Designing Faster CMOS Subthreshold Circuits Using Transistor Sizing and Parallel Transistor Stacks

by

Morteza Nabavi

A Thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Master of Applied Science
in

Electrical and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
September 2012

Your file Votre référence
ISBN: 978-0-494-93513-2

Our file Notre référence
ISBN: 978-0-494-93513-2

Abstract

Subthreshold region of operation in digital CMOS circuits provides an ideal low-power solution for many applications that need tremendously low-energy operation. However, this advantage comes at the cost of speed, so enhancing the speed of subthreshold circuits can expand their application spectrum. This thesis deals with the techniques for speed improvement of subthreshold circuits to expand the domain of subthreshold circuits at no or minimal energy cost.

The first part of the thesis examines the effect of PMOS-to-NMOS width ratio on frequency of operation in the subthreshold region. Analytical and simulation results illustrate that in this region the frequency attains its maximum at the optimum PMOS-to-NMOS ratio independent of the supply voltage. Using this optimum value leads to designing circuits for highest speed and nearly lowest energy in the subthreshold region. These results reveal that the minimum sizing doesn't give the minimum energy per cycle in all cases.

Parallel Transistor Stacks (PTS) technique has been shown to be effective for improving the speed of digital circuits operating in the subthreshold region, which comes at the cost of power consumption and area. However, our experience shows that using PTS is not beneficial in all cases. In the second part of the thesis, we present a methodology to identify whether or not using PTS is beneficial (or not) in a particular CMOS technology and to determine what transistor sizing can be employed to maximize the circuit speed. Our technique is based on analyzing the

current-over-capacitance (COC) ratio of PMOS and NMOS transistors. The results of incorporating the proposed methodology in a 4-bit comparator and a 19-stage inverter ring oscillator, using 90 nm CMOS technology, illustrate 26% and 40% extra improvement compared with the blind use of PTS, respectively.

Acknowledgments

I would like to thank

Professor Maitham Shams, thesis supervisor, for his guidance and patience

Professors Emad Gad, Garry Tarr, Steve McGarry, Tom Smy for their careful review and contributions to the thesis

All my friends at school, Professor Mojtaba Ahmadi, Farhad Ramezankhani, Behzad Yadegari, Kimia Ansari for their technical assistance

CMC Microsystems and their technology partners for access to the design tools and technology kits used in the research for this thesis

Professor Pavan Gunupudi, Anna Lee and Blazenka Power for their help in bringing the thesis process to an end

And my family, Professor Abdolreza Nabavi, Masoumeh Mirzaeepour, and Fatemeh Nabavi for their moral support.

Table of Contents

List of Acronyms

Acronyms	Definition
CMOS	Complementary metal-oxide-semiconductor
COC	Current Over Capacitance
COC_{opt}	COC ratio corresponding to W_{opt}
COC_{INWE}	COC ratio corresponding to W_{INWE}
DIBL	Drain induced barrier lowering
EDP	Energy-delay product
GP	General purpose transistor variant
GND	Ground
INV	Logic gate that implements negation
INWE	Inverse-narrow-width effect
LOCOS	Local oxidation of silicon
LP	Low-power transistor variant
MOSFET	Metal-oxide-semiconductor field-effect transistor

NAND	Logic gate that implements negated AND
NMOS	n-channel MOSFET
NOR	Logic gate that implements negated OR
PDP	Power-delay product
PMOS	p-channel MOSFET
PTS	Parallel transistor stacks
PVT	Process-voltage-temperature
STI	Shallow trench isolation
VLSI	Very-large-scale integration

List of Symbols

Symbols	Definition
a	Activity factor
α	velocity-saturation index
β	Beta ratio, or PMOS-to-NMOS width ratio
β_{opt}	Optimum Beta ratio
C	Capacitance
C_{bottom}	Diffusion bottom-plate capacitance
C_{DB}	Drain-to-body diffusion capacitance
C_{diff}	Diffusion capacitance
C_G	Gate capacitance
C_{GB}	Gate-to-bulk capacitance
C_{GD}	Gate-to-drain capacitance
C_{GS}	Gate-to-source capacitance
C_j	Junction capacitance per unit area

C_{jsw}	Side-Wall junction capacitance per unit parameter
C_{JD}	Junction-drain capacitance
C_{JS}	Junction-source capacitance
C_L	Load capacitance
C_{ox}	Gate oxide capacitance per unit area
C_{SB}	Source-to-body diffusion capacitance
C_{sw}	Diffusion side-wall capacitance
ΔF	Difference in operational frequencies at β_{opt} obtained from analytical model and β_{opt} obtained from the 19-stage inverter ring oscillator
ΔE	Difference in Energy consumption operational frequencies at β_{opt} obtained from analytical model and β_{opt} obtained from the 19-stage inverter ring oscillator
E	Energy
E_A	Energy consumption obtained from analytical model
E_S	Energy consumption obtained from simulation results
F_A	Operating frequency obtained from analytical model
F_S	Operating frequency obtained from simulation results
f	(i) Frequency; (ii) tapering ratio
$fmax$	Maximum frequency

γ	body effect factor
Γ	$\frac{\mu_n}{\mu_p}$; Superthreshold β_{opt} factor
I	Current
I_1	PN junction reverse-bias current
I_2	Subthreshold current
I_3	Current tunnelling into and through gate-oxide
I_4	Gate current due to hot-carrier-injection
I_5	Current Gate induced drain leakage
I_6	Current Channel punch-through current
I_N	Average drain current of an NMOS transistor
I_P	Average drain current of a PMOS transistor
I_D	Drain current
$I_{leakage}$	Total Leakage current
L	Transistor length
L_n	Channel length of an NMOS transistor
L_p	Channel length of a PMOS transistor
λ	DIBL coefficient
Λ	Subthreshold β_{opt} modification factor
L_S	Junction side-wall length

m	Empirical parameter
μ_n	Electron mobility
μ_n	Hole mobility
N	Number of parallel transistors in PTS structures
n	Subthreshold slope factor
P	Power consumption
P_V	Fitting parameter
P_C	Fitting parameter
$Pavg$	Average power
$P_{dynamic}$	Dynamic power
P_{static}	Static power
T	Period of oscillation
t_r	Rise time
t_f	Fall time
t_p	Propagation delay
t_{PHL}	High-to-low propagation delay
t_{PLH}	Low-to-high propagation delay
V_{DD}	Supply voltage
V_{D0}	drain-saturation voltage at $V_{GS} = V_{DD}$

V_{DS}	Drain-to-source voltage
V_{GS}	Gate-to-source voltage
V_T	Thermal voltage
V_{th}	Threshold voltage
V_{th0}	Threshold voltage at large widths
V_{tp}	PMOS Threshold voltage
V_{tn}	NMOS Threshold voltage
W	Transistor width
W_{INWE}	The optimum width corresponding to the INWE
W_n	NMOS channel width
W_{opt}	Optimum MOSFET width obtained from Logical Effort
W_p	PMOS channel width
xd	Lateral diffusion

Chapter 1

Introduction

Ultra-low power applications such as micro-sensor networks, pacemakers, and many portable devices require extreme energy constraints for longer battery life. Increasing the battery life can provide a competitive advantage in the marketplace. Traditionally, reducing the power supply voltage is regarded as the most effective means of reducing power consumption [2]. It is shown that the minimum energy is achieved when power supply is scaled below the threshold voltage which is known as subtreshld operation [3]. Therefore, digital circuits operating in the subthreshold region offer a promising solution for emerging portable applications that require tremendously low-energy consumption. However, this does come at the cost of very slow operational speed due to the extremely scaled-down supply voltage. Despite very high energy efficiency of subthreshold circuits, the subthreshold design has been only applied in niche markets because of low performance. I believe the application domain of subthreshold circuits may be extended by establishing techniques to enhance their performance.

To address the challenging issue of enhancing the speed of subthreshold circuits, this thesis investigates: (1) the optimum PMOS-to-NMOS width ratio which results in the maximum frequency of operation for logic gates operating in the subthreshold region and, (2) a methodology to identify when using the PTS (parallel transistor

stacks) technique [4] is beneficial (or not) in a particular CMOS technology and what transistor sizing can be employed to maximize the circuit speed. Our approach is based on analyzing the current-over-capacitance (COC) ratio of PMOS and NMOS transistors in a given CMOS technology. I demonstrate that one may improve the performance of subthreshold circuits by utilizing the new design methodologies proposed in this dissertation.

In the following I present the thesis motivation and previous work in this area. The objective and thesis organization also follows.

1.1 Motivation

The subthreshold region in digital CMOS circuits provides an ideal low-power solution for many applications where power consumption is the main concern. However, low operating speed is a drawback of subthreshold circuits and they may not provide enough speed for applications that require higher speeds in addition to low power consumption. Improving the operational frequency of superthreshold circuits can expand the application spectrum of these circuits.

This motivation has led to exploring the state-of-the-art of designing digital circuits operating in the subthreshold region and proposing techniques for improving their speed. This is achieved through proper transistor sizing and using Parallel Transistor Stacks (PTS) when it is beneficial. The results of incorporating the proposed techniques lead to designing subthreshold circuits with higher speed and lower energy costs compared to the conventional minimum-sized circuits and blind use of PTS.

1.2 Previous Work

The main research in the area of designing ultra-low-power circuits operating in the subthreshold region have been focused on reducing the supply voltage, optimizing energy and power consumption, transistor sizing, and delay compensation techniques. In the following I present the previous work done in these areas and applications of subthreshold circuits.

1.2.1 Reducing Power Supply Voltage

In 1997, the effect of lowering the supply and threshold voltages on the energy efficiency of CMOS circuits was explored [5]. It is shown that lowering the supply and threshold voltage is generally advantageous.

The reduction of the power supply voltage offers the most direct and dramatic means of reducing the power consumption [6]. It is also shown that subthreshold operation, where the supply voltage is below the transistor's threshold voltage, can be regarded as the most energy-efficient solution for low-power applications [3]. In this paper subthreshold logic and memory design methodologies were developed and their effectiveness were demonstrated on a fast Fourier transform (FFT) processor.

The amount of energy per cycle is important since it determines the battery life cycle [2]. In 2002 a simple characterization circuit is introduced, by which the performance and energy dissipation for a given process is evaluated [7]. Results show that operation at the near threshold supply voltage levels can lead to energy savings of an order of magnitude.

In [8], energy minimization for circuits operating in the subthreshold region was explored. It is shown that the optimum supply voltage depends on design characteristics and operating conditions.

An analytical solution for the optimum supply voltage and threshold voltage to

minimize the energy for a given frequency is presented in [9]. The optimum supply voltage for the FFT designed and fabricated is reported at 0.25 V for both simulation and fabrication results. It also examines the effect of transistor sizing on energy consumption for subthreshold circuits. The Authors show that the minimum-sized devices are theoretically optimal for reducing energy in subthreshold circuits. They suggest that instead of upsizing PMOS transistors to equalize the rising and falling delays in a logic gate, one can use minimum size devices and increase the supply voltage to minimize the energy in the subthreshold operation. However, in certain designs the supply voltage is fixed and cannot be changed. Besides, minimum-size devices may not have the driving capability to operate in the subthreshold region and require the increase of the supply voltage beyond the threshold. Moreover, there are examples shown in this thesis that minimum-size devices are not always the best in terms of energy consumption in the subthreshold region.

Bol et al. [10] [11] show that the minimum energy consumption of subthreshold logic circuits dramatically increases towards 45 nm technology. They demonstrated, by circuit simulation and analytical modelling, that this increase comes from the combined effects of process variation, gate leakage and Drain-Induced-Barrier-lowering (DIBL).

1.2.2 Transistor Sizing and Delay Compensation Techniques

Several papers have explored the effect of sizing transistors on the subthreshold operation. Due to the inverse narrow width effect (INWE), the current of a MOSFET transistor in the subthreshold region, unlike in the superthreshold region, is not linearly proportional to the transistor width. Therefore, transistor sizing in the subthreshold region is different than that in the superthreshold region.

In [12], a subthreshold sizing method to balance the rising and falling delays by taking into account the influence of INWE while minimizing the area is proposed.

Based on this sizing, the delay and power-delay-product (PDP) are reduced by up to 35.4% and 73.4%, respectively, with up to 57% saving in the area compared the conventional sizing method. Further, the minimum operating voltage can be lowered by 8% due to the symmetric rising and falling delays. However in this paper, the authors compare their circuits with minimum-sized circuits with higher supply voltages as suggested in [9]. Besides, they only reported simulations results for one logic stage. I show in Chapter 6 that the rising and falling propagation delays can be made equal after the second stage in a logic path even if they are not equal in the first stage. Thus it is our belief that equalizing the rising and falling delays should not be a primary concern.

In [13], the increase of the MOSFET channel length in the conventional 6T SRAM cell to operate safely in the subthreshold region is proposed. The two channel-length upsizing schemes proposed in this paper show an efficient increase in robustness with a minimum area overhead.

In [14], a framework for choosing the optimal transistor-stack sizing factors in terms of current drivability is proposed for subthreshold design. A closed-form solution for the proposed sizing of transistors in a stack is also derived. In [15], the use of Logical Effort in the subthreshold region was explored. However, the impact of INWE on transistor sizing for the subthreshold operation was not considered in [14] and [15].

Subthreshold CMOS circuits are slow and inadequate for high performance applications. Hence, subthreshold circuits with higher speeds are in high demand. To achieve this goal, delay-compensation techniques for subthreshold digital circuits have been proposed in [16]. Since the delay in these circuits changes exponentially with variations of the threshold voltage, threshold voltage monitoring and relative supply voltage scaling techniques are adopted [16].

An accurate and fully analytical model of the delay in subthreshold CMOS inverters is presented in [17]. This model is capable of predicting the signal slew at the inverter output which can then be used for reducing and estimating the logic-gate delays. The concept of PTS is proposed in [4]. It is shown that using PTS and transistor widths that maximize the Current-Over-Capacitance (COC) ratio, either individually or in parallel stacks, in subthreshold circuits leads to circuits that are up to three times faster. Even though PTS has shown improvements in terms of speed in the subthreshold operation, there are circumstances that are not reported in [4] where PTS should be avoided. In this thesis I identify when to use PTS and when to avoid it. The concept of PTS in [4] was explored in the 65 nm and 90 nm ST CMOS technologies. In this thesis I verify the concept of PTS and its usability in the 65 nm ST, 90 nm TSMC, 130 nm IBM, and 180 nm TSMC CMOS technologies. All simulations are done using Cadence and the BISIM4 model. I have also verified the model provided in the 180 nm CMOS technology kit with Minimos and Supreme4.

1.2.3 Subthreshold Applications and Related Designs

Several efforts have been made in designing circuits such as memory, microprocessors, and adders, that employ the above techniques for the subthreshold operation. In the following I describe these applications.

In [18], authors reported a 13-bit ultra-low-power subthreshold memory fabricated in a 130 nm process technology. The read operation is performed with a 190 mV power supply at 28 KHz, and the write operation occurs at 216 mV with the same speed.

In [19], a full adder circuit optimized for ultra low-power operation is proposed. The circuit is based on modified XOR gates designed for the subthreshold operation to minimize the power consumption.

In [20], the design of a subthreshold processor for use in ultra low-energy sensor systems is explored. An 8-bit subthreshold processor is designed with energy efficiency

as the primary constraint.

In [21] and [22], a logic style called ultra low-power (ULP) logic is proposed. This logic style benefits from the small area and low dynamic power of silicon-on-insulator (SOI) deep submicron technologies while keeping ultra-low leakage.

In [24], a new computational design automation that tests every cell in a standard cell library for proper operation in the subthreshold region is proposed. The conventional method to improve digital circuit operation in the subthreshold region is to design every logic cell manually, requiring complete re-design and re-characterization for every process node.

In [25], the performance of single wall carbon nanotube (SWCNT), Cu, and mixed carbon-nanotubes (CNT) bundle interconnects for different interconnect lengths and biasing levels under subthreshold conditions are compared. It proposes that for short and intermediate length interconnects at different bias points in the subthreshold region individual SWCNT can be used. Furthermore, it claims that in the moderate subthreshold region, scaled Cu interconnect performs better than individual SWCNT and mixed CNT bundle, whereas in deep subthreshold region individual SWCNT is still better.

Near-threshold computing (NTC), a designing space where the supply voltage is near to the threshold voltage, is explored in [26]. This design space preserves much of the power savings of the subthreshold region, but benefits more in terms of speed. This characteristic makes it applicable to broader range of applications such as sensors and high-performance servers.

Markovic et al. [27] has also explored the near-threshold operation. It is shown that a 20% increase in the energy from the minimum-energy point results in 10 times increase in performance. Also they have introduced a pass-transistor based logic operating in this region. Time multiplexing is also shown as an approach yielding not only area, but also energy due to the lower leakage.

Improving the speed in the subthreshold region by deploying parallel architectures is addressed in [28]. In this paper, clock-less designs are also explored. The authors demonstrate that the timing issues associated with significant process-voltage-temperature (PVT) variations that occur in the subthreshold operation can be mollified by using clock-less logic.

1.3 Objective

The objectives of this thesis are summarized as follows.

1. Investigating the transistor subthreshold behavior, in particular with regard to the currents and capacitances
2. Exploring the effect of transistor sizing in performance and energy consumption of subthreshold circuits
3. Developing an analytical model to determine the optimum PMOS-to-NMOS width ratio to achieve the maximum speed in subthreshold region
4. Investigating CMOS technologies to identify where using Parallel Transistor Stacks are beneficial for subthreshold operation circuits and which sizing achieves the highest performance
5. Developing a methodology to design subthreshold circuits based on a modified Parallel Transistor Stacks technique
6. Demonstrating the effectiveness of the proposed design methodology through the use in some application circuits

1.4 Contributions

The followings are contributions of this thesis

1. Expanding the domain of applications of subthreshold circuits by increasing the speed of these circuits with no or minimal energy cost.

2. Obtaining the optimum PMOS-to-NMOS width ratio (β_{opt}) by plotting Current-Over-Capacitance ratio versus width.

3. Developing an analytical model for β_{opt} verifies its independence of the supply voltage.

4. Developing a methodology to identify when using PTS is beneficial (or not) in a particular CMOS technology and what transistor sizing can be employed to maximize the circuit speed in a given CMOS technology.

5. Demonstrating that using minimum-size devices are not always the best in terms of energy consumption in subthreshold circuits.

6. Illustrating that minimum energy operation depends on β. As V_{DD} decreases, minimum energy operation occurs at higher values of β.

1.5 Thesis Organization

Chapter 1 presented the motivation for designing digital subthreshold circuits and addressed a selection of related previous research work on the topic. Chapter 2 focuses on the properties of MOSFET transistors such as the current and capacitances. It shows how these properties are affected by altering the transistor's width. A qualitative description of the cause of the INWE is also provided. Leakage currents and the relevant capacitances are also examined. Chapter 3 describes the behaviour of

CMOS circuits in the subhreshold region. Chapter 4 provides an analytical model for optimal PMOS-to-NMOS width ratio for subthreshold operation. Chapter 5 presents a methodology for designing circuits based on a modified PTS technique. Chapter 6 reports the impact of PTS on driving large loads. Chapter 7 reports the results of applying the proposed methodologies on different application circuits. Chapter 8 provides a summary of this thesis and its contributions.

Chapter 2

MOSFET Currents and Capacitances

In this chapter, the relevant transistor properties to the delay and energy such as the current and capacitances are discussed, sources of the leakage current are addressed, and a qualitative explanation of the inverse-narrow-width-effect (INWE) is presented. The delay, power consumption, energy consumption of a logic gate is also presented in this chapter.

2.1 MOSFET Transistors' Currents

Figure 2.1 depicts an NMOS transistor with its channel length (L) and channel width (W). The current between the drain and the source of the transistor, I_D, is also shown in this figure.

