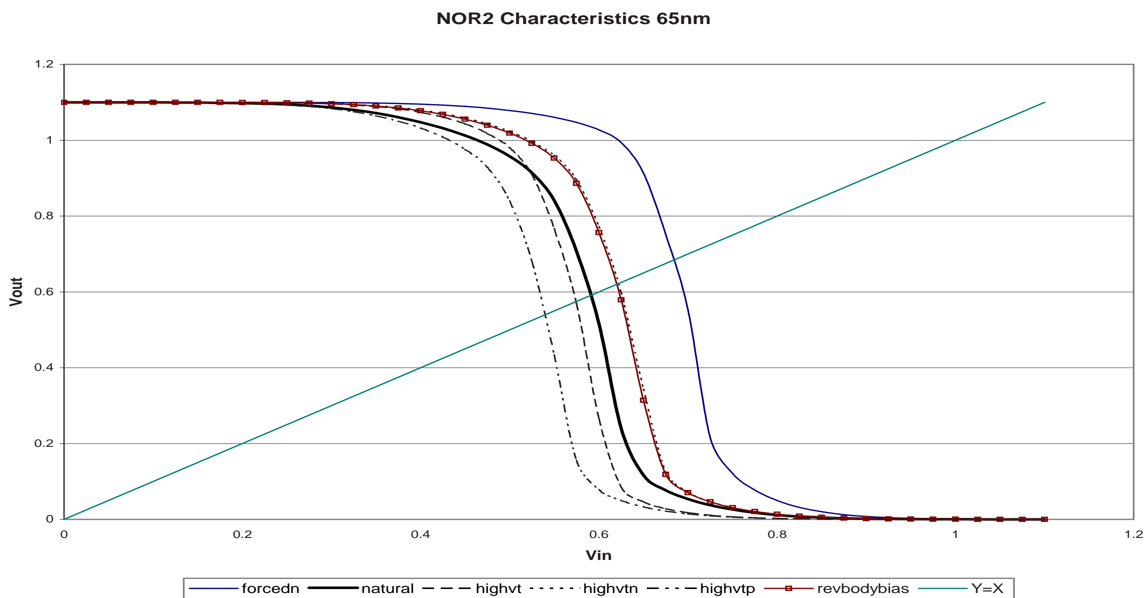
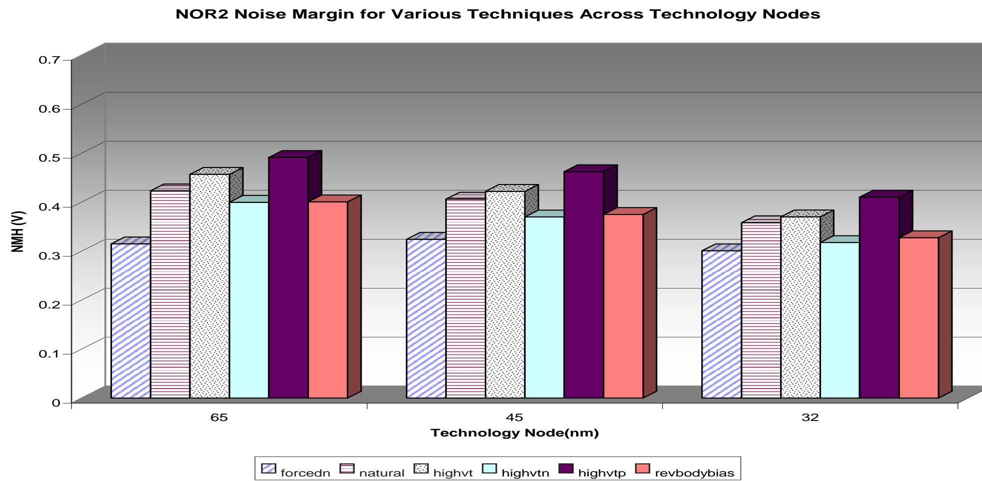
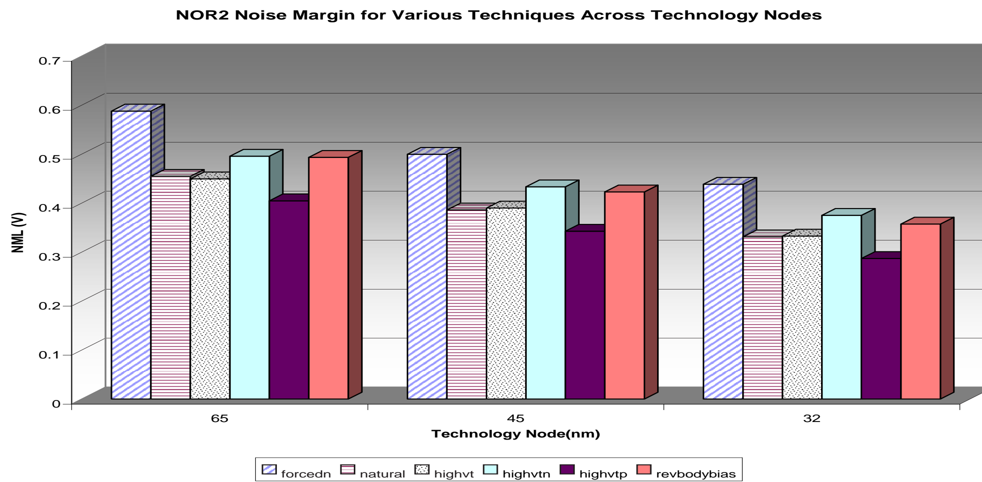


Figure 4.23: NOR2 DC characteristics 65nm @ 25 °C

Figure 4.25 reflects the switching voltage for each technique at different technology nodes. The change in the switching voltage is proportional to the change in the noise margins for the particular technique and technology node. This change also seems to be equal in each different technology node. The other shift in total reduction of the switching voltage from one technology node to another, is due to the reduction of power supply voltage.



(a) NMH



(b) NML

Figure 4.24: NOR2 noise margins @ 25 °C

4.6.2 Change in Rise and Fall Times

Table 4.6 lists the rise time and fall time of the two-input NOR gate for each applied technique for all three technology nodes. By looking at the percentage change in rise and fall times of each technique compared to the original NOR gate the technique that has the most or the least impact on performance of the gate can be deduced.

As expected, the results shown in the table are in line with what is observed from the

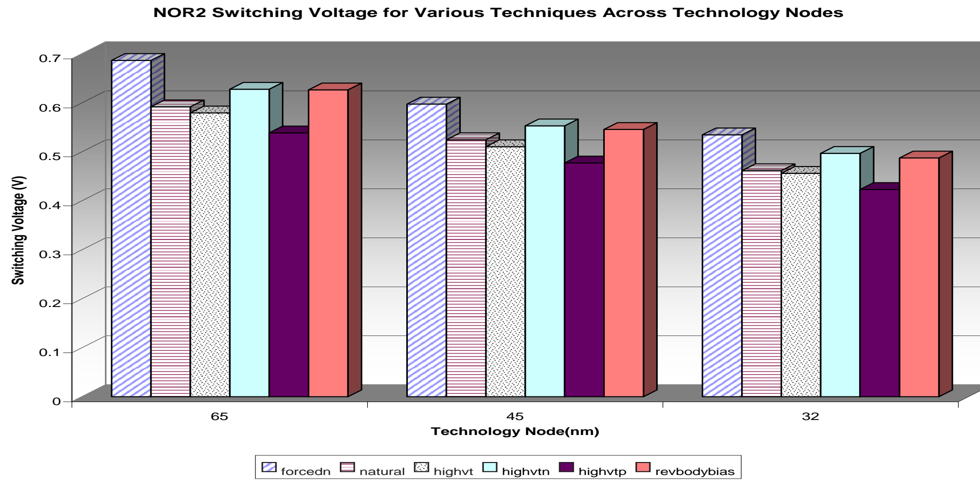
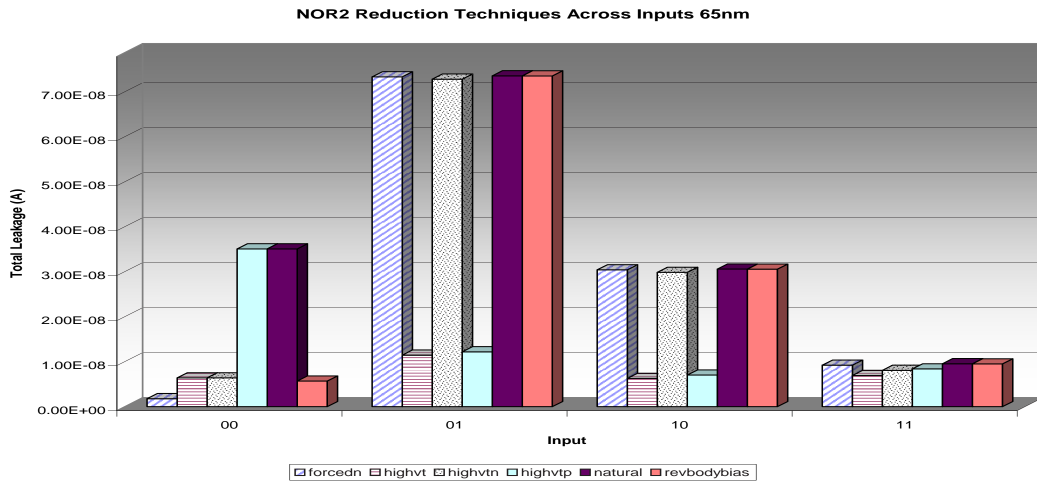


Figure 4.25: NOR2 switching voltage across technologies @ 25 °C

inverter. Meaning that reducing leakage in the pull-up network will impact the rise time and reducing leakage in the pull-down network will impact the fall time. The most impact is observed from stack forcing, as it reduces the drive of the transistor. The application of RBB, high V_T NMOS, and high V_T PMOS transistor seem to have less effect on the rise and fall times respectively.

Table 4.6: NOR2 rise and fall times @ 25 °C

Tech(nm)	**Method**	**tRise(ps)**	**%change**	**tFall(ps)**	**%change**
65	natural	15.5	0.0	11.1	0.0
	forcedn	15.2	-2.3	35.4	217.6
	highvt	17.2	10.6	12.2	9.7
	highvtn	15.5	0.0	12.2	9.5
	highvtp	17.2	10.8	11.1	0.0
	revbodybias	15.5	-0.3	12.1	8.7
45	natural	16.9	0.0	10.3	0.0
	forcedn	16.6	-1.8	34.7	236.2
	highvt	19.3	13.7	12.3	18.8
	highvtn	16.9	-0.1	12.2	18.6
	highvtp	19.3	13.8	10.3	0.0
	revbodybias	16.9	-0.3	11.9	15.5
32	natural	16.2	0.0	9.8	0.0
	forcedn	16.5	1.7	34.4	252.3
	highvt	20.1	23.8	11.8	20.7
	highvtn	16.7	3.0	11.8	20.4
	highvtp	20.1	24.0	9.8	0.2
	revbodybias	16.7	2.8	10.7	9.6



(a) 65nm

Figure 4.26: NOR2 leakage vs. inputs @ 25 °C

4.6.3 Total Leakage versus Inputs

The two-input NOR gate total leakage for all the different inputs for 65nm technology node, is illustrated in Figure 4.26. It is evident that the inputs 01, 00, 10, and 11 are sorted in the order of the highest leakage to the lowest.

In Figure 4.27, the gate and subthreshold leakages are shown for different input. An input of 11 has the least leakage since P2 experiences negative V_{gs}, body effect, and reduced V_{ds}, indicating very little subthreshold leakage. Therefore, the main source of leakage is gate leakage through the PMOS and NMOS, which makes this the lowest leakage state. Input of 01 experiences the most leakage, since one PMOS experiences maximum subthreshold leakage and due to its sizing it is the most amount of leakage. In input 00, only the two NMOS experience maximum subthreshold leakage, but their size is about a fifth of the PMOS the leakage is less than the 01 input. Input of 10 has more leakage than case of 11 since it a PMOS conducts all the leakage from the top PMOS.

It is evident in Figure 4.26 that the principal contributors to the leakage are the pull-up transistors for inputs of 01 and 11. Therefore using high V_T PMOS helps reduce the leakage for these cases and the pull-down network techniques are only effective for 00 input case.

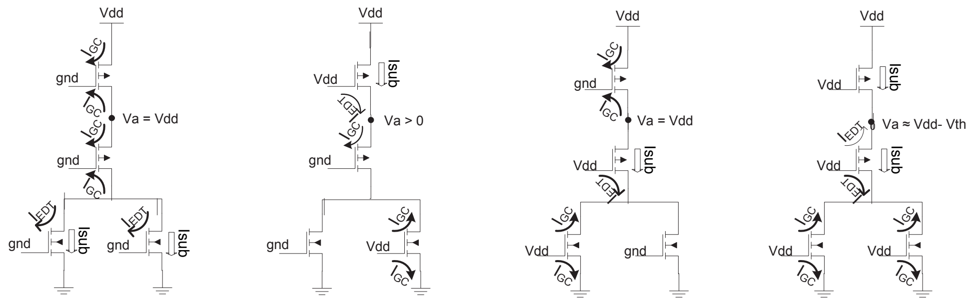


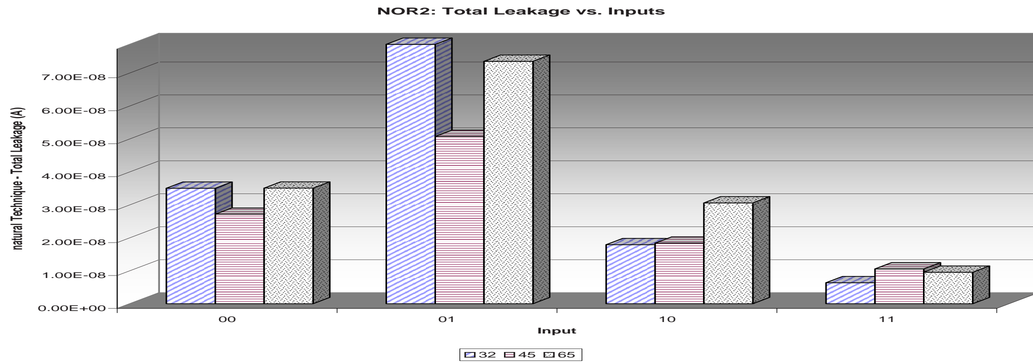
Figure 4.27: NOR2 inputs and leakages in steady state

Figure 4.28 (a) portrays the two input NOR gate's total leakage in its natural form across technologies. The figure shows a leakage reduction in the 45nm technology node for both inputs of 00, 01. This behavior can be explained by the fact that in the PTM, the threshold voltage of the NMOS and PMOS device is a little higher than the 65nm and 32nm technology models as explained in Section 3.3.1.

