RICE UNIVERSITY

Novel dual-threshold voltage FinFETs for circuit design and optimization

by

Masoud Rostami

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

APPROVED, THESIS COMMITTEE:

Kartik Mohanram, Chairman
Assistant Professor of Electrical and
Computer Engineering
Rice University

Farinaz Koushanfar
Assistant Professor of Electrical and
Computer Engineering
Rice University

Yehia Massoud
Associate Professor of Electrical and
Computer Engineering
Rice university

Peter Varman
Professor of Electrical and Computer
Engineering
Rice University

HOUSTON, TEXAS

NOVEMBER, 2010

Abstract

Novel dual-threshold voltage FinFETs for circuit design and optimization

by

Masoud Rostami

A great research effort has been invested on finding alternatives to CMOS that have better process variation and subthreshold leakage. From possible candidates, FinFET is the most compatible with respect to CMOS and it has shown promising leakage and speed performance. This thesis introduces basic characteristics of FinFETs and the effects of FinFET physical parameters on their performance are explained quantitatively. I show how dual-V_{th} independent-gate FinFETs can be fabricated by optimizing their physical parameters. Optimum values for these physical parameters are derived using the physics-based University of Florida SPICE model for double-gate devices, and the optimized FinFETs are simulated and validated using Sentaurus TCAD simulations. Dual-V_{th} FinFETs with independent gates enable series and parallel merge transformations in logic gates, realizing compact low power alternative gates with competitive performance and reduced input capacitance in comparison to conventional FinFET gates. Furthermore, they also enable the design of a new class of compact logic gates with higher expressive power and flexibility than CMOS gates. Synthesis results for 16 benchmark circuits from the ISCAS and OpenSPARC suites indicate that on average at 2GHz and 75 °C, the library that contains the novel gates reduces total power and the number of fins by 36% and 37% respectively, over a conventional library that does not have novel gates in the 32nm technology.

Acknowledgements

I cannot count how many times this text has been returned to me by my advisor Prof. Kartik Mohanram with constructive comments about the grammar and technical issues. I thank him for his astonishing patience and valuable support during the project.

I am grateful to my committee members Prof. Farinaz Koushanfar, Prof. Yehia Massoud and Prof. Peter Varman for their time and comments on this thesis.

I acknowledge Leo Mathew at Applied Novel Devices, Murshed Chowdhury at IBM, and Professor Jerry Fossum at the University of Florida for helpful discussions and support with the UFDG simulator. I also acknowledge Ajay Bhoj and Professor Niraj Jha at Princeton University for helpful discussions and suggestions over aspects of FinFET TCAD simulation. I thank Mihir Choudhury at Rice University for the simulator used for dynamic power estimation. I would also like to thank Dr. Janice Hewitt for proof-reading the manuscript. Finally, I thank my parents for their love throughout these years.

Contents

Abstract ii

Acknowledgements iii

List of Figures vi

List of Tables vii

1 Introduction 1

2 Background 5

3 Dual-V_{th} independent-gate FinFETs 9

3.1 Design of High-V_{th} devices 10

3.2 The optimum gate underlap 12

3.3 Characteristics of low and high-V_{th} devices 13

3.4 Fabrication issues of high-V_{th} devices 16

3.5 Asymmetric double-gate devices 16

4 Logic design with dual-V_{th} FinFETs 18

4.1 Merging and back-gate disabling 19

4.2 Novel dual-V_{th} logic gates 21

4.3 Case study of Boolean networks with four inputs 23

4.4 Novel gates by defactoring the Boolean function 25

4.5 Sequential elements with high–V_{th} devices 28
4.6 The effects of process variation 28

5 Results and conclusions **31**
5.1 Discussions about temperature and frequency 34

6 Conclusions **36**

List of Figures

2.1 2-D cross section of a typical FinFET 7

3.1 I-V curves of (a) n-type and (b) p-type high-V_{th} and low-V_{th} FinFETs in shorted-gate and disabled back-gate modes 14

3.2 UFDG (dotted lines) and TCAD (solid lines) simulations of n-type devices are compared. 15

4.1 Symbols for independent-gate (IG) and shorted-gate (SG) low-V_{th} and high-V_{th} n-type and p-type double-gate FinFETs. The dotted-X sign in high-V_{th} devices denotes their AND-like behavior. 18

4.2 NAND2 gates designed by disabling the back-gates and merging parallel or series transistors 19

4.3 Novel implementation of $[(a+b)*(c+d)]'$ 22

4.4 Four possible implementations of $[(a+b)*c*d]'$ 24

4.5 Novel logic gates by defactoring the Boolean function by using (a) low-V_{th} (b) high-V_{th} FinFETs 26

List of Tables

2.1 Physical parameters of 32nm FinFETs 7

3.1 V_{th}, t_{ox}, and electrode work-function (ϕ) of high-V_{th} (H) and low-V_{th} (L) devices in shorted-gate (SG) and disabled back-gate (IG) modes 13

4.1 Characteristics of conventional and novel NAND gates. 19

4.2 Characteristics of conventional and novel implementations of $[(a+b)*c*d]'$ 24

4.3 Characteristics of conventional and defactored implementations of $[a*(b+c)]'$ 26

5.1 Static power (nW), dynamic power (μW), and number of fins of sixteen benchmarks from the ISCAS and OpenSPARC benchmarks are listed. They are mapped using three different technology libraries: basic, previous work, and complete. 32

5.2 The relationship between frequency and the total power savings is compared at different frequencies for the previous work and complete libraries. 35

Chapter 1

Introduction

Difficulties in the state of the art CMOS devices, such as high levels of process variations and leakage power, have been the motivation for discovery of novel device structures. ITRS 2009 has mentioned multi-gate devices, such as FinFET, along with ultra-thin SOI devices as possible scaling paths for low power digital CMOS technologies [2, 8]. ITRS speculates that FinFET research and development will result in a successful double-gate chip in 2013. FinFET is a slab (fin) of undoped silicon perpendicular to the substrate. FinFET is compatible with standard CMOS over most of its processing steps [14]. At least two sides of the fin are wrapped around by oxide simultaneously, which breaks up the active regions into several fins. As a result, an additional gate increases the electrostatic control of the gate over the channel and makes very high I_{on}/I_{off} ratios achievable. FinFETs have also shown excellent scalability, suppression of short channel effects, and limited parametric variations.

A FinFET with two independent gates is a novel variant of double-gate devices. Two isolated gates are formed by removing the gate regions at the top of the fin. Although the gates are electrically isolated, their electrostatics is highly coupled. In an independent-gate FinFET, the threshold voltage of either gate is easily influenced by applying an appropriate voltage to the other gate. This technology is called multiple independent-gate FET (MIGFET) [22] and can be integrated with regular double-gate devices on the same

chip. A successful implementation of a FinFET device with InGaAs material and a FinFET with three independent gates has also been reported in [39] and [23], respectively.

Many innovative circuit styles exploiting the extra gate(s) in these devices have been proposed in literature [4,9,25,34]. In [9], the authors showed that a pair of parallel transistors in the pull-down or pull-up network of gates can be merged into a single independent-gate FinFET to get a compact low power implementation of the same Boolean function. In [25], four variants for the same function were designed: conventional shorted-gate (SG) mode, independent-gate (IG) mode with merged parallel transistors driven by independent inputs, low power (LP) mode with a reverse-biased back-gate, and an IG/LP mode that combined the LP and IG modes. The use of an independent-gate voltage keeper to improve the reliability of dynamic logic has also been proposed in [27,34]. However, no published work based on FinFETs has extensively explored the possibility of merging series transistors to reduce power and area.

This thesis proposes two innovations in FinFET circuit design. The first innovation is the realization of dual-V_{th} independent-gate FinFETs to enable the merging of pairs of series transistors in logic gates. I show that a dual-V_{th} FinFET can be realized by tuning the electrode work-function, oxide thickness, gate underlap, and silicon thickness without any additional biasing scheme. New high-V_{th} transistors are realized in addition to the regular low-V_{th} ones by tuning these parameters. The high-V_{th} devices will have low resistance *iff* both independent gates are simultaneously activated. The high-V_{th} behavior complements the behavior of low-V_{th} independent-gate FinFETs. The low-V_{th} devices will have a low resistance when either of the gates is activated.

The optimum values of the design parameters for both the low-V_{th} and the high-V_{th} devices were determined using the University of Florida double-gate (UFDG) SPICE model [41]. The UFDG model is a physics-based model that has shown excellent agreement with physical measurements of fabricated FinFETs [41]. It allows several design parameters such as the fin width, channel length, gate-source/drain underlap, and work-function to be varied simultaneously. UFDG enables fast and accurate exploration of the best techno-

logically feasible parameters that are required to realize independent-gate dual-V_{th} FinFETs for the 32nm node. The threshold voltage of high-V_{th} devices is engineered by tuning their silicon thickness and electrode work-function. It is also shown that increasing the oxide thickness of high-V_{th} devices by a factor of two ensures low current when only one of the gates is activated and boosts the current when both the gates are activated. Finally, all the designed devices were simulated and validated using the Sentaurus design suite [3]. The results show excellent agreement in I-V behavior, thereby verifying the integrity of the proposed design methodology.