In long-channel devices, the Shockley model was used to describe the MOSFET current behaviour. However, the Shockley model fails to characterize the behaviour of the modern MOSFET. To improve the accuracy of Shockley's model, Sakurai et

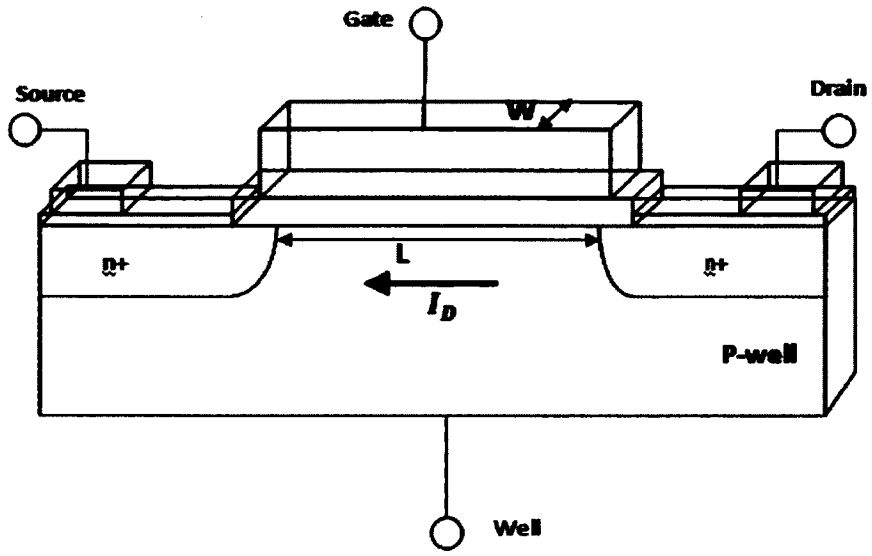

Figure 2.1: NMOS transistor

al. [29] proposed the alpha-power law model, as follows

$$I_D = \begin{cases} 0 & V_{GS} < V_{th} \text{ cut-off region} \\ (\frac{I'_{D0}}{V'_{D0}})V_{DS} & V_{DS} < V'_{D0} \text{ triode region} \\ I'_{D0} & V_{DS} > V'_{D0} \text{ saturation region} \end{cases} \quad (1)$$

where

$$I'_{D0} = I_{D0}(\frac{V_{GS} - V_{th}}{V_{DD} - V_{th}})^{\alpha} \left(= \frac{W}{L} P_C (V_{GS} - V_{th})^{\alpha} \right) \quad (2)$$

$$V'_{D0} = V_{D0}(\frac{V_{GS} - V_{th}}{V_{DD} - V_{th}})^{\alpha} \left(= P_V (V_{GS} - V_{th})^{\frac{\alpha}{2}} \right), \quad (3)$$

where V_{DD} is the supply voltage, V_{GS} is the gate-source voltage, V_{th} is the threshold voltage, P_C and P_V are fitting parameters, W is the channel width, and L is the

effective channel length. Parameter α is the velocity-saturation index that ranges from 2 to 1, as the carrier velocity-saturation becomes more severe; V_{D0} is the drain-saturation voltage at $V_{GS} = V_{DD}$; and I_{D0} is the drain-current at $V_{GS} = V_{DS} = V_{DD}$. As shown in Eq. (1), MOSFETs operate in three different regions: the cut-off, the triode, and the saturation regions [29]. To turn on a MOSFET, the gate-source voltage, V_{GS}, has to be greater than the threshold voltage, V_{th}. When V_{GS} is less than V_{th}, the transistor is operating in the cut-off (subthreshold) region and the drain current is limited to a very small leakage current (discussed in the next section) that is usually considered to be zero ($I_D = 0$). In the triode region where $V_{DS} < V'_{D0}$, a relatively small voltage V_{DS} is applied between the drain and the source, resulting in a current I_D to flow between the drain and the source. In the saturation region, where $V_{DS} > V'_{D0}$, the current is further increased compared to the triode region.

Although the leakage current is very small compared to the saturation current, it is exploited in designing subthreshold circuits. The various leakage currents are described in the next section.

2.2 Leakage Currents

In this section, different sources of the leakage currents are introduced. As later explained, the main leakage current is the subthreshold current, which is exploited in designing subthreshold circuits.

Six leakage currents are shown in Figure 2.2 and are described as follows [30]

- PN junction reverse-bias current (I_1): This current is produced by the drain-to-body and source-to-body junctions, which are typically reverse biased.

- Subthreshold current (I_2): This current flows between the drain and the source of a MOSFET when the gate-source voltage is below the threshold voltage.

- Tunnelling into and through gate-oxide (I_3): The oxide thickness has become smaller, and the channel length is reduced in modern technologies. The reduction of the oxide thickness results in tunnelling of electrons from the gate to the substrate and from the substrate to the gate, which is called the gate-oxide tunnelling current.

- Gate current due to hot-carrier-injection (I_4): Due to the high electric field near the Si-SiO_2 interface, electrons and holes can gain sufficient energy from the electric field to cross the interface potential barrier and inject from Si to SiO_2.

- Gate induced drain leakage (GIDL), (I_5): This current is due to a high electric field in the drain junction of a MOSFET. Thinner oxide thickness and higher supply voltage increase the potential between the gate and drain, which results in higher GIDL leakage.

- Channel punch-through current (I_6): The depletion regions at the drain-substrate and source-substrate junctions extends into the channel in short-channel devices. Therefore, the separation between these two depletion region boundaries decreases. The punch-through occurs when the depletion region is formed by the combination of channel length and the depletion regions of the drain-substrate and source-substrate junctions.

I simulated the currents through the drain (I_2), gate (I_3 and I_4), and the bulk (I_1 and I_5) nodes of a PMOS and an NMOS transistor each biased at the supply voltage of V_{DD}=0.2 V for four technologies. These currents are represented in Table 2.1. As shown in the table, the leakage current through the drain node (subthreshold current) contributes the most significant portion of the total leakage current.

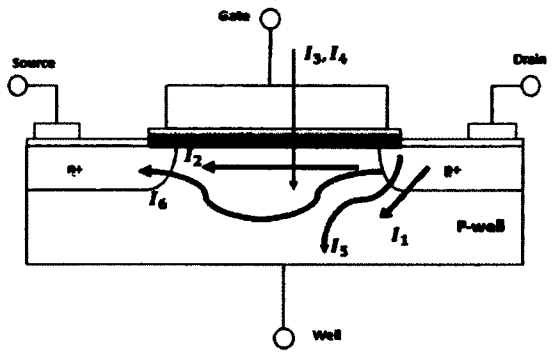

Figure 2.2: Summary of leakage currents in a MOS transistor

Table 2.1: Leakage Currents of PMOS and NMOS Transistors Biased at $V_{DD} = 0.2$ V

Device	Size	Suthreshold Current (I_2)	Gate Leakage ($I_3 + I_4$)	Junction Leakage ($I_1 + I_5$)
180 nm PMOS	220 nm	61.78 pA	negligible	4.75 aA
180 nm NMOS	220 nm	4.16 nA	negligible	3.39 aA
130 nm PMOS	160 nm	2.56 nA	17.9 aA	65 zA
130 nm NMOS	160 nm	11.71 nA	473.1 aA	156 zA
90 nm PMOS	120 nm	15.68 nA	701.4 fA	840 zA
90 nm NMOS	120 nm	154.9 nA	574.8 fA	2 aA
65 nm PMOS	120 nm	947.2 pA	1.674 fA	230 zA
65 nm NMOS	120 nm	2.366 nA	767 aA	350 zA

The drain current represented in the alpha-power-law model in the superthreshold region does not account for the drain current of the MOSFET when it is operating in the subthreshold region (subthreshold current). The basic equation for modelling subthreshold current is expressed as [9]

$$I_D = I_S e^{\frac{(V_{GS}-V_{th}+\lambda V_{DS})}{nV_T}}(1 - e^{\frac{-V_{DS}}{V_T}}) \tag{4}$$

$$I_S = \mu C_{ox}\frac{W}{L}(n-1)V_T^2 \tag{5}$$

$$V_{th} = V_{th0} + \gamma V_{SB}, \tag{6}$$

μ is the charge carriers' mobilities of MOSFET; C_{ox} is the gate-oxide capacitance; V_T is the thermal voltage ($= K_BT/q$), the value of which is 26 mV at 300 K; and n is the subthreshold slope factor. This factor is typically in the range of 1.3 to 1.5 [2], and for convenience it can be assumed to be equal for PMOS and NMOS transistors [31]. λ is the drain induced barrier lowering (DIBL) effect, and γ is the body effect factor.

Since the voltages across the gate and the drain of a MOSFET in the subthreshold region are low, leakage components such as the gate currents, GIDL, and PN junction leakage become negligible. Since γ is near zero, V_{DS} is also small, and λ is between 0.01 and 0.2, V_{DS} can be neglected as it is not comparable with the threshold voltage, V_{th}. Therefore, by sacrificing accuracy for simplicity, Eq. (4) can be approximated by [9]

$$I_{DS} = I_S e^{\frac{V_{GS}-V_{th}}{nV_T}} \tag{7}$$

2.3 MOSFET Capacitances

Figure 2.3 shows different capacitances of a MOSFET. As shown in the figure, the main capacitances of a MOSFET are the gate (C_G) and the junction (C_J) capacitances [2].

2.3.1 Gate Capacitance

The gate capacitance consists of three components: gate-drain (C_{GD}), gate-source (C_{GS}), and gate-bulk (C_{GB}) capacitances. The gate-bulk capacitance is estimated by

$$C_{GB} = C_{ox} W L, \tag{8}$$

The other components of the gate capacitance, gate-drain and gate-source capacitances are caused by the source and drain extensions of the MOSFET under the gate oxide. They are expressed as

$$C_{GS} = C_{GD} = C_{ox} x_d W, \tag{9}$$

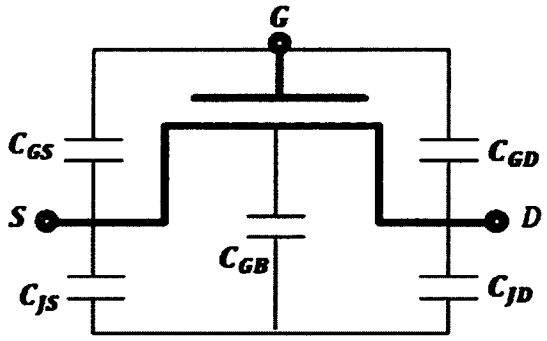

Figure 2.3: MOSFET capacitances

where x_d is the extension of the source or drain under the gate oxide. Summing Eqs. (8) and (9), the total gate capacitance of a MOSFET is expressed as

$$C_G = C_{ox}WL + 2C_{ox}x_dW. \tag{10}$$

2.3.2 Junction Capacitance

The source and the drain diffusions into the substrate create reversed PN junctions, which contribute to the junction-drain (C_{JD}), and the junction-source (C_{JS}) capacitances. Each junction capacitance consists of a bottom-plate and a side-wall capacitance. The bottom-plate capacitance is expressed as

$$C_{bottom} = C_jWL_s, \tag{11}$$

where C_j is the junction capacitance per unit area, and L_s is the side-wall length. The side-wall capacitance is expressed as

$$C_{sw} = C_{jsw}(W + 2L_s), \tag{12}$$

where C_{jsw} is the capacitance per unit perimeter. Note that the diffusion-to-channel capacitance is ignored. The total diffusion capacitance is then

$$C_J = C_{bottom} + C_{sw} = C_jWL_s + C_{jsw}(W + 2L_s). \tag{13}$$

Considering Eqs. (10) and (13), a linear relationship exists between the total capacitance of a MOSFET with its width.

2.4 Subthreshold Operation

Considering an inverter connected to an capacitance, if the input signal node is equal to zero, the PMOS transistor is conducting and the NMOS transistor is off. Since the I_{on}/I_{off} is at the order of magnitude, the PMOS transistor is much stronger than the NMOS transistor. Therefore the PMOS transistor charges the load capacitance up to the supply voltage. If the input signal is hooked up to the supply voltage, the NMOS transistor discharge the load capacitance to zero.

2.5 Inverse-Narrow-Width-Effect

2.5.1 Narrow-Channel Device

Figure 2.4 shows a cross section of a MOSFET's channel along its width. The figure depicts the local oxidation of silicon (LOCOS) isolation, which results in the gradual transition from the thick (field) oxide to the thin (gate) oxide. This transition region from the thick to thin oxide is called the bird's-beak region.

Due to the lateral oxidation, LOCOS technology caused problems because of the so called birds-beak phenomenon. Shallow Trench Isolation (STI) with a vertical field oxide, improves the area efficiency in device isolation. In STI, extensive gathering of the fringing field lines appear on the side of the depletion region under the gate. This phenomenon can be modeled as an effective increase in gate oxide capacitance [32]. This increase in gate oxide causes a reduction in the threshold voltage as the transistor width becomes narrower. Therefore, LOCOS causes a threshold roll-up while STI causes a threshold roll-off as the channel width decreases.

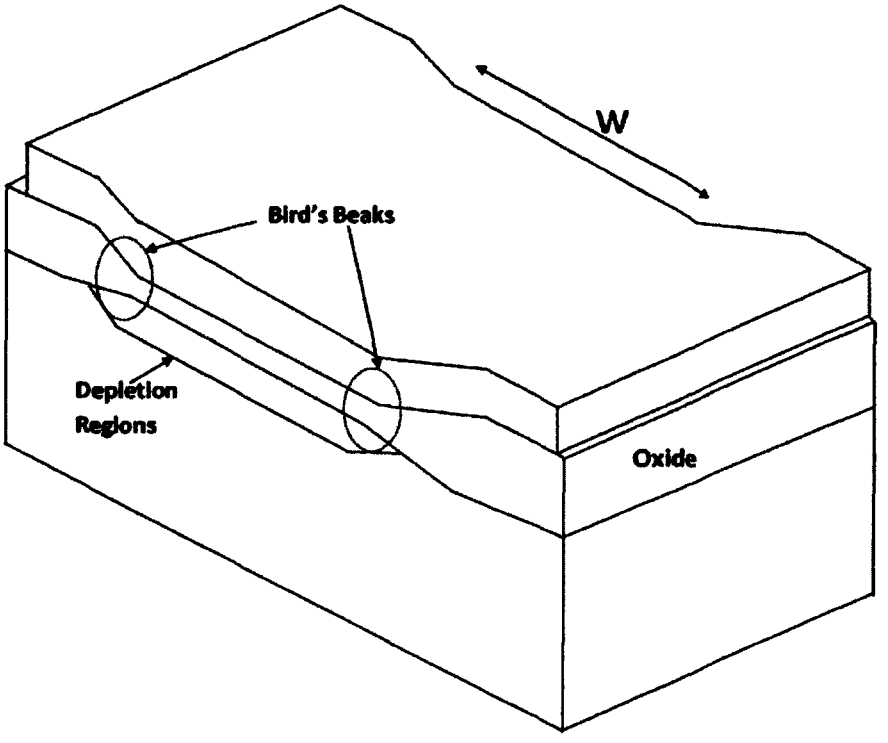

Figure 2.4: A cross section of a MOSFET along its channel width, with a representation of a LOCOS isolated device with a bird's-beak in the width of the gate

2.5.2 LOCOS Isolation

As shown in Figure 2.4, the depletion region is not only limited to existing beneath the thin oxide region, but can extend to the sides. This can be extended by the fact that some of the fringing field lines come out from the gate charges and terminate on ionized acceptor atoms on the sides. If the transistor's width is large, the extended depletion regions on both sides will be a small percentage of the whole depletion region and can be neglected. However, in devices with small width values, the side parts become relatively large. As a result, the threshold voltage in the LOCOS isolation process increases while the transistor's width decreases [32].

2.5.3 Shallow Trench Isolation

As shown in Figure 2.5, the depletion region for a STI MOSFET is deeper compared to a similar device in a LOCOS process. STI helps increase the surface potential, therefore, a lower V_{GS} is needed to deplete the channel before the inversion layer is created. STI results in a lower threshold voltage at lower widths , which gives rise to what is called the inverse-narrow-width-effect.

The threshold voltage versus width of a MOSFET for both LOCOS and STI isolation processes are plotted in Figure 2.6. As shown in this figure, the threshold voltage increases at narrower widths in the LOCOS isolation process while it decreases in the STI isolation process. Since the subthreshold current (Eq. (7)) is exponentially dependent on the threshold voltage, increasing the transistor width causes the current to decrease at narrower widths. In the superthreshold region, the drain current is linearly proportional to the transistor width.

In Chapter 3, I show that in modern nano meter technologies with STI isolation, using minimum-width devices is beneficial. This is true because they attain the minimum threshold voltage thus yielding maximum current.

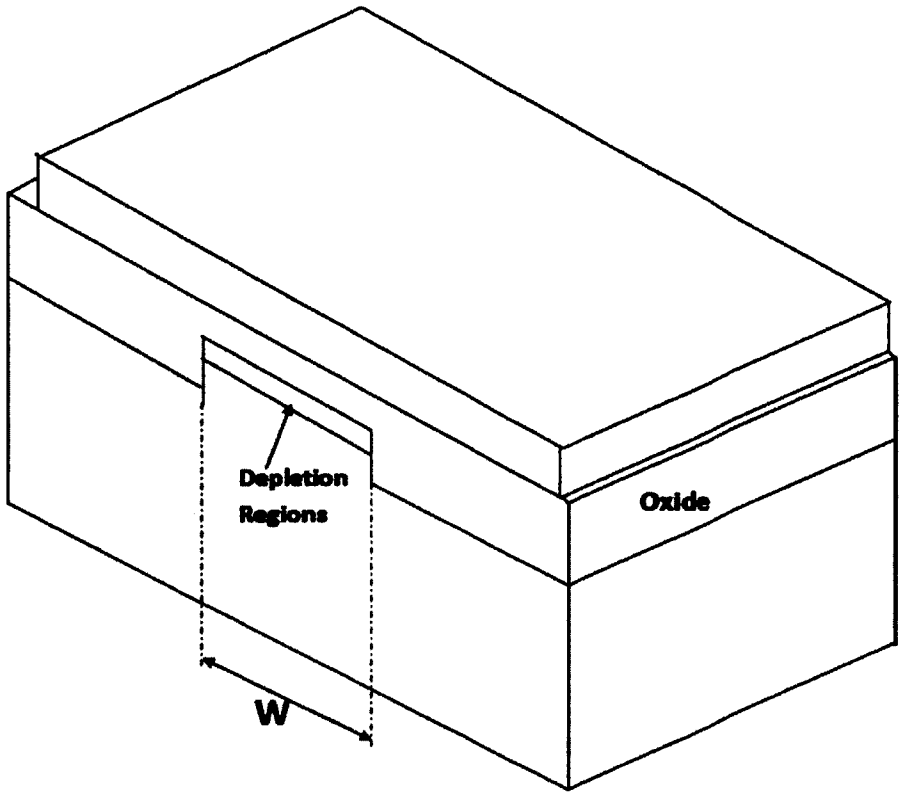

Figure 2.5: Cross section of a MOSFET along its width with a representation of shallow trench isolation (STI)

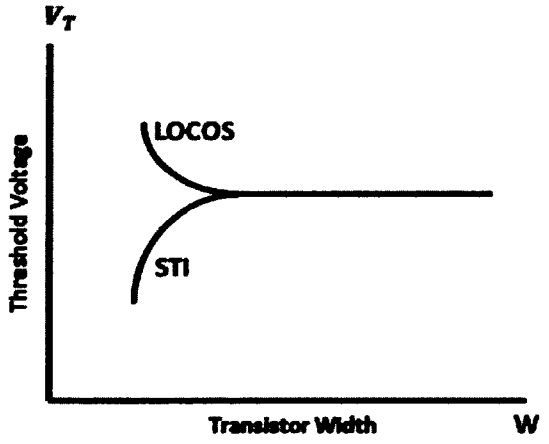

Figure 2.6: Threshold voltage vs. width for LOCOS and STI isolation processes

2.6 Propagation Delay

Propagation delay is the time interval from the input signal crossing 50% to the output signal crossing the 50% point. t_{PHL} is the the time from the input crossing 50% to the output crossing 50% when the output is making a transition from high to low. t_{PLH} is the the time from the input crossing 50% to the output crossing 50% when the output is making a transition from low to high [33]. The propagation delay is the average of the two, as follows [2]

$$t_p = \frac{t_{PHL} + t_{PLH}}{2} = \frac{t_{PHL} + t_{PLH}}{2} = \frac{C_L V_{DD}}{2}\left(\frac{1}{I_N} + \frac{1}{I_P}\right) \tag{14}$$

where I_N and I_P are the average currents of the NMOS and PMOS transistors, respectively, and V_{DD} is the supply voltage. In Chapter 3, I discuss how to size transistors to obtain the lowest capacitance and maximum current.