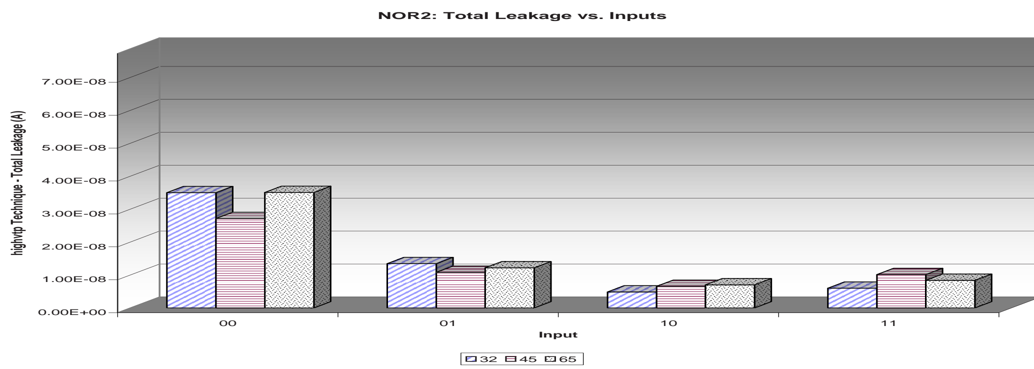
Figure 4.28 (b) shows the two input NOR gate's total leakage, when high V_T PMOS is used in its structure. The effectiveness of this technique in leakage reduction seems to be consistent across all the technology nodes.

Figure 4.28 (c) signifies the two input NOR gate's total leakage, when a forced NMOS stack is applied. This technique effects the leakage mainly for the 00 input. The effectiveness of this technique in leakage reduction seems to be consistent across all the technology nodes.

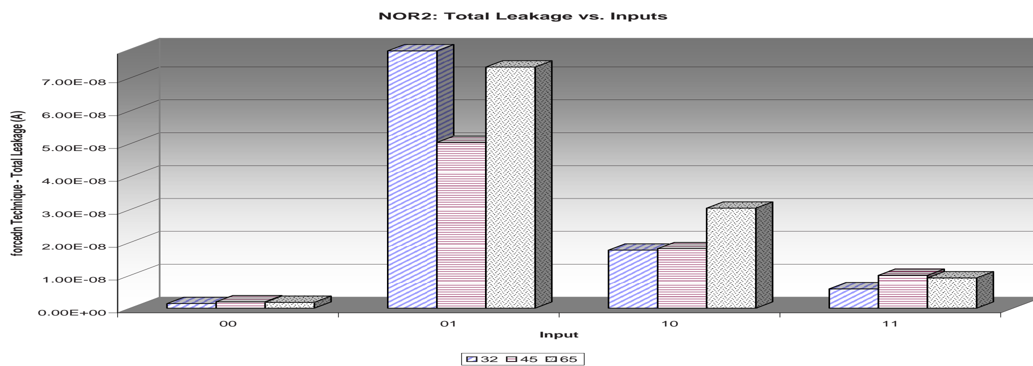
Figure 4.29 (a,b,c) show the two input NOR gate's total leakage, when using both high V_T PMOS and NMOS, high V_T NMOS and RBB technique in the NOR gates structure. They all exhibit the same behavior in all the technology nodes according to their expected leakage savings. In the case of RBB technique, for the input of 00 the leakage reduction seems to decrease as technology scales. This can be attributed to the fact that the amount of RBB applied might have not been enough since the optimum value for RBB could not be simulated using these models.



(a) Natural



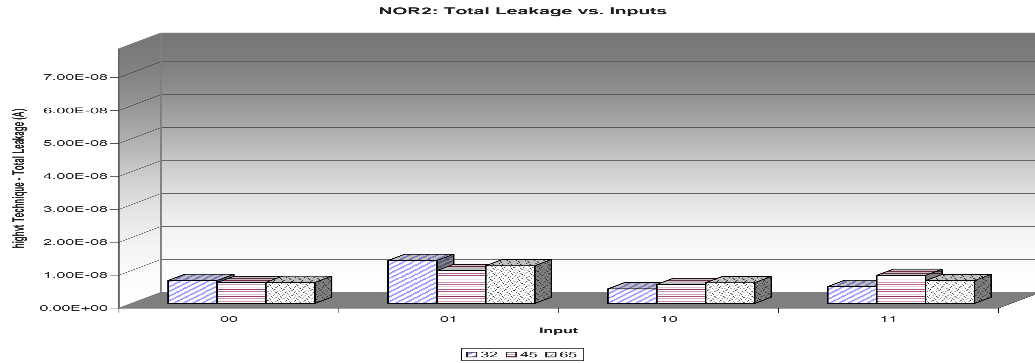
(b) High V_T PMOS



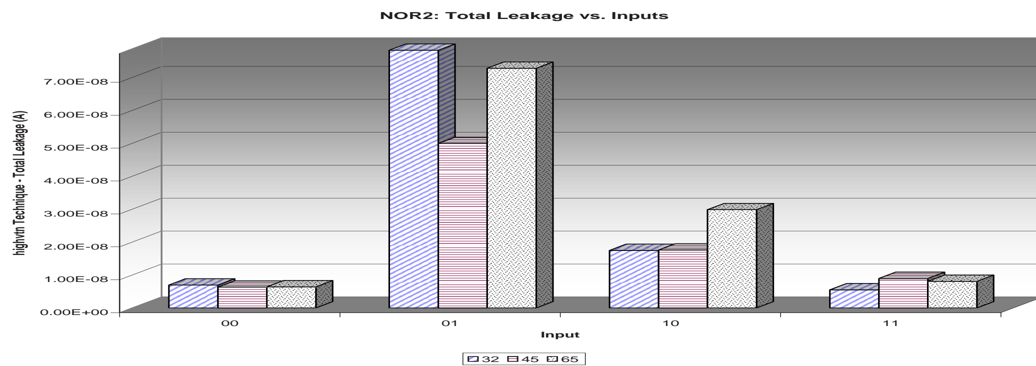
(c) Forced NMOS stack

Figure 4.28: NOR2 leakage vs. input for various leakage reduction techniques @ 25 °C

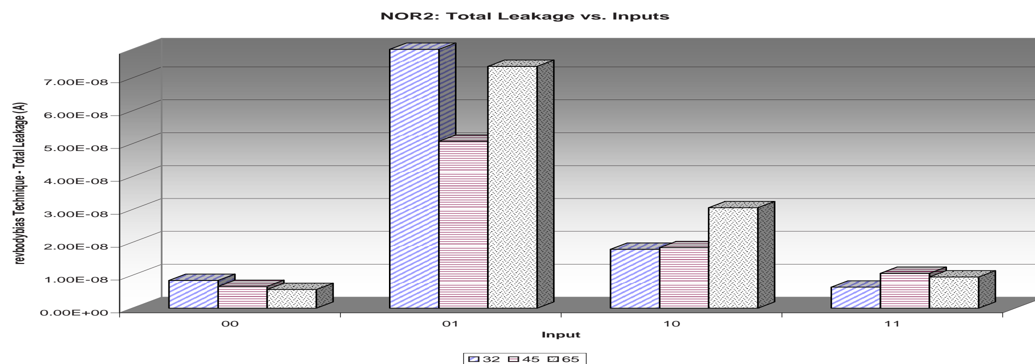
Figure 4.30 shows the average total leakage experienced by the two input NOR for an equiprobable input. Here again reduced leakage is observed in the 45nm node due to the



(a) High V_T CMOS



(b) High V_T NMOS



(c) Reverse Body Bias

Figure 4.29: NOR2 leakage vs. input for various leakage reduction techniques @ 25 °C

slightly higher threshold voltage in PTM.

The savings achieved relative to the natural form of the two input NOR2 are shown in

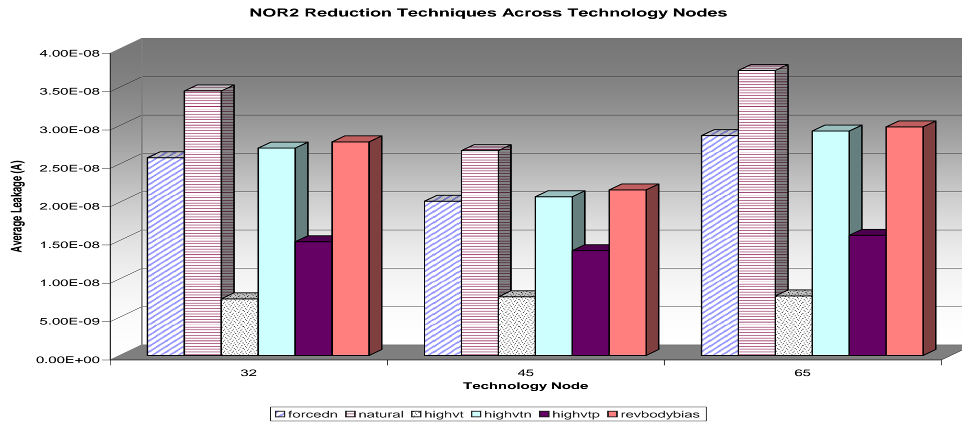


Figure 4.30: NOR2 average total leakage @ 25 °C

Figure 4.31. It is evident that the best method to reduce leakage current is to use high V_T NMOS and PMOS. It is interesting to note that there are more leakage savings when a high V_T PMOS is used, rather than a high V_T NMOS for equiprobable inputs. Due to the size of the PMOS relative to the NMOS for equal rise time and fall time the leakage through PMOS become much more significant. Since the transistors are even larger than the case of two input NAND the savings achieved using high V_T PMOS is even more.

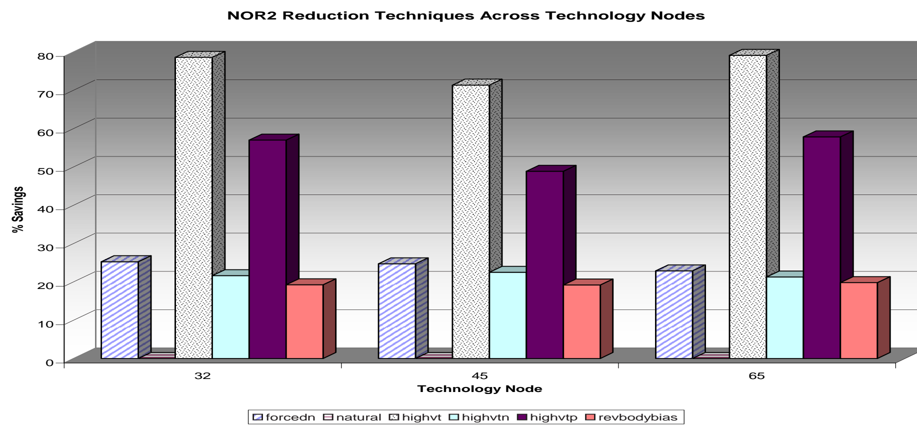


Figure 4.31: NOR2 % leakage savings @ 25 °C

4.6.4 Effect of Temperature

The percentage leakage savings using different techniques relative to the natural form of the NOR gate at 90 °C, are presented in Figure 4.32. Comparing this figure with Figure 4.31 the effectiveness of these reduction techniques are observed with respect to temperature. The figure shows that using forced stack NMOS becomes more effective in higher temperature. The reason for that is the increase in the proportion of NMOS leakage with respect to the total leakage at higher temperatures. This is due to the doping level of PMOS and NMOS which directly effect the change in threshold voltage with respect to temperature.

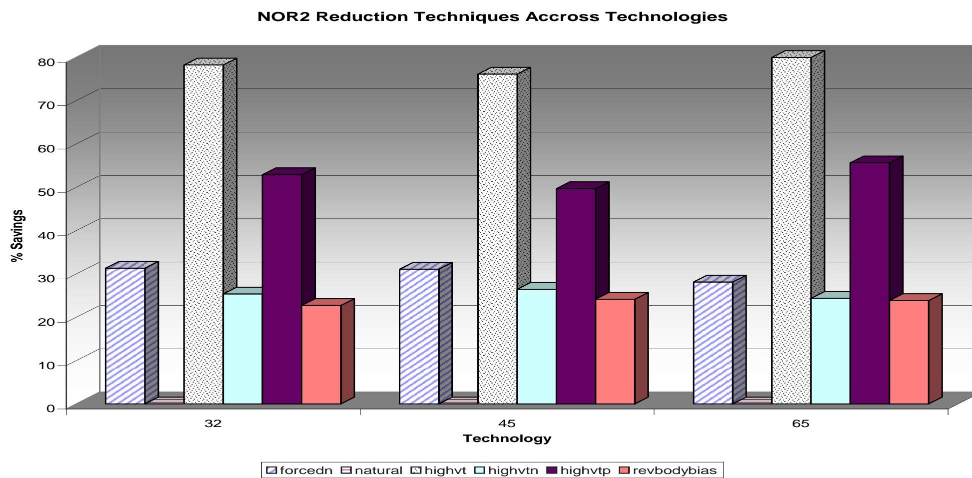


Figure 4.32: NOR2 % leakage savings @ 90 °C

4.7 Three Input NAND Gate

4.7.1 DC Characteristics

Here a three input NAND gate sized relative to the inverter in Section 3.2 is simulated. Figure 4.33 shows the DC simulation results for 65nm technology node and the noise margin low and high can be found in Figure 4.34(a,b) respectively. By comparing the graph with that of the inverter in Figure 4.4, similar results are observed. Therefore, the same argument as the inverter can be used to explain the graph.