The second innovation described in this report, based on dual-V_{th} FinFETs, is the design of new classes of compact logic gates with higher expressive power and flexibility than conventional forms. Dual-V_{th} FinFETs with independent gates make it possible to merge series transistors, and simultaneously merging series and parallel transistors allows the realization of compact low power logic gates. By performing series or parallel mergers, logic gates with lower input capacitance and area footprint can be obtained. Although these fin mergers come with a slight deterioration in gate delay, it is shown that reducing the number of stacked devices by series mergers and moving high-V_{th} devices closer to the output pin is a good strategy to mitigate the loss in performance. Further, it is proposed to use the independent back-gate as an independent input, effectively doubling the number of inputs to a logic gate. Using the rules for static logic, if a high-V_{th} transistor is used in the pull-down network, the corresponding transistor in the pull-up network is a low-V_{th} transistor, and vice versa, respectively. These transformations enable us to to implement 12 (56) unique logic gates using only 4 (6) transistors. Finally, I also illustrate how defactoring Boolean expressions can be used to convert the pull-up and pull-down networks into equivalent forms where series/parallel transistors can be merged effectively using dual-V_{th} transistors. The defactoring transformation not only reduces the number of devices, but also the number of stacked transistors in the optimized logic gates, which can potentially increase the speed of the gates.

The logical effort parameters of the basic and the optimized logic gates were extracted

into conventional and enhanced technology libraries. 16 benchmark circuits from the ISCAS and OpenSPARC suites were synthesized to operate at a frequency of 2.5 GHz, and their dynamic power was estimated at 2 GHz. The results show that on average, the complete library reduces the total power by 36% and the number of fins by 37%, over a conventional library based on shorted-gate FinFETs in 32nm technology. On the other hand, the library that is built using only parallel mergers proposed in literature results in a 20% reduction in the total power and 21% reduction in the number of fins, over a conventional library based on shorted-gate FinFETs in 32nm technology.

Chapter 2 provides a basic review of FinFETs. Chapter 3 describes the design of dual-V_{th} independent-gate FinFETs based on electrode work-function, gate oxide thickness, silicon thickness, and gate-source/drain underlap tuning. Possibility of asymmetric double-gate devices is also explored in this chapter. Chapter 4 describes new circuit styles based on these FinFETs. Chapter 5 presents the results and chapter 6 is a conclusion.

Chapter 2

Background

Double-gate devices were first investigated because intuitively, an additional gate is expected to suppress short channel effects and improve I_{on}/I_{off} ratios by increasing electrostatic stability. The electric potential along the undoped channel (x direction in Fig. 2.1) can be approximated by

$$\phi = C_0 \cdot \exp\left(\pm\frac{x}{\lambda}\right), \tag{2.1}$$

where C_0 is a constant and λ is the natural length of the device. λ is given by the following expression [8]:

$$\lambda = \sqrt{\frac{\varepsilon_{Si}}{n \cdot \varepsilon_{ox}} t_{ox} t_{Si}}. \tag{2.2}$$

λ should be as small as possible to quickly damp the influence of drain potential on the channel. Reducing λ is possible by using high-κ dielectric materials, decreasing oxide thickness t_{ox} and/or silicon thickness t_{Si}, or by increasing the relative control of the gate through the coefficient n. n is one for single-gate devices and two for double-gate devices. Thus, using double-gate devices not only helps suppress short channel effects, but also relaxes the physical requirements on t_{Si} and t_{ox}.

Early double-gate devices were manufactured using planar technology and suffered from several manufacturing hurdles, such as self-alignment of the front-gate and back-gate

and the lack of an area efficient contact to the back-gate. Each of these physical challenges effectively creates new parasitic elements that counterbalance the main benefits of the double-gate device. FinFET devices have been proposed to overcome the manufacturing hurdles of double-gate devices. In FinFETs, the gate oxide is formed on both sides of the fin simultaneously, which solves alignment issues of source and drain junctions and simplifies the manufacturing process.

The FinFET channel is a tiny slab (fin) of undoped silicon perpendicular to the device substrate. The cross-section of a typical FinFET is presented in Fig. 2.1, where t_{gf}, t_{gb}, t_{Si}, and L_u are the thickness of front-gate, the thickness of the back-gate, the fin thickness, and the gate-source/drain underlap, respectively. The height of the fin (h_{fin}) is perpendicular to this cross-section and is not shown. The fin height, h_{fin}, acts as the width of the channel. If the front-gate and the back-gate are shorted (tied), the effective channel width is twice the fin height. h_{fin} cannot be changed across the chip, but stronger devices can be built by using an appropriate number of parallel fins in each transistor. Thus, the channel width of a FinFET is given by $W = n_{fin} \times h_{fin}$, where n_{fin} is the number of parallel fins. Since the distance between the parallel fins must be greater than or equal to a technology-specified fin pitch, the fins must be high enough to make the FinFET I_{on} competitive with planar CMOS; i.e., FinFETs should be able to deliver the same I_{on} for an equal area. However, taller fins come at the cost of granularity in the gate strength. In other words, the smallest gates that are usually used in non-critical paths would be too big, which may increase the leakage power of circuits.

The FinFET structure has several advantages over planar CMOS. Although phonon and surface scattering is higher than planar CMOS, the undoped channel of the FinFET eliminates Coulomb scattering due to impurities, resulting in higher electron and hole mobilities overall [26]. Furthermore, the ratio of p-type to n-type mobility is better than CMOS. Unlike CMOS, the threshold voltage is not altered by variations in the source-to-body voltage. This, along with improvement in mobility, paves the way for a longer series of stacked transistors in the pull-up or pull-down networks of logic gates.

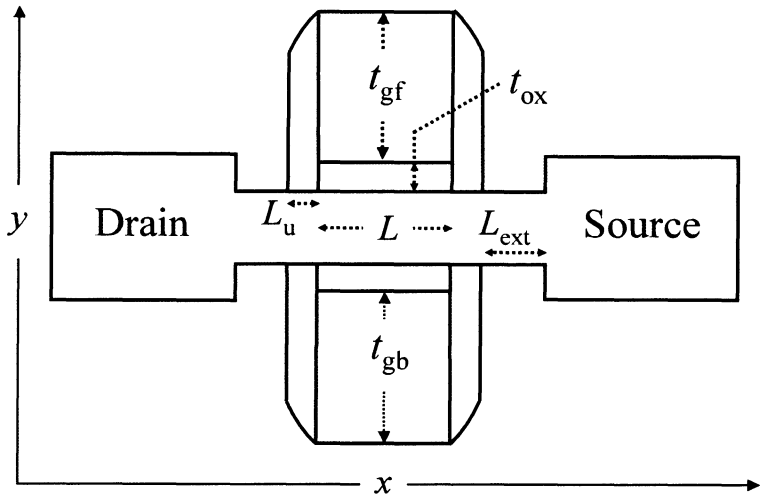

Figure 2.1: 2-D cross section of a typical FinFET

Table 2.1: Physical parameters of 32nm FinFETs

Parameter	Range
t_{ox} of front and back	1-2 nm
source/drain doping	$2 \cdot 10^{20}$
work-function n-type	4.5-4.8 eV
work-function p-type	4.5-4.85 eV
L_u	3-5 nm
gate length (L)	32 nm
h_{fin}	40 nm
t_{Si}	6-12 nm
V_{DD}	0.9 V
t_{gf}	28 nm
t_{gb}	28 nm

Three available models exist for FinFETs: the predictive technology model (PTM) [5], BSIM-MG model [10], and the University of Florida double-gate (UFDG) model. Excellent agreement with physical measurements have been reported for the UFDG model [41]. The UFDG model successfully accounts for quantum mechanical carrier distribution in the

body and channel in both the sub-threshold and strong inversion regions of operation. Furthermore, the UFDG model is a physical model that allows designers to change several design parameters such as fin width, channel length, gate-source/drain underlap, and work-function simultaneously. Subthreshold leakage that is the dominant component of leakage in FinFETs, is rigorously treated within the UFDG model. Note that the UFDG model does not account for the gate leakage in FinFETs. This is not a significant drawback since gate leakage is not the dominant leakage component in FinFETs owing to the presence of a low electric field across the gate.

All simulations reported in this project are performed with the UFDG model. In table I, I report the typical ranges of physical parameters for a 32 nm FinFET technology used in our simulations. Note that all the parameters are in the acceptable range for this technology node. Note also that the designed FinFETs are validated with Sentaurus TCAD simulations to ensure the integrity of the designed FinFETs, as reported in chapter 3.

Chapter 3

Dual-V_{th} independent-gate FinFETs

Independent-gate (IG) FinFETs can be fabricated along with conventional shorted-gate (SG) devices on the same die by removing the top gate region of the FinFET. Since the thickness of the silicon fin is small (1-2 nm), the electrostatic coupling between the gates is high, and the channel formation in one gate is highly dependent on the state of the other gate. In other words, channel formation under a gate is easier if the other gate is already turned on. Furthermore, if the back-gate of an IG FinFET is disabled, not only is no channel formed near the disabled gate, but the threshold voltage of the other gate is also increased. Hence, disabling one gate reduces the drive strength of the transistor by more than half. However, the disabling of one gate may speed up the circuit indirectly, because the input capacitance of devices with disabled back-gates is roughly half of conventional shorted-gate devices. The reduction in the input capacitances reduces the load on the gate that drives them, which makes disabled back-gate FinFETs an attractive option for non-critical circuit paths. Note that the back-gate of n-type and p-type devices are disabled by applying zero and V_{DD}, respectively.

In conventional IG FinFET devices, a channel will be formed if either of the gates is activated. In other words, the device behaves like the OR function; so, they are suitable for merging parallel transistors in pull-up or pull-down logic networks. However, in order to merge series transistors, devices that behave like the AND function are needed. Such a

device is required to have a higher threshold voltage than the regular devices. In IG devices with AND-like behavior, if just one gate is activated, the threshold voltage must be high enough to prevent meaningful channel formation. But, if the other gate is also turned on, fast electrostatic coupling between the two gates must decrease the threshold voltage and enable channel formation. In other words, these high-V_{th} devices must be activated *iff* both their gates are activated in order to be suitable for merging series transistors. Note that high-V_{th} FinFETs cannot be realized by engineering the channel dopant concentration, like [7], because the FinFET channel should be kept undoped to avoid excessive random dopant fluctuations. I will show that high-V_{th} IG FinFETs can be realized by careful selection of FinFET physical parameters without the use of any additional bias voltages. Tuning the gate oxide thickness, the electrode work-function, the silicon thickness, and the gate-source/drain underlap to realize dual-V_{th} devices is thoroughly explored in this chapter. The chapter is concluded with a brief introduction of asymmetric double-gate devices.