2.7 Rise and Fall Time

Rise time, t_r, is defined as the time for a waveform to rise from 20% to 80% of its steady-state value. Fall time, t_f, is the time for a waveform to fall from 80% to 20% of its steady-state value [33].

2.8 Power and Energy Dissipation

The power consumption of a circuit determines the energy consumed per unit of time, and it also determines the amount of heat dissipated by the circuit [2]. The power dissipation in CMOS circuits can be expressed as the sum of three main sources:

static power consumption, dynamic power consumption, and short-circuit power consumption. The static power consumption is due to leakage currents, and the dynamic power consumption is the power dissipated in the switching capacitive loads. The short-circuit power consumption is due to the current that flows between the supply voltage and ground when both the pull-up and pull-down networks in a CMOS gate are both on simultaneously. The short-circuit power consumption is typically small (10% of the total power consumption), and therefore it is negligible [5]. The static and dynamic power consumptions are described by Eqs. (15) and (16), respectively [2].

$$P_{static} = I_{leakage} V_{DD} \tag{15}$$

$$P_{dynamic} = a C_L f V_{DD}^2, \tag{16}$$

$$P_{dynamic} = a C_L f V_{DD}^2,$$

where $I_{leakage}$ is the total leakage current, a is the activity factor, f is the switching frequency. The energy dissipated in a circuit is important because it determines the battery life. The energy/cycle consumed by a device is the sum of the dynamic and static energy. The total energy/cycle is calculated by the total power dissipated in a cycle multiplied by the cycle time (T). This is shown by [2]

$$E_{total} = P_{total} T \tag{17}$$

The static and dynamic energy dissipated in a circuit are expressed by Eqs. (18) and (19), respectively [2].

$$E_{static} = I_{leakage} V_{DD} T \tag{18}$$

$$E_{dynamic} = a\, C_L\, V_{DD}^2, \tag{19}$$

where $I_{leakage}$ is the total leakage current, T is the cycle time (period).

2.9 Power-Delay-Product and Energy-Delay-Product

The power-delay-product (PDP) is the energy consumed by the gate per switching event. The ring oscillator is the circuit of choice for measuring the PDP of a logic family [2]. The PDP as a quality measure for a logic gate is expressed by [2]

$$PDP = P\, t_p = C_L\, V_{DD}^2\, f\, t_p = \frac{C_L\, V_{DD}^2}{2} \tag{20}$$

The PDP measures the energy needed to switch a gate. This number can be made low by reducing the supply voltage. To keep the PDP as low as possible, I can reduce the supply voltage to its minimum value as long as it ensures the functionality. This comes at the cost of performance.

Another metric that involves both the energy consumption and the performance is the energy-delay-product (EDP) [2]:

$$EDP = PDP\, t_p = \frac{C_L V_{DD}^2}{2}\, t_p. \tag{21}$$

The EDP is a combined metric that brings the energy and the performance together, and it is often used as the ultimate quality metric.

Chapter 3

MOSFET Behaviour in Subthreshold Region

As discussed in Chapter 2, due to the INWE, the threshold voltage of the MOSFET varies with the variation of the transistor width. Therefore, due to the exponential relationship of the drain current with the threshold voltage in the subthreshold region, the drain current in the subthreshold region is not linearly proportional to the transistor width. In other words in a typical MOSFET, increasing the MOSFET's channel width increases its threshold voltage and, hence, the current initially decreases showing a minimum point. In this chapter, I explore the effect of current and threshold voltage behaviour versus width of a MOSFET in the subthreshold region. The concept of current-over-capacitance ratio versus width is also introduced in this chapter.

3.1 Threshold Voltage in Subthreshold Region

As discussed in Chapter 2, modern nano-meter technologies, that are processed based on STI, are affected by the INWE. Because of the INWE, the threshold voltage increases with the increases of in the widths. The threshold voltage behaviour versus

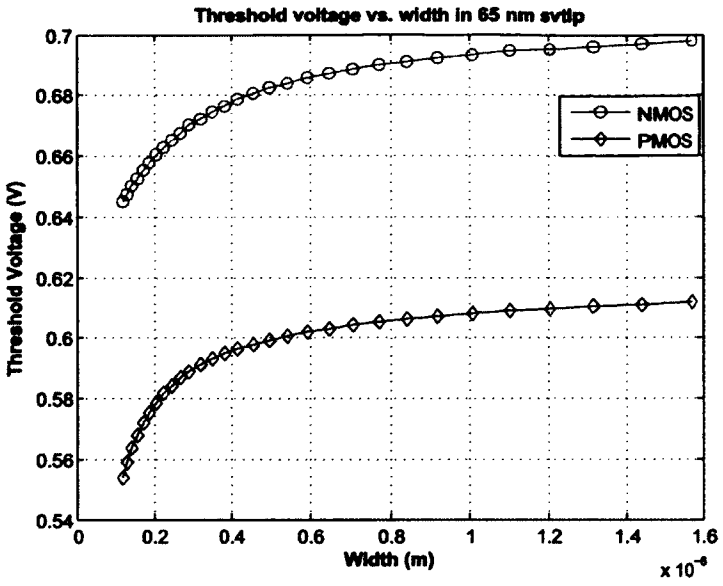

Figure 3.1: Threshold voltage vs. width for NMOS and PMOS transistors in 65 nm CMOS technology

width for an NMOS and a PMOS transistor in several CMOS technologies biased at the supply voltage of 0.2 V are plotted in Figures (3.1-3.4). As illustrated in these figures, the threshold voltage initially increases with the increase in the width before reaching a fixed value.

3.2 MOSFET Currents in Subthreshold Region

Quoting Eq. 4 here again shows that the subthreshold current is linearly and exponentially proportional to the width (W) and the threshold voltage (V_{th}), respectively.

$$I_{DS} = \mu C_{ox} \frac{W}{L} (n-1) V_T^2 e^{(\frac{V_{GS}-V_{th}}{nV_T})} \tag{22}$$

In Figures (3.5-3.8), the subthreshold current versus width for NMOS and PMOS transistors biased at the supply voltage of 0.2 V for the four technologies are plotted.

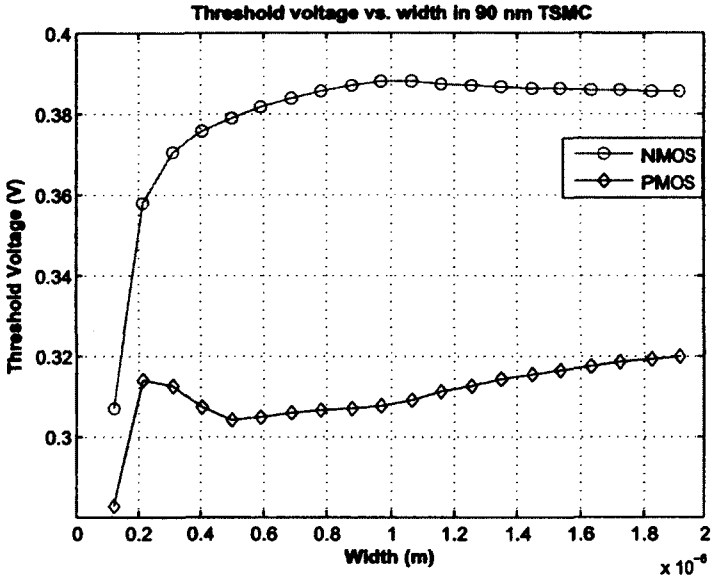

Figure 3.2: Threshold voltage vs. width for NMOS and PMOS transistors in 90 nm CMOS technology

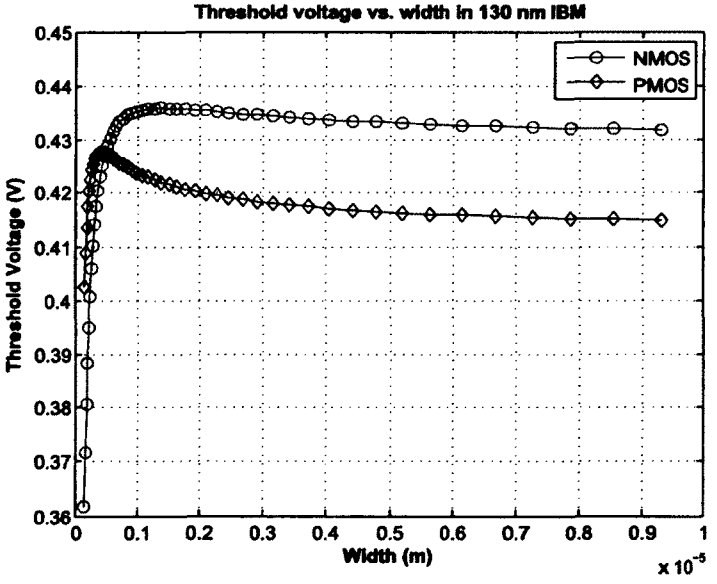

Figure 3.3: Threshold voltage vs. width for NMOS and PMOS transistors in 130 nm CMOS technology

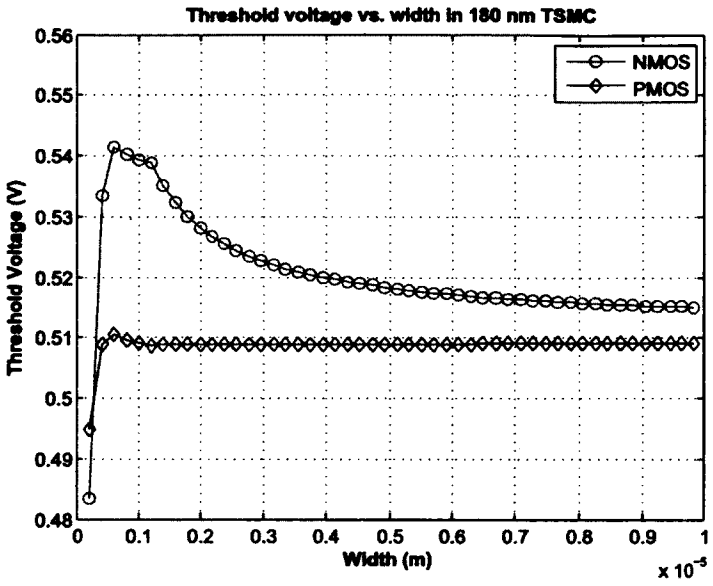

Figure 3.4: Threshold voltage vs. width for NMOS and PMOS transistors in 180 nm CMOS technology

As shown in these figures, due to the INWE, the current is not linearly proportional to the width at narrow-width sizes. In some cases (e.g., NMOS transistor in 130 nm CMOS technology), the MOSFET current decreases while the width increases at the narrow widths. As discussed in the previous section, the threshold voltage changes while the width of a transistor increases. Since the subthreshold current is exponentially related to the thresold voltage it decreases while the width increases at narrower widths. At larger widths, where the threshold voltage reaches its final value, the subthreshold current increases linearly as the width increases. As shown in Figure 3.8, the current of the NMOS transistor starts decreasing at the minimum width size (220 nm), but it then starts increasing at the width of 0.5 μm. At the width of 2.2 μm, it reaches the same current value of the minimum-width, and above 2.2 μm it continues to increase.

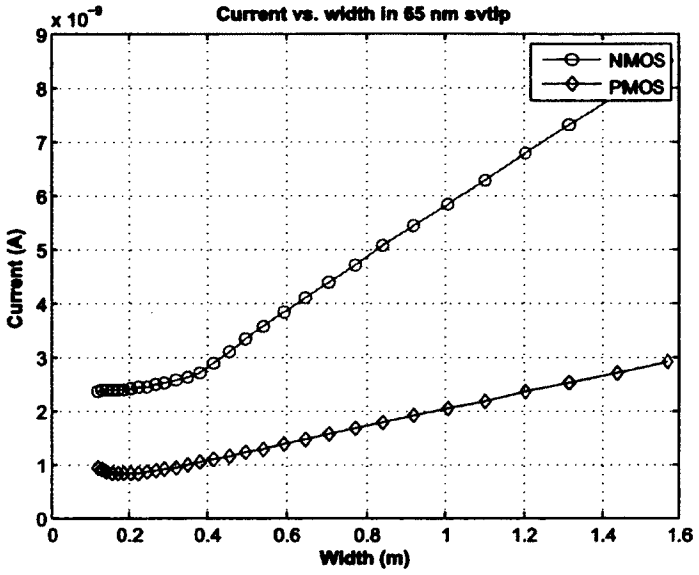

Figure 3.5: Current vs. width for NMOS and PMOS transistors in 65 nm CMOS technology

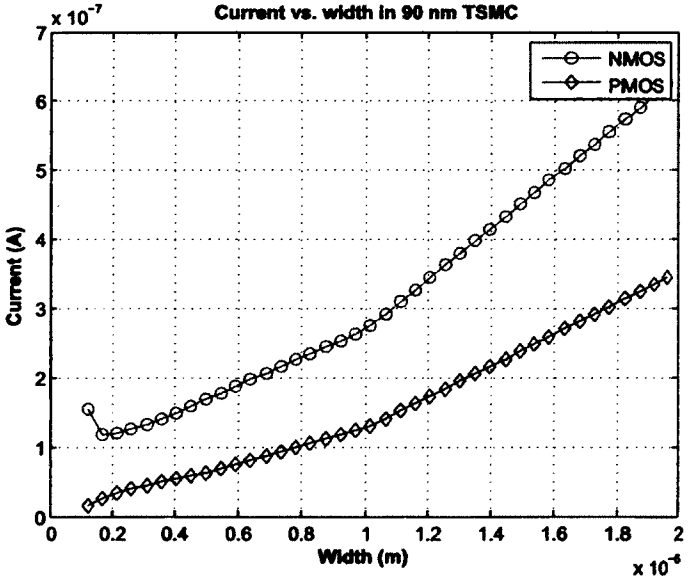

Figure 3.6: Current vs. width for NMOS and PMOS transistors in 90 nm CMOS technology

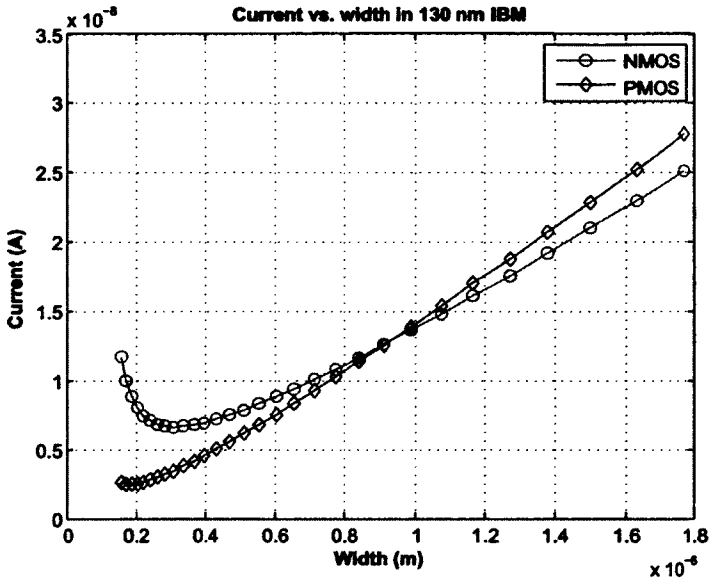

Figure 3.7: Current vs. width for NMOS and PMOS transistors in 130 nm CMOS technology

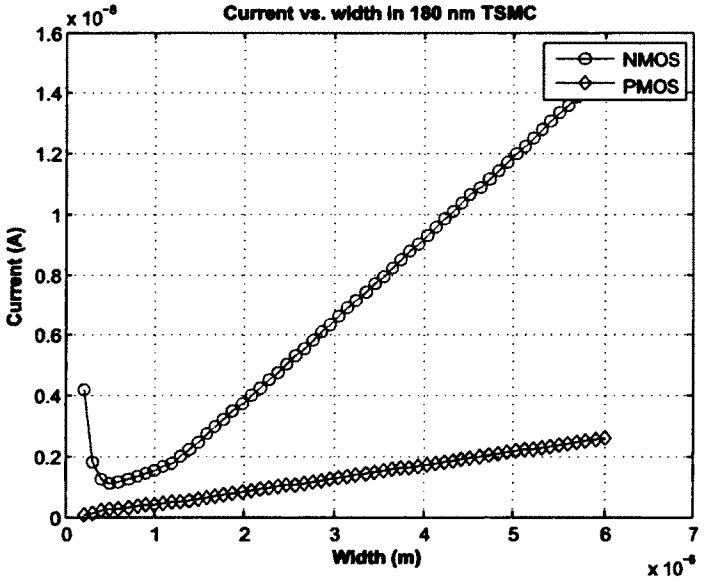

Figure 3.8: Current vs. width for NMOS and PMOS transistors in 180 nm CMOS technology

3.3 Capacitances in Subthreshold Region

The speed of a circuit is inversely proportional to the total effective capacitance involved in the switching activity as shown in Eq. (14) [2]. To find out the total capacitance behaviour of a MOSFET versus its channel-width in the subthreshold region, I simulated the total capacitance of each NMOS and PMOS transistor in the four CMOS technologies as the sum of the total gate capacitances and one of the diffusion capacitances (This models the cases, where one of the diffusion capacitances is connected to a DC voltage, and is not involved in the switching activity, like the transistors in an inverter). Figures (3.9-3.12) show the total capacitance versus width for NMOS and PMOS transistors for the four CMOS technologies. As shown in these figures, the total capacitance in the subthreshold region is linearly proportional to the transistor width.