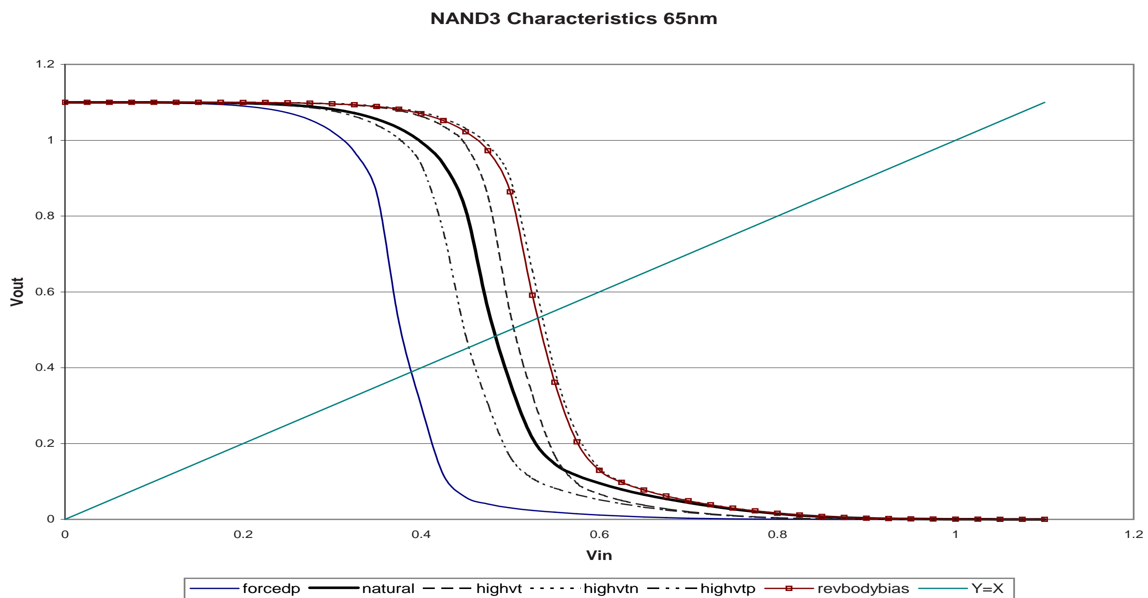
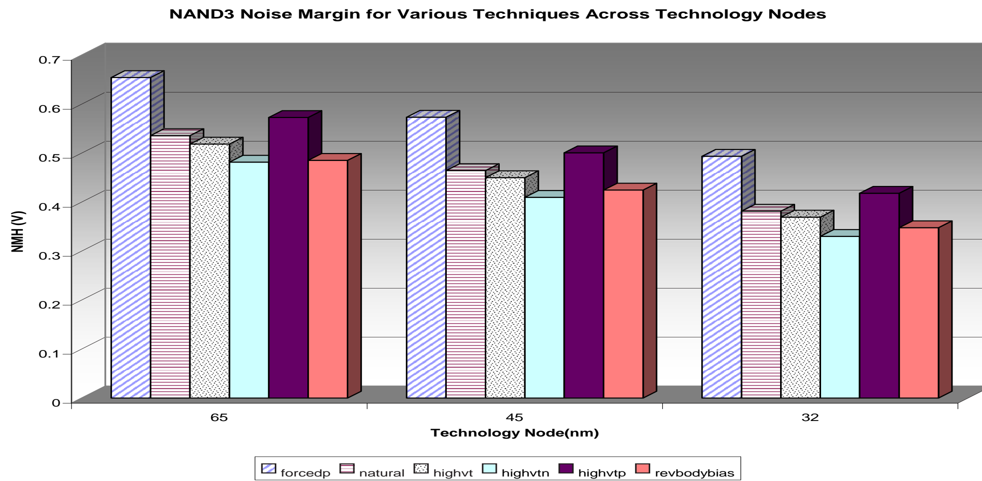
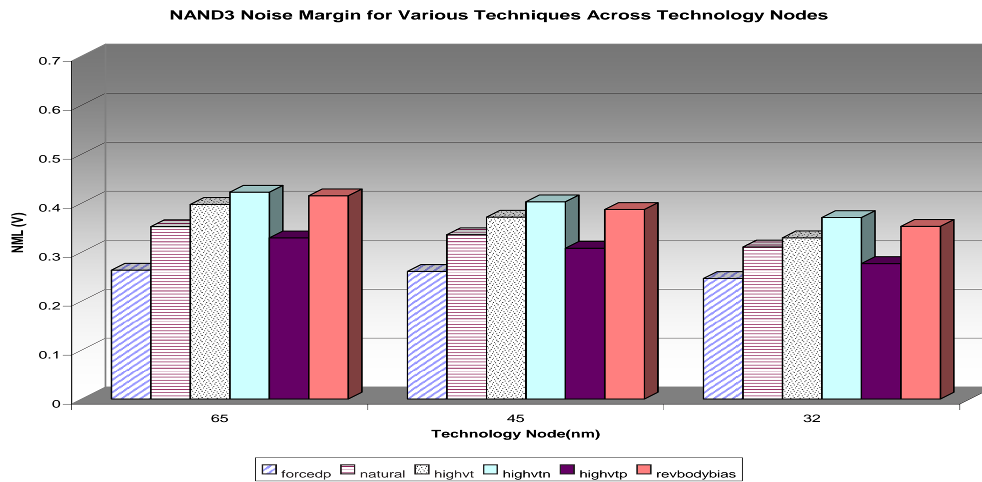


Figure 4.33: NAND3 DC characteristics 65nm @ 25 °C

Figure 4.35 shows the switching voltage for each technique at different technology nodes. The change in the switching voltage is proportional to the change in the noise margins for the particular technique and technology node. This change also seems to be equal in each technology node. The other shift in total reduction of the switching voltage from one technology node to another is due to the reduction of power supply voltage.



(a) NMH



(b) NML

Figure 4.34: NAND3 noise margins @ 25 °C

4.7.2 Change in Rise and Fall Times

In Table 4.7 the rise time and fall time of the three input NAND gate is summarized for each applied technique for all three technology nodes. By looking at the percentage change in rise and fall times of each technique compared to the original NAND gate one can deduce which technique has the most or the least impact on performance of the gate.

As expected, the results in the table are in agreement with those observed from the

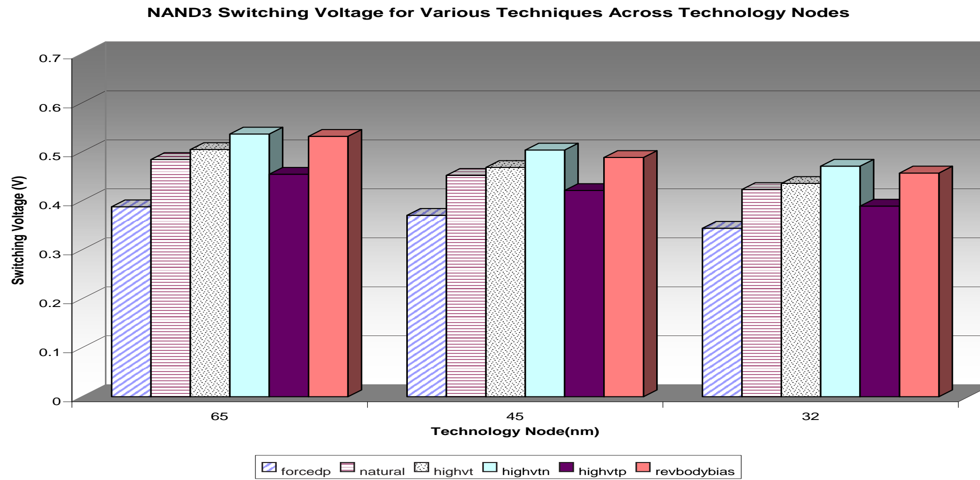
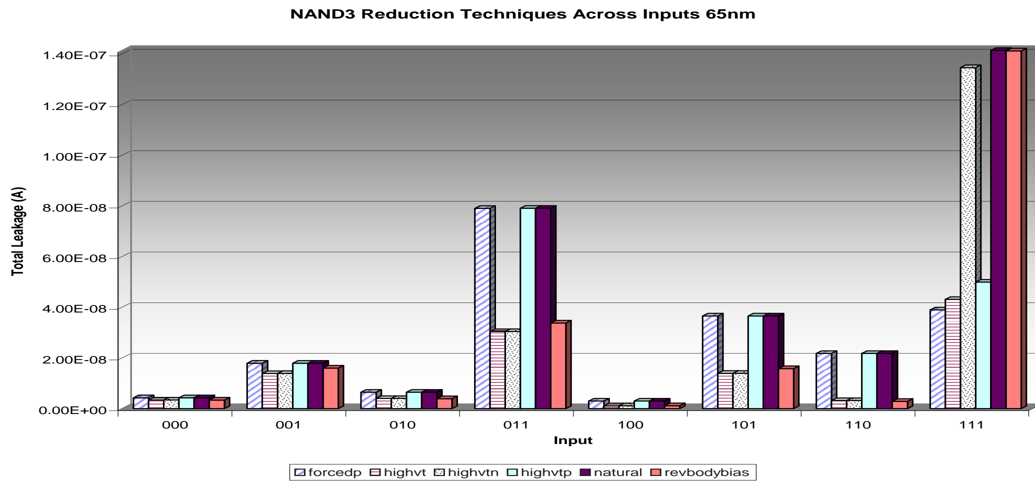


Figure 4.35: NAND3 switching voltage across technologies @ 25 °C

Table 4.7: NAND3 rise and fall times @ 25 °C

Tech(nm)	Method	tRise(ps)	%change	tFall(ps)	%change
65	natural	28.5	0.0	17.3	0.0
	forcedp	87.9	208.9	12.6	-27.0
	highvt	32.0	12.3	18.9	9.1
	highvtn	27.3	-4.1	18.9	9.2
	highvtp	33.3	17.0	17.3	-0.1
	revbodybias	27.0	-5.3	18.5	7.0
45	natural	29.4	0.0	16.6	0.0
	forcedp	94.6	221.4	12.0	-27.7
	highvt	34.3	16.5	18.8	12.8
	highvtn	28.3	-4.0	18.8	12.8
	highvtp	35.8	21.7	16.6	0.0
	revbodybias	28.1	-4.7	18.0	8.0
32	natural	27.5	0.0	16.2	0.0
	forcedp	92.4	236.2	11.5	-28.9
	highvt	33.1	20.4	19.5	20.3
	highvtn	26.4	-4.0	19.5	20.3
	highvtp	34.5	25.6	16.2	0.0
	revbodybias	26.3	-4.4	18.0	11.3

inverter and the two input NAND gate. Indicating that the reduction in the leakage in the pull-up network impacts the rise time and reducing leakage in the pull-down network impacts the fall time. The most impact is observed from stack forcing, as it reduces the drive of the transistor. The performance gain on the opposite network is more sensible than the case of the inverter and the two input NAND gate. This results from the additional NMOS transistor in the stack, which makes the NMOS transistors are very large, thus, more leakage. Therefore, a little leakage savings has a more pronounced effect on the rise time since the PMOSs are affected by less leakage current, when charging the capacitor.



(a) 65nm

Figure 4.36: NAND3 leakage vs. inputs @ 25 °C

Using high V_T PMOS technique increases the rise time more than the case of two input NAND gate. This can be due to the sizing of the transistor, which was just modeled based on the inverter. Because using three times the size of NMOS in the pull-down network, does not accurately cause equal rise and fall times. A similar effect of this sizing is observed, when using high V_T NMOS technique. The fall time is less compared to the case of two input NAND gate. Similarly, as explained above this is due to extra transistor in the pull-down network of the three input NAND gate, which makes the sizes of the transistors in the pull-down network much larger. Therefore, less performance penalty is experienced when high V_T NMOSs are used.

4.7.3 Total Leakage versus Inputs

Figure 4.36 shows the three input NAND gate's total leakage for all the different inputs for 65nm technology node. It can be observed that the inputs 100, 000, 010, 001, 110, 101, 011, and 111 are sorted in the order of lowest leakage to highest.

In figure 4.37 the gate and subthreshold leakages are shown for different input. An input of 100 has the least leakage since the middle NMOS experiences negative V_{gs}, body

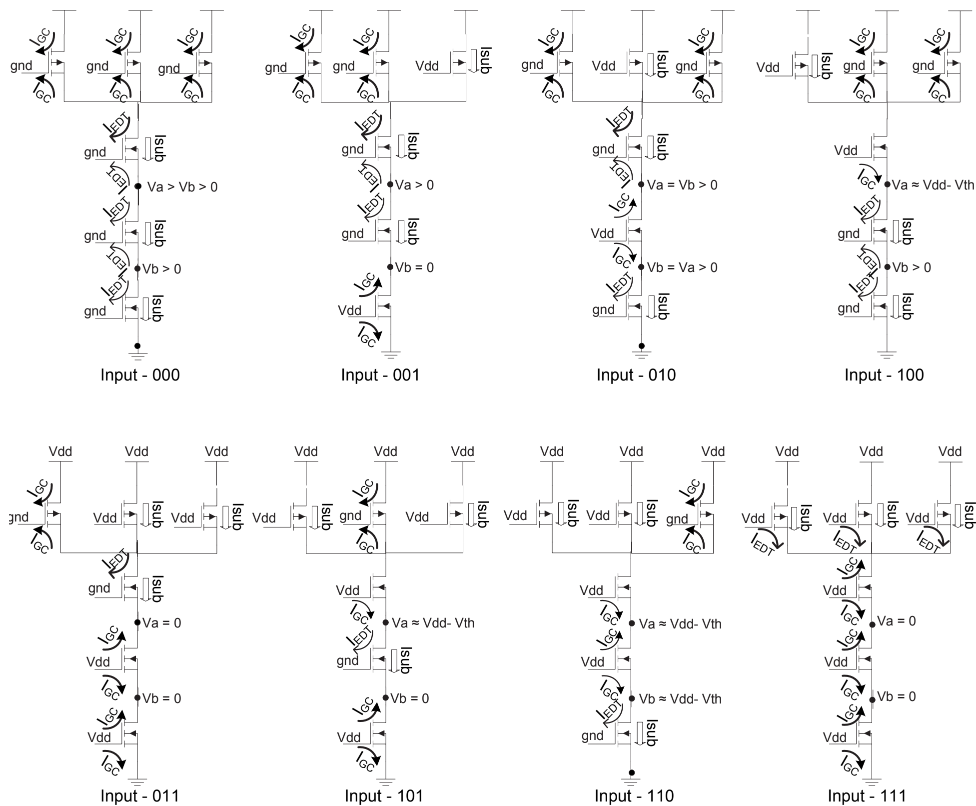


Figure 4.37: NAND3 inputs and leakages in steady state

effect, and reduced V_{ds}. Hence, very little subthreshold leakage is experienced and the gate leakage through PMOSs is not significant. Given an input of 111 the most amount of leakage is experienced since three PMOSs contribute to subthreshold leakage and gate leakage is also contributed by all six transistors.