3.1 Design of High-V_{th} devices

The physical parameters of high-V_{th} devices must be selected to achieve the following two objectives simultaneously: (a) if only one gate is activated, the current must be as low as possible and (b) if both gates are activated, the current must be as high as possible. The first objective necessitates that the device have a high-threshold voltage. The threshold voltage of a FinFET threshold voltage is approximated by

$$V_{th} = -\phi_{ms} + \frac{Q_D}{C_{ox}} + V_{inv} + V^{QM} - V^{SCE}, \tag{3.1}$$

where ϕ_{ms} is the difference between work-function of electrode and silicon, Q_D is the depletion charge in the channel, C_{ox} is the gate capacitance, V_{inv} is a constant that represents the limited availability of inversion charges in the undoped channel, V_{QM} models the quantum-mechanical increase in the threshold voltage, and V_{SCE} models the short channel effect [8]. Since the transverse electric field is quite low in undoped FinFETs with silicon thickness greater than five nanometers [12], V_{QM} is negligible for the FinFETs considered in this

project with t_{Si} in the 6-12nm range. Q_D is relatively small in undoped or slightly doped channels, hence increasing t_{ox} ($\propto C_{ox}^{-1}$) does not have much effect on threshold voltage. In summary, a high threshold voltage can be achieved only by manipulating the ϕ_{ms} and V_{SCE} terms. Since V_{SCE} is mainly governed by the thickness of the silicon, decreasing t_{Si} improves the short channel effects and hence increases the threshold voltage.

Increasing the threshold voltage is not sufficient to simultaneously achieve objectives (a) and (b). Besides the threshold voltage, it is imperative to manipulate the subthreshold slope in modes (a) and (b). The subthreshold slope S is the logarithm of the slope of the device I-V curve in the subthreshold region and is given by the following equation:

$$S = \frac{\partial V_{GS}}{\partial \log I_{DS}} = \ln 10 \cdot \frac{kT}{q} \cdot \frac{\triangle V_{GS}}{\triangle \psi_{Si}} = 60 \cdot \frac{\triangle V_{GS}}{\triangle \psi_{Si}}, \tag{3.2}$$

where ψ_{Si} is the surface potential at the gate of interest. For the case when one of the gates is deactivated and the other is turned on, meeting (a) requires that S must be as high as possible to decrease I_{on}. The subthreshold slope can be approximated by the following equation in this mode of operation [21]:

$$S = 60 \cdot \frac{t_{Si} + 6t_{ox}}{t_{Si} + 3t_{ox}}. \tag{3.3}$$

Differentiating this equation with respect to t_{ox} yields

$$\frac{\eta_1}{(t_{Si} + 3t_{ox})^2}, \tag{3.4}$$

where η_1 is a positive constant. Since this derivative is always positive, the subthreshold slope S can be increased in this mode by increasing t_{ox} for the device.

For the case when one of the gates is already activated and the other gate is to be turned on, (b) requires that S must be as low as possible to increase the I_{on}. S can be approximated in this mode by [21]:

$$S = 60 \cdot \frac{t_{Si} + 6t_{ox}}{3t_{ox}}. \tag{3.5}$$

Differentiating this equation with respect to t_{ox} yields

$$\frac{-\eta_2}{(3t_{ox})^2}, \tag{3.6}$$

where η_2 is a positive constant. Since the derivative in this mode is always negative, the subthreshold slope S can be decreased in this mode by increasing t_{ox} for the device.

Thus, higher t_{ox} increases S in mode (a), decreases it in mode (b), and helps achieve both objectives simultaneously. However, as Eq. 3.4 and Eq. 3.6 show, the gain from increasing t_{ox} quickly diminishes as t_{ox} increases. In undoped devices, the gate quickly loses control over the channel if t_{ox} is increased aggressively [38]. In fact, the overall leakage first decreases as t_{ox} is increased. Beyond a certain point, however, this trend reverses and leakage current increases due to severe drain-induced barrier lowering (DIBL) effects. Thus, there exists an optimum t_{ox} to obtain minimum leakage, while trying to achieve both objectives (a) and (b).

3.2 The optimum gate underlap

In addition to the work-function, the silicon thickness, and the oxide thickness, it is also necessary to consider the effects of gate-source/drain underlap on the performance of low and high-V_{th} devices. As described in the previous chapters, an optimum underlap is imperative for efficient suppression of short channel effects. Optimizing the amount of underlap has been used in literature to enhance the performance of FinFETs [33, 35]. The effect of underlap on performance can be modeled by a bias-dependent effective channel length. Under weak inversion, the underlap is added to the gate length, which causes a drastic reduction in I_{off}. At high drain-source voltages, the effective channel length is almost the same as the physical channel length resulting in a small reduction in I_{on}. Hence, the amount of underlap must be carefully selected to achieve the highest possible suppression of short channel effects, while keeping I_{on} in its acceptable range.

Besides I_{on}, I_{off}, and drain/source contact resistances, the parasitic gate-source/drain capacitances ($C_{GS/D}$) also strongly depend on the amount of underlap. These parasitic capacitances are caused by inner and outer fringing electric fields and are important in performance optimization of FinFETs [12]. Increasing the underlap separates the gate and

source/drain region further from each other, which reduces the gate parasitic capacitances. Therefore, modifying the gate capacitance enables a trade-off between the power and speed of logic gates. The delay of a logic gate depends on I_{on} and the gate capacitance as

$$t_d \propto \frac{I_{on}}{C_{GS/D}}. \tag{3.7}$$

Hence, increasing the underlap may improve the speed of gates, while counter-intuitively decreasing I_{on}. In the following paragraphs, the electrical characteristics of these devices will be explored.

3.3 Characteristics of low and high-V_{th} devices

Table 3.1: V_{th}, t_{ox}, and electrode work-function (ϕ) of high-V_{th} (H) and low-V_{th} (L) devices in shorted-gate (SG) and disabled back-gate (IG) modes

	t_{ox} (nm)		ϕ (eV)		t_{Si}(nm)		l_u(nm)		V_{th} (V)			
									SG		IG	
	L	H	L	H	L	H	L	H	L	H	L	H
n-type	1	2	4.5	4.8	12	6	3	5	0.18	0.3	0.54	0.97
p-type	1	2	4.85	4.5	12	6	3	5	0.09	0.16	0.5	0.95

The t_{ox}, t_{Si}, L_U, and electrode work-function (ϕ) of p-type and n-type FinFETs were swept over their ranges in UFDG to obtain the optimum combination of these parameters, summarized in Table 3.1. The threshold voltage is defined as the gate-source voltage necessary to obtain $I_{DS} = 100\text{nA}/\mu\text{m}$, when $V_{DS} = 50\text{mV}$ [6]. Threshold voltage of both high-V_{th} and low-V_{th} FinFETs in shorted-gate (SG) and disabled back-gate modes (IG) are also listed in Table 3.1. As expected, the threshold voltage difference between SG and IG modes is considerably higher in high-V_{th} devices than low-V_{th} devices. This difference is explained by the fact that in the IG mode of low-V_{th} FinFETs, the inversion layer can be easily formed. This channel shields further gate-to-gate coupling, and hence a huge drop in

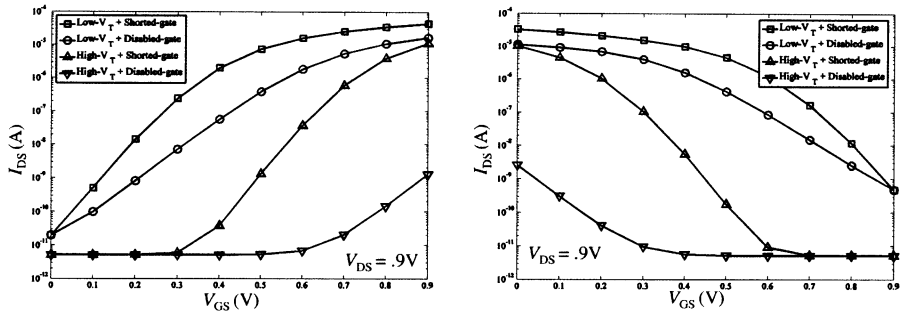

Figure 3.1: I-V curves of (a) n-type and (b) p-type high-V_{th} and low-V_{th} FinFETs in shorted-gate and disabled back-gate modes

threshold voltage is not seen in this mode [7]. In contrast to low-V_{th} devices, no inversion layer can be formed in the IG mode of high-V_{th} FinFETs. Thus, when both gates in a high-V_{th} FinFET are simultaneously on, the strong electrostatic coupling between them creates an inversion layer and produces an acceptable I_{on}. Further, the t_{Si} of high-V_{th} devices is chosen to be smaller to enhance this effect.

SPICE simulations with the UFDG model have shown that using the physical parameters in Table 3.1 results in acceptable performance with minimum static leakage in both high-V_{th} and low-V_{th} devices. I-V curves of n-type and p-type FinFETs for four configurations: low-V_{th} shorted-gate, low-V_{th} disabled back-gate, high-V_{th} shorted-gate, and high-V_{th} disabled back-gate are shown in Fig. 3.1. Static leakage of these modes is also in the range of a recently manufactured FinFET [20].