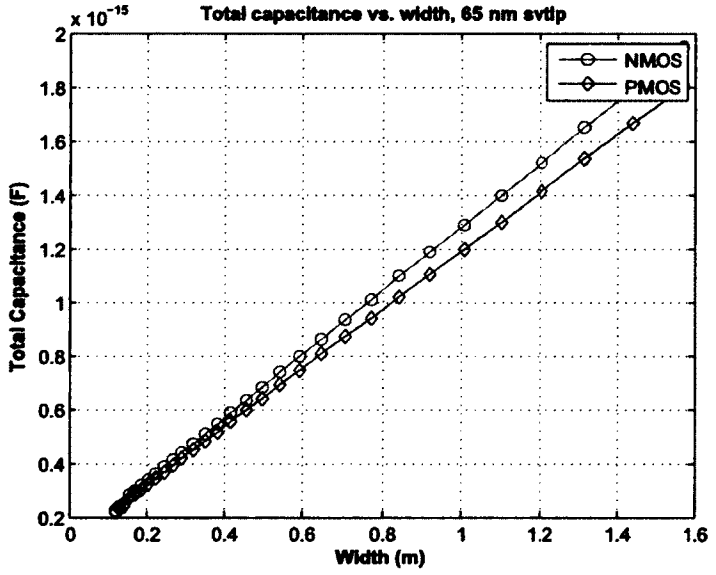

Figure 3.9: Total capacitance vs. width for NMOS and PMOS transistors in 65 nm CMOS technology

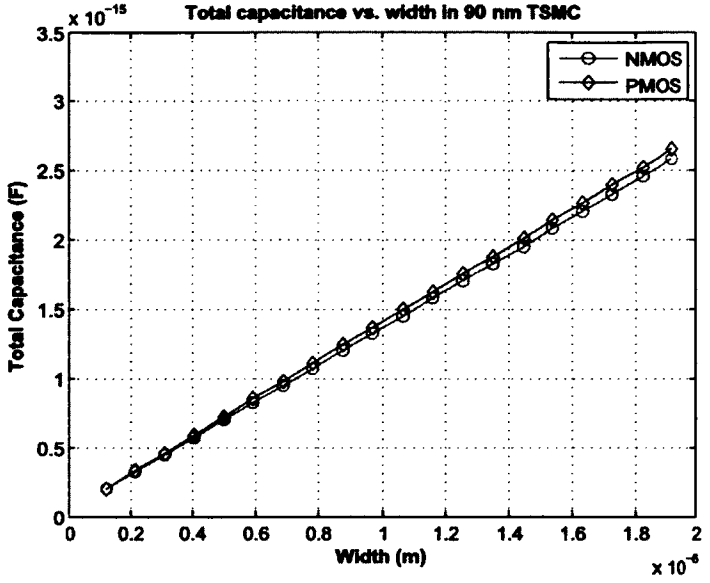

Figure 3.10: Total capacitance vs. width for NMOS and PMOS transistors in 90 nm CMOS technology

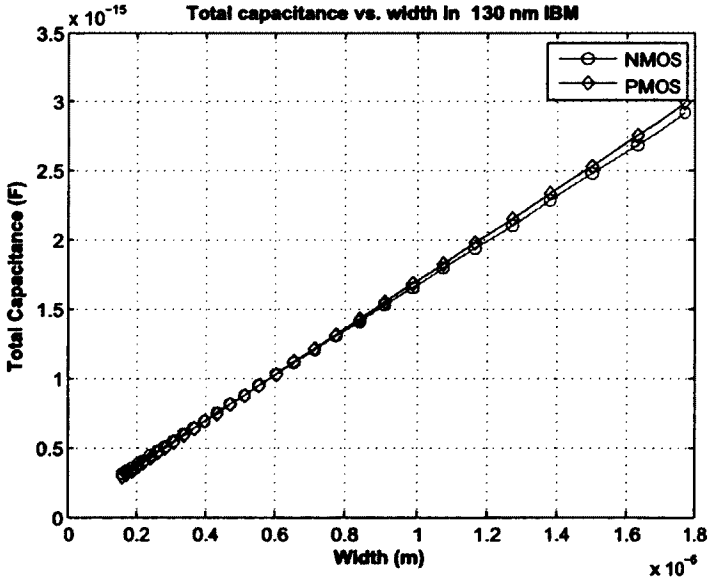

Figure 3.11: Total capacitance vs. width for NMOS and PMOS transistors in 130 nm CMOS technology

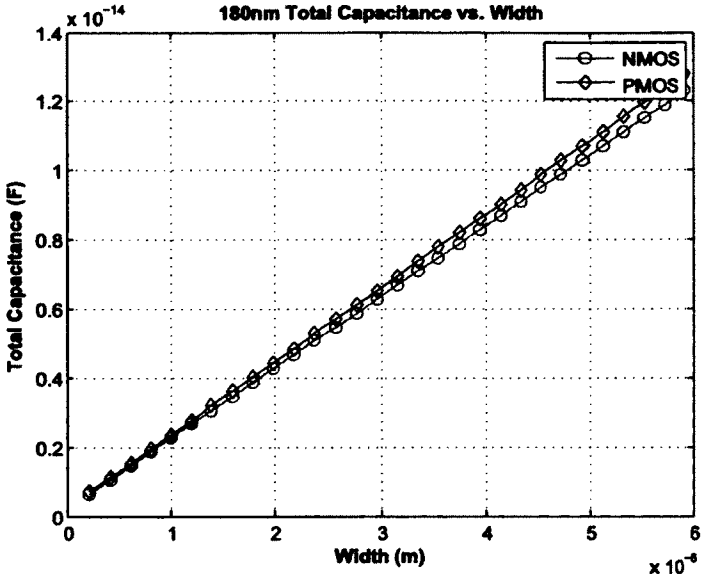

Figure 3.12: Total capacitance vs. width for NMOS and PMOS transistors in 180 nm CMOS technology

3.4 Current-Over-Capacitance Ratio in Sub-threshold Region

Quoting Eq. 14 here again, the minimum propagation delay (maximum speed) is achieved by lowering the effective switching capacitance and increasing the average current.

$$t_p = \frac{C_L V_{DD}}{2}\left(\frac{1}{I_N} + \frac{1}{I_P}\right)$$

As discussed in Section 3.3, the total MOSFET capacitance increases linearly as the width increases in CMOS technologies. As also discussed in Section 3.2, due to the INWE, some minimum-width devices have higher currents compared to the moderately wider devices. As a result, designing logic gates with devices close to the minimum-width size tends to have higher currents and lower capacitances [4]. To

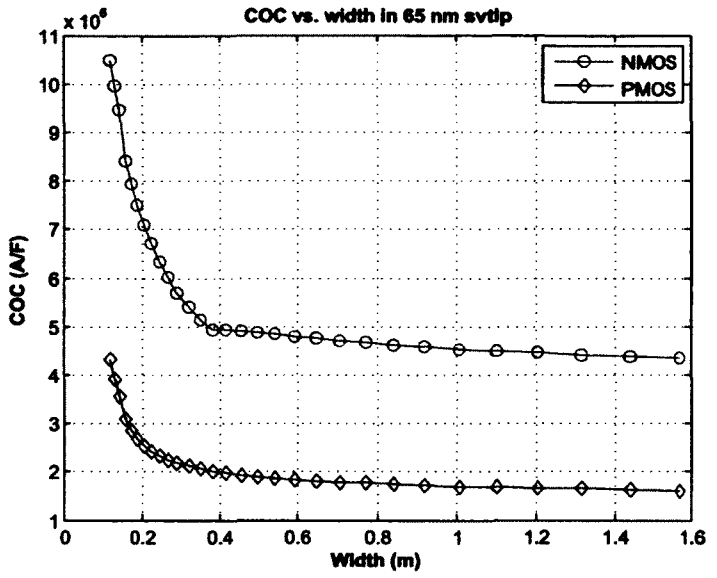

Figure 3.13: Current-over-capacitance ratio vs. width for NMOS and PMOS transistors in 65 nm CMOS technology

obtain the width (W_{INWE}) at which the maximum COC ratio occurs for each device, I plot the COC ratio for NMOS and PMOS transistors in four CMOS technologies in Figures (3.13-3.16).

As shown in Figure 3.13, the maximum COC ratio occurs at the minimum width in the 65 nm (LP) CMOS technology for both NMOS and PMOS transistors. In the 90 nm CMOS technology (Figure 3.14), the maximum COC ratio for the NMOS transistor occurs at the minimum width, and the COC ratio for the PMOS transistor has a local maximum at the width of 240 nm. However, the COC ratio for widths larger than 1.2 μm have a higher values compared with the local maximum at the width of 240 nm. In the 130 nm CMOS technology (Figure 3.15), the maximum COC ratio occurs at the minimum width for the NMOS transistor, and for the PMOS transistor the maximum COC ratio for widths less than 1.6 μm occurs at the minimum width. In the 180 nm CMOS technology (Figure 3.16), the maximum COC ratio for the NMOS transistor occurs at the minimum width, and for the PMOS transistor it

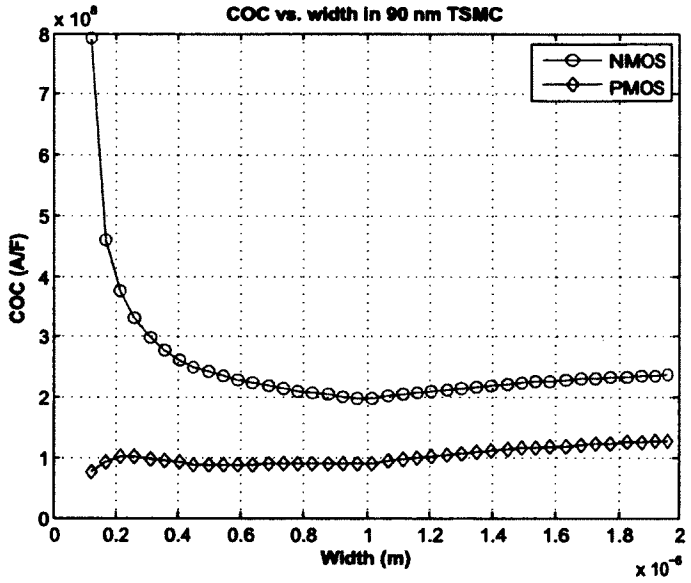

Figure 3.14: Current-over-capacitance ratio vs. width for NMOS and PMOS transistors in 90 nm CMOS technology

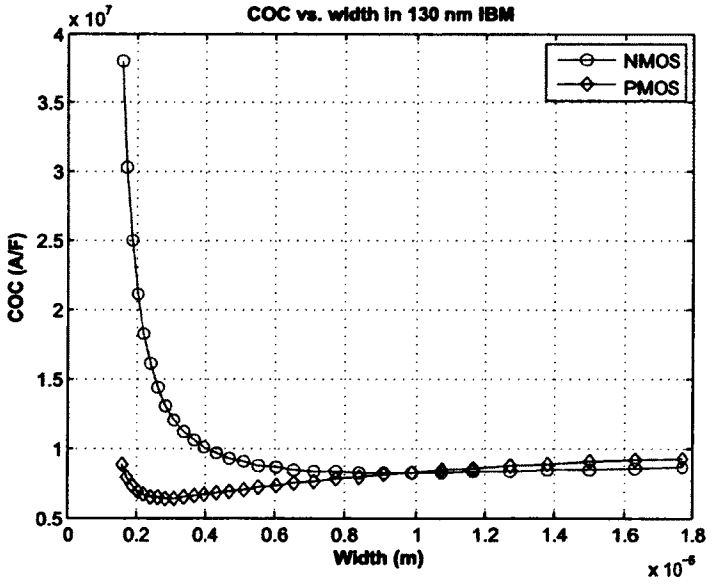

Figure 3.15: Current-over-capacitance ratio vs. width for NMOS and PMOS transistors in 130 nm CMOS technology

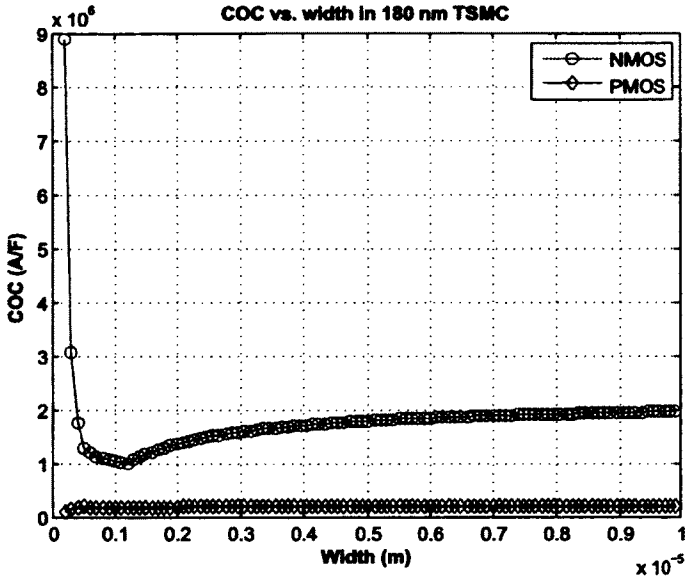

Figure 3.16: Current-over-capacitance ratio vs. width for NMOS and PMOS transistors in 180 nm CMOS technology

occurs at 500 nm. Table 3.1 presents PMOS-to-NMOS width ratio (β) and W_{INWE} of the NMOS and PMOS transistors. In Chapter 4, I develope an analytical model to obtain the optimum β at which the maximum COC ratio occurs in each CMOS technology. In Chapter 5, I show how to exploit the optimum β in PTS to design faster subthreshold circuits.

Table 3.1: The NMOS and PMOS Transistor Width Sizes for the Maximum COC Ratio

Technology Kit	W_n	W_p	$\beta = \frac{W_p}{W_n}$
180 nm, TSMC	220 nm	500 nm	2.3
130 nm, IBM	160 nm	160 nm	1
90 nm, TSMC	120 nm	240 nm	2
65 nm, ST(LP)	120 nm	120 nm	1

Chapter 4

Delay Optimization in Subthreshold Operation

In Chapter 3, I reported the β at which the maximum COC ratio occurs in the subthreshold region using simulations for each CMOS technology. In this chapter, I develop an analytical expression for obtaining the optimum β that minimizes the delay (β_{opt}) in subthreshold CMOS circuits. This is done by modelling the delay of an inverter driving an identical inverter. In later chapters I incorporate this β to design faster subthreshold circuits.

4.1 Delay Modelling in Subthreshold Region

In this section, I model the propagation delay (t_p) of an inverter driving an identical inverter, as shown in Figure 4.1, taking the β into account. The sizing results from the modelled inverter can then be applied on any CMOS logic gate in order to design faster subthreshold circuits.

In Figure 4.1, W_{n1} (W_{n2}) and W_{p1} (W_{p2}) denote the channel widths of the NMOS and PMOS transistors in the first and second inverters, respectively. L_n and L_p are the channel lengths of NMOS and PMOS transistors, respectively. C_L represents all

Node B capacitances, with the exception of wire capacitances. The width of each PMOS transistor is sized β times larger than that of the NMOS transistor in the inverters. In addition, the minimum technology feature size is used for the channel length of the NMOS and PMOS transistors in each inverter.

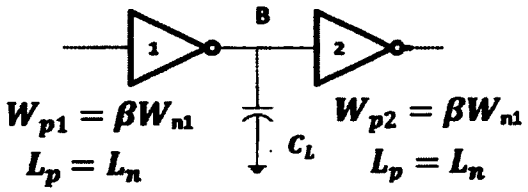

Figure 4.1: An inverter driving an identical inverter

As described in Chapter 2, the propagation delay of the first inverter in Figure 4.1 can be expressed as

$$t_p = \frac{t_{PHL} + t_{PLH}}{2} = \frac{C_L V_{DD}}{2}\left(\frac{1}{I_N} + \frac{1}{I_P}\right) \tag{23}$$

where t_{PHL} and t_{PLH} are the propagation delays of the high-to-low and low-to-high transitions, respectively. Parameters I_N and I_P are the subthreshold currents of the NMOS and PMOS transistors of the first inverter, respectively, and according to Chapter 2, can be expressed as

$$I_N = \mu_n C_{ox} \frac{W_n}{L}(n-1)V_T^2 e^{\frac{(V_{GS}-V_{tn})}{nV_T}} \tag{24}$$

$$I_P = \mu_p C_{ox} \frac{W_p}{L}(n-1)V_T^2 e^{\frac{(V_{GS}-V_{tp})}{nV_T}}, \tag{25}$$

Although n varies between 1.3 to 1.5, for convenience it can be assumed to be equal for NMOS and PMOS transistors.

By substituting Eqs. (24) and (25) in Eq. (23), the propagation delay of the first

inverter in Figure 4.1 can be expressed as

$$t_p = \frac{C_L V_{DD}}{2}\left(\frac{1}{\frac{C_{ox}}{L}(n-1)V_T^2}\right)\left(\frac{1}{\mu_n W_n e^{\frac{(V_{GS}-V_{tn})}{nV_{th}}}} + \frac{1}{\mu_p W_p e^{\frac{(V_{GS}-|V_{tp}|)}{nV_T}}}\right). \tag{26}$$

By assuming an ideal rail-to-rail step input signal (i.e. $V_{GS} = V_{DD}$) and defining the following parameters,

$$\alpha_n = e^{\frac{-V_{tn}}{nV_T}}, \alpha_p = e^{\frac{-|V_{tp}|}{nV_T}}, \Gamma = \frac{\alpha_n}{\alpha_p} = e^{\frac{|V_{tp}|-V_{tn}}{nV_T}}, \Lambda = \frac{\mu_n}{\mu_p}, \beta = \frac{W_p}{W_n}, \tag{27}$$

Eq. 26 can be reduced to

$$t_p = \frac{C_L V_{DD}}{2} \frac{1}{e^{\frac{V_{DD}}{nV_T}} \mu_n W_n \frac{C_{ox}}{L}(n-1)V_T^2} \frac{\Lambda}{\alpha_n \alpha_p}\left(\frac{\alpha_n + \frac{1}{\Lambda}\beta\alpha_p}{\beta}\right). \tag{28}$$

Let C_N and C_P denote the total capacitance at Node B associated with all NMOS and PMOS transistors, respectively. These capacitances include the drain capacitances of the NMOS and PMOS transistors of the first inverter, and the total gate capacitance of the succeeding inverter. Therefore, the total load capacitance at Node B is $C_L = (1+\beta)C_N$, which by substituting into Eq. (28), leads to

$$t_p = \frac{(1+\beta)C_N V_{DD}}{2} \frac{1}{e^{\frac{V_{DD}}{nV_T}} \mu_n W_n \frac{C_{ox}}{L}(n-1)V_T^2} \frac{1}{\frac{1}{\Lambda}\alpha_n \alpha_p}\left(\frac{\alpha_n + \frac{1}{\Lambda}\beta\alpha_p}{\beta}\right). \tag{29}$$

Note that although this result is obtained for an inverter, any complex CMOS logic gate can be modelled by an equivalent inverter.

4.2 Maximum Frequency of Operation

Differentiating Eq. (29) and equating it to 0 allows us to solve for β_{opt} at which the maximum operating frequency of the inverter in the subthreshold region occurs

$$\partial t_p/\partial\beta = \frac{\partial[\frac{(1+\beta)C_N V_{DD}}{2}\frac{1}{\frac{V_{DD}}{e^{nV_T}}\mu_n W_n\frac{C_{ox}}{L}(n-1)V_T^2}\frac{1}{\frac{1}{\Lambda}\alpha_n\alpha_p}(\frac{\alpha_n+\frac{1}{\Lambda}\beta\alpha_p}{\beta})]}{\partial\beta} \tag{30}$$

$$\partial t_p/\partial\beta = (\frac{C_N V_{DD}}{2})(\frac{1}{\frac{V_{DD}}{e^{nV_T}}\mu_n W_n\frac{C_{ox}}{L}(n-1)V_T^2})(\frac{1}{\frac{1}{\Lambda}\alpha_n\alpha_p})\frac{\partial[(1+\beta)(\frac{\alpha_n+\frac{1}{\Lambda}\beta\alpha_p}{\beta})]}{\partial\beta} \tag{31}$$

$$\partial t_p/\partial\beta = \overbrace{(\frac{C_N V_{DD}}{2})(\frac{1}{\frac{V_{DD}}{e^{nV_T}}\mu_n W_n\frac{C_{ox}}{L}(n-1)V_T^2})(\frac{1}{\frac{1}{\Lambda}\alpha_n\alpha_p})}^{A}\frac{\partial[\frac{\alpha_n+\frac{1}{\Lambda}\beta\alpha_p+\beta\alpha_n+\frac{1}{\Lambda}\beta^2\alpha_p}{\beta})]}{\partial\beta} \tag{32}$$

$$\partial t_p/\partial\beta = A\frac{\partial[\frac{\frac{1}{\Lambda}\beta\alpha_p+\beta\alpha_n+2\frac{1}{\Lambda}\beta^2\alpha_p-\alpha_n-\frac{1}{\Lambda}\beta\alpha_p-\beta\alpha_n-\frac{1}{\Lambda}\beta^2\alpha_p}{\beta}]}{\partial\beta} \tag{33}$$

$$\partial t_p/\partial\beta = A(\frac{\frac{1}{\Lambda}\beta^2\alpha_p-\alpha_n}{\beta^2}). \tag{34}$$

By equating Eq. (34) to 0:

$$\partial t_p/\partial\beta = \frac{\frac{1}{\Lambda}\beta^2\alpha_p - \alpha_n}{\beta^2} = 0 \tag{35}$$

$$\frac{1}{\Lambda}\beta^2\alpha_p - \alpha_n = 0 \tag{36}$$

$$\beta_{opt} = \sqrt{\frac{\Lambda\alpha_n}{\alpha_p}} = \sqrt{\Lambda}.\sqrt{\Gamma} \tag{37}$$

In [2], β_{opt} in the superthreshold region is reported equal to $\sqrt{\Lambda}$ which typically is between 1.5 to 2. However, as shown in Eq. (37), β_{opt} in the subthreshold region is the superthreshold β_{opt} multiplied by $\sqrt{\Gamma}$. I refer to $\sqrt{\Gamma}$ the subthreshold β_{opt}

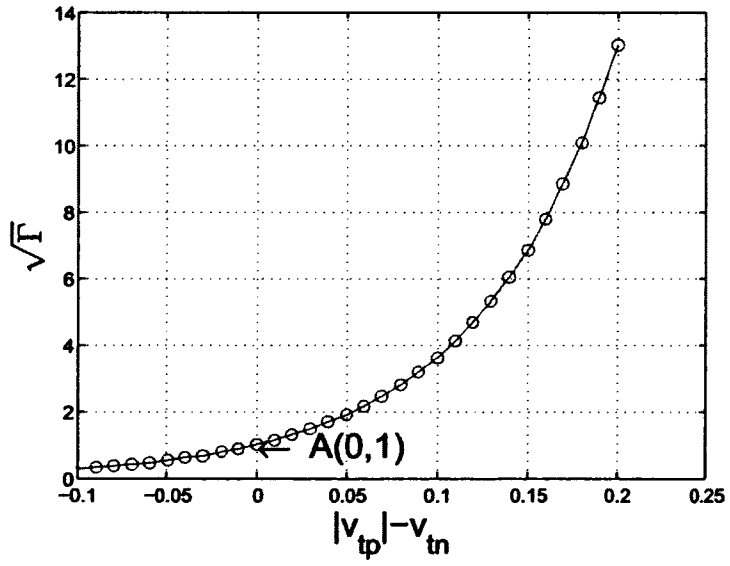

Figure 4.2: Subthreshold B_{opt} modification factor vs. difference between PMOS and NMOS threshold voltages

modification factor. Factor $\sqrt{\Gamma}$ is exponentially related to the difference of PMOS and NMOS threshold voltages, as shown in Eq. 27. Figures. 3.1-3.4 illustrate that the difference between the PMOS threshold voltage and NMOS threshold voltage ($|V_{tp}|$-V_{tn}) in different technologies varies between -0.1 to 0.2. Factor $\sqrt{\Gamma}$ versus ($|V_{tp}|$-V_{tn}) is plotted in Figure 4.2, assuming $n = 1.5$ for both PMOS and NMOS transistors. As shown in this figure, the range of this factor is between 0.27 to 13. The following comments can be raised based on the results.