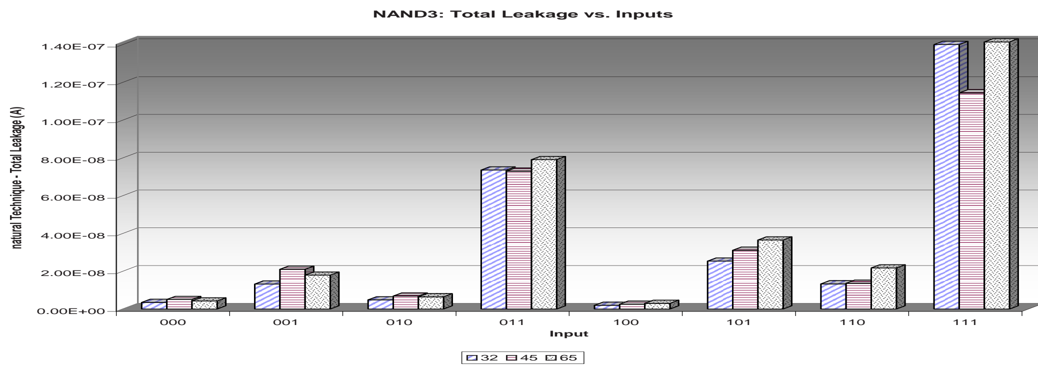
It is evident in Figure 4.36 that the primary contributors to the leakage are the pull-down transistor for all inputs except input of 111. Hence only using high V_T NMOS or RBB techniques are the effective techniques for those inputs. For input 111, the techniques that effect the pull-up network are effective, since the PMOSs are the main sources of the subthreshold current.

Figure 4.38 (a) illustrates the three input NAND gate's total leakage in its natural form across technologies. The figure displays a leakage reduction in the 45nm technology node for both inputs of 111. This behavior for input of 111 can be explained by the fact that in the PTM, the threshold voltage of the NMOS and PMOS device is a little higher than the 65nm and 32nm technology models as explained in Section 3.3.1.

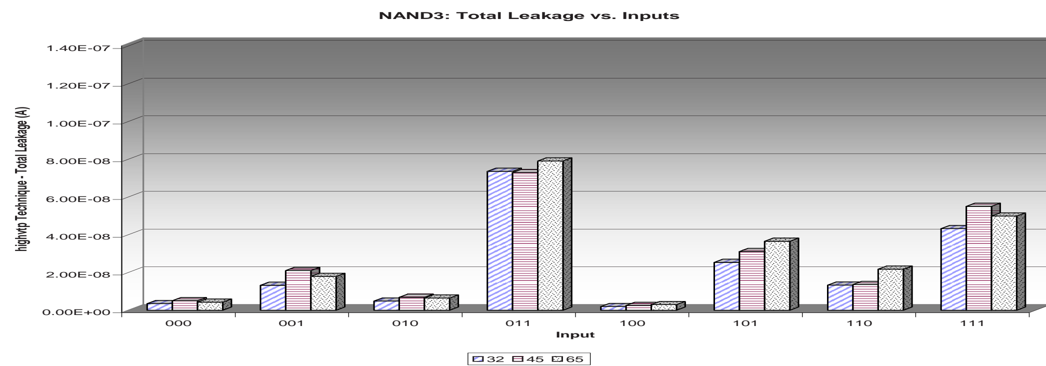
Figure 4.38 (b) shows the three input NAND gate's total leakage when high V_T PMOS is used in its structure. The effectiveness of this technique in leakage reduction seems to be consistent across all the technology nodes.

Figure 4.38 (c) shows the three input NAND gate's total leakage when a forced PMOS stack is applied. This technique effects the leakage mainly for the 111 input. The effectiveness of this technique in leakage reduction seems to be consistent across all the technology nodes.

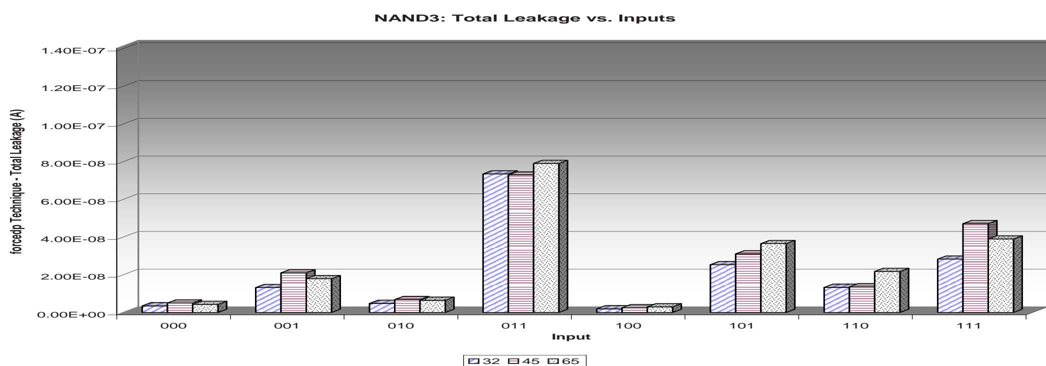
Figure 4.39 (a,b,c) show the three input NAND gate's total leakage when using high V_T PMOS and NMOS,a high V_T NMOS and RBB technique in the NAND gates structure. They all exhibit the same behavior in all the technology nodes in agreement with the expected leakage savings.



(a) Natural



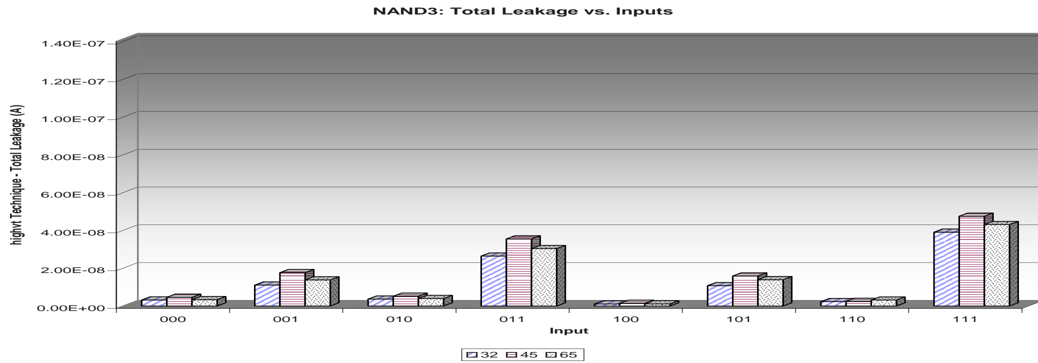
(b) High V_T PMOS



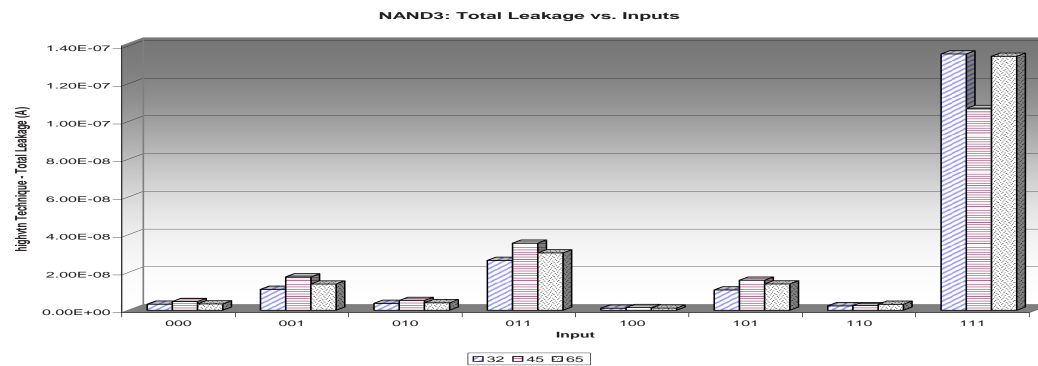
(c) Forced PMOS stack

Figure 4.38: NAND3 leakage vs. input for various leakage reduction techniques @ 25 °C

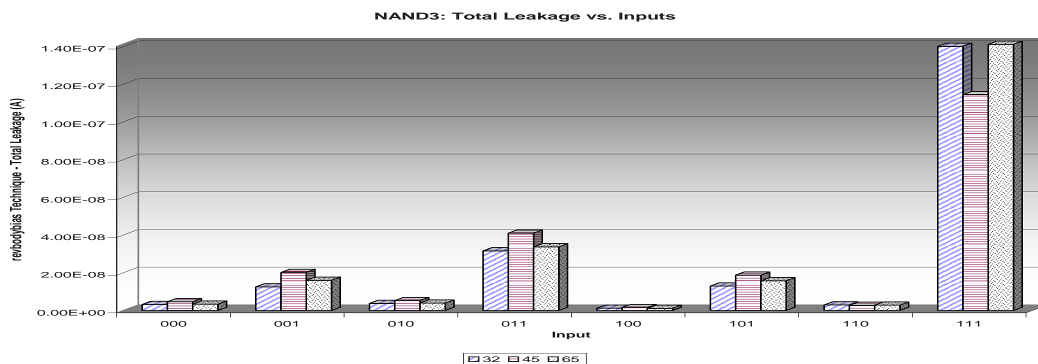
Figure 4.40 shows the average total leakage experienced by the three input NAND for an equiprobable input. Here again, a reduction in leakage at the 45nm node is observed,



(a) High V_T CMOS



(b) High V_T NMOS



(c) Reverse Body Bias

Figure 4.39: NAND3 leakage vs. input for various leakage reduction techniques @ 25 °C

and as it was explained above it is due to the slightly higher threshold voltage in the model.

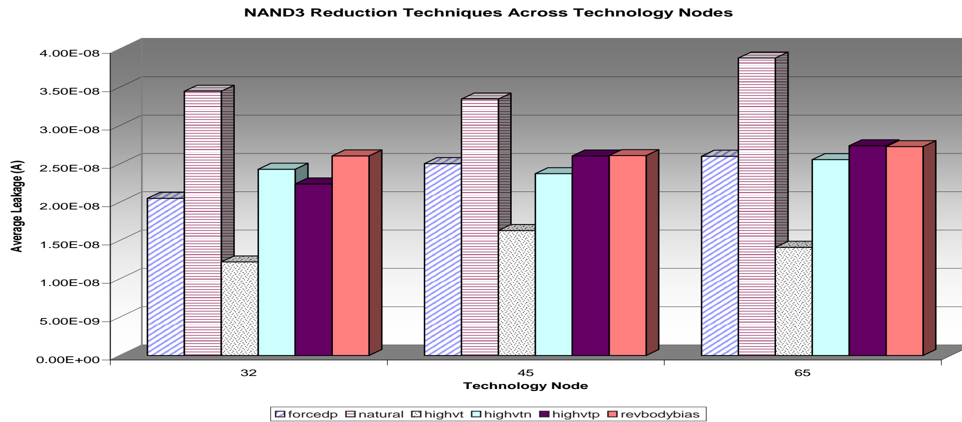


Figure 4.40: NAND3 average total leakage @ 25 °C

The savings achieved relative to the natural form of the three input NAND is portrayed in Figure 4.41. It is evident that the best method to reduce leakage current is to use high V_T NMOS and PMOS. It is interesting to note that there are more leakage savings when a high V_T NMOS is used than a high V_T PMOS for equiprobable inputs unlike the two input NAND gate. This is because the size of the NMOS has increased to keep equal rise and fall time, hence increasing the amount of subthreshold leakage in the pull-down network.

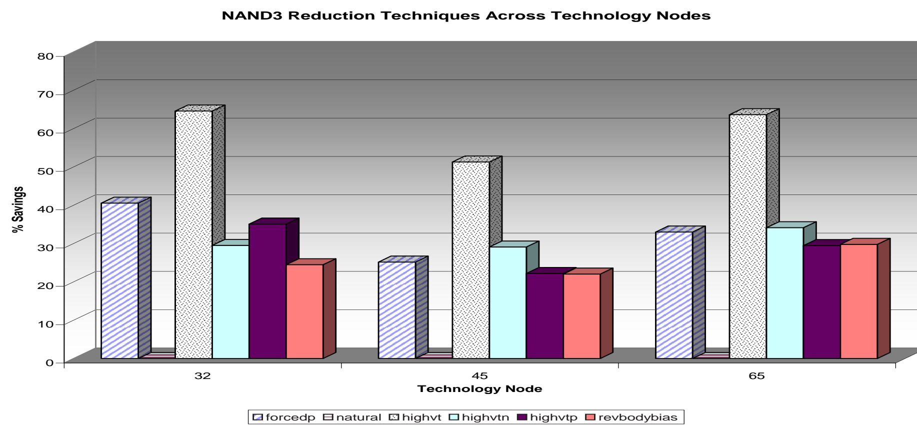


Figure 4.41: NAND3 % leakage savings @ 25 °C

4.7.4 Effect of Temperature

Figure 4.42 shows the percentage leakage savings using different techniques relative to the natural form of the NAND gate at 90 °C. By comparing this figure with Figure 4.41 the effectiveness of reduction techniques are observed with respect to temperature. From the graph it is observed that the techniques effecting the pull-down network create more savings at higher temperatures. The reason for that is the increase in the proportion of NMOS leakage with respect to the total leakage at higher temperatures. This is due to the doping level of PMOS and NMOS ,which directly effect the change in threshold voltage with respect to temperature.

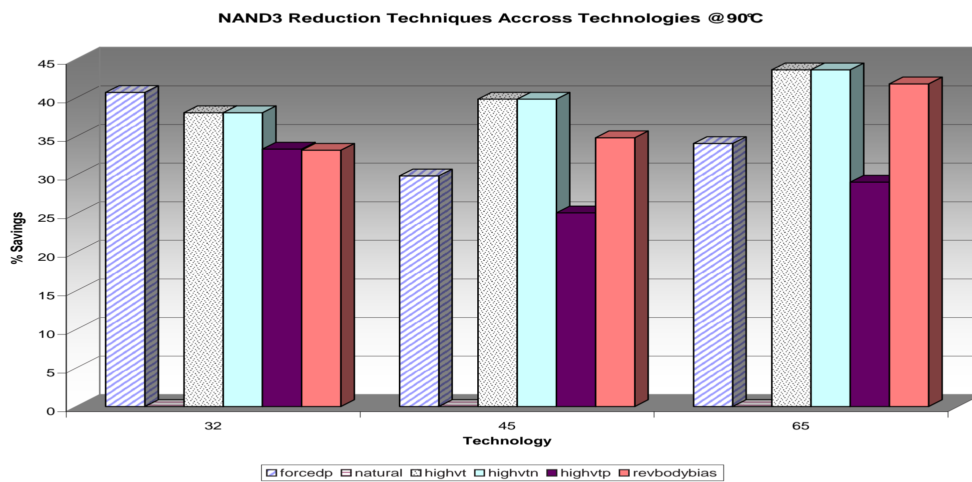


Figure 4.42: NAND3 % leakage savings @ 90 °C

4.8 Three Input NOR Gate

4.8.1 DC Characteristics

Here a three input NOR gate is simulated, which is sized relative to the inverter explained in Section 3.2. Figure 4.43 shows the DC simulation results for 65nm technology node and the noise margin low and high can be found in Figure 4.44(a,b). By comparing the graph with the one of the inverter in Figure 4.4 very similar results are obtained. Therefore, the same argument as the inverter can be used to explain the graph.