All the n-type and p-type devices were simulated and validated with the Sentaurus design suite [3] to verify the integrity of the proposed methodology. The 2-D FinFET structure shown in Fig. 2.1 [28] was used for the simulations. In Sentaurus, the drift-diffusion mobility and density-gradient quantum correction models were enabled. Since FinFETs consist of ultra-thin slabs, quantum correction is also necessary and this feature was enabled. The mobility models also include mobility degradation due to scattering and high lateral and perpendicular electric fields. Additional steps to calibrate the Sentaurus

tools for a completely accurate simulation of FinFETs is discussed in [19]. The results of simulations are compared with UFDG in Fig. 3.2 for n-type devices. The figure confirms the underlying hypothesis that high-V_{th} devices with AND-like behavior and manageable leakage is physically possible in FinFETs.

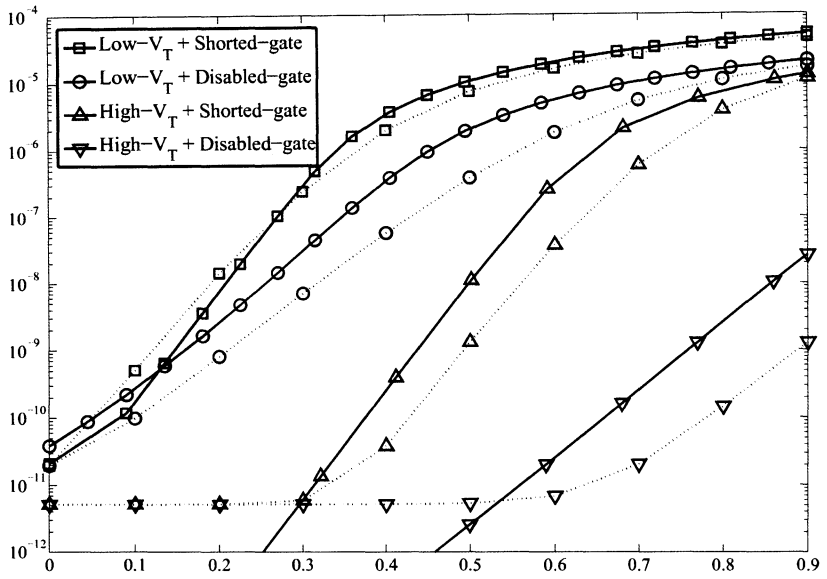

Figure 3.2: UFDG (dotted lines) and TCAD (solid lines) simulations of n-type devices are compared.

From the *I-V* curves, it is clear that if just one gate is activated in high-V_{th} transistors, the current is low enough that the transistor can be considered to be in the off-state. Thus, these devices will still have low static leakage. In the case of low-V_{th} devices, if just one of the gates is activated, the device can be considered to be in the on-state. However, the device current drive is around 60% less than the current drive of shorted-gate devices. Lower current drive makes the gates with merged series or parallel transistors slower than gates with conventional shorted-gate transistors and limits their use to non-critical paths.

3.4 Fabrication issues of high-V_{th} devices

Note that technologically, fabricating multiple work-functions requires two additional steps to mask and etch the gate material. It has been reported [17] that the work-function of TiN gate on HfO_2 oxide is 4.83 eV and the work-function of TiN gate on SiO_2/HfO_2 can be set to 4.54 eV by modulating the the SiO_2 thickness. These values are very close to the selected work-functions in table 3.1. It is also possible to have two values for t_{ox}; even FinFETs with asymmetric front and back t_{ox} have been recently reported [20]. Gate underlap engineering has also been considered as an attractive design option in FinFETs [16].

The proposed high-V_{th} IG devices are robust to parametric variations in oxide thickness and do not lose their AND-type functionality. Variations in oxide thickness degrades sub-threshold slope and changes the gate capacitance, but does not have a huge impact on the V_{th} of these devices due to negligible inversion charge Q_D (see Eq. 3.1). Further, FinFETs are known to be less susceptible to variations in physical parameters in comparison to planar CMOS, with the exception of variations in t_{Si} [40]. Process variations in t_{Si} influence the device characteristics by means of quantum-mechanical effects. However, the values of t_{Si} used in this project are high enough to render the conversion probability of a high-V_{th} device to a low-V_{th} device negligible.

In the next chapter, I describe new circuit styles and logic gates based on these dual-V_{th} FinFETs.

3.5 Asymmetric double-gate devices

Many innovative circuits can be designed with asymmetric independent-gate double-gate FinFETs. A successful fabrication of these devices are reported in [20]. The authors fabricated asymmetric double-gate devices with additional masks, etching and polishing steps.

In asymmetric FinFETs, the back-gate should have a very high threshold voltage, in such a way that no channel is formed under the back-gate. In this configuration, the back-gate can be used to modulate the threshold voltage of the front-gate, statically or dynam-

ically. Authors in [11] have exploited this feature of asymmetric devices to increase the manufacturing yield of SRAM cells.

The effectiveness of asymmetric independent-gate FinFETs depends on how much the front-gate V_{th} can be changed by modulating the back-gate voltage. In other words, the sensitivity of front-gate threshold voltage (V_{th_f}) in respect to the back-gate voltage V_{bg} should be as high as possible. Front-gate threshold voltage can be approximated to have a linear dependence on V_{bg} [15]:

$$V_{th\_f} = V_{th} - r\dot{V}_{bg}, \tag{3.8}$$

where r is the gate-to-gate coupling factor and is given by:

$$r = \frac{3t_{ox\_f}}{3t_{ox\_b} + t_{Si}}, \tag{3.9}$$

where t_{ox_f} and t_{ox_b} are the front-gate and back-gate oxide thicknesses, respectively.

From Eq. 3.9, a high sensitivity to V_{bg} necessitates a high t_{ox_f} and a low t_{ox_b}. But, recall from Eq. 3.3 that t_{ox_b} should be high enough to increase the subthreshold slope and reduce the leakage current. This contradiction puts yet another constraint on t_{ox_b}, which calls for a careful calibration of this value.

Chapter 4

Logic design with dual-V_{th} FinFETs

In this chapter, the effects of merging series and parallel devices are first analyzed. Without loss of generality, two special cases will be further investigated: logic gates with two devices in either pull-down or pull-up networks and Boolean series-parallel networks with four inputs. Then, novel logic gates are introduced by defactoring the Boolean equations in either pull-down or pull-up networks. All experiments in this chapter have been performed with $V_{DD} = 0.9$V. The circuit symbols of dual-V_{th} FinFETs in SG and IG configurations are shown in Fig. 4.1.

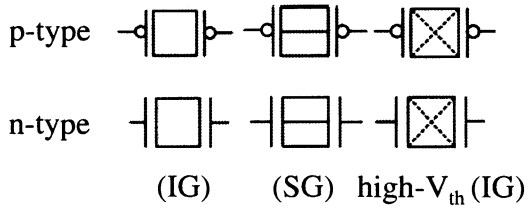

Figure 4.1: Symbols for independent-gate (IG) and shorted-gate (SG) low-V_{th} and high-V_{th} n-type and p-type double-gate FinFETs. The dotted-X sign in high-V_{th} devices denotes their AND-like behavior.

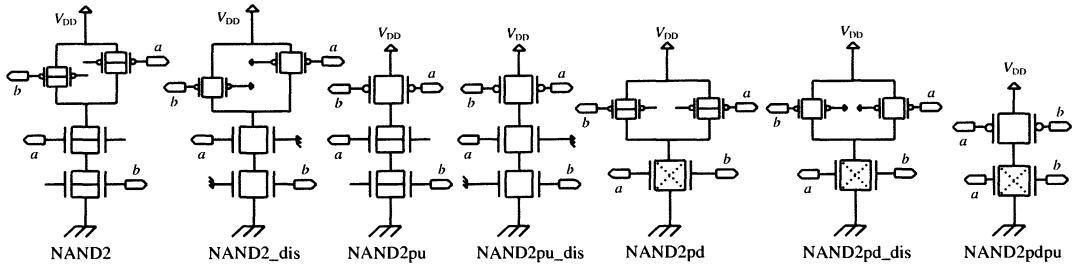

Figure 4.2: NAND2 gates designed by disabling the back-gates and merging parallel or series transistors

Table 4.1: Characteristics of conventional and novel NAND gates.

Gate	Intrinsic (ps)		FO4 (ps)		I_{off} (pA); *ba*, *b* is the MSB					C_{in}(aF)	No.
	T_{phl}	T_{plh}	T_{phl}	T_{plh}	00	01	10	11	Avg.		Trans.
NAND2	3.9	2.2	8.4	7.3	6.3	19	19.7	943	246	83	4
NAND2_dis	5.6	4.1	13.7	14	6.3	14.4	19.7	943	245	48	4
NAND2pu	2.5	4.2	6.3	12.9	6	19	19.7	471	129	61	3
NAND2pu_dis	5	3.4	13	14.2	6	14.4	19.7	471	128	46	3
NAND2pd	5.1	2.2	12.5	7.1	5	1284	1284	942	878	52	3
NAND2pd_dis	4.7	3.7	9.5	11.7	5	1284	1284	942	878	33	3
NAND2pdpu	4.1	2.9	9	11.6	5	1284	1284	761	761	31	2

4.1 Merging and back-gate disabling

Fig. 4.2 presents all possible realizations of a NAND gate with two inputs. NAND2 is the conventional 2-input gate that uses low-V_{th} FinFETs in shorted-gate configuration. NAND2_dis is derived by disabling the back-gates of all devices in the conventional NAND2 gate. NAND2pu is the result of merging two parallel transistors and replacing it by one low-V_{th} FinFET in the pull-up network of NAND2. NAND2pu_dis is derived by disabling the back-gates of pull-down devices of NAND2pu. The two series transistors in the pull-down network of the conventional NAND2 gate can be replaced by one high-V_{th} transistor to realize NAND2pd. NAND2pd_dis is derived by disabling the back-gates of

pull-up devices in NAND2pd. Finally, one can merge both series and parallel transistors in the conventional NAND2 gate to realize NAND2pdpu. The first four figures of fig. 4.2 have been proposed in literature [9,25] for FinFET devices with some minor modifications. The last three gates can only be realized with the proposed high-V_{th} devices. In Table 4.1, low-to-high (T_{plh}) and high-to-low (T_{phl}) transition delays, average input capacitance (C_{in})*, and the static power consumption of these gates in four possible input configurations are reported. It should be noted that the static leakage current can vary by more than one order of magnitude depending on the input to the gates. For example, the static leakage of NAND2 in its four input configurations is 6.3pA, 19pA, 19.7pA, and 943pA, and the average as recorded in Table 4.1 is 245pA. Thus, it is necessary to simulate the gates in all input configurations in order to estimate static power.