1. The effect of $|V_{tp}|$-V_{tn} on β_{opt} in the subthreshold region is much more than the case of superthreshold. For example, if in a technology, $|V_{tp}|$-$V_{tn} = 0.2(0.1)$, which is not usual, $\beta_{opt} = 20(5)$. That is, the frequency of operation is maximum when the PMOS transistor is made 20(5) times larger than the NMOS transistor.

2. If $|V_{tp}| = V_{tn}$, then the subthreshold β_{opt} and the superthreshold β_{opt} would be identical (point A in Figure 4.2).

3. The most suitable technologies for subthreshold operation are those with $|V_{tp}| < V_{tn}$ (i.e. $\sqrt{\Gamma} < 1$). The optimum circuits implemented in these technologies are smaller and, hence, consume lower amounts of energy.

4. β_{opt} is independent of power supply voltage, V_{DD}. This is also verified later through simulations (Figures. 4.3 and 4.4)

To calculate β_{opt} for each technology, I have extracted the model parameters for the four CMOS technologies and presented in Table 4.1. Since all NMOS transistors are affected by the INWE, I choose the threshold voltage at the minimum width. However, since some of the PMOS transistors are affected by the INWE and some not, I choose $V_t = V_{th0}$ (i.e. the zero-biased threshold voltage for large widths).

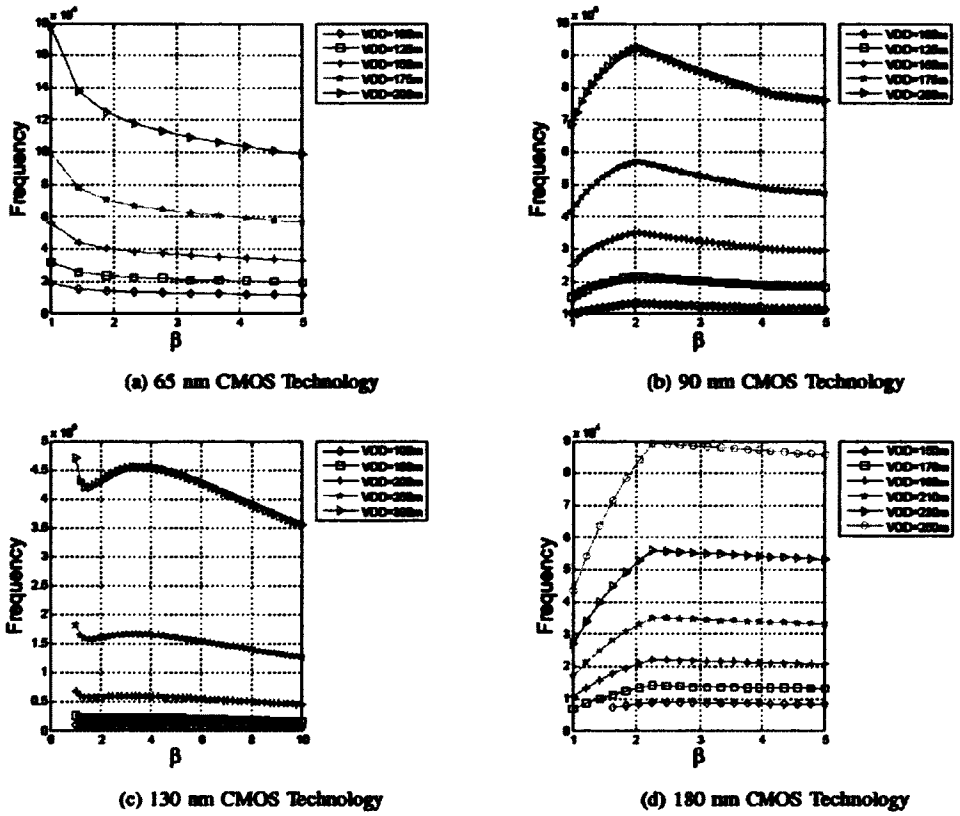

Figure 4.3: Frequency of a 19-stage inverter ring oscillator vs. β

Next, I compare the theoretical β_{opt} value to those obtained through simulations. The simulaton results for a 19-stage inverter ring oscillator is shown in Figure 4.3 depicting the operational frequency versus β for different supply voltages for the four CMOS technologies operating in the subthreshold region. Note that in all simulations the minimum width is selected for all NMOS transistors. Doing so has two advantages. First, choosing minimum-width NMOS devices causes the area to be smaller. Second, since the maximum COC ratio of NMOS transistors in each technology is attained at the minimum width, the maximum speed is also achieved.

Table 4.2 shows β_{opt} obtained from the theoretical analysis and simulations results. Also in column COC I show the β_{opt} obtained from the COC smimulations presented in Chapter 3. The table results show that β_{opt} from COC simulations are exactly

the same as the β_{opt} obtained from the 19-stage inverter ring oscillator simulations. In other words, one can find the β_{opt} from only COC simulations. The predicted values from the analytical model and the simulation results are very close in the 180 nm, 90 nm, and 65 nm CMOS technologies. However, there are discrepancies between these two values for the 130 nm CMOS technology. The reason is that the subthreshold current does not predict the INWE. It assumes a linear relationship between the current and the transistor's width. Moreover, the threshold voltage is assumed to be constant, that is, not prone to the INWE. Note that the PMOS transistors in the 180 nm and 90 nm CMOS technologies are not affected by the INWE. However, in the 130 nm and, to a lesser extent, the 65 nm CMOS technologies both PMOS and NMOS transistors are affected by the INWE. The readers may refer to Chapter 3 for verification.

For the 130 nm CMOS technology, there are two maximums, as shown in Figure 4.3. The maximum frequency of operation occurs at $\beta = 1$, which is caused by the INWE, and a local maximum occurs at $\beta = 3.6$. Our model predicts a value of $\beta_{opt} = 4.1$. As presented in Table 4.2, in the 65 nm technology the β_{opt} obtained from our model is 1.1, whereas the simulation results show $\beta = 1$. I have also simulated the difference in operational frequencies (ΔF) and energy consumption (ΔE) at β_{opt} obtained from our analytical model and β_{opt} obtained from the 19-stage inverter ring oscillator simulations at $V_{DD} = 0.2$. I have calculated the (ΔF) and (ΔE) as

$$\Delta F = \frac{F_S - F_A}{F_A} \tag{38}$$

$$\Delta E = \frac{E_S - E_A}{E_A} \tag{39}$$

where $F_S(E_S)$ and $F_A(E_A)$ are the operating frequencies (energy consumptions) obtained from simulation results and analytical model. As shown in Table 4.2, the ΔF

in the 180 nm and 90 nm CMOS technologies are negligible and for the 130 nm and 65 nm CMOS technologies they are about 10%. The ΔE shows -4%, -60%, -0.02%, -4% in the 180 nm, 130 nm, 90 nm, and 65 nm CMOS technologies, respectively. In 130 nm CMOS technology $\beta_{opt} = 1$ should be chosen instead of $\beta = 4.1$ since the ΔE shows a 60% difference.

The following strategy is proposed to obtain β_{opt} for a given technology based on whether a transistor type in that technology experiences INWE or not.

1. choose the minimum width for NMOS transistors.
2. If the PMOS transistor is affected by the INWE, choose $\beta_{opt} = 1$.
3. If the PMOS transistor is not affected by the INWE, β_{opt} can be calculated using $\sqrt{\Lambda\Gamma}$.

Table 4.1: Technology Parameters for Calculating Analytical β_{opt}

Technology Kit	μ_n ($\frac{cm^2}{V.S}$)	μ_p ($\frac{cm^2}{V.S}$)	V_{tn} (mV)	$\lvert V_{tp}\rvert$ (mV)	$\sqrt{\Lambda}$	$\sqrt{\Gamma}$	β_{opt}
180 nm TSMC	37.1e-3	10.8e-3	484	510	1.8	1.38	2.5
130 nm IBM	440e-3	94e-3	362	415	2.1	1.93	4.1
90 nm TSMC	29.65e-3	8.787e-3	309.8	320	1.8	1.13	2.0
65 nm (svtlp) ST	398.7e-3	137e-3	646.2	610	1.7	0.63	1.0

Table 4.2: β_{opt} Obtained from Ring Oscillator Simulation Results, Analytical Model, and COC Simulations

Technology Kit	β_{opt} (Simulation)	β_{opt} (Analytical)	β_{opt} (COC)	ΔF	ΔE
180 nm TSMC	2.33	2.5	2.3	0.2%	-4%
130 nm IBM	1	4	1	10%	-60%
90 nm TSMC	2.05	2.02	2.02	0.02%	-0.02%
65 nm (svtlp) ST	1	1.1	1	10%	-4%

4.2.1 Complex Gates

I have also verified our analytical-model results through simulations of simple and complex logic gates. The simple gates chosen are NAND and NOR, and the complex gate is AND-OR-NOT (AORN). The simulations are performed for best-case (bc) and worst-case (wc) delay scenarios by changing the gate's input connection. The operational frequency of the simulated logic gates versus β at V_{DD} = 250 mV are plotted in Figure 4.4

The important conclusion is that the same β_{opt} obtained for the inverter is also found to be the most suitable β for simple and complex CMOS logic gates.

4.3 Energy consumption at β_{opt}

Since the energy consumption of subthreshold circuits is an important quality metric, I explore the energy consumption of the simulated circuits versus β. Calhoun et. al [9] have reported that to reduce the energy consumption, minimum-width devices should be used and to achieve the desirable performance (assuming it is achievable in the subthreshold region) the supply voltage should be increased. However, our work shows that this is not always correct. First, in some applications the supply voltage is fixed at a certain value and cannot be modified. Besides, I present some examples

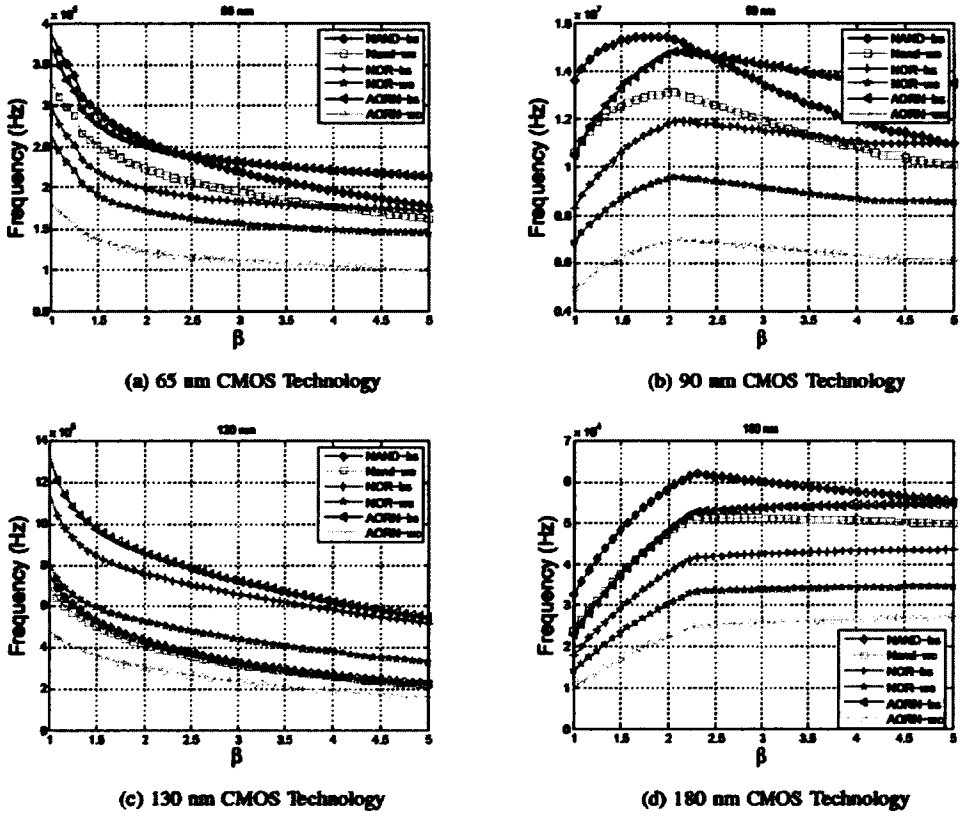

Figure 4.4: Frequency of a 19-stage NAND,NOR, and AND-OR-NOT ring oscillators for best case (bc) and worst case (wc) delay scenarios vs. β in the a) 65 nm b) 90 nm c) 130 nm d) 180 nm CMOS technologies

in this section that the minimum energy operation can be achieved by sizing the logic gates larger than the minimum size.

The results for energy versus β for different supply voltages in a 19-stage inverter ring oscillator are shown in Figure 4.5 in the four CMOS technologies. The energy versus β for the NAND, NOR, AND-OR-NOT logic gates are illustrated in Figure 4.6 at the supply voltage of $V_{DD} = 0.25$ V. As shown in Figure 4.5, using $\beta_{opt} = 1$ in the 65 nm and 130 nm CMOS technologies results in minimum energy at all voltages in the subthreshold region. In the 90 nm CMOS technology $\beta = 1.9$ results in the minimum energy operation at $V_{DD} = 0.1$. In the 180 nm technology, choosing $\beta = 2$ results in the minimum energy for the 19-stage inverter chain ring oscillator at $V_{DD} = 0.18$ V, and also this β minimizes the energy for the NOR and AND-OR-NOT circuits (both worst case and best case delay scenarios) as shown in Figure 4.6.

I have also compared the frequency of operation and energy consumption for the 19-stage ring oscillator implemented with minimum size devices and β_{opt}. The simulation results are shown in Tables 4.3 and 4.4. As shown in these tables, by designing circuits at $\beta = 1$ in the 65 nm and 130 nm CMOS technologies both the minimum energy and maximum speed is achieved. I have performed the simulations in these technologies at the supply voltage of 0.25 V. However, since the β_{opt} is not at the minimum size in the 180 nm and 90 nm CMOS technologies I performed simulation at $V_{DD} = V_{tmax}/2$ (half of the maximum threshold voltage) in each technology. I make sure that the simulations are done in the deep subthreshold region and not near threshold.

In the 90 nm CMOS technology, by using β_{opt}, the inverter ring oscillator shows 34.3% speed improvements compared to $\beta = 1$. This comes at the cost of 22.2% energy consumption. The speed improves by 19.5% and 32.1% for the NAND ring oscillator for the best case and worst case delay scenarios, respectively. This is while the energy consumption increases by 34.2% and 24.4% for the best case and worst

case delay scenarios, respectively. The NOR ring oscillators show 53.7% and 49.2% speed improvements for the best cases and worst case delay scenarios. This is while the energy consumption decreases by 6.43% and for the best case delay scenario and for the worst case delay scenario the energy remains the same. For the AND-OR-NOT ring oscillator the speed increases by 50.6% and 52.8% for the best case and worst case scenarios, respectively. The energy consumption increases by 3.2% for the best case delay scenario and it gets decreased by 2.79% for the worst case delay scenario.

In the 180 nm CMOS technology, by using β_{opt}, the inverter ring oscillator shows 106% speed improvements compared to $\beta = 1$. This comes at the cost of 21.5% energy consumption. The speed improves by 88.7% and 115% for the NAND ring oscillator for the best case and worst case scenarios. This is while the energy consumption increases by 29.3% and 17.2% for the best case and worst case delay scenarios, respectively. The NOR ring oscillators show 134% and 138% speed improvements for the best case and worst case delay scenarios, respectively. This is while the energy consumption decreases by 18% and 14.2% for the best case and worst case delay scenarios, respectively. For the AND-OR-NOT ring oscillator the speed increases by 50.6% and 52.8% for the best case and worst case scenarios, respectively. The energy consumption decreases by 3.59% and 19% for the best case and worst case delay scenarios.

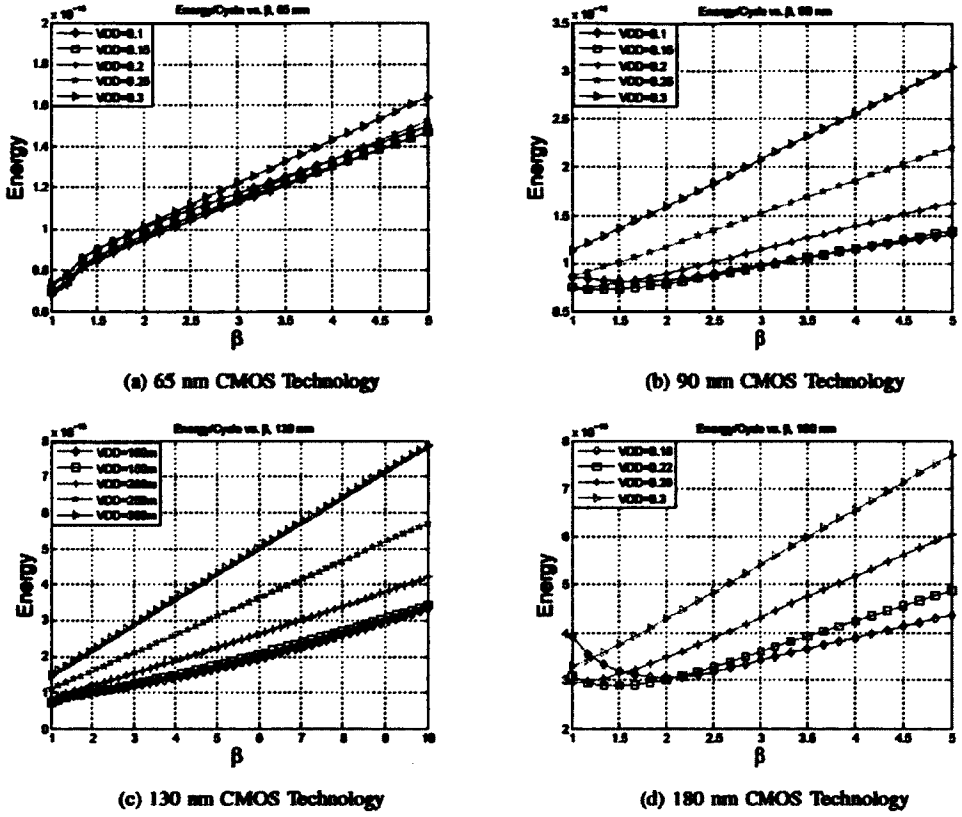

Figure 4.5: Energy of a 19-stage inverter ring oscillator vs. β

The following items can be concluded from the simulations results:

1. minimum-size devices are not always the best in terms of energy consumption.