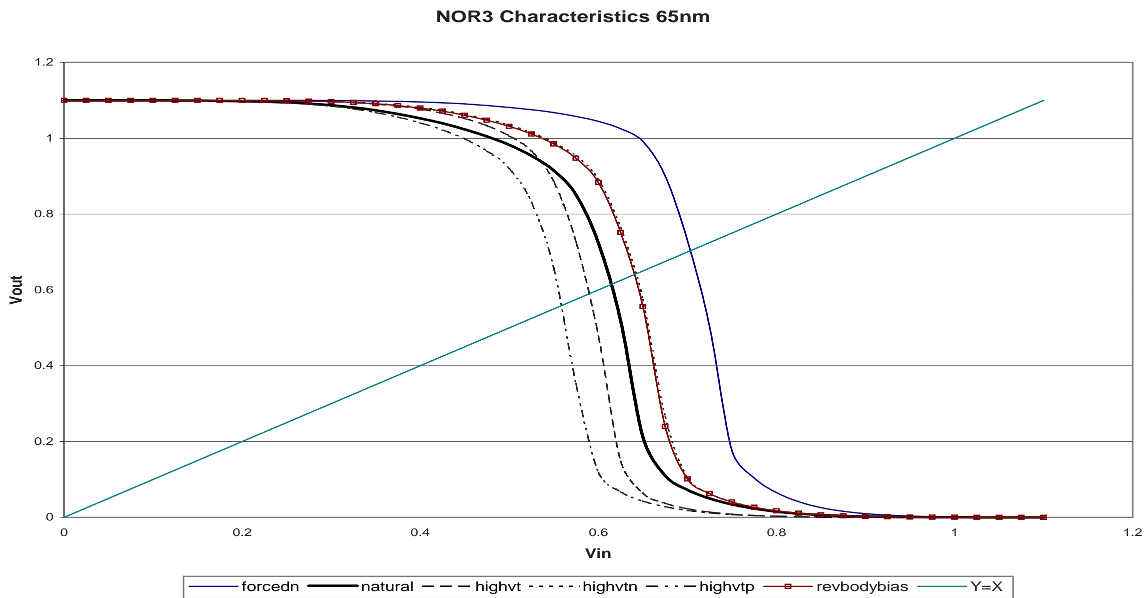
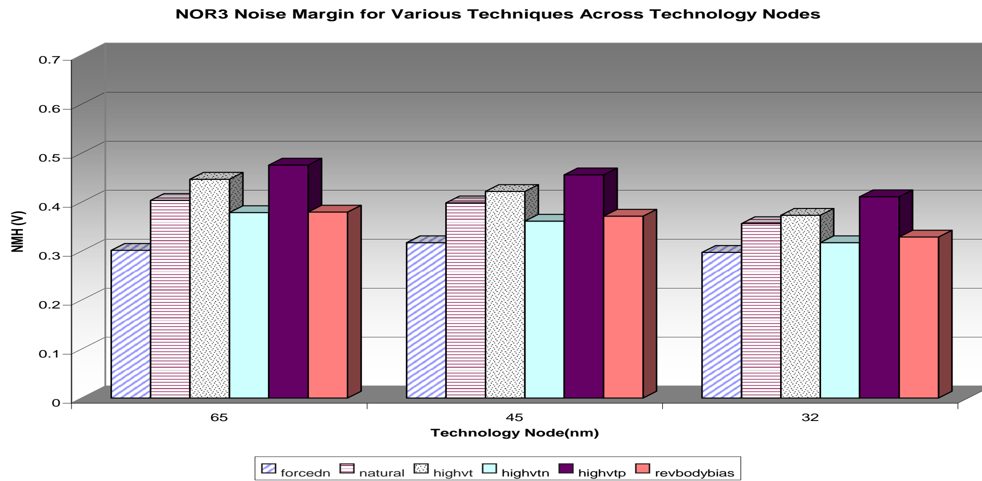
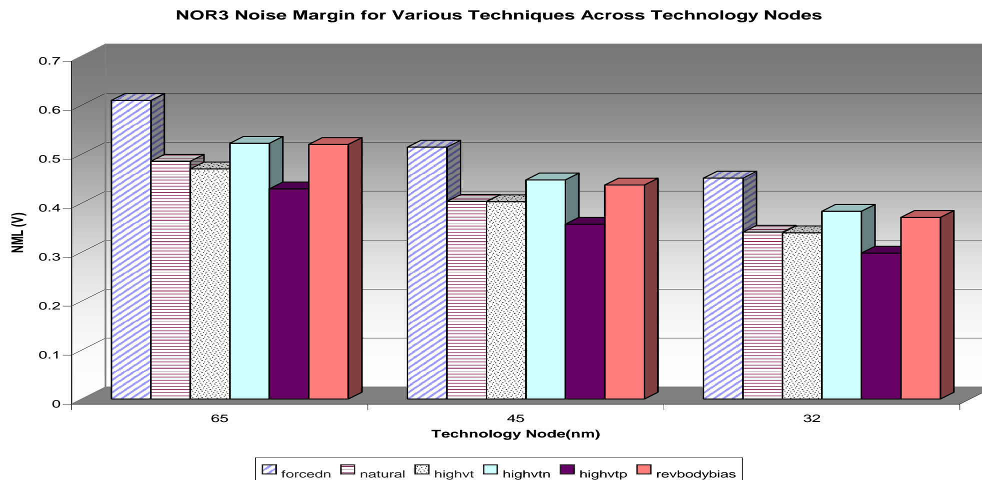


Figure 4.43: NOR3 DC characteristics 65nm @ 25 °C

Figure 4.45 shows the switching voltage for each technique at different technology nodes. The change in the switching voltage is proportional to the change in the noise margins for the particular technique and technology node. This change also seems to be equal for each different technology node. The other shift in the total reduction of the switching voltage from one technology node to another is due to the reduction of power supply voltage.



(a) NMH



(b) NML

Figure 4.44: NOR3 noise margins @ 25 °C

4.8.2 Change in Rise and Fall Times

Table 4.8 lists the rise time and fall time of the three input NOR gate for each applied technique for all three technology nodes. By looking at the percentage change in rise and fall times of each technique compared to the original NOR gate the technique that has the most or the least impact on performance of the gate is obtained.

As expected the results shown in the table are in line with what is observed from the

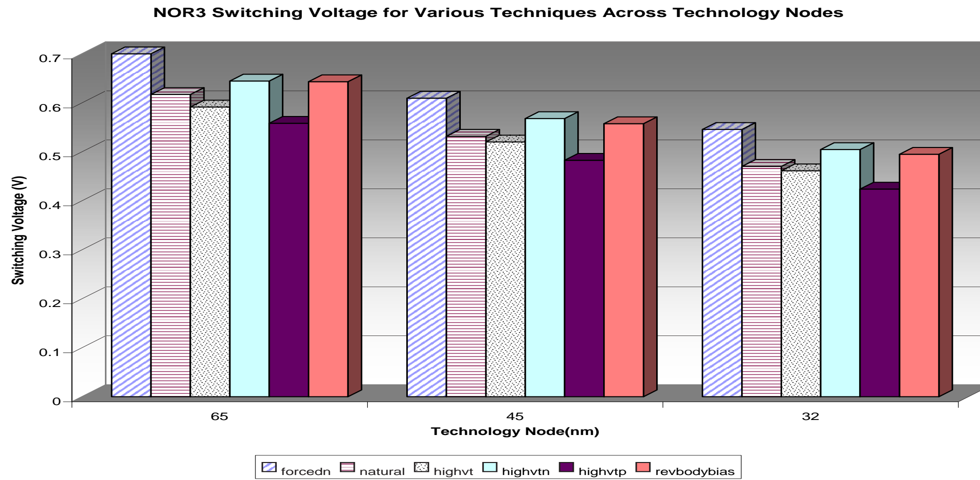


Figure 4.45: NOR3 switching voltage across technology nodes @ 25 °C

inverter and the two input NOR gate. Confirming that reducing leakage in the pull-up network impacts the rise time, and reducing leakage in the pull-down network impacts the fall time. The most impact is observed from stack forcing, almost twice as the two input NOR gate, since it reduces the drive of the transistor and there is a larger capacitance that has to be discharged. The performance gain on the opposite network is more sensible than the case of the inverter or the two input NOR gate. This can be due to the fact that there is one more PMOS transistor in the stack, resulting in large PMOS transistors. Therefore, a little leakage savings has a more pronounced effect on the fall time since the NMOSs require to fight less leakage current to discharge the capacitor.

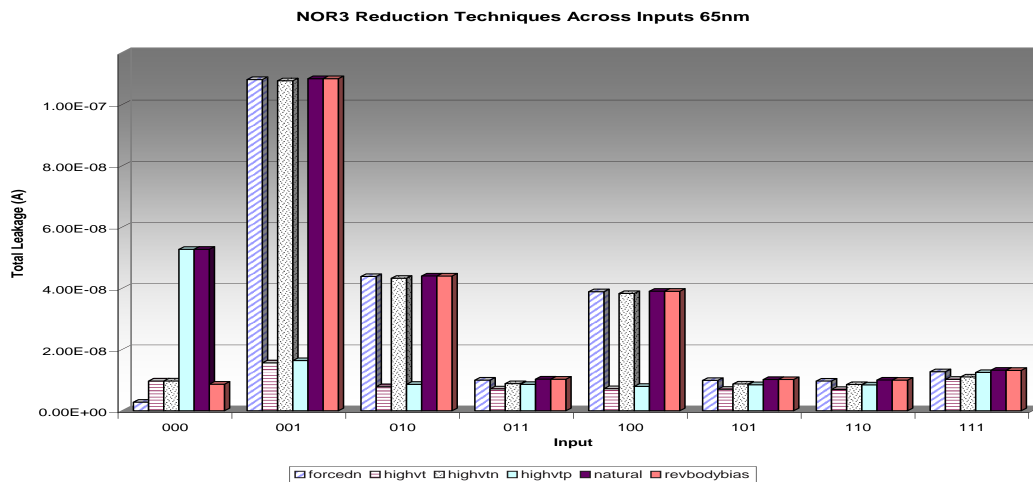
4.8.3 Total Leakage versus Inputs

Figure 4.46 shows the three input NOR gate total leakage for all different inputs for 65nm technology node. It can be observed that the inputs 110, 101, 011, 111, 100, 010, 000, and 001 are sorted in the order of lowest leakage to highest.

In Figure 4.47 the gate and subthreshold leakages are shown for the different inputs. An input of 110 has the least leakage, since the middle PMOS experiences negative V_{gs}, body effect, and reduced V_{ds}, resulting in very little subthreshold leakage. Given an input of 001 the most leakage occurs, since two top PMOSs are conducting, and the lower PMOS

Table 4.8: NOR3 rise and fall times @ 25 °C

Tech(nm)	Method	tRise(ps)	%change	tFall(ps)	%change
65	natural	28.7	0.0	9.5	0.0
	forcedn	28.7	0.2	33.6	253.3
	highvt	32.7	14.0	10.0	4.9
	highvtn	28.6	-0.1	10.5	10.2
	highvtp	32.6	13.9	9.1	-4.0
	revbodybias	28.6	-0.3	10.4	8.8
45	natural	30.7	0.0	9.3	0.0
	forcedn	30.4	-0.8	33.3	259.7
	highvt	36.7	19.8	9.8	6.3
	highvtn	30.7	0.1	10.4	11.9
	highvtp	36.7	19.7	8.1	-12.1
	revbodybias	30.6	-0.2	10.1	9.0
32	natural	30.6	0.0	8.3	0.0
	forcedn	30.4	-0.6	34.0	307.7
	highvt	39.2	28.2	9.7	16.7
	highvtn	30.6	0.1	10.4	24.2
	highvtp	39.1	28.1	8.1	-3.1
	revbodybias	30.5	-0.2	9.8	17.7



(a) 65nm

Figure 4.46: NOR3 leakage vs. inputs @ 25 °C

experiences maximum subthreshold current. Since the size of the PMOS is almost 8 times (2.7*3) of the NMOS, the leakage of this single PMOS is more than three NMOSs as for input 000.

It can be seen in Figure 4.46 that the key contributors to the leakage are the pull-up transistors for all the inputs except 000. Therefore the techniques effecting the pull-down network only effect the 000 input.