From the table, it is seen that merging parallel transistors has a negligible effect on static power consumption. However, merging series transistors with an IG high-V_{th} FinFET increases average static power by an order of magnitude. This increase is because for some input patterns one of the gates is active while the other gate is inactive. Although the high-V_{th} FinFET is supposed to be in the off-state, the activation of one of its gates reduces the threshold voltage and results in an increase in static power consumption. Since the FinFETs were engineered with adequate L_U and t_{Si}:L ratios, the worst-case leakage current of 0.88nA is still comparable to 2.9nA for an equivalent planar 32nm CMOS technology [5]. Also, note that both series and parallel transistor merging and back-gate disabling results in a circuit with higher worst-case transition delay.

The gates realized by merging parallel transistors or disabling the back-gate generally have less input capacitance, leakage power, and gate overdrive. The input capacitance of the gate can also be further reduced by merging the series transistors. The series merger may even help to balance the relative drive strength of the pull-down and pull-up networks,

*UFDG is based on Berkeley SPICE3 and does not have a command for capacitance extraction. An AC voltage source should be placed at the node of interest to measure the imaginary component of current at the node. The capacitance is calculated using the following equation: $C = \frac{I}{2\pi f V}$, where f is the frequency of the voltage source.

which results in the reduction in the worst-case delay of the gate. The worst-case delay of NAND2_pu is 4.5ps, while it is 4.1ps for NAND2_pdpu. The T_{plh} and T_{phl} of NAND2pu are not balanced and a race exists between the pull-up and pull-down networks while it switches. On the other hand, merging of cascaded n-type devices lessens the drive power of the pull-down network and mitigates this problem [25].

4.2 Novel dual-V_{th} logic gates

The availability of dual-V_{th} IG FinFETs motivates design of a new class of compact logic gates with higher expressive power and flexibility. Both high-V_{th} and low-V_{th} transistors are utilized in both the pull-up and pull-down networks. High-V_{th} IG devices inherently act as an AND function. They will have low resistance if both their inputs are on. Thus, they can be considered as a network with two series transistors. With the same reasoning, low-V_{th} IG FinFETs can be represented by two parallel transistors in the Boolean network. The rules for static logic require that the pull-down network should be the dual of the pull-up network. Hence, if a high-V_{th} transistor is used in pull-down network with inputs a and b, the corresponding device in the pull-up network is a low-V_{th} device with inputs a and b, and vice versa.

Starting from a structure that resembles the NAND2 gate in Fig. 4.3, low-V_{th} transistors are used in the pull-down network and high-V_{th} transistors in the pull-up network. The stacked devices show higher resistance than the parallel devices. Therefore, it is preferable to use the stronger low-V_{th} devices in series structures. This consideration makes balancing the pull-up and pull-down networks easier during design. For the logic gate shown in Fig. 4.3, the pull-down network will be activated *iff* the Boolean function of Eq. 4.1 holds:

$$\text{PD} = (a + b) * (c + d). \tag{4.1}$$

Similarly, the pull-up network will be activated *iff* Eq. 4.2 holds:

$$\text{PU} = (a' * b') + (c' * d'). \tag{4.2}$$

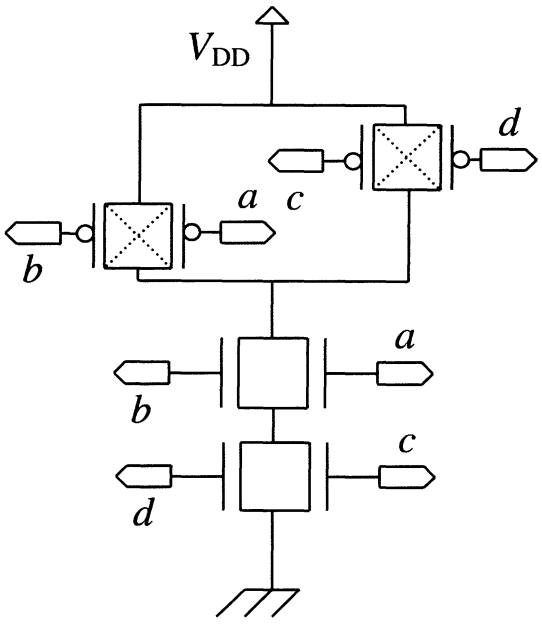

Figure 4.3: Novel implementation of $[(a + b) * (c + d)]'$

These two equations are Boolean complements and they will never be true simultaneously. Thus, the logic gate represented in Fig. 4.3 is a static logic gate. Other compact Boolean functions can be realized from this structure. For example, if the inputs c and d are replaced by the complements of the inputs a and b, (i.e., $c = a'$ and $d = b'$), the gate becomes one of the most compact implementations of XNOR logic. This structure is flexible and can easily realize the XOR function when b, c, and d are replaced by b', a', and b.

Independent-gate dual-V_{th} FinFETs increase the available options in logic circuit design. For example, it is possible to implement 12 unique Boolean functions using only four transistors as follows. Since the pull-up network is the dual of the pull-down network, it is sufficient to enumerate all the unique configurations in the pull-down network. A logic gate with two IG transistors in the pull-down network can have two, three, or four inputs. With two inputs, all the devices should be SG low-V_{th} devices; i.e., there is only *one* option. With three inputs, one of the FinFETs must be an IG FinFET and the other must be a SG FinFET. *Two* options exist for the IG device: a high-V_{th} or a low-V_{th} device. Finally, with

four inputs, all devices must be IG, and *three* possible options exist: both low-V_{th}, both high-V_{th}, and a low-V_{th} along with a high-V_{th} FinFET. Thus, six unique combinations of dual-V_{th} FinFETs exist. Finally, since the two transistors in the pull-down network can be in series or in parallel, a total of 12 unique Boolean functions can be realized using four IG dual-V_{th} FinFETs.

The number of logic gates that can be implemented using dual-V_{th} FinFETs increases exponentially with the number of transistors used in the gate. For example, if the gate has six transistors (three each in the pull-down and pull-up network), 56 unique gates can be realized. Although some of the 56 gates are functionally equivalent, they are structurally different. Some of them are not as competitive in performance as other members of this logic family. This lower performance is mostly due to a large difference between low-to-high and high-to-low transition delay that occurs when high-V_{th} devices are stacked in either the pull-down or pull-up network.

Since static CMOS logic is inverting, the delay where several gates are cascaded usually reduces skew between T_{phl} and T_{plh}. This inverting nature enables the synthesis tool to use skewed gates during its optimization. It is also possible to address the skew by increasing the number of fins in the stacked high-V_{th} devices. However, it may result in a large increase in input capacitance of the gate, such that the fanout-of-four delay may remain almost unchanged. In the next part, I will use an example to illustrate design rules that can be used to further optimize the performance of dual-V_{th} logic gates.

4.3 Case study of Boolean networks with four inputs

The number of possible non-isomorphic series-parallel networks in the pull-down network that can be implemented using four devices is ten. For the rest of this discussion, I assume that both the pull-up and the pull-down networks are simultaneously modified; i.e., a series (parallel) merger in the pull-up (pull-down) network is mirrored by a parallel (series) merger in the pull-down (pull-up) network. More than one merging can be performed on

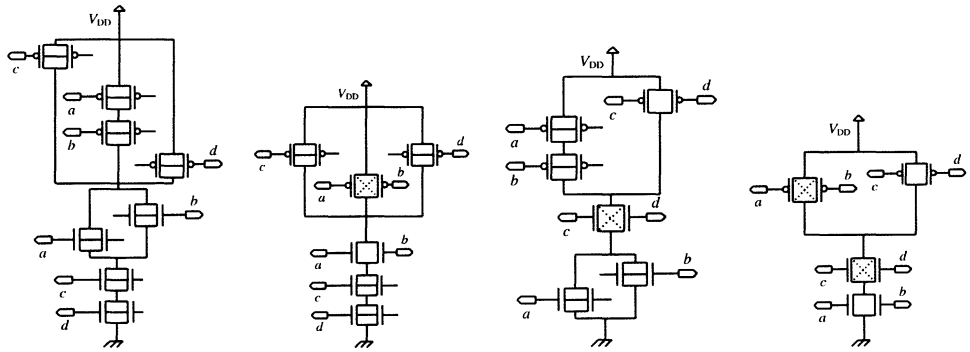

Figure 4.4: Four possible implementations of $[(a + b) * c * d]'$

Table 4.2: Characteristics of conventional and novel implementations of $[(a + b) * c * d]'$

Gate	Intrinsic (ps)		FO4 (ps)		Average I_{off}(pA)	C_{in}(aF)	No. Trans.
	T_{phl}	T_{plh}	T_{phl}	T_{plh}			
(a)	5.2	4.8	12.5	13.8	256	73	8
(b)	6.2	5.1	12.8	13.2	508	50	6
(c)	7.5	4.4	16.9	13.6	643	50	6
(d)	6.1	5.1	17.5	12.1	894	32	4

some of these networks, thereby increasing the available flexibility in logic design. Without loss of generality, I investigate the available options for implementing the network that implements $[(a + b) * c * d]'$. Fig. 4.4 shows four possible implementations of this logic function. Worst case T_{phl} and T_{plh} with average I_{off} and input capacitance of these implementations are also listed in table 4.2. The first implementation only uses shorted-gate low-V_{th} devices. In the second and third implementation, only one parallel or series merger is performed on the pull-up and pull-down networks, respectively. The last implementation applies one series and one parallel merger in both the pull-up and pull-down network and requires only four transistors.