2. β_{opt} is not always equal to one.

3. In some cases, energy consumption becomes minimum when $\beta > 1$.

4. The minimum energy of operation depends on β.

5. As V_{DD} decreases, minimum energy operation occurs at higher values of β.

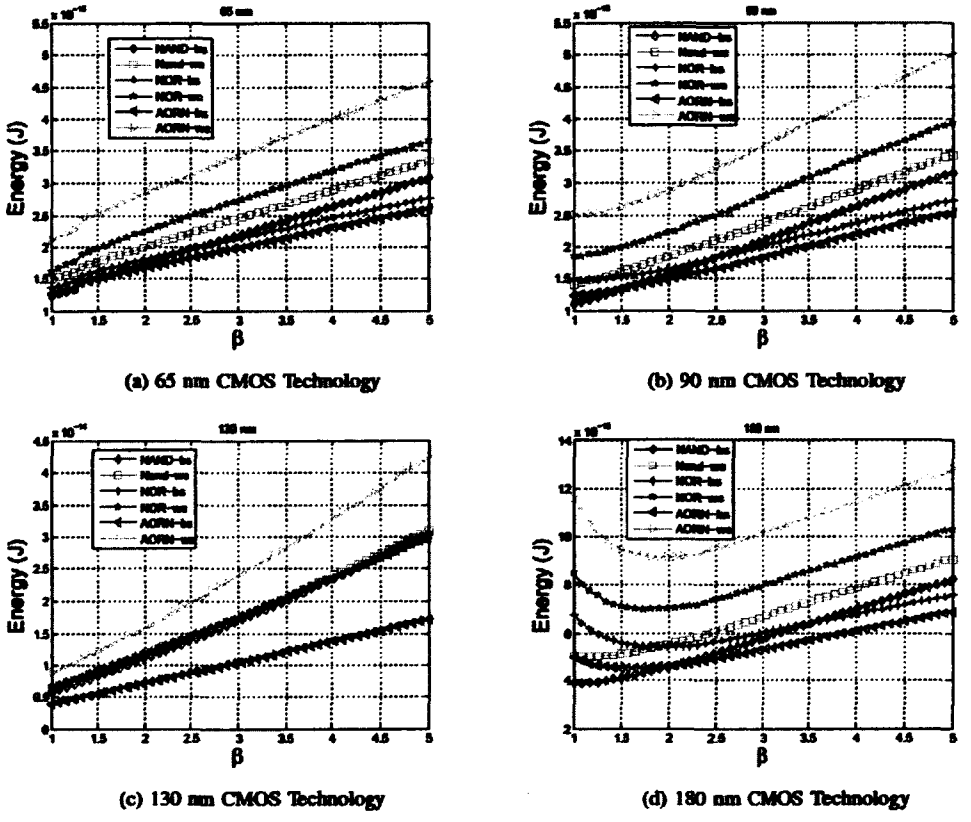

Figure 4.6: Frequency of a 19-stage NAND, NOR, and AND-OR-NOT ring oscillators for best case (bc) and worst case (wc) delay scenarios vs. β in the a) 65 nm b) 90 nm c) 130 nm d) 180 nm CMOS technologies

Table 4.3: Comparison of Frequency of Operation and Energy Consumption for 19-stage Ring Oscillator implemented for Minimum Size and Optimum β in the 65 and 90 nm CMOS Technologies

65 nm CMOS Technology, $V_{DD} = 0.25V$						
Logic Gates	Frequency			Energy		
	$\beta = 1$	β_{opt}	**ΔF**	$\beta = 1$	β_{opt}	**ΔE**
Inv	6.72E+05	6.72E+05	0%	1.06E-15	1.06E-15	0%
NAND-bc	3.86E+05	3.86E+05	0%	1.24E-15	1.24E-15	0%
NAND-wc	3.28E+05	3.28E+05	0%	1.50E-15	1.50E-15	0%
NOR-bc	2.96E+05	2.96E+05	0%	1.36E-15	1.36E-15	0%
NOR-wc	2.62E+05	2.62E+05	0%	1.62E-15	1.62E-15	0%
AORN-bc	3.71E+05	3.71E+05	0%	1.27E-15	1.27E-15	0%
AORN-wc	1.80E+05	1.80E+05	0%	2.10E-15	2.10E-15	0%
90 nm CMOS Technology, $V_{DD} = 0.2V$						
Logic Gates	Frequency			Energy		
	$\beta = 1$	β_{opt}	**ΔF**	$\beta = 1$	β_{opt}	**ΔE**
Inv	6.88E+06	9.24E+06	34.3%	7.30E-16	8.92E-16	22.2%
NAND-bc	5.13E+06	6.13E+06	19.5%	9.54E-16	1.28E-15	34.2%
NAND-wc	3.96E+06	5.23E+06	32.1%	1.19E-15	1.48E-15	24.4%
NOR-bc	3.07E+06	4.72E+06	53.7%	1.71E-15	1.60E-15	-6.43%
NOR-wc	2.50E+06	3.73E+06	49.2%	2.10E-15	2.10E-15	0.00%
AORN-bc	3.87E+06	5.83E+06	50.6%	1.25E-15	1.29E-15	3.20%
AORN-wc	1.80E+06	2.75E+06	52.8%	2.87E-15	2.79E-15	-2.79%

Table 4.4: Comparison of Frequency of Operation and Energy Consumption for 19-stage Ring Oscillator implemented for Minimum Size and Optimum β in the 130 and 180 nm CMOS Technologies

130 nm CMOS Technology, $V_{DD} = 0.25V$						
Logic Gates	Frequency			Energy		
	$\beta = 1$	β_{opt}	**ΔF**	$\beta = 1$	β_{opt}	**ΔE**
Inv	1.82E+06	1.82E+06	0%	1.09E-15	1.09E-15	0%
NAND-bc	7.46E+05	7.46E+05	0%	5.79E-15	5.79E-15	0%
NAND-wc	6.90E+05	6.90E+05	0%	6.27E-15	6.27E-15	0%
NOR-bc	1.13E+06	1.13E+06	0%	4.08E-15	4.08E-15	0%
NOR-wc	7.98E+05	7.98E+05	0%	6.64E-15	6.64E-15	0%
AORN-bc	1.31E+06	1.31E+06	0%	3.98E-15	3.98E-15	0%
AORN-wc	4.84E+05	4.84E+05	0%	8.93E-15	8.93E-15	0%
180 nm CMOS Technology, $V_{DD} = 0.25V$						
Logic Gates	Frequency			Energy		
	$\beta = 1$	β_{opt}	**ΔF**	$\beta = 1$	β_{opt}	**ΔE**
Inv	4.34E+04	8.93E+04	106%	2.93E-15	3.56E-15	21.5%
NAND-bc	3.27E+04	6.17E+04	88.7%	3.27E+04	5.03E-15	29.3%
NAND-wc	2.39E+04	5.14E+04	115%	5.07E-15	5.94E-15	17.2%
NOR-bc	1.79E+04	4.19E+04	134%	6.77E-15	5.55E-15	-18%
NOR-wc	1.41E+04	3.35E+04	138%	8.46E-15	7.26E-15	-14.2%
AORN-bc	2.30E+04	5.29E+04	130%	5.01E-15	4.83E-15	-3.59%
AORN-wc	9.98E+03	2.52E+04	153%	1.16E-14	9.40E-15	-19%

Chapter 5

Modified Parallel Transistor Stacks Technique

Digital circuits operating in the subthreshold region benefit from very low power consumption at the cost of speed [9]. Therefore, subthreshold circuits may not be practical in some industrial applications that require higher speeds. In order to improve the speed of subthreshold circuits, the method of parallel transistor stacks (PTS) was proposed in [4]. It is shown that using PTS and transistor widths that maximize COC, either individually or in parallel stacks, in subthreshold circuits leads to up to three times faster circuits. The transistor sizing presented for the four CMOS technologies in Chapters 3 and 4 can be applied to designing faster circuits based on PTS. Even though PTS has shown improvements in terms of speed in the subthreshold region, there are circumstances where PTS should be avoided. In this chapter first we introduce PTS followed by a methodology to identify when using PTS is beneficial in subthreshold circuits and when to avoid it.

5.1 Parallel Transistor Stacks

PTS is a technique in which large transistors are replaced with multiple parallel and smaller transistors with higher COC ratio [4]. Figures 5.1 and 5.2 show an inverter in the simple and PTS structures, respectively, in the 65 nm CMOS technology.

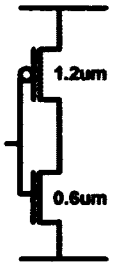

Figure 5.1: An inverter in simple structure in 65 nm CMOS technology

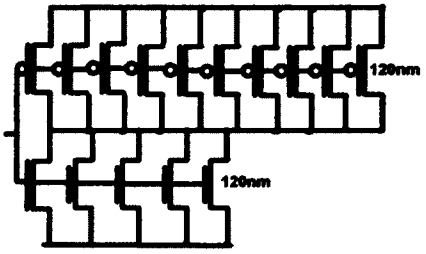

Figure 5.2: An inverter in PTS structure in 65 nm CMOS technology

As mentioned earlier in Chapters 3 and 4, higher speeds (i.e. lower propagation delays) can be achieved by increasing the current and lowering the capacitance. Therefore, the highest operating frequency is attained when the COC ratio is maximized. Due to the INWE, the maximum COC ratio occurs exactly, or close, to the transistor's minimum width (W_{INWE}) in modern CMOS technologies [4]. For example, in the 65 nm CMOS technology the maximum COC ratio occurs at $\beta_{opt} = 1$. Therefore, higher performance is achieved by splitting large transistors into multiple parallel transistors at the minimum width. The frequency of operation of a 9-stage inverter ring oscillator shown in Figure 5.3 increases by 2.63 times when the simple

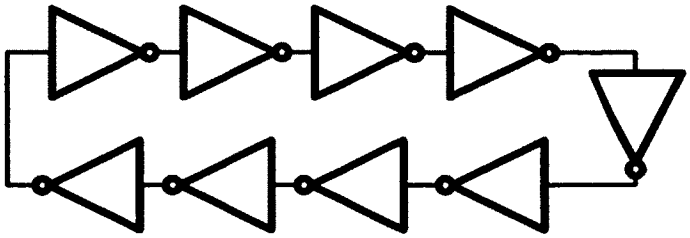

Figure 5.3: 9-stage inverter ring oscillator

inverter shown in Figure 5.1 is replaced with the PTS inverter shown in Figure 5.2 [4].

Based on PTS, an NMOS (PMOS) transistor at the width of 1.2 μm should be split into 5 (2) transistors of 220 nm (540 nm) wide in the 180 nm CMOS technology, 7 (7) transistors of 160 nm (160 nm) wide in the 130 nm CMOS technology, 10 (5) transistors of 120 nm (240 nm) wide in the 90 nm CMOS technology, and 10 (10) transistors of 120 nm (120 nm) wide in the 65 nm CMOS technology.

5.2 PTS and Subthreshold Current

In the previous section we introduced PTS and a methodology to increase the speed of digital subthreshold circuits. Another very important advantage of using PTS in addition to the speed improvement is that a linear current-width relationship is established by applying PTS on digital circuits in the subthreshold region. In fact, a linear relationship between the current and the number of parallel transistors is established. This is shown in Figure 5.4 and Eq. (40) as follows

$$I_n = \mu C_{ox} \frac{NW_{INWE}}{L} (n-1) V_T^2 e^{\frac{(V_{GS}-V_{th})}{nV_T}} \tag{40}$$

$$= KNW_{INWE} \tag{41}$$

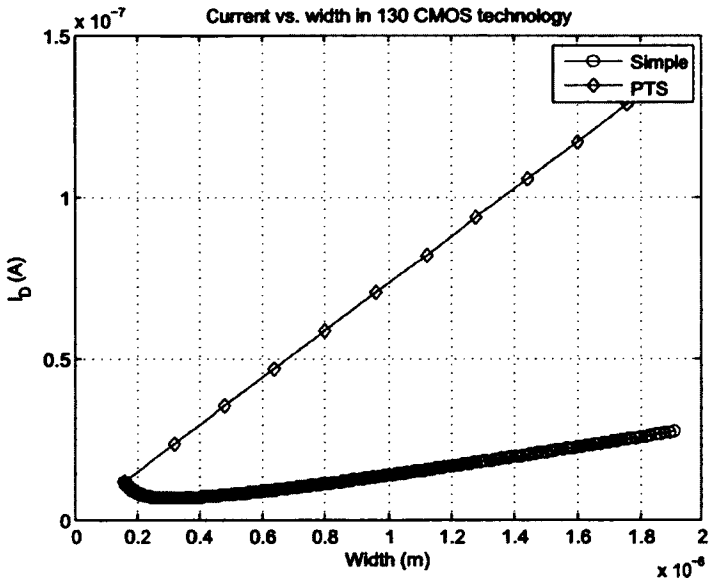

Figure 5.4: Drain current vs. width for an NMOS transistor for simple and PTS structures in 130 nm CMOS technology

where K is the proportionality constant, N is the number of parallel transistors at the width of W_{INWE}; V_{th} is now constant since the width of each transistor is now fixed.

Figure 5.4 shows the drain current versus width of an NMOS transistor for both simple and PTS structures in 130 nm CMOS technology. As shown in this figure a linear relationship is established between the current and the width of the transistor after applying PTS on the NMOS transistor. As we shall see in the next chapter, this enables the designer to apply transistor sizing method such as logical effort for subthreshold circuits.

5.3 When Using Parallel Transistor Stacks is Beneficial in Subthreshold Region

Although using PTS has generally result in speed improvements, there are CMOS technologies in which using PTS blindly may not improve the speed of a subthreshold circuit. In this section I provide a general design methodology that can be used to design circuits with higher speeds in the subthreshold region. I use 130 nm technology as an example to illustrate the design techniques and concepts. The proposed methodology in this chapter can be applied on any other CMOS technology. Figures 5.5 and 5.6 show NMOS and PMOS transistor's COC ratio vs. width in 130 nm CMOS technology, respectively. The transistors are biased for maximum current drive in the subthreshold region (i.e., ($V_{DS} = |V_{GS}|$). The capacitance includes the total gate and junction capacitance of the transistor. Note that I only interested in the values of the capacitances and current in the configurations shown. The COC ratio for the 130 nm NMOS transistor, shown in Figure 5.5, has a sharp decay as the width increases. Thus, in this technology, the maximum COC ratio affected by INWE is achieved at the minimum width for the NMOS transistor. Based on this figure, in order to achieve a higher COC ratio we either use the minimum transistor width (point A) or use PTS locked to the minimum width (i.e. we use multiple transistors at the minimum width instead of a large transistor). Nevertheless, the maximum COC ratio is not always associated with the minimum transistor width, as shown in Figure 5.6. The COC ratio for the PMOS transistor shown in this figure consists of two regions. In Region 1, the COC ratio corresponding to the widths larger than the minimum width have a lower value compared to the COC ratio at the minimum width. On the other hand, the widths located in Region 2 have a COC ratio larger than the minimum width. Based on this figure, we use PTS for transistor widths in Region 1 and avoid PTS for transistor widths in Region 2. Applying PTS on the

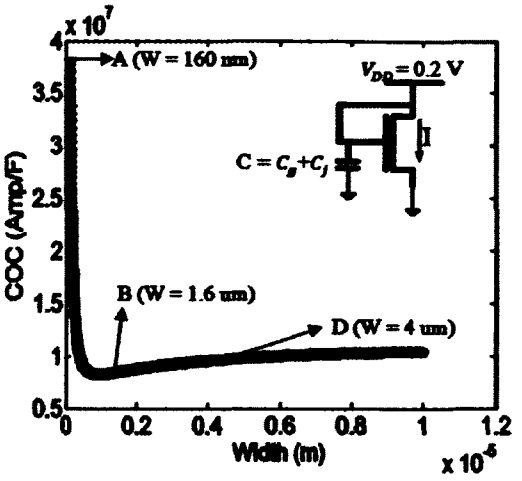

Figure 5.5: COC ratio vs. width for NMOS transistor in 130 nm CMOS technology

transistors in Region 2 will result in a lower circuit speed.

The following steps explain our methodology on how to identify the cases when to use and when not to use PTS in designing a logic circuit:

1. Plot the COC ratio vs. width for each transistor type (PMOS and NMOS) for the given technology kit (e.g., Figures 5.5 and 5.6).

2. Based on Step 1, identify the transistor width (W_{INWE}) associated with the local maximum COC ratio, which typically appears near the minimum width due to the INWE. As shown in Figures 5.5 and 5.6, point A denotes W_{INWE} for NMOS transistor and point C denotes W_{INWE} for PMOS transistor.

3. Apply a transistor sizing optimization method such as Logical Effort to optimize the circuit performance.

4. Find the optimum width (W_{opt}) for each transistor in the circuit based on Step 3.

5. Compare COC_{opt}, the COC ratio corresponding to W_{opt} (Step 4), with COC_{INWE}, the COC ratio corresponding to W_{INWE} (Step 2).

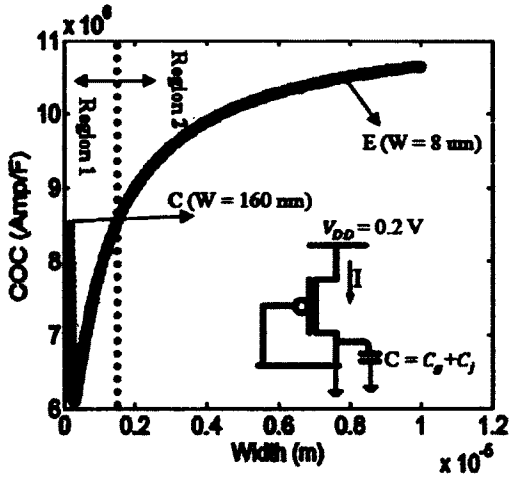

Figure 5.6: COC ratio vs. width for PMOS transistor in 130 nm CMOS technology

(a) If $COC_{opt} \leq COC_{INWE}$, use PTS, and split the transistor of width W_{opt} into multiple (K) transistors of width W_{INWE}, where $K = W_{opt}/W_{INWE}$ rounded to the nearest digit. For example, if Logical Effort finds that an NMOS transistor should have a width of 1.6 μm (point B in Figure 5.5) we split it into $K = W_{opt}/W_{INWE} = 10$ transistors at the width of 160 nm (point A in Figure 5.5).

(b) Else If $COC_{opt} > COC_{INWE}$, using PTS results in a slower circuit. In this case we keep the transistor width at W_{opt}. For example, if Logical Effort finds a transistor of width 8 μm (point E in Figure 5.6), we don't use PTS, as the COC ratio at point E has already a higher value compared to the COC ratio at 160 nm (point C in Figure 5.6).

Chapter 6

Impact of PTS on Driving Large Loads and Popagation Delay

To have a symmetric voltage transfer characteristic, equal rising and falling propagation delays, and better noise margins (for both high and low levels), it is desired to have the switching threshold, also called the logic threshold, of a logic gate at $V_{DD}/2$ (i.e. midway between supply rails) [2], [9]. To do this, both pull-up and pull-down transistors should have the same driving strengths. Since PMOS transistors have a lower carrier mobility, and hence a lower driving capability than NMOS transistors (i.e. holes are slower than electrons), we usually make PMOS transistors larger than NMOS transistors. In superthreshold circuits, PMOS transistors are usually sized two times larger than NMOS transistors. For subthreshold operation, the β to achieve equal driving strength in PMOS and NMOS transistors depends on the particular technology being used. For example, simulation results show that PMOS transistors should be 2 and 5.6 times larger than the NMOS transistors in 130 nm and 65 nm CMOS technologies, respectively.