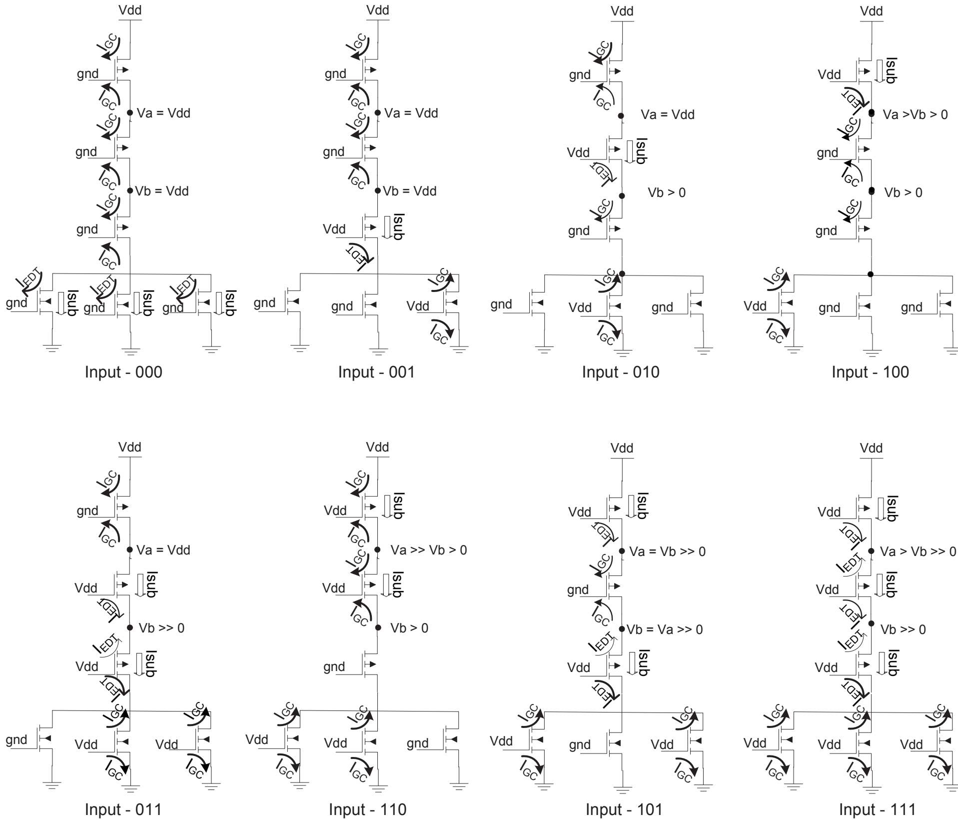


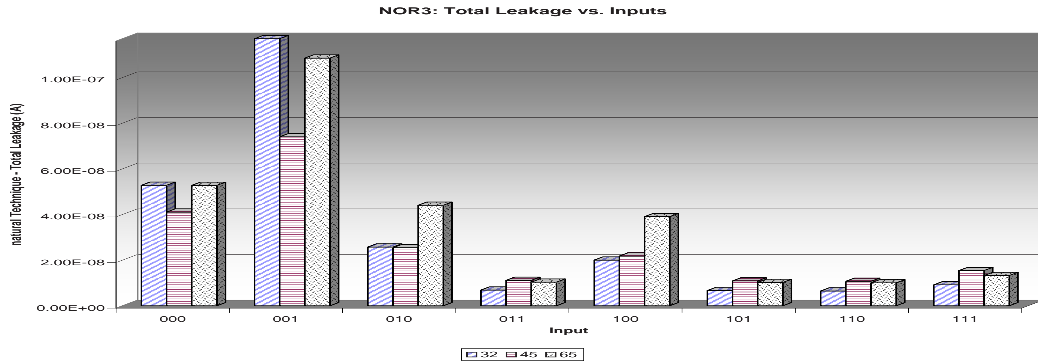
Figure 4.47: NOR3 inputs and leakages in steady state

Figure 4.48 (a) shows the three input NOR gate's total leakage in its natural form across technologies. The figure shows a leakage reduction in the 45nm technology node for both inputs of 000 and 001. This behavior can be explained by the fact that in the PTM, the threshold voltage of the NMOS and PMOS device is a little higher than the 65nm and 32nm technology models as explained in Section 3.3.1.

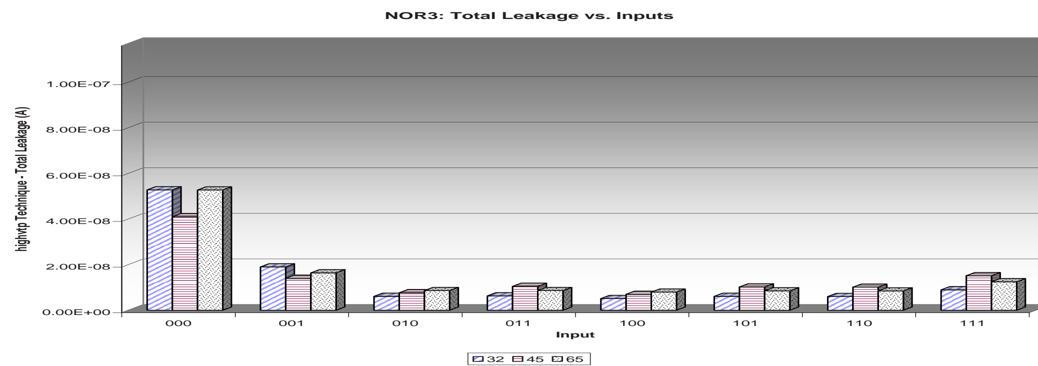
Figure 4.48 (b) shows the three input NOR gate's total leakage when high V_T PMOS is used in its structure. The effectiveness of this technique in leakage reduction seems to be consistent across all the technology nodes.

Figure 4.48 (c) shows the three input NOR gate's total leakage when a forced NMOS stack is applied. This technique effects the leakage mainly for the 000 input. The effectiveness of this technique in leakage reduction seems to be consistent across all the technology nodes.

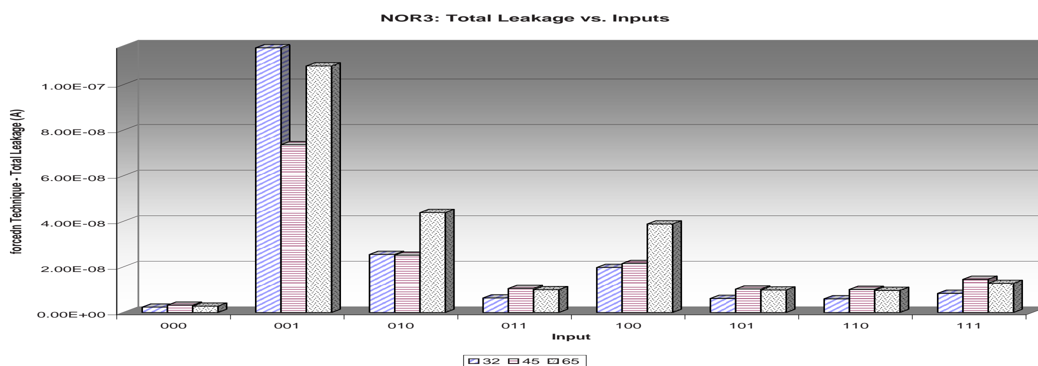
Figure 4.49 (a,b,c) show the three input NOR gate's total leakage when using high V_T PMOS and NMOS, high V_T NMOS, and RBB technique in the NOR gates structure. They all exhibit the same behavior in all technology nodes.



(a) Natural



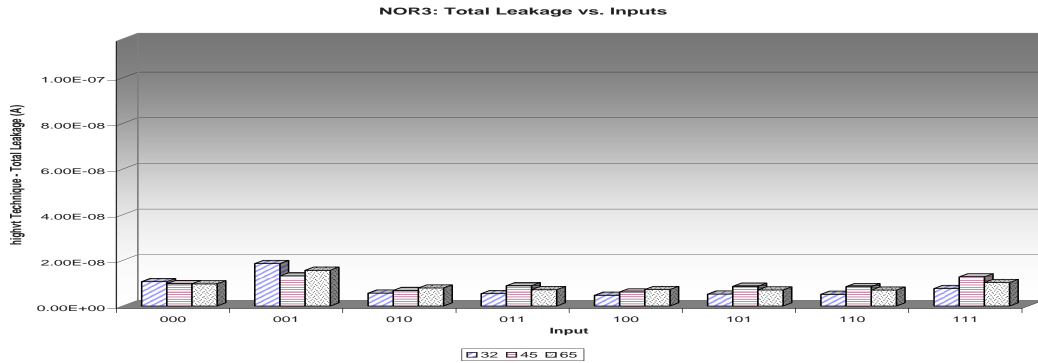
(b) High V_T PMOS



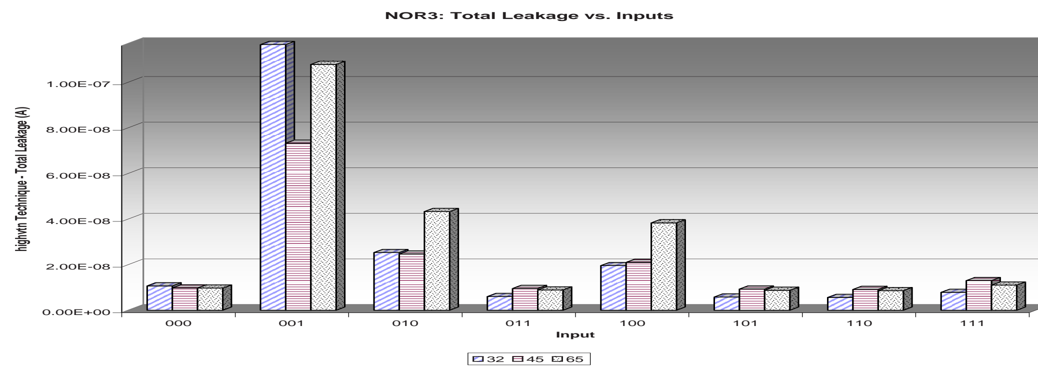
(c) Forced NMOS stack

Figure 4.48: NOR3 leakage vs. input for various leakage reduction techniques @ 25 °C

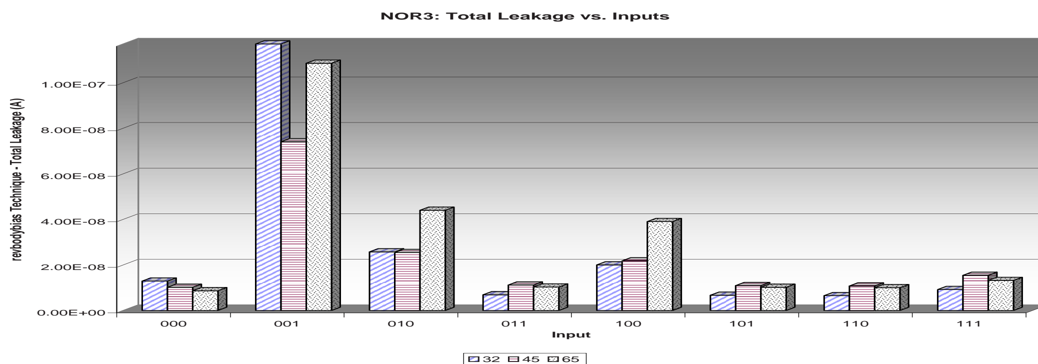
Figure 4.50 illustrates the average total leakage experienced by the three input NOR gate for an equiprobable input. Here again, a reduced leakage in the 45nm node is observed



(a) High V_T CMOS



(b) High V_T NMOS



(c) Reverse Body Bias

Figure 4.49: NOR3 leakage vs. input for various leakage reduction techniques @ 25 °C

due to the slightly higher threshold voltage in the model. The savings achieved relative to the natural form of the three input NOR are shown in Figure 4.51. It is obvious that the

best method to reduce leakage current is to use high V_T NMOS and PMOS. Due to the sizing of the PMOS, there are more leakage savings when a high V_T PMOS is used, than a high V_T NMOS for equiprobable inputs.

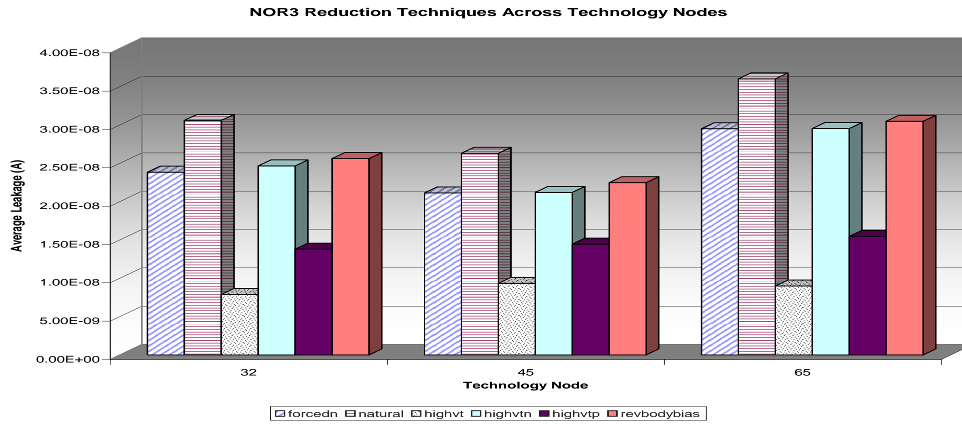


Figure 4.50: NOR3 average total leakage @ 25 °C

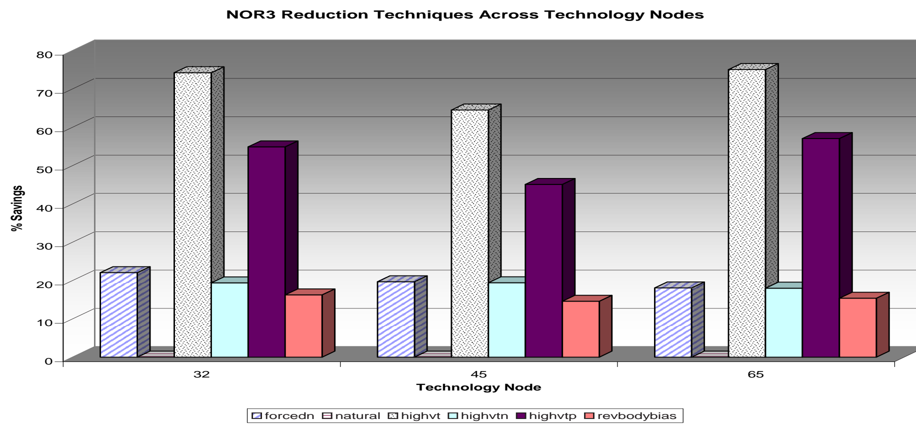


Figure 4.51: NOR3 % leakage savings @ 25 °C

4.8.4 Effect of Temperature

Figure 4.52 shows the percentage leakage savings using different techniques relative to the natural form of the NOR gate at 90 °C. Comparing this figure with Figure 4.51 the

effectiveness of these reduction techniques is observed with respect to temperature. The amount of leakage savings for pull-down network techniques increases a little at 90 degrees. The reason for that is the increase in the proportion of NMOS leakage with respect to the total leakage at higher temperatures. This is due to the doping level of PMOS and NMOS, which directly effect the change in threshold voltage with respect to temperature.