Table 4.2 shows that considerable reduction in input capacitance of gates can be

achieved by merging series or parallel devices. The reduction in input capacitance comes with a slight deterioration in transition delays, which can be tolerated if the gate is not on a critical path. Despite the fact that all devices in the fourth configuration have been merged, this configuration still has better intrinsic T_{phl} than the second and third configurations, because the pull-up and pull-down networks have become more balanced in this configuration. Also, the high-V_{th} device in the pull-down network of the third and fourth configurations has been moved up closer to the output pin. This design rule helps reduce the worst-case T_{phl} and T_{plh} delay of the third configuration from 10.6ps and 7.2ps to 7.5ps and 4.4ps, respectively. The next chapter discusses a method to realize a new class of logic gates by defactoring the Boolean equations that govern the pull-down or pull-up networks.

4.4 Novel gates by defactoring the Boolean function

It is also possible to use dual-V_{th} FinFETs to realize compact logic gates by using defactorization of Boolean expressions. Consider the logic network in Fig. 4.5(a) that conducts between nodes x and y *iff* $[a + (b * c)]$ holds true. The logic network on the left in the figure is realized using conventional FinFETs, whereas the logic network on the right is realized using dual-V_{th} independent-gate FinFETs. The Boolean function of the logic network on the right, $[(a + b) * (a + c)]$, is derived by defactoring the original Boolean equation $[a + (b * c)]$. Similarly, Fig. 4.5(b) illustrates the application of the same defactoring procedure to $[a * (b + c)]$. The defactored logic $[(a * b) + (a * c)]$ is implemented on the right in the figure by using high-V_{th} devices.

Although these new realizations may increase the worst-case transition delays, the new gates will require fewer fins and the input capacitance seen from inputs b and c is reduced by roughly 50%. As a result, defactoring can be used to realize novel logic gates based on dual-V_{th} FinFETs. These gates have the advantages of low power and low area, and they find ready use on non-critical paths. Furthermore, as illustrated in Fig. 4.5(b), defactoring allows the reduction of the number of series-stacked transistors from two to one. This

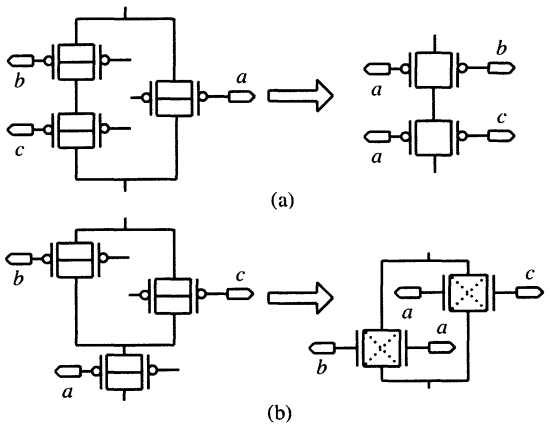

Figure 4.5: Novel logic gates by defactoring the Boolean function by using (a) low-V_{th} (b) high-V_{th} FinFETs

Table 4.3: Characteristics of conventional and defactored implementations of $[a * (b + c)]'$

Gate	Intrinsic (ps)		FO4 (ps)		I_{off} (pA); cba, c is the MSB									C_{in}(aF)	No. Trans.
	T_{phl}	T_{plh}	T_{phl}	T_{plh}	000	001	010	011	100	101	110	111	Avg.		
Conventional	5.4	4.6	10	13.5	8	39	19	933	19	942	19	628	326	78	6
Pull-down	7.7	3.3	15.2	12	10	2565	1289	933	1289	943	2569	628	1278	53	5
Pull-up	4.3	15.2	7.2	29.2	8	39	19	377	19	471	19	156	138	65	5
Both	5.9	6	10.5	19.5	10	2565	1289	378	1290	471	2570	157	1091	40	4

cannot be achieved using the conventional parallel merge transformation of the transistors b and c using a low-V_{th} FinFET, as described in literature [4].

The tradeoffs of defactoring are discussed using the following example. If the Boolean function $[a * (b + c)]'$ is implemented with conventional shorted-gate FinFETs, its pull-up and pull-down networks are illustrated by the figures on the the left in Fig. 4.5(a) and Fig. 4.5(b), respectively. Note that the n-type FinFETs will have to be replaced by p-type FinFETs in Fig. 4.5(a). The defactoring procedure described above can be applied to either

its pull-down network, its pull-up network, or both. Table 4.3 compares the characteristics of the conventional implementation of $[a * (b + c)]'$ with the implementations obtained by defactoring transformations. The table shows that the full defactoring transformation can reduce input capacitance by up to 47%.

Intrinsic T_{phl} increases from 5.4ps to 7.7ps when only the pull-down network is defactored, as illustrated in Fig. 4.5(b), since the independent-gate FinFETs in the pull-down network are replaced by high-V_{th} FinFETs. On the other hand, intrinsic T_{plh} increases from 4.6ps to 15.2ps when only the pull-up network is defactored, as illustrated in Fig. 4.5(a). It is observed that defactoring only the pull-up network has a more adverse effect on the worst-case transition delay. The reason can be attributed to the fact that the number of stacked devices remains the same when only the pull-up network is defactored. However, the number of series-stacked transistors is reduced from two to one when only the pull-down network is defactored, which has a mitigating effect on transition delays.

It is also observed that defactoring only the pull-up (pull-down) network has a positive impact on the transition delay of the pull-down (pull-up) network. For example, the T_{plh} of the gate where only the pull-down network is defactored is reduced from 4.6ps to 3.3ps. This reduction is explained by the fact that the pull-up network is relatively stronger than the defactored pull-down network. Similarly, the T_{phl} of the gate where only the pull-up network is defactored is reduced from 5.4ps to 4.3ps. The pull-down network is relatively stronger than the defactored pull-up network, which explains the reduction in delay. This effect can be mitigated by defactoring both the networks simultaneously to balance their strength and reduce contention during switching. When the pull-up and pull-down networks are simultaneously defactored, the T_{phl} and T_{plh} increase from 5.4ps to 5.9ps and 4.6 ps to 6ps, respectively, over the conventional gate with independent-gate FinFETs.

It is also observed that the effect of defactoring on FO4 delays is less than its effect on intrinsic delays. For example, defactoring only the pull-up network increases the intrinsic T_{plh} by 230% (from 4.6ps to 15.2ps), while it increases the FO4 T_{plh} by 123% (from 13.5 ps to 29.2ps). This difference is to be expected because FO4 delay is estimated by simu-

lating gates that drive four identical copies. In this case, the fanout gates have lower input capacitance after the defactoring transformation. The possible application of these gates in sequential elements is explored next.

4.5 Sequential elements with high–V_{th} devices

Sequential elements are one of the most sensitive elements of integrated circuits. I introduce novel high-V_{th} devices with the goal of providing more flexibility in design of low-power combinational circuits. Gates realized with dual-V_{th} FinFETs are inherently slower, but their noise and parametric variation is not fundamentally different from gates based on conventional FinFETs. Although there are some works [32, 36] that report improved performance in sequential elements with the use of low-V_{th} independent-gate FinFETs, they are mostly used to weaken "the feedback loop" in flip-flops and latches. As a result, it is the position of the authors that the application of the dual-V_{th} devices will remain limited in the design of sequential elements.

4.6 The effects of process variation

It is important to quantitatively measure the effects of process variation and faults on the performance of FinFET devices. Because, these effects are very important in the state of the art CMOS process and have motivated the search for alternatives to planar CMOS devices in the first place. Parametric variations on chip are classified to catastrophic and process variations. First, I look into the first category. In catastrophic variations or faults, part of the device fails to work. For example, a node may permanently stuck at zero or one. It has been shown in literature [29] that the nature of majority of faults in FinFETs are the same as in faults in planar CMOS. In other words, usual CMOS stuck at fault models and tests can still be of use in modeling of FinFET faults. However, double-gate FinFETs have some faults that cannot be captured by available models. These unique faults happen whenever

the connection between the back-gate and the front-gate in the shorted-gate FinFET is cut open. In this case, the device may remain operational although with a different leakage and delay performance. The performance of a FinFET with a floating back-gate depends on the voltage of floating gate. The voltage of a floating wire or gate connection is a random variable and cannot be predicted beforehand. This random voltage on the floating node makes the FinFET to have an undetermined performance. If the voltage is higher than the threshold voltage, the device will stuck open. In meanwhile, if it is less than the threshold voltage, the device remains operational but with a different leakage and delay.

Without loss of generality, I confine my discussion to n-type devices. In n-type device, if the voltage on the floating node is positive and less than the threshold voltage, the device operates with a higher T_{plh} and a lower T_{phl} delay. In the case of a negative voltage, the delay is higher no matter what the transition is. This random behavior of stuck open faults in FinFETs necessitates ammendments to the CMOS faults models in order to have full coverage in test of FinFET chips.

Not all variations are catastrophic, most of them just make the delay and leakage performance of devices to deviate from their nominal values. In the rest of this section, I discuss methods that quantitatively measure the effects of process variations on leakage and delay.

Since the device on-current can be approximated to have a linear dependence on its physical parameters, the statistical average of the on-current will be the same as its nominal value under process variations. However, this approximation does not hold for the off-current (leakage) of the device, since, the leakage current has an exponential dependence on its physical parameters. In other words, leaky devices contribute to the bulk of the statistical average, and hence the average leakage becomes higher than the nominal leakage.