Calhoun et al. [9], have reported that minimum-size devices are theoretically optimal in terms of energy consumption in the subthreshold region. Therefore, in order to design minimum-energy devices instead of up-sizing PMOS transistors to make the

rising and falling propagation delays equal, one can use minimum-size devices and increase the supply voltage to achieve the required performance. This design approach is not always optimum. First, in some applications the supply voltage is fixed and cannot be changed. Besides, in Section 6.1 we show that by using Logical Effort and PTS both propagation delay and energy can be improved compared to the minimum sized circuits. Moreover, in section 6.2, we show that by using PTS the rising and falling propagation delays in a chain of logic gates get fairly close after the second gate stage.

6.1 Driving Large Loads

In this section we show that minimum-size device are not always optimum in terms of energy consumption in the subthreshold region. To demonstrate this, we have compared the propagation delay and the energy consumption of three 4-stage inverter chains each connected to a load of 64x (a load of 64 times larger than a minimum-size inverter) in the 130 nm CMOS technology. The first inverter chain is designed with all inverters being of the minimum size (Min), the second inverter chain is sized with Logical Effort (with tapering ratio of $f = \sqrt[3]{64} = 4$) shown in Figure 6.1 and the third inverter is restructured with PTS after applying Logical Effort (LE & PTS).

The propagation delay (from node A to node B shown in Figure 6.1) and the energy consumption of all three inverter chains are shown in Figures 6.2 and 6.3, respectively. As shown in Figure 6.2, the minimum size inverter chain has the maximum propagation delay at all supply voltages in the subthreshold region. The inverter chain that is sized using Logical Effort has a lower propagation delay compared to the minimum size inverter chain, and the PTS inverter chain has the minimum delay at all supply voltages.

Figure 6.3 shows the energy consumption of all three inverter chains. As shown in

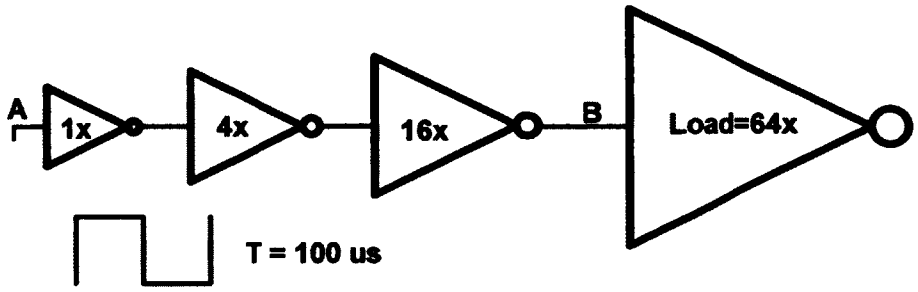

Figure 6.1: Inverter chain sized with Logical Effort and tapering ratio of 4

this figure, the minimum size inverter chain has the maximum energy consumption. The inverter chain sized with Logical Effort consumes lower energy than the minimum size inverter chain. The inverter chain with PTS structure has the minimum energy consumption among the three.

Logical Effort assumes that the current of a transistor is linearly proportional to its width. This assumption is not valid for MOSFETs operating in the subthreshold region in most modern CMOS technologies due to the INWE [4]. By using PTS, a linear-relationship between the current of the transistor and its width is established, enabling Logical Effort to be used for speed improvements in the subthreshold region [4].

The inverter chain sized using LE & PTS has a propagation delay of 2.67 μs whereas the minimum-size inverter chain has a propagation delay of 10 μs at 0.1 V (The speed of the LE & PTS inverter chain is 2.86 times more than the minimum-size inverter chain). According to the suggestion in [9], to achieve the same propagation delay of 2.67 μs for the minimum-size inverter chain, the supply voltage has to increase from 0.1 V to 0.15 V. However, the simulation results show 519 aJ of energy consumption for the minimum-size inverter at $V_{DD} = 0.15$ whereas the energy consumption of the (LE & PTS) inverter chain shows 404 aJ at $V_{DD} = 0.1$ (i.e 23% lower energy for LE & PTS inverter chain compared to the minimum-size inverter chain).

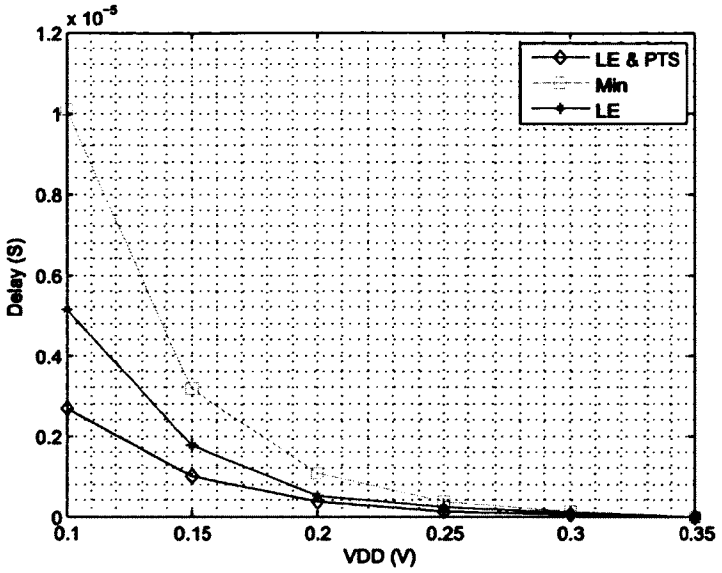

Figure 6.2: Propagation delay of three 4-stage inverter chains connected to a 64x load with: 1)all minimum-size inverters (Min), 2) sized using Logical Effort with a tapering ratio of 3(LE), and 3) with Logical Effort and PTS (LE PTS)

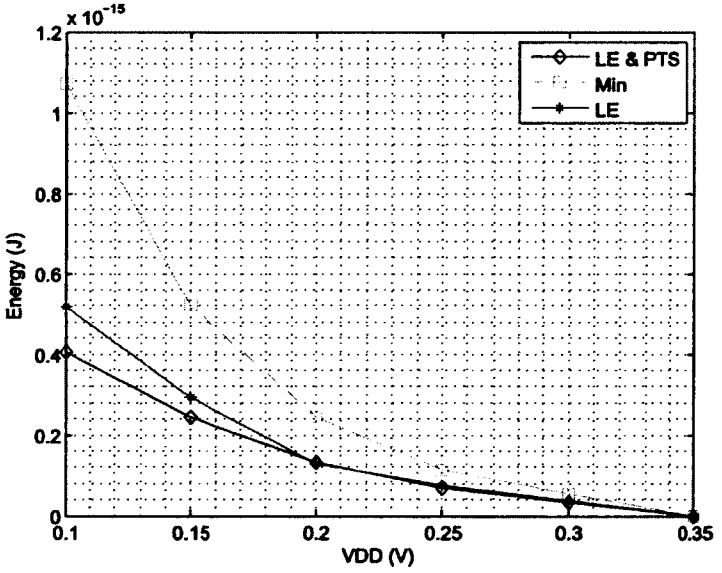

Figure 6.3: Energy per cycle of three 4-stage inverter chains connected to a 64x load with: 1)all minimum-size inverters (Min), 2) sized using Logical Effort with a tapering ratio of 3(LE), and 3) with Logical Effort and PTS (LE PTS)

6.2 Impact of PTS on Rising and Falling propagation Delays

To explore the effect of PTS on rising and falling propagation delays, we simulated the propagation delays of an inverter chain shown in Figure 6.4 in the 130 nm CMOS technology. To have equal rising and falling propagation delays each inverter is sized at $\frac{W_P}{W_N} = \frac{640\,\text{nm}}{320\,\text{nm}}$ running at the supply voltage of 0.3 V. The inverter chain is supplied with a input signal of 10 μs with a rise and fall time of 800 fs. The rise and fall time, that is calculated by the time that the output signal is between its 10% to 90% of its final value, is presented in the second and third row of Table 6.1, respectively. The 4th and 5th row of Table 6.1 presents the rising and falling propagation delays of each node from its previous node. The last two rows present the absolute rising and falling propagation delays of each node from the input node. Table 6.2 presents the same parameters presented in Table 6.1, for the inverter chain shown in Figure 6.4 when PTS is applied on all inverters. By applying PTS on each inverter, the NMOS transistor at the width of 320 nm is replaced by two NMOS transistors at the width of 160 nm, and each PMOS transistor at the width of 640 nm is replaced by four PMOS transistors at the width of 160 nm.

The simulation results presented in Table 6.2 show that applying PTS results in a non-equal rising and falling propagation delays (rows 4 and 5) at the first-gate stage (node A1). However, the rising and falling propagation delays get fairly close at the second (node A2) or at most the third gate stage (node A3). The reason is that the pull-down and pull-up network of each stage will charge the pull-up and pull-down network of the following stage, respectively.

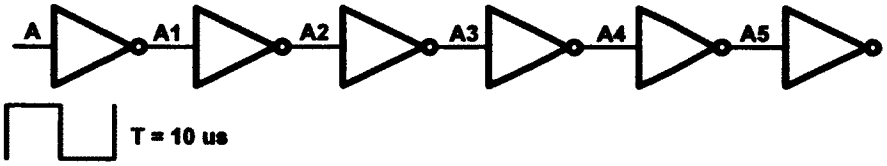

Figure 6.4: Inverter chain

Table 6.1: Propagation Delays of the Inverter Chain Shown in Figure 6.4 With Simple Structure

Parameter	Input	A1	A2	A3	A4	A5
Rise Time	800fs	6.672ns	7.807ns	7.77ns	7.903ns	7.903ns
Fall Time	800fs	7.068ns	8.303ns	8.39ns	8.453ns	8.43ns
Rise Delay		4.199ns	8.268ns	8.684ns	8.675ns	8.679ns
Fall Delay		4.478ns	8.53ns	8.897ns	8.88ns	8.885ns
Absolute Rise Delay		4.199ns	12.75ns	21.42ns	30.33ns	39ns
Absolute Fall Delay		4.478ns	12.73ns	21.64ns	30.32ns	39.22ns

Table 6.2: Propagation Delays of the Inverter Chain Shown in Figure 6.4 With PTS Structure

Parameter	Input	A1	A2	A3	A4	A5
Rise Time	800fs	6.225ns	6.225ns	6.242ns	6.39ns	6.35ns
fall Time	800fs	2.156ns	3.641ns	3.671ns	3.616ns	3.66ns
Rise Delay		3.65ns	4.77ns	5.129ns	5.13ns	5.126ns
Fall Delay		1.33ns	4.523ns	4.629ns	4.633ns	4.62ns
Absolute Rise Delay		3.65ns	6.103ns	13.2ns	15.87ns	23.06ns
Absolute Fall Delay		1.33ns	8.173ns	10.73ns	17.94ns	20.5ns

Chapter 7

Applications

The design methodology developed in Chapter 5 and the findings of Chapter 6 will be put to the test in some application circuits. The first test circuit is a 19-stage inverter ring oscillator. The second circuit is a 4-bit comparator shown in Figure 7.3.

7.1 Ring Oscillator

A 19-stage inverter ring oscillator shown in Figure 7.1 running at 0.2 V is simulated to verify the correctness of the proposed methodology in the 130 nm and 90 nm CMOS technologies. We explored the effect of incorporating PTS on each inverter as shown in Figure 7.2.

Considering a standard inverter shown in Figure 7.2(a), the simulation results for the 130 nm CMOS technology shown in Table 7.1 reveals that applying PTS on only PMOS transistors (column P-PTS) of each inverter in the ring oscillator reduces the

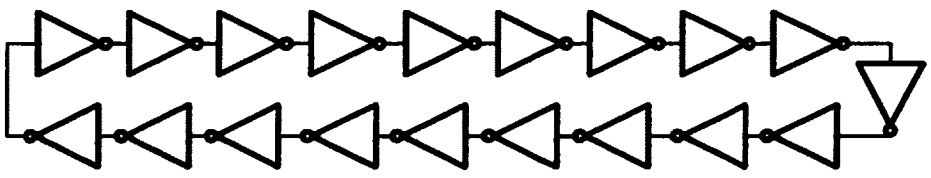

Figure 7.1: 19-stage ring oscillator

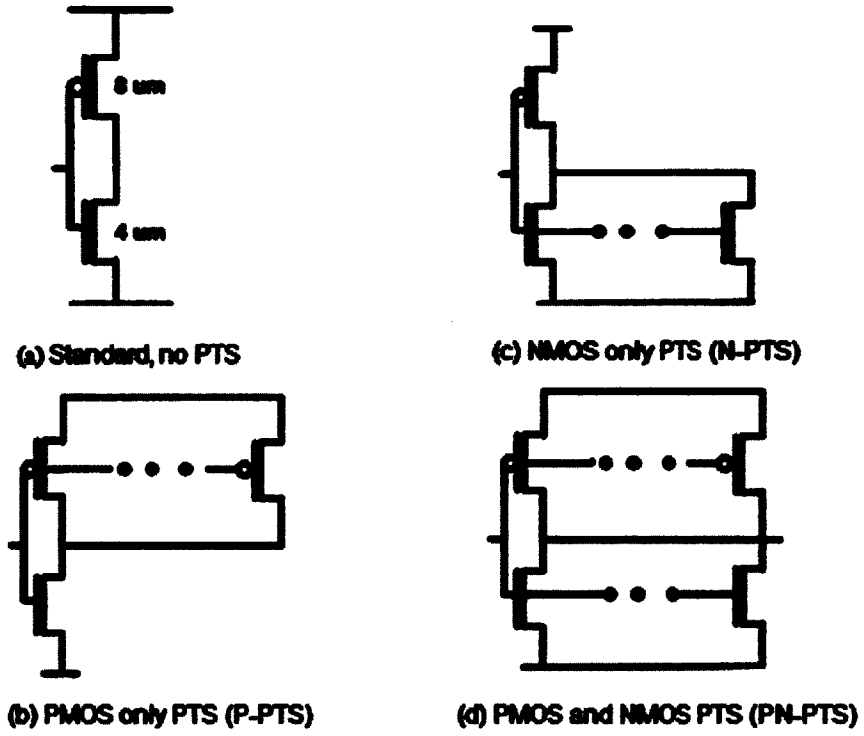

Figure 7.2: Inverter configurations for a 19-inverter ring oscillator: (a) Standard, (b) PMOS PTS, (c) NMOS PTS, (d) PMOS and NMOS PTS

frequency of oscillation by 3%. On the other hand, applying PTS on only NMOS transistors (column N-PTS) improves the frequency by 120%, while applying PTS on both NMOS and PMOS transistors increases the frequency only by 100%. This result is expected as the COC ratio corresponding to the PMOS transistor of 8 μm width (point E in Figure 5.6) is more than the COC ratio compared to 160 nm (point C in Figure 5.6), and the NMOS transistor at the width of 4 μm (point D in Figure 5.5) has much lower COC ratio compared to point A. Therefore, in this case, the optimum sizing (column OPT-PTS) is achieved by applying PTS on NMOS transistors, and avoiding applying PTS on PMOS transistors.

In the 90 nm CMOS technology, the COC ratio for both transistor types follow the same behavior as that of the 130 nm technology. Simulation results illustrate

that the corresponding COC ratio of the NMOS transistor versus width has a sharp decay, whereas the COC ratio corresponding to the PMOS transistor increases after a certain width (1 μm). The simulation results on a 19-stage inverter ring oscillator shown in Table 7.2 for the 90 nm CMOS technology confirm that applying PTS on all PMOS transistors of 8 μm width results in 16% lower frequency compared to that of the standard sizing. On the other hand, applying PTS on only NMOS transistors improves the speed almost by 81%. Applying PTS on both PMOS and NMOS transistors increases the speed only by 41% compared to the standard case. Again, for achieving the maximum speed we advise applying PTS just on NMOS transistors and not PMOS transistors.

In the 65 nm CMOS technology, the COC ratio for both transistors versus width have a sharp decay. Therefore, to achieve the maximum speed both PMOS and NMOS transistors should be sized at the minimum width. The simulation results on a 19-stage inverter ring oscillator shown in Table 7.3 for the 65 nm CMOS technology confirm that by applying PTS on both transistors improves the speed by 210% compared to the standard size. However, the speed improves by 41% and 43% when applying PTS on only NMOS transistors and PMOS transistors, respectively. Therefore in 65 nm CMOS technology applying PTS on both transistors are advised.

Table 7.1: Simulation Results of a 19-stage Ring Oscillator in 130 nm CMOS Technology

Metrics	**Standard**	**P-PTS**	**N-PTS**	**PN-PTS**	**Opt-PTS**
Frequency (KHz)	364.8	357.2	806.1	753.9	806.1
Power (nW)	10.3	9.821	23.45	21.73	23.45
Energy (fJ)	28.24	27.49	29.09	28.82	29.0
EDP (fJ.μs)	77	76	36	38	36

Table 7.2: Simulation Results of a 19-stage Ring Oscillator in 90 nm CMOS Technology

Metrics	Standard	P-PTS	N-PTS	PN-PTS	Opt-PTS
Frequency (MHz)	6.4	5.3	11.6	9.15	11.6
Power (nW)	165.9	138.4	327.6	269.8	327.6
Energy (fJ)	25.92	26.11	28.24	29.48	28.24
EDP (fJ.μs)	4.0	4.7	2.4	3.2	2.4

Table 7.3: Simulation Results of a 19-stage Ring Oscillator in 65 nm CMOS Technology

Metrics	Standard	P-PTS	N-PTS	PN-PTS	Opt-PTS
Frequency (KHz)	59.1	84.6	83.99	181.7	181.7
Power (nW)	1.44	2.817	2.766	6.618	6.618
Energy (fJ)	24.37	33.29	32.93	36.42	36.42
EDP (fJ.μs)	412	393	392	200	200

7.2 4-bit Comparator

We also applied our methodology on implementing a 4-bit comparator. The simulation results after Logical Effort optimization are shown in Table 7.4 (column Standard). By running Logical Effort on this complex circuit, it suggests PMOS transistors with widths in both regions shown in Figure 5.6. According to our methodology, for NMOS transistors, as all suggested widths have lower COC ratio than the minimum width, we apply PTS on all NMOS transistors in both CMOS technologies. Applying PTS on PMOS transistors depends on whether the Logical Effort locates the width

sizes in Region 1 or Region 2. Regardless the fact that PTS should be applied on certain transistors, in Table 7.4 we list the simulation results of applying PTS only on all PMOS and only on all NMOS transistors in the P-PTS, and N-PTS columns, respectively. Column PN-PTS lists the simulation results for applying PTS on both NMOS and PMOS transistors. Considering the results using 130 nm CMOS technology, the reason that applying PTS on all transistors in 4-bit comparator gains speed benefit compared to the standard sizing, with the delay of 862.0 ns vs. 908.3 ns, is that the number of PMOS transistors which benefit from PTS (Region 1) are dominant. The last column, OPT-PTS, illustrates the results of optimum sizing (maximum speed) based on our methodology. This column shows 41% increase of speed compared to the standard sizing, which is 3% more than applying PTS on all transistors blindly. The reason that our methodology in this case doesn't show considerable improvement compared to the blind PTS is that most of the PMOS transistors are in Region 1.

The result of applying the proposed methodology, i.e. selective application of PTS, on the 4-bit comparator in the 90 nm CMOS technology is also shown in OPT-PTS column of Table 7.5. The speed is increased by around 39% compared to standard sizing, while applying blind PTS only improves the speed by 13%. Note that applying PTS only on all PMOS transistors (P-PTS) degrades the speed by 33%, while applying PTS only on all NMOS transistors (N-PTS) improves the speed by 35% compared to the standard sizing.