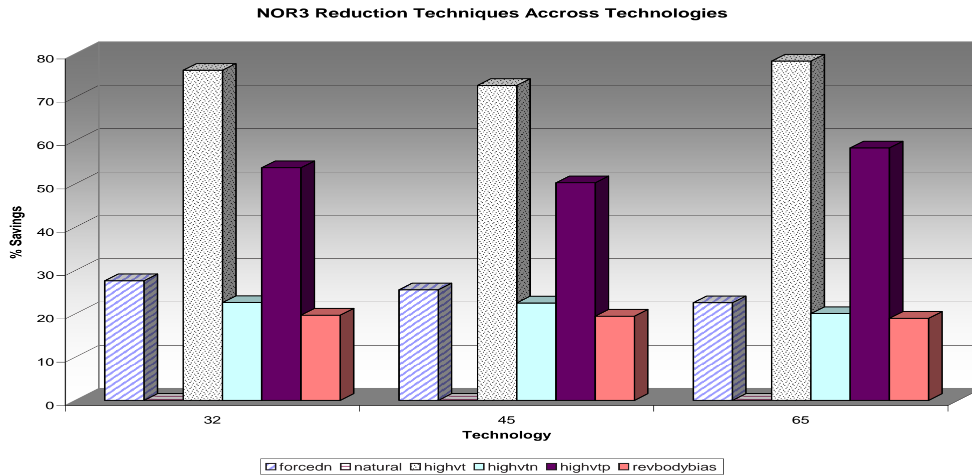


Figure 4.52: NOR3 % leakage savings @ 90 °C

4.9 Low Leakage Logic Gate Selection

It is important to choose the basic gates which have the lowest leakage to compose the building blocks of a design. Here, the two logic gates are compared for different input combinations. Then, a simple algebraic function is implemented in three ways by using a mix of NAND and NOR gates, only NAND gates, and only NOR gates. This example shows that according to the specific application, mathematical conversions and optimizations can be made so that a lower leakage realization is chosen at the beginning, then as a second step, the other techniques are applied to reduce the power even further.

Table 4.9 provides the total leakage for a two input NAND and NOR gates for all possible inputs. By comparing the NAND gate and NOR gate more closely, the input combinations that result in the least amount of leakage in each gate are readily observed.

The leakage values in bold indicate the gate that has the lowest leakage for that par-

Table 4.9: Comparison of the total leakage of two input NAND and NOR gates - 65nm @ 25 °C

Method	Input	NOR2 (nA)	NAND2 (nA)
Natural	00	35.1	**3.5**
	01	73.6	**44.0**
	10	30.6	**15.9**
	11	**9.6**	85.7
Forced N/P	00	**1.8**	3.5
	01	73.3	**44.0**
	10	30.4	**15.8**
	11	**9.2**	17.4
High Vt	00	6.4	**2.3**
	01	**11.5**	13.3
	10	6.3	**2.4**
	11	**6.9**	21.7
High Vtn	00	6.4	**2.3**
	01	72.8	**13.3**
	10	29.9	**2.4**
	11	**8.1**	82.7
High Vtp	00	35.1	**3.5**
	01	**12.2**	44.0
	10	**7.0**	15.8
	11	**8.4**	24.7
RBB	00	5.7	**2.3**
	01	73.5	**14.0**
	10	30.6	**2.1**
	11	**9.5**	85.6

ticular input. Using these values and the probabilities of the each input of the gate, allow one to choose the gate that has the lowest leakage for the most probable input.

Here the leakage savings of NAND and NOR gate are contrasted by looking at the values of the total leakage for implementing a simple function. The implementations can be advantageous, depending on the availability of inverted inputs or whether an inverted output is acceptable. Because these considerations result in fewer gates being used. For example, Figure 4.53 shows three different implementations for the function $\tilde{(}a+bc)$ by assuming inverters are available. The first implantation, utilizes both NAND and NOR gates to realize the function. The second implementation, uses only NAND gates and finally the last implementation uses only NOR gates.

Clearly, Figure 4.53(a) consumes less power than the other implantations, since it uses one less inverter. However, depending on availability of the inputs and their complements, the other designs can use less number of gates than Figure 4.53(a). As it can be seen, there are two main gates (G1,G2) that are employed in all three designs, and they only differ in the number of inverters used. The total leakage for all these implementations, and in addition, the same implementations while ignoring the power of inverters, have been

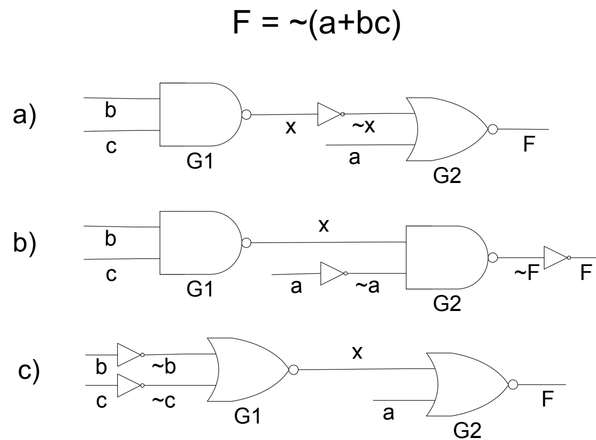


Figure 4.53: Different implantations of the function

listed in Table 4.10. This way the total leakage can be judged independent of availability of inverted or non-inverted inputs.

To calculate the total leakage for this function, inputs a,b, and c are assumed to be equiprobable. However, since the inputs to gate G2 are not equiprobable the order which they are assigned affects leakage. Therefore, each implementation has its total power calculated in two different ways. In Table 4.10 the lowest leakage for each low power technique is in bold. The total leakage for the cases of a1 and a2 using inverters at 25 °C and 90 °C are shaded, and not considered to determine the lowest leakage implementation, as it was discussed previously.

It is noteworthy that by ignoring the power of the inverter, the total leakage would be lower for all the low power techniques using implementation (b) (only NAND gates used in the design) except for the case of using high V_T or high V_{tP} low power technique at 25 °C. But, implementation (b) also has less leakage in the high V_T case at 90 °C. Which is again consistent with our observations in the previous chapters on NOR and NAND gates. Since NOR has larger PMOS used in its pull-up network using a high V_{tP} PMOS will lower its leakage significantly, but it will not impact the NAND gate as much.

In a 65nm process technology, when inputs do not limit the implementation, using only NAND gates even with no leakage reduction techniques, yields more leakage savings than the case of mixed or NOR only implementations. Also the pin reordering results in about

Table 4.10: Total leakages for the implementations in Figure 4.53

Circuit	Natural	HighVtp	HighVtn	HighVt	RBB	ForceP/N
Assume no inverter is needed - 65nm @ 90 ℃						
a1	201.0E-9	108.6E-9	133.8E-9	41.4E-9	134.6E-9	123.8E-9
a2	178.3E-9	104.8E-9	111.2E-9	37.7E-9	111.9E-9	101.3E-9
b1	**164.2E-9**	113.6E-9	**88.4E-9**	**37.8E-9**	**89.4E-9**	**104.9E-9**
b2	**148.9E-9**	98.4E-9	**83.9E-9**	**33.3E-9**	**84.7E-9**	**89.7E-9**
c1	201.7E-9	**93.2E-9**	147.8E-9	39.3E-9	148.3E-9	138.8E-9
c2	179.1E-9	**89.4E-9**	125.2E-9	35.5E-9	125.7E-9	116.3E-9
Assume inverter is used - 65nm @ 90 ℃						
a1	269.3E-9	143.1E-9	180.7E-9	54.6E-9	163.4E-9	170.7E-9
a2	246.6E-9	139.4E-9	158.1E-9	50.9E-9	140.8E-9	148.2E-9
b1	**300.7E-9**	182.7E-9	**182.2E-9**	**64.2E-9**	**147.0E-9**	**198.7E-9**
b2	**285.5E-9**	167.5E-9	**177.7E-9**	**59.7E-9**	**142.4E-9**	**183.5E-9**
c1	338.3E-9	**162.3E-9**	241.6E-9	65.6E-9	206.0E-9	232.7E-9
c2	315.7E-9	**158.6E-9**	218.9E-9	61.9E-9	183.3E-9	210.1E-9
Assume no inverter is needed - 65nm @ 25 ℃						
a1	83.0E-9	41.7E-9	59.6E-9	18.3E-9	60.7E-9	53.3E-9
a2	72.3E-9	40.4E-9	48.9E-9	17.0E-9	50.0E-9	42.6E-9
b1	**67.8E-9**	44.9E-9	**41.7E-9**	18.8E-9	**43.1E-9**	**42.2E-9**
b2	**60.8E-9**	37.9E-9	**38.9E-9**	16.0E-9	**40.1E-9**	**35.1E-9**
c1	83.0E-9	**35.4E-9**	63.7E-9	**16.1E-9**	64.5E-9	61.8E-9
c2	72.2E-9	**34.1E-9**	53.0E-9	**14.8E-9**	53.8E-9	51.1E-9
Assume inverter is used - 65nm @ 25 ℃						
a1	111.0E-9	54.5E-9	80.1E-9	23.5E-9	81.4E-9	64.3E-9
a2	100.3E-9	53.2E-9	69.4E-9	22.2E-9	70.7E-9	53.6E-9
b1	**123.8E-9**	70.5E-9	**82.7E-9**	29.3E-9	**84.5E-9**	**64.2E-9**
b2	**116.8E-9**	63.5E-9	**79.9E-9**	26.5E-9	**81.5E-9**	**57.1E-9**
c1	139.0E-9	**61.0E-9**	104.7E-9	**26.6E-9**	105.9E-9	83.8E-9
c2	128.2E-9	**59.7E-9**	94.0E-9	**25.3E-9**	95.2E-9	73.1E-9

8%-15% savings which is contributed to both gate and subthreshold leakage reduction. These saving can be more significant if the gate oxide thickness is small or high-K material is not used.

By observing Table 4.10 more closely some design guidelines can be extracted. If the designer is limited to only choose between high V_{tP} or high V_{tN} transistors, or RBB for a

65nm technology node, at room temperature, using a high V_{tP} device with a design that has large pull-up networks (NORs) would result in lower leakages. But, at high temperatures using high V_{tN} or RBB for a design with large pull-down structure (NANDs) will result in slightly more leakage savings.

However, the conclusions above will change when other technology nodes are used, since the total leakage of these gates changes with the process technology parameters. Table 4.9 summarizes the average leakage for two input NAND and NOR gate for different technology nodes. The bold numbers indicate the gate that has the lower leakage. As depicted in the table, NAND and NOR's total leakages for 65 and 32nm technologies exhibit the same behavior relative to each other. However, due to the differences of the 45nm technology parameters the pattern observed above is different at this node.

Table 4.11: Comparison of the average total leakage two input NAND and NOR gates @ 25 °C

Method	65nm		45nm		32nm	
	NOR2 (A)	NAND2 (nA)	NOR2 (nA)	NAND2 (nA)	NOR2 (nA)	NAND2 (nA)
Natural	**37.2**	**37.3**	**26.7**	29.4	**34.5**	**35.6**
High Vt	**7.77**	9.92	**7.66**	10.7	**7.39**	9.22
High Vtn	29.3	**25.2**	**20.7**	**20.6**	27.1	**25.4**
High Vtp	**15.7**	22	**13.7**	19.5	**14.8**	19.5
RBB	29.8	**26**	**21.6**	**22**	27.9	**26.6**
Forced N/P	28.7	**20.2**	20	**18.2**	25.8	**17**

4.10 Power Aware Technology Mapping Tools

Power aware technology mapping tools can use the following guidelines as part of the low power design flow to achieve the required maximum leakage in a circuit. These guidelines are expressed in general terms for use in any application and process technology. Algorithm 1 is the main algorithm that calls algorithm 2 and algorithm 3 to choose the gate leakage reduction techniques and subthreshold leakage reduction techniques, respectively.

```
Inputs :
  • Leakage Tolerance := MaxLeakage
  • circuit's operating temperature := operatingT

Compute alternate realizations of the circuit foreach alternate realization do
    if operatingT >> room temperature then
        Apply subthreshold reduction technique:
        returnValue1 ← Choose subthreshold reduction technique;
        if returnValue1 == Failed Timing then
            return Failed;
        else
            returnValue2 ← Choose gate leakage reduction techniques;
            if returnValue1 == Failed Leakage && (returnValue2 == Failed Timing || returnValue2 == Failed
            Leakage) then
                return Failed;
            end
            return Success;
        end
    else
        if gate leakage > 20% of total leakage then
            returnValue1 ← Choose gate leakage reduction techniques;
            if returnValue1 == Failed Timing then
                return Failed;
            else
                returnValue2 ← Choose subthreshold leakage reduction techniques;
                if returnValue1 == Failed Leakage && (returnValue2 == Failed Timing || returnValue2 ==
                Failed Leakage ) then
                    return Failed;
                end
                return Success;
            end
        else
            goto: Apply subthreshold reduction technique;
        end
    end
end
```

Algorithm 1: Applying leakage reduction techniques

Choose gate leakage reduction techniques

```
if Thicker oxide transistor available then
    Apply thicker oxide gates to the transistors that are not in the critical path
    if timing is met then
        Pin reordering:
        total-lkg leftarrow current total leakage
        Apply pin reordering to reduce gate leakage
        if current total leakage > total-lkg then
            Undo pin reordering
        else if current total leakage < MaxLeakage then
            return Success
        end
        return Failed Leakage
    else
        if Circuit is composed mostly series pull-up devices then
            Remove thicker oxide gates from PMOSs if timing is met then
                goto: Pin reordering
            else
                Remove thicker oxide gates from NMOSs if timing is met then
                    goto: Pin reordering
                end
                return Failed Timing
            end
        Remove thicker oxide gates from NMOSs if timing is met then
            goto: Pin reordering
        else
            Remove thicker oxide gates from PMOSs if timing is met then
                goto: Pin reordering
            end
            return Failed Timing
        end
    end
end
```