I simulated the leakage current of the proposed devices using Monte-Carlo simulations. The main sources of performance variations in FinFETs are thickness of silicon, thickness of oxide, fin height, and channel length [18]. Since leakage has a linear dependence on fin height, variations in the fin height are not considered in this project. The variations in the remaining variables are approximated to have Gaussian distributions in which their 3σ

equals to 10% of their corresponding nominal values[†]. In undoped devices with a length of less than 15 nm, the unwanted presence of a few dopants in the channel is enough to effectively influence the threshold voltage, and the resulting distribution of the threshold voltage would not be even Gaussian [37]. Since, the channel length considered in this work is 32 nm, I do not consider the random dopant fluctuations in our simulations.

From the Monte-Carlo simulations, I observed that the average leakage current of the n-type and p-type devices is roughly 5% higher than their nominal values. The statistical average of leakage of FinFETs with multiple fins is the average leakage one fin multiplied by the number of fins. This leakage calculation follows from the fact that a FinFET with multiple fins is multiple single-fin FinFETs in parallel, and statistical average operation is linear. So, there is no need to resort to the methods advocated in [13] to measure the statistical average of FinFETs with multiple fins.

In logic gates, the leakage path from V_{DD} to the ground consists of two or more n-type or p-type devices in which some of them are in linear mode while the rest are in their non-linear mode. Therefore, the effects of non-linearity is less pronounced in the leakage of the logic cells. This observation has been confirmed by Monte-Carlo simulations over all combinations of inputs to the logic gates, which show that the statistical average of leakage of the cells is higher than their nominal value by 2% to 3%.

In the next chapter, the libraries provided to the synthesis tool use the statistical average for leakage power of each cell, and not the nominal value. Using the statistical average makes the leakage analysis more accurate. The savings in total power consumption and number of fins that can be achieved by using these optimized gates in combinational circuits are summarized in the next chapter.

[†]It was observed that UFDG becomes unstable if different values are selected for the back-gate and front-gate oxide thicknesses. Thus, I assumed that oxide thicknesses of back-gate and front-gate are perfectly correlated.

Chapter 5

Results and conclusions

This chapter presents the results for improvements in the number of fins and power consumption that the proposed circuit innovations offer and compares these results to previously published work. In the first step of implementation, logical effort [30] parameters of all novel and conventional gates are extracted using rigorous UFDG SPICE simulations. They consist of input and output capacitances, intrinsic delay, fanout-of-four delay, rise and fall resistance, and statistical average of leakage power over all input vector permutations. In the next step, three technology libraries are generated using the extracted parameters. They are called basic, previous work, and complete libraries:

1. Basic library: It is the simplest library and contains only the conventional gates, i.e., shorted-gate NOT, NAND2, NOR2, NAND3, NOR3, AND_OR, OR_AND, etc.

2. Previous work library: In addition to the gates from the basic library, this library with 41 cells contains the logic gates that are realized by merging parallel transistors or disabling the back-gate as proposed in prior work [9,25].

3. Complete library: This library with 135 cells uses high-V_{th} devices along with regular low-V_{th} devices, and contains all the gates that are realized by merge series or parallel transformation, along with the gates realized by defactoring the Boolean equations. This library is a super-set of the two previous libraries.

Table 5.1: Static power (nW), dynamic power (μW), and number of fins of sixteen benchmarks from the ISCAS and OpenSPARC benchmarks are listed. They are mapped using three different technology libraries: basic, previous work, and complete.

Circuit	No. Cells	Basic			Previous work			Complete		
		Power		No. Fins	Power		No. Fins	Power		No. Fins
		Dyn‡ (μW)	Stat (nW)		Dyn‡ (μW)	Stat (nW)		Dyn‡ (μW)	Stat (nW)	
b9	66	1.4	96.2	296	1.2	84.3	255	0.9	157.4	203
C880	232	6.0	361.9	1124	4.3	307.3	896	3.3	715.2	711
C1908	262	6.7	426.9	1340	5.3	388.5	1093	3.8	816.1	881
C499	310	9.9	508.6	1414	7.8	391.0	1128	6.6	899.8	901
C1355	314	6.6	396.1	1128	5.3	315.6	899	3.7	630.4	760
dalu	365	6.7	668.7	2202	5.4	530.9	1785	3.5	1331.0	1363
C3540	493	10.8	773.5	2660	8.9	629.4	2128	6.2	1652.0	1658
sparc_ifu_errctl	1208	22.1	1764.0	5452	18.0	1522.0	4424	14.2	2901.0	3658
C7552	1210	37.3	2058.0	5874	29.7	1635.0	4673	20.1	4378.0	3595
tlu_hyperv	1302	28.8	2117.0	6436	23.3	1988.0	5108	16.9	4617.0	3974
sparc_ifu_fcl	1548	32.0	2317.0	7368	25.5	1935.0	5881	19.4	3767.0	4737
sparc_exu_ecl	1761	36.2	2685.0	7854	29.2	2222.0	6240	23.9	4387.0	5221
sparc_ifu_ifqdp	2158	51.4	3343.0	10492	40.6	2293.0	8135	30.5	6135.0	6470
sparc_ifu_errdp	2979	61.8	4776.0	14872	48.7	3759.0	11611	35.6	9023.0	8923
C6288	3223	131.9	6424.0	17668	117.4	5575.0	16216	65.5	6287.0	10266
sparc_exu_byp	4482	116.2	7628.0	25140	91.8	5934.0	19504	64.6	15650.0	14681
Average	–	35.4	2271.5	6957.5	28.9	1844.4	5623.5	19.9	3959.2	4250.1

‡ Dynamic power of all circuits is estimated at 2GHz. Simulations are performed at 75 °C.

Each gate is represented in the libraries by four different strengths, i.e., 1X, 2X, 3X, and 4X. The strength of FinFET gates can be increased by adding parallel fins in each of its transistors. Therefore, FinFET gate sizing is inherently a discrete optimization problem, and heuristics have been proposed in [31] to tackle this problem. Synopsys Design Compiler was used to synthesize and map 16 ISCAS and OpenSPARC benchmarks using

these three libraries. It is necessary to estimate the dynamic frequency of all circuits at the same frequency in order to have a meaningful comparison between them. Thus, all circuits are synthesized to meet a timing goal of 2.5 GHz, and the dynamic power of all circuits is estimated at a frequency of 2 GHz. This difference between the frequency of synthesis and power calculation was adopted to mirror the common practice of guard-banding against process variations. In the absence of input traces, dynamic power is estimated by assuming that the the signal activity factor at all the primary inputs is 10%. From the primary inputs, the activity factor of all other gates in the circuit is estimated by Monte-Carlo logic simulations. This is implemented by adding modules to ABC [1]. As mentioned earlier, the static power consumption can differ by more than one order of magnitude depending on the input signals applied to the gate. Thus, each cell is simulated in all its input configurations and the average over all configurations is recorded in the Synopsys libraries.

Since there is no available tool to place and route the FinFET circuits, the number of fins is selected as an indicator of cell area. If an independent-gate FinFET is used in a cell, the cell area will be increased due to routing complexity incurred by additional contacts. However, since the fin count is reduced substantially by using the complete library and from previous similar works [9], I predict that the area improvement will still hold true for the place-and-routed circuits. The first and second columns of table 5 give the name of the circuit and the number of cells in the circuit when it is synthesized with the basic library. This number gives a good estimate of the original circuit size. The number of fins, leakage power, and dynamic power are listed in Table 5 for each circuit after technology mapping with the basic, previous work, and complete libraries.

The overall trend of results indicates that the previous work library provides limited reduction in dynamic power or number of fins. However, the complete library provides larger reductions in dynamic power. This reduction is due to inclusion of novel logic gates designed with both low-V_{th} and high-V_{th} devices in the complete library. The table shows that the static power consumption of circuits synthesized with the complete library is 2–3$\times$ higher than the circuits synthesized with the basic library. This increase in static power

comes from higher leakage of high-V_{th} gates in some of their input configurations. However, the reduction in dynamic power consumption in circuits synthesized with the complete library easily compensates for this increase in static power. On average, the complete library reduces total power and number of fins by 36% and 37%, respectively, over the basic library based on conventional shorted-gate FinFETs in 32nm technology. On the other hand, the previous work library achieves 20% and 21% reduction in total power and number of fins, respectively, over the basic library based on shorted-gate FinFETs in 32nm technology.

5.1 Discussions about temperature and frequency

In this report, new logic gates are proposed to achieve lower dynamic power and area consumption. This improvement comes at the cost of additional leakage power. Therefore, the effective usage of these novel gates depends on the relative contribution of leakage power to the total power consumption. One of the important factors determining leakage current is the operating temperature. As temperature increases, the leakage power increases exponentially, which potentially reduces the effectiveness of the proposed gates. For example, the simulations in Table 5 were performed at 75 °C, but if they had been performed at a lower temperature of 27 °C (the SPICE default), the reduction in total power consumption would have increased from 36% to 39%. Thus, it is recommended to simulate the circuits at a higher temperature to capture the worst case leakage power. Increasing the temperature also has a negative effect on dynamic power. The gates become slower at the higher temperature, and the synthesis tool picks slightly larger logic gates for critical paths. Synthesizing the circuit with larger gates increases the dynamic power, nevertheless, the dominant effect at higher temperatures is the increase in leakage power.

The savings in the power consumption also depend on the operating frequency, since dynamic power has a linear dependence on frequency. Having novel logic gates in the synthesis libraries results in a higher leakage power and lower dynamic power, thus the

effectiveness of novel gates depends on the relative contribution of the dynamic power to the total power consumption. As frequency decreases, the contribution of dynamic power is reduced, thus the novel dual-V_{th} gates will be less effective in reducing the total power consumption. Table 5.2 compares the relative total power savings of the previous work and complete libraries at three different frequencies. As the frequency decreases, the power savings from the previous work library remain almost constant, while the savings from the complete library decrease. The table shows that the complete library will lose its competitive edge in terms of the total power consumption at low frequencies. Synthesis tool will not use any of the novel gates in very low frequencies and the total power savings will eventually be the same as the savings from the previous work library.