Table 7.4: Simulation Results of a 4-bit Comparator in 130 nm CMOS Technology

Metrics	Standard	P-PTS	N-PTS	PN-PTS	Opt-PTS
Delay (ns)	908.3	862.0	573.5	548.3	531.0
Power (pW)	772	775.2	1401	1403	1390
Energy (aJ)	701	668	803	769	738
EDP (fJ.μs)	636	576	607	555	517

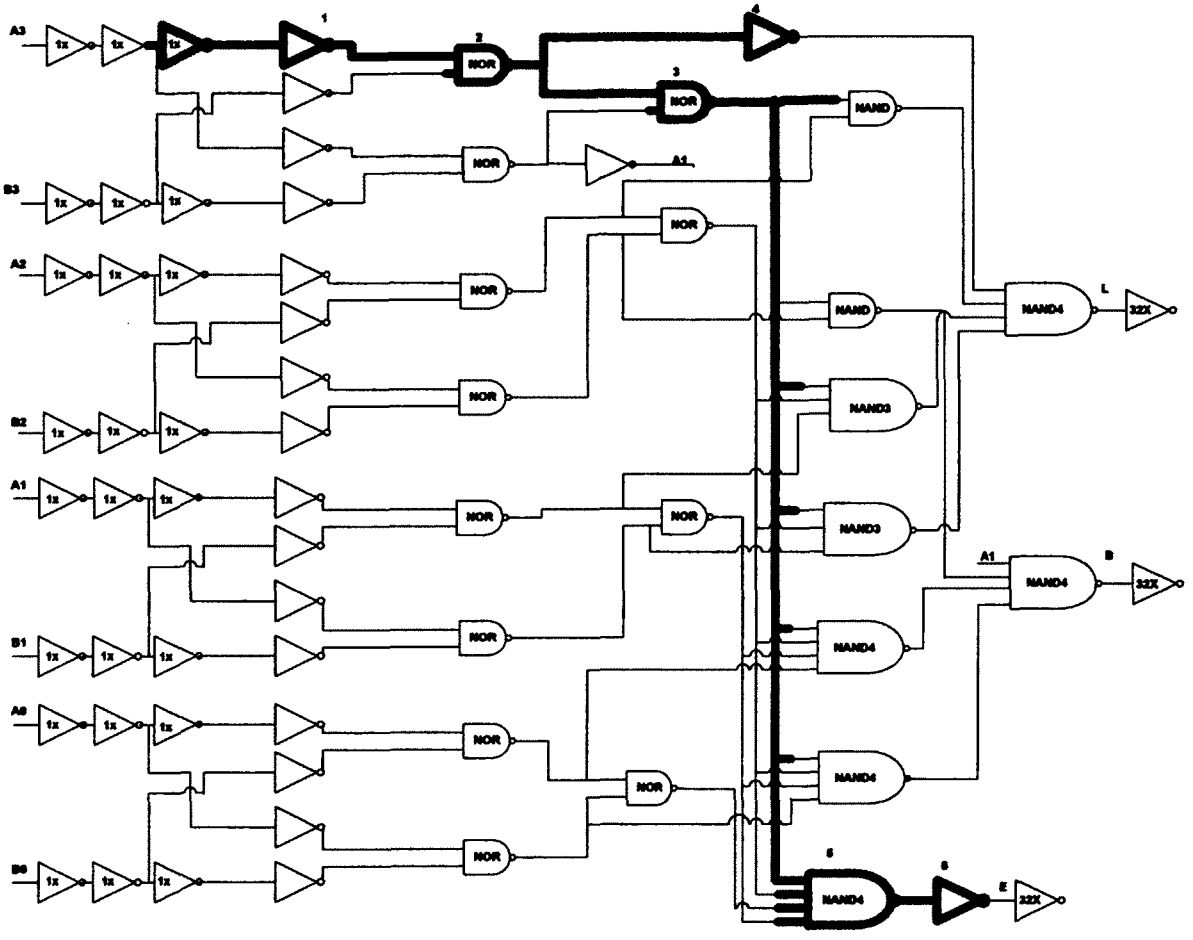

Figure 7.3: 4-bit comparator

Table 7.5: Simulation Results of a 4-bit Comparator in 90 nm CMOS Technology

Metrics	Standard	P-PTS	N-PTS	PN-PTS	Opt-PTS
Delay (ns)	56.0	74.7	36.2	48.6	34.2
Power (pW)	9.76	8.32	17.12	15.2	16.79
Energy (zJ)	547	622	620	739	574
EDP (fJ.μs)	31	46	22	36	20

As shown in Tables (7.1-7.5), the improvement in the delay (frequency) comes at the cost of power consumption. This follows the linear relationship between power and frequency $P = CV_{DD}^2 f$. By increasing the frequency the power consumption linearly increases. By applying PTS on the ring oscillator in both the 130 nm and 90 nm technology kits the power increases by 127% and 97% respectively. This is while the results show 121% and 81% improvement in terms of speed respectively. Also by applying PTS on the 4-bit comparator the power increases by 80% in the 130 nm CMOS technology and by 72% in the 90 nm technology kit compared to the standard case. The differences between the speed improvement and the power consumption is due to the extra contacts and RCs created when applying PTS on logic gates. This will also cause the area to increases by 50% when using PTS compared to the conventional method. In other words, if we keep the total MOSFET width constant, the speed improvement that PTS results comes at the cost of power consumption and area. We have included the power consumption results from our simulations in the Tables (7.1-7.5) shown above. However, in general by applying PTS, the energy consumption marginally differs and as shown in the next section on the extracted layout of a 32-bit look-ahead adder the energy consumption is reduced.

The Energy-Delay-Product (EDP) which is often used as an ultimate quality metric [2] exhibits considerable reduction when incorporating the proposed methodology on the test circuits. For example, the EDP shown in column OPT-PTS (maximum speed) in Tables (7.1-7.5) shows 18% to 53% decrease compared to those of the standard sizing technique (Logical Effort), and 5% to 45% reduction compared to the case when blind PTS is applied (PN-PTS column).

7.3 32-bit Carry Look Ahead Adder

We applied our methodology to a 32-bit carry-lookahead adder developed in our group [1]. The adder is made from eight valency-4 blocks. One block is shown in Figure 7.4. Bitwise propagate and generate (PG) signals are evaluated for the four bits of A and B. The bitwise PG signals are then combined with Cin to produce four output bits of S. In the first block, Cin refers to the Cin input to the adder, otherwise Cin refers to the signal Cout from a previous block. Two versions of this adder are implemented in the 65 nm LP CMOS technology and the extracted layouts are compared. The first version is a minimum-size design where all gates are made from minimum-sized transistors. The second is designed using Logical Effort followed by PTS. For the minimum design, the full layout area is 1869 μm^2 and the PTS design occupies 2887 μm^2. All inputs are driven through two successive inverters that have roughly equal rising and falling propagation delays. Every output is loaded with an inverter with PMOS and NMOS widths of 0.24 and 0.12 μm, respectively. We simulated the adder with a test case of all A inputs fixed at 0, and all bits of B fixed at 1. We hooked up Cin input to a square wave with a period of 300 μs. The results for the two adders are shown in Table 7.6. As shown in this table, the PTS adder is as twice as fast as the minimum-sized adder at different supply voltages. In addition, the energy consumed by the PTS adder is marginally lower than its counter

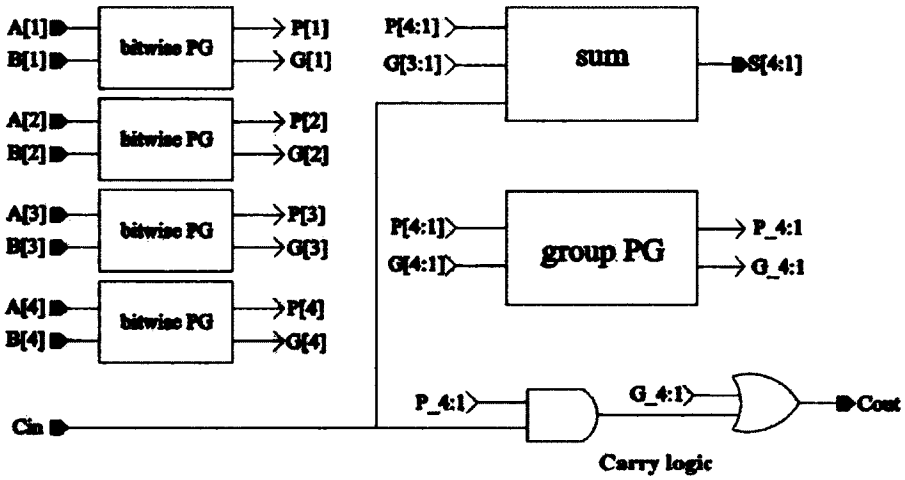

Figure 7.4: Valency-4 carry-lookahead adder block [1]

Table 7.6: Simulation Results of the Extracted Layout of 32-bit Adder in 65 nm CMOS Technology

V_{DD} (V)	Delay			Energy		
	Min	PTS	**Improvement**	Min	PTS	**Improvement**
0.3	5.64E-06	2.87E-06	49.1%	1.57E-14	1.46E-14	7%
0.25	1.78E-05	9.47E-06	46.8%	3.62E-14	3.52E-14	2.7%
0.2	5.68E-05	3.04E-05	46.4%	7.93E-14	7.78E-14	1.8%
0.15	1.84E-04	9.43E-05	48.7%	1.60E-13	1.52E-13	5%

part at different supply voltages.

Chapter 8

Conclusion and Future Work

With rigid power budget in ultra-low power applications such as micro-sensor networks, pacemakers, and many portable devices, circuits operating with a supply voltage below the threshold voltage (subthreshold) of a transistor provides an ideal low-power solution while increasing the battery life [9]. However, the benefit comes at the cost of performance due to the extremely scaled down supply voltage. This has limited the applications of the subthreshold design to specific markets. Therefore, improving the performance of subthreshold circuits can expand the application spectrum of these circuits.

The main contribution of this thesis is expanding the domain of applications of subthreshold circuits. This is attained by developing methodologies for designing faster subthreshold circuits at no or minimal energy cost.

8.1 Summary

Chapter 1 presented the previous work in the area of subthreshold circuits. In Chapter 2, the relevant transistor properties to the delay and energy consumption such as the current and capacitances are discussed, sources of the leakage current are addressed, and a qualitative explanation of the inverse-narrow-width-effect (INWE) is

also presented.

Chapter 3 presented the Current-Over-Capaciatance (COC) ratio and the corresponding simulations to examine the optimum PMOS-to-NMOS width ratio (β) for improving the speed of CMOS logic gates operating in the subthreshold region.

In Chapter 4, analytical model for obtaining the maximum COC ratio was developed, and simulation results illustrate that the frequency attains its maximum for optimum β independent of the supply voltage.

In Chapter 5, we introduced the Parallel Transistor Stacks (PTS) technique. We show how β_{opt} obtained from simulations and analytical model in Chapters 3 and 4 can be exploited in designing circuits with PTS. PTS has been shown in [4] as an effective technique for improving the speed of digital circuits operating in the subthreshold region, but there are circumstances where PTS would not improve the speed of a subthreshold circuit. Chapter 5 of this thesis deals with a methodology to identify when using PTS is beneficial (or not) in a particular CMOS technology and what transistor sizing can be employed to maximize the circuit speed. This methodology is based on analyzing the COC ratio of PMOS and NMOS transistors.

In Chapter 6, we explore the effect of PTS on driving large loads and propagation delays. We show that by using PTS both the delay and energy can be minimized in some cases, and using minimum-size devices are not always the best solution in terms of energy consumption, as reported in the literature [4]. The impact of PTS on the rising and falling delays of a logic chain is also explored in this chapter. It is shown that PTS doesn't disturb the rising and falling propagation delays. The rising and falling propagation delays become equal after the second stage of the chain.

In Chapter 7, we put the developed methodology and the findings of Chapter 6 to the test in some application circuits. We show the superiority of the developed methodology in Chapter 5 over the blind use of PTS. We demonstrate that using PTS is not beneficial in all cases. Incorporating the proposed methodology in a 4-bit

comparator and a 19-stage inverter ring oscillator, in the 90 nm CMOS technology, results in 26% and 40% extra improvement (over the minimum-size circuits) compared to the blind use of PTS, respectively. Applying our methodology on the extracted layout of a 32-bit carry-look ahead adder in the 65 nm CMOS technology illustrates that the PTS adder is as twice as fast as the minimum-sized adder at different supply voltages. In addition, the energy consumed by the PTS adder is marginally lower than its counter part at different supply voltages.

8.2 Future Work

The design methodology developed in this thesis is geared towards performance. However, another measure such as power and energy consumption could be the ultimate goal.

The PTS technique can be improved by the fingering technique. In other words, instead of connecting the transistors in parallel we can use fingers. By doing so, the capacitance and the area well become less than that of a PTS structure. Therefore, more speed improvements can be achieved.

Circuits operating in the subthreshold region are prone to PVT variations. Optimizing this parameter for minimum-energy operation could be a topic of further investigation.

List of References

[1] M. Muker. *Subthreshold CMOS logic design using parallel transistor stacks.* Masters thesis, Dept. of Electronics, Carleton Universisty (2010).

[2] J. Rabaey, A. Chandrakasan, and B. Nikolic. *Digital Integrated Circuits: A Design Perspective.* NJ: Prentice Hall/Pearson Education (2003).

[3] A. Wang and A. Chandrakasan. "A 180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology." *IEEE Journal of Solid-State Circuits* **40**(1), 310–319 (2005).

[4] M. Muker and M. Shams. "Designing Digital Subthreshold CMOS Circuits Using Parallel Transistor Stacks." *Electronics Letters* **47**(6), 372–374. ISSN 00135194 (2011).

[5] R. Gonzalez, B. Gordon, and M. Horowitz. "Supply and Threshold Voltage Scaling for Low Power CMOS." *IEEE Journal of Solid-State Circuits* **32**(8), 1210–1216 (1997).

[6] D. Liu and C. Svensson. "Trading Speed for Low Power by Choice of Supply and Threshold Voltages." *IEEE Journal of Solid-State Circuits* **28**(1), 10–17 (1993).

[7] A. Wang, A. Chandrakasan, and S. Kosonocky. "Optimal Supply and Threshold Scaling for Subthreshold CMOS Circuits." *Proceedings of IEEE Computer Society Annual Symposium on VLSI* pages 5–9 (2002).

[8] B. H. Calhoun and A. Chandrakasan. "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits." *Proceedings of the 2004 International Symposium on Low Power Electronics and Design - ISLPED '04* pages 90–95 (2004).

[9] B. Calhoun, A. Wang, and A. Chandrakasan. "Modeling and Sizing for Minimum Energy Operation in Subthreshold Circuits." *IEEE Journal of Solid-State Circuits* **40**(9), 1778–1786 (2005).

[10] D. Bol, D. Kamel, D. Flandre, and J. D. Legat. "Nanometer MOSFET effects on the minimum-energy point of 45nm subthreshold logic." *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design (ISLPED '09)* pages 3–8 (2009).

[11] D. Bol and D. Flandre. "Nanometer MOSFET Effects on the Minimum-Energy Point of Sub-45nm Subthreshold Logic—Mitigation at Technology and Circuit Levels." *ACM Transactions on Design Automation* **16**(1), 1–26 (2010).

[12] J. Zhou, S. Jayapal, B. Busze, L. Huang, and J. Stuyt. "Digital Computation in Subthreshold Region for Ultralow-Power Operation : A Device Circuit Architecture Codesign Perspective." *48th ACM/EDAC/IEEE Design Automation Conference (DAC)* **98**(2), 441–446 (2011).

[13] D. Bol, R. Ambroise, D. Flandre, and J. Legat. "Channel Length Upsize for Robust and Compact Subthreshold SRAM." *in Proc. Workshop Faible Tension Faible Consommation (FTFC)* **16**(5), 117–120 (2008).

[14] J. Keane, H. Eom, T.-h. Kim, S. Sapatnekar, and C. Kim. "Stack Sizing for Optimal Current Drivability in Subthreshold Circuits." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **16**(5), 598–602 (2008).

[15] J. Keane, H. Eom, T. Kim, S. Sapatnekar, and C. Kim. "Subthreshold Logical Effort: A Systematic Framework for Optimal Subthreshold Device Sizing." *Proceedings of the 43rd annual Design Automation Conference* pages 425–428 (2006).

[16] Y. Osaki, T. Hirose, K. Matsumoto, N. Kuroki, and M. Numa. "Delay-Compensation Techniques for Ultra-Low-Power Subthreshold CMOS Digital LSIs." *Proceeding of 52nd IEEE International Midwest Symposium on Circuits and Systems, MWSCAS'09* (1), 503–506 (2009).

[17] J. Tolbert and S. Mukhopadhyay. "Accurate Buffer Modeling with Slew Propagation in Subthreshold Circuits." *Proceedings of the 2009 International Symposium on Quality of Electronic Design* pages 91–96 (2009).

[18] J. Chen, L. T. Clark, and T.-h. Chen. "An Ultra-Low-Power Memory With a Subthreshold Power Supply Voltage." *IEEE Journal of Solid-State Circuits* **41**(41), 2344–2353 (2006).

[19] F. Moradi, D. Wisland, A. Peiravi, and H. Mahmoodi. "1-bit sub threshold full adders in 65nm CMOS technology." *International Conference on Microelectronics (ICM)* pages 268–271 (2008).

[20] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw. "Exploring Variability and Performance in a Sub-200-mV Processor." *IEEE Journal of Solid-State Circuits* **43**(4), 881–891 (2008).

[21] D. Bol, D. Flandre, and J.-D. Legat. "Ultra-Low-Power Logic Style for Low-Frequency High-Temperature Applications." *Proceedings of the Thematic Network on Silicon on Insulator Technology, Devices and Circuits (EuroSOI)* pages 33–34 (2008).

[22] D. Bol, J. De Vos, R. Ambroise, D. Flandre, and J.-D. Legat. "Building Ultra-Low-Power High-Temperature Digital Circuits in Standard High-Performance SOI Technology." *Elsevier Journal of Solid-State Electronics* **52**(12), 1939–1945. ISSN 00381101 (2008).

[23] D. Bol, D. Flandre, and J.-d. Legat. "Robustness-Aware Sleep Transistor Engineering for Power-Gated Nanometer Subthreshold Circuits." *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)* pages 1484–1487 (2010).

[24] J. Crop, R. Pawlowski, N. Moezzi-Madani, J. Jackson, and P. Chaing. "Design Automation Methodology for Improving the Variability of Synthesized Digital Circuits Operating in the Sub/Near-Threshold Regime." *International Green Computing Conference and Workshops (IGCC)* pages 1–6 (2011).

[25] S. D. Pable and M. Hasan. "Interconnect Design For Subthreshold Circuits." *IEEE Transactions on Nanotechnology* **11**(3), 633–639 (2012).

[26] R. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge. "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits." *Proceedings of the IEEE* (98), 253–266 (2010).

[27] D. Markovic, C. C. Wang, L. Alarcon, T.-T. Liu, and J. Rabaey. "Ultralow-Power Design in Near-Threshold Region." *Proceedings of the IEEE* **98**(2), 237–252. ISSN 0018-9219 (2010).

[28] R. D. Jorgenson, L. Sorensen, D. Leet, M. S. Hagedorn, D. R. Lamb, T. H. Friddell, and W. P. Snapp. "Ultralow-Power Operation in Subthreshold Regimes Applying Clockless Logic." *Proceedings of the IEEE* **98**(2), 299–314. ISSN 0018-9219 (2010).

[29] T. Sakurai and A. R. Newton. "Alpha-power law MOSFET Model and its Applications to CMOS Inverter Delay and other Formulas." *IEEE Journal of Solid-State Circuits* **25**(2), 584–594 (1990).

[30] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand. "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits." *Proceedings of the IEEE* **91**(2), 305–327 (2003).

[31] P. Gryboś, M. Idzik, and A. Skoczen. "Design of Low Noise Charge Amplifier in Sub-Micron Technology for Fast Shaping Time." *Analog Integrated Circuits and Signal Processing* **49**(2), 107–114 (2006).

[32] Y. Tisividis. *Operation Modeling MOS Transistor.* 2ed ed, Oxford University Press (1999).

[33] N. H. E. Weste and D. Harris. *CMOS VLSI Design: A Circuits and Systems Perspective.* 3rd ed, Pearson Addison-Wesley (2005).