Algorithm 2: Choosing gate leakage reduction scheme

Choose subthreshold reduction technique

```
if Circuit is composed mostly series pull-up devices then
    if High V_T PMOS available then
        Apply high V_T PMOS to transistor on non-critical path;
        if timing is met then
            Check Leakage 1:
            if total leakage > MaxLeakage then
                Apply high V_T NMOS to transistors on non-critical path;
                if timing is not met then
                    Remove high V_T NMOS transistors;
                    if Reverse body biasing (RBB) is available then
                        Apply reverse body biasing to the same transistors ;
                    end
                    Check Timing 1:
                    if timing is not met then
                        if RBB is not available then
                            return Failed Timing;
                        end
                        Reduce RBB voltage by δ ;
                        goto: Check Timing 1;
                    else
                        goto: Check Leakage 2;
                    end
                else
                    Check Leakage 2:
                    if total leakage > MaxLeakage then
                        return Failed Leakage
                    else
                        if noise margin and rise and fall time equality are not as desired then
                            Change pull-up to pull-down ratio by δV to reduce rise time or by -δV to reduce fall time;
                            goto: Check Timing 1;
                        else
                            goto: Pin Reordering;
                        end
                    end
                end
            else
                if noise margin and rise and fall time equality are not as desired then
                    Change pull-up to pull-down ratio by δV to reduce rise time or by -δV to reduce fall time;
                    goto: Check Timing 1;
                else
                    if Pin reordering is already applied then
                        return Success;
                    end
                    goto: Pin Reordering;
                end
            end
        else
            Pin reordering:
            if Pin reordering is already applied then
                return Success;
            end
            Apply pin reordering to mitigate subthreshold leakage;
            if timing is not met then
                return Failed Timing;
            else
                goto: Check Leakage 1;
            end
        end
    else
        Apply pin reordering to mitigate subthreshold leakage;
        if timing is not met then
            return Failed
        else
            if RBB is not available then
                if total leakage > MaxLeakage then
                    return Failed Leakage
                end
                if noise margin and rise and fall time equality are not as desired then
                    Change pull-up to pull-down ratio by δV to reduce rise time or by -δV to reduce fall time;
                    goto: Check Timing 1;
                end
                return Success
            end
            Apply RBB to the pull-down network transistors ;
            goto: Check Timing 1;
        end
    end
```

Algorithm: continued on next page ...

```
else
    if High V_T NMOS available then
        Apply high V_T NMOS to transistor on non-critical path;
        if timing is met then
            Check Leakage 3:
            if total leakage > MaxLeakage then
                Apply high V_T PMOS to transistors on non-critical path;
                if timing is not met then
                    Remove high V_T PMOS transistors;
                    Apply pin reordering to mitigate subthreshold leakage;
                    Check Timing 2:
                    if timing is not met then
                        if RBB is not available then
                            return Failed Timing;
                        Reduce RBB voltage by δ ;
                        goto: Check Timing 2;
                    else
                        goto: Check Leakage 4;
                    end
                else
                    Check Leakage 4:
                    if total leakage > MaxLeakage then
                        return Failed Leakage
                    else
                        if noise margin and rise and fall time equality are not as desired then
                            Change pull-up to pull-down ratio by δV to reduce rise time or by -δV to reduce fall time;
                            goto: Check Timing 2;
                        else
                            return Success;
                        end
                    end
                end
            else
                if noise margin and rise and fall time equality are not as desired then
                    Change pull-up to pull-down ratio by δV to reduce rise time or by -δV to reduce fall time;
                    goto: Check Timing 2;
                else
                    if Pin reordering is already applied then
                        return Success;
                    goto: Pin Reordering;
                end
            end
        else
            Remove high V_T NMOS transistors;
            Pin reordering:
            if Pin reordering is already applied then
                return Success;
            Apply pin reordering to mitigate subthreshold leakage;
            if timing is not met then
                return Failed Timing
            else
                goto: Check Leakage 4;
            end
        end
    else
        Apply pin reordering to mitigate subthreshold leakage;
        if timing is not met then
            return Failed Timing
        else
            Apply high V_T PMOS to transistors on non-critical path;
            if timing is not met then
                Remove high V_T PMOS transistors;
            if RBB is not available then
                if total leakage > MaxLeakage then
                    return Failed Leakage
                if noise margin and rise and fall time equality are not as desired then
                    Change pull-up to pull-down ratio by δV to reduce rise time or by -δV to reduce fall time;
                    goto: Check Timing 4;
                end
                return Success
            end
            Apply RBB to the pull-down network transistors ;
            goto: Check Timing 1;
        end
    end
end
```

Algorithm 3: Choosing subthreshold leakage reduction scheme

Chapter 5

Conclusions

In this thesis leakage reduction techniques are explored that mitigate leakage in circuits, operating in the active mode at various temperatures. Also, implications of technology scaling on the choice of techniques to mitigate total leakage are closely examined. The result, is guidelines for designing low-leakage circuits in nanometer technology nodes. Logic gates in the 65nm, 45nm, and 32nm technology nodes are simulated and analyzed. The techniques that are selected for comparison in this dissertation affect both gate leakage and subthreshold leakage, namely, stack forcing, pin reordering, reverse body biasing, and high V_T MOS. Aside from leakage, the analysis also highlights the impact of each technique on the circuit's performance and noise margins.

The Reverse Body Biasing (RBB) scheme tends to be less effective as the technology scales, since body effect coefficient is scaling and also RBB increases the Band To Band Tunneling (BTBT) current. BTBT will be significant in nanometer technologies and further increases, when a RRB is applied. Therefore, there exists an optimum RBB value that yields the lowest combined subthreshold and BTBT leakage, and this RBB value decreases in magnitude as the technology scales [23]. However, some works such as [25] have successfully implemented a RRB technique in a 65nm technology node by optimizing the device parameters.

As shown in the last chapter, employing high V_T MOS is one of the most effective techniques for reducing leakage with small performance degradation. Depending on the choice of structure, a parallel pull-up or parallel pull-down network, employing high V_T in

that network yields in more leakage savings. In this investigation, this technique proves to be attractive, since it scales well with technology, and reduces the leakage of the circuit for any input in both active and standby modes of operation with only a small performance penalty.

Pin reordering and natural stacks are techniques that reduce leakage, yet they do not influence the performance of the device. However, it is demonstrated that they are not as effective in all types of logic, since the input values might switch only between the highly leaky states. As technology scales, the supply voltage is reduced, which renders pin reordering a more attractive technique, due to its minimal impact on circuit performance. If the gate leakage is not controlled as technology scales (higher effective gate oxide thickness) and/or the circuit operates at lower temperatures, the high and low leaky states are shown to differ from one technology to another, in accordance with the contribution of the gate leakage to the total leakage. For example, in the case of a two stacked NMOSs, an input of 00 gives the lowest total leakage, if subthreshold leakage is the dominant component of leakage, and 10 provides the lowest total leakage, if gate leakage is the dominant component. Similar analysis is conducted for the 45nm and 32nm nodes.

It is confirmed that according to the design requirements of the circuit, one technique can result in a better performance or leakage savings than another. This observation is validated by analyzing logic structures that are inherently different: having parallel PMOS devices in the pull-up network (e.g., 2-3 input NAND gates) or parallel NMOS devices in the pull-down network (e.g., 2-3 input NOR gates). It is concluded that the leakage of the pull-up transistors are more substantial than their counterparts in the pull-down network, since PMOS devices need to be enlarged (2-3 times) to achieve equal rise and fall times, consequently increasing leakage. In addition, to attain equal rise and fall times, a stacked pull-up network requires larger transistors to charge the node capacitance than the same network with parallel transistors. Typically at room temperatures due to the PMOS sizing, leakage reduction techniques that affect the pull-up network will result in higher leakage savings, than those affecting pull-down networks. However, if this technique is used only in a pull-up or pull-down network, it will result in unequal rise and fall times, so the gate has to be resized to achieve equal times.

From the above, it is obvious that transistor sizing plays vital role in leakage reduction. Because high V_T transistors are an effective way for reducing subthreshold leakage, transistor sizing should be considered as another technique that should accompany the high V_T transistor technique. Since smaller transistors produce less gate and subthreshold leakage, they are appropriate for reducing leakage, since gate leakage is more pronounced in lower technology nodes. Transistor sizing can be used similar to the high V_T assignment of transistors that are in the non-critical path of the circuit. This technique is embellished, when gate leakage is a significant part of the total leakage, that is, when the non-critical path is operating at lower temperatures.

It is also demonstrated that another effective method to tackle power reduction, is modifying the realization of the circuit by choosing the gates with the lowest leakage to implement the circuit. This approach gives a head start on reducing the amount of leakage the circuit. After this step, further leakage reduction techniques can be applied to achieve an optimum solution.

In conclusion, stack forcing, pin reordering, RBB, and high V_T MOS are explored as leakage reduction techniques across the 65, 45, and 32nm technology nodes. Since a single technique cannot address all of the various leakage components and performance requirements, a combination/hybrid of techniques must be chosen depending on the design requirements of the circuit. Therefore, unlike in the past, it is not sufficient for power reduction tools to use only one leakage reduction technique to create a low power design. Power sensitive technology mapping tools that incorporate a variety of low power design techniques must be created for optimizing a circuit to achieve maximum leakage savings. The guidelines presented in Section 4.10 should be part of the low power design flow to meet the required maximum leakage in a circuit.

5.1 Future Work

The techniques previously addressed, primarily influence active leakage even though they intrinsically reduce standby leakage. Also, the analysis in this work is focused on logic gates only, whereas memory circuits have different requirements and behaviors. This research can

be easily extended to encompass standby leakage reduction techniques and apply similar analysis to memory circuits to analyze the impact of technology scaling on them.

In the past, various low power techniques have been applied to mitigate subthreshold leakage. However, power aware technology mapping tools must be created or updated to consider a variety of low power design techniques to optimize a circuit. For instance, in [26] technology mapping has been addressed by looking only at pin reordering and hot-carrier effect. Such tools can be created that can include a variety of techniques to reduce the total leakage. However, more complete tools and guides should be created to employ various techniques as a mean of reducing total leakage according to the type of application and mode of operation.

Bibliography

[1] R. Datta, M. Doczy, J. Kavalieros, and M. Metz, "Gate dielectric scaling for high-performance cmos: from siol to high-k," *IWGI*, pp. 124–127, 2003.

[2] K. Roy, S. Mukhopadhyay, and H. MahmoodiMeimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits," *Proceedings of the IEEE Transactions*, vol. 91, pp. 305–327, 2003.

[3] Y. Taur, "Cmos design near the limit of scaling," *IBM Journal of Research and Development*, vol. 46, pp. 213–222, 2002.

[4] S. G. Narendra and A. Chandrakasan, *Leakage in Nanometer CMOS technologies*. John Wiley & Sons, 1st ed., 2005.

[5] A. Corporation, "Startix iii programmable power," *White Papers*, 2006.

[6] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration," *IEEE International Symposiom on Quality Electronic Design (ISQED)*, pp. 585–590, 2006.

[7] G. Moore, "Progress in digital integrated circuits," *International Electron Devices Meeting*, pp. 11–13, 1975.

[8] R. Dennard, F. Gaensslen, H. Yu, V. Rideout, E. Bassous, and A. LeBlanc, "Design of ion-implanted mosfets with very small dimensions," *IEEE Journal of Solid-State Circuits*, pp. 256–268, 1974.

[9] P. Chatterjee, W. Hunter, T. Holloway, and Y. Lin, “The impact of scaling laws on the choice of n-channel or p-channel for mos vlsi,” *IEEE Electron Device Letters*, vol. 1, pp. 220–223, 1980.

[10] Intel-Inc., “Intel microsporcessor quick reference guide,” 2004.

[11] K. M. Cao, W.-C. Lee, W. Liu, X. Jin, P. Su, S. K. H. Fung, J. X. An, B. Yu, and C. Hu, “Bsim4 gate leakage model including source-drain partition,” *IEEE Electron Devices Meeting*, 2000.

[12] N. R. Mohapatra, M. P. Desai, S. Narendra, and V. R. Rao, “The effect of high-k gate dielectrics on deep submicrometer cmos device and circuit performance,” *IEEE Transactions on Electron Devices*, pp. 826–831, 2002.

[13] R. V. Langevelde, A. Scholten, and D. Klaassen, “Physical background of mos model 11,” unclassified report, Koninklijke Philips Electronics N.V. 2003, 2003.

[14] Y.-S. Lin, C.-C. Wu, C.-S. Chang, R.-P. Yang, W.-M. Chen, J.-J. Liaw, and C. Diaz, “Leakage scaling in deep submicron cmos for soc,” *IEEE Journal of Solid-State Circuits*, vol. 49, pp. 1034–1041, 2002.

[15] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits.* Tom Robbins, 2nd ed., 2003.

[16] A. Keshavarzi, K. Roy, and C. F. Hawkins, “Intrinsic leakage in low power deep submicron cmos ics,” *International Test Conference*, pp. 146–155, 1997.

[17] N. R. J. F. S. S. Ogura, C. F. Codella and J. Riseman, “Half micron mosfet using double implanted ldd,” *IEEE Electron Devices Meeting*, 1982.

[18] K. F. Y. C. M. Osburn, I. De and A. Srivastava, “Design and integration considerations for end-of-the roadmap ultrashallow junctions,” *Journal of Vacuum Science and Technology B - Microelectronics and Nanometer Structures*, vol. 18, 2000.

[19] M. Y. Kwong, R. Kasnavi, P. Griffin, J. D. Plummer, and R. W. Dutton, "Impact of lateral source/drain abruptness on device performance," *IEEE Transactions on Electron Devices*, pp. 1882–1991, 2002.

[20] S. Venkatesan, J. W. Lutze, C. Lage, and W. J. Taylor, "Device drive current degradation obsereved with retrograde channel profiles," *IEEE Electron Devices Meeting*, p. 419422, 1995.

[21] S. E. Thompson, P. A. Packan, and M. T. Bohr, "Linear versus saturated drive current: Tradeoffs in super steep retrograde well engineering," *IEEE Symp. VLSl Tech. Digest*, p. 154155, 1996.

[22] O. Mizuki, N. Akira, and K. Masato, "Degradation of current drivability of schottky barrier source/drain transistors induced by high-k gate dielectrics and possible measures to suppress the phenomenon," *Solid-State Electronics*, vol. 50, pp. 788–794, 2006.

[23] A. Keshavarzi, S. Narendra, S. Borkar, C. Hawkind, K. Roy, and V. De, "Technology scaling behavior of optimum reverse body bias for standby leakage power reduction in cmos ic's," *ISLPED*, pp. 252–255, 1999.

[24] "International technology roadmap for semiconductors," 2005. `http://public.itrs.net`.

[25] K. Imai, Y. Yamagata, S. Masuoka, N. Kimuzuka, Y. Yasuda, M. Togo, M. Ikeda, and Y. Nakashiba, "Device technology for body biasing scheme," *ISCAS*, pp. 13–15, 2005.

[26] C. W. Kang and M. Pedram, "Technology mapping for low leakage power and high speed with hot-carrier effect consideration," *DAC*, pp. 203–208, 2003.