Table 5.2: The relationship between frequency and the total power savings is compared at different frequencies for the previous work and complete libraries.

Frequency	Previous work library	Complete library
2 GHz	20%	36 %
1500 MHz	19.1%	30.2%
1 GHz	19.7%	25.6 %

Finally, the savings in power consumption is approximated without placement and routing of the circuits. Introduction of the novel gates also reduces the area consumption, which reduces the distance between the gates and hence their corresponding parasitic wire capacitances. Therefore, it is expected that the savings in the total power consumption will increase once the placement and routing step is performed. An attractive option called orientation engineering [24] exists for layout design of FinFETs. The optimum orientation of n-type and p-type devices is different, hence it is possible to decrease the gate delay by rotating its p-type devices by 45 degrees. This rotation substantially increases the area and routing complexity while offering savings in the power consumption. In this paper, it was assumed that n-type and p-type devices have similar orientations.

Chapter 6

Conclusions

I proposed the design of dual-V_{th} independent-gate FinFETs by optimizing the oxide thickness, electrode work-function, silicon thickness, and gate-source/drain underlap. It is shown that the dual-V_{th} independent-gate FinFETs enable merging of series and parallel transistors, with efficient realization of logic gates. Complex functions were also implemented using dual-V_{th} independent-gate devices in pull-down or pull-up networks of gates. The gates have lower input capacitance and number of fins, and comparable performance to conventional implementations. A class of novel logic gates has also been proposed by defactoring the Boolean functions in the pull-down and/or the pull-up networks. Results on several benchmark circuits demonstrate that significant savings in the number of fins and the total power consumption can be achieved by incorporating these novel gates into the technology library. The effects of the frequency of operation and temperature on the relative performance of the proposed logic gates are also explored and reported.

Detailed investigation of the process variation effects on the performance of FinFET logic gates are the focus of my future research in this field. The noise figure of FinFETs devices are not as good as conventional CMOS devices, hence designing analog and RF circuits with FinFETs becomes challenging. Methodologies to alleviate this shortcoming of FinFET devices will be explored.

Bibliography

[1] Abc synthesis tools. http://www.eecs.berkeley.edu/ alanmi/abc/, 2008.

[2] International technology roadmap for semiconductors. http://www.itrs.net/links/2009ITRS/Home2009.htm, 2009.

[3] Synopsys sentaurus design suite, 2009.

[4] R.T. Cakici and K. Roy. Analysis of options in double-gate MOS technology: A circuit perspective. *IEEE Trans. on Electron Devices*, 54(12):3361–3368, 2007.

[5] Yu Cao and Wei Zhao. Predictive technology model for nano-CMOS design exploration. In *Intl. Conference on Nano-Networks*, pages 1–5, 2006.

[6] L. Chang et al. Gate length scaling and threshold voltage control of double-gateMOSFETs. In *IEDM Technical Digest*, pages 719–722, 2000.

[7] Meng-Hsueh Chiang et al. High-density reduced-stack logic circuit techniques using independent-gate controlled double-gate devices. *IEEE Trans. on Electron Devices*, 53(9):2370–2377, 2006.

[8] J.-P. Colinge. *FinFETs and other multi-gate transistors*. Springer, November 2007.

[9] A. Datta et al. Modeling and circuit synthesis for independently controlled double gate FinFET devices. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 26(11):1957–1966, 2007.

[10] M.V. Dunga et al. BSIM-MG: A versatile multi-gate FET model for mixed-signal design. In *IEEE Symposium on VLSI Tech.*, pages 60–61, 2007.

[11] B. Ebrahimi, M. Rostami, A. Afzali-Kusha, and M. Pedram. Statistical design optimization of FinFET SRAM using back-gate voltage. *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, (99), 2010.

[12] Jerry G. Fossum. Physical insights on nanoscale multi-gate CMOS design. *Solid-State Electronics*, 51(2):188194, February 2007.

[13] J. Gu, J. Keane, S. Sapatnekar, and C.H. Kim. Statistical leakage estimation of double gate FinFET devices considering the width quantization property. *IEEE Transactions on Very Large Scale Integration Systems*, 16(2):206–209, 2008.

[14] Xuejue Huang et al. Sub-50nm p-channel FinFET. *IEEE Trans. on Electron Devices*, 48(5):880–886, 2001.

[15] K. Kim and J.G. Fossum. Double-gate CMOS: Symmetrical-versus asymmetrical-gate devices. *IEEE Trans. on Electron Devices*, 48(2):294–299, 2002.

[16] S.H. Kim and JG Fossum. Design Optimization and Performance Projections of Double-Gate FinFETs With Gate–Source/Drain Underlap for SRAM Application. *IEEE Trans. on Electron Devices*, 54(8):1934–1942, 2007.

[17] A. Kuriyama et al. A systematic investigation of work function in advanced metal gate-HfO2-SiO2 structures with bevel oxide. *Solid-State Electronics*, 51(11-12):1515–1522, 2007.

[18] D.D. Lu et al. Design of FinFET SRAM cells using a statistical compact model. In *Proc. Int'l Conference on Simulation of Semiconductor Devices and Processes*, 2009.

[19] CR Manoj et al. Device design and optimization considerations for bulk FinFETs. *IEEE Trans. on Electron Devices*, 55(2):609–615, 2008.

[20] M. Masahara et al. Demonstration of asymmetric gate-oxide thickness four-terminal FinFETs having flexible threshold voltage and good subthreshold slope. *IEEE Electron Device Letters*, 28(3):217–219, 2007.

[21] M. Masahara et al. Experimental Investigation of Optimum Gate Workfunction for CMOS Four-Terminal Multigate MOSFETs (MUGFETs). *IEEE Trans. on Electron Devices*, 54(6):1431–1437, 2007.

[22] L. Mathew et al. CMOS vertical multiple independent gate field effect transistor (MIGFET). In *Intl. SOI Conference*, pages 187–189, 2004.

[23] L. Mathew et al. Multiple independent gate field effect transistor (MIGFET) - multi-fin RF mixer architecture, three independent gates (MIGFET-T) operation and temperature characteristics. In *Intl. Symposium on VLSI technology*, pages 200–201, 2005.

[24] P. Mishra and N.K. Jha. Low-power FinFET circuit synthesis using surface orientation optimization. In *Design, Automation and Test in Europe (DATE)*, pages 311–314, 2010.

[25] A. Muttreja et al. CMOS logic design with independent-gate FinFETs. In *Intl. Conference on Computer Design*, pages 560–567, 2007.

[26] Sebastien Nuttinck et al. Double-gate FinFETs as a CMOS technology downscaling option: An RF perspective. *IEEE Trans. on Electron Devices*, 54(2):279–283, 2007.

[27] S.H. Rasouli et al. High-speed low-power FinFET based domino logic. In *Proceedings of the Asia and South Pacific Design Automation Conference*, pages 829–834, 2009.

[28] AB Sachid et al. Gate Fringe-Induced Barrier Lowering in Underlap FinFET Structures and Its Optimization. *IEEE Electron Device Letters*, 29(1):128–130, 2008.

[29] M.O. Simsir, A. Bhoj, and N.K. Jha. Fault modeling for FinFET circuits. In *Proceedings of Intl. Symp. on Nanoscale Architectures*, pages 41–46, 2010.

[30] Robert F. Sproull and David Harris. *Logical effort: Designing fast CMOS circuits*. Morgan Kaufmann, 1999.

[31] B. Swahn and Soha Hassoun. Gate sizing: FinFETs vs 32nm bulk MOSFETs. In *Proceedings of Design Automation Conference*, pages 528–531, 2006.

[32] S.A. Tawfik and V. Kursun. Low-power and compact sequential circuits with independent-gate FinFETs. *IEEE Trans. on Electron Devices*, 55(1):60–70, 2008.

[33] S.A. Tawfik and V. Kursun. Portfolio of FinFET memories: Innovative techniques for an emerging technology. pages 101–104, 2008.

[34] S.A. Tawfik and V. Kursun. FinFET domino logic with independent gate keepers. *Microelectronics Journal*, 40(11):1531–1540, 2009.

[35] S.A. Tawfik and V. Kursun. Low-power and robust six-FinFET memory cell using selective gate-drain/source overlap engineering. In *Proc. of the Intl. Symp. on Integrated Circuits*, pages 244 – 247, 2009.

[36] S.A. Tawfik and V. Kursun. Multi-threshold boltage FinFET sequential circuits. *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, 99:1–5, 2009.

[37] S. Toriyama and N. Sano. Probability distribution functions of threshold voltage fluctuations due to random impurities in deca-nano MOSFETs. *Low-dimensional Systems and Nanostructures*, 19:44–47, 2003.

[38] A. Tsormpatzoglou et al. Threshold voltage model for short-channel undoped symmetrical double-gate MOSFETs. *IEEE Trans. on Electron Devices*, 55(9):2512–2516, 2008.

[39] YQ Wu et al. First experimental demonstration of 100nm inversion-mode InGaAs FinFET through damage-free sidewall etching. In *International Electron Devices Meeting*, 2009.

[40] S. Xiong and J. Bokor. Sensitivity of double-gate and FinFET devices to process variations. *IEEE Trans. on Electron Devices*, 50(11):2255–2261, 2003.

[41] Weimin Zhang et al. Physical insights regarding design and performance of independent-gate FinFETs. *IEEE Trans. on Electron Devices*, 52(10):2198–2206, 2